VOICE INTERACTION GAME DESIGN AND GAMEPLAY

Fraser Allison

ORCID: 0000-0002-7005-7390

Submitted in total fulfilment of the requirements
of the degree of Doctor of Philosophy

January 2020

School of Computing & Information Systems
Melbourne School of Engineering
The University of Melbourne

Abstract

This thesis is concerned with the phenomenon of voice-operated interaction with characters and environments in videogames. Voice interaction with virtual characters has become common in recent years, due to the proliferation of conversational user interfaces that respond to speech or text input through the persona of an intelligent personal assistant. Previous studies have shown that users experience a strong sense of social presence when speaking aloud to a virtual character, and that voice interaction can facilitate playful, social and imaginative experiences of the type that are often experienced when playing a videogame. Despite this, the user experience of voice interaction is frequently marred by frustration, embarrassment and unmet expectations.

The aim of this thesis is to understand how voice interaction can be used in videogames to support more enjoyable and meaningful player experiences. Voice-operated videogames have existed for more than three decades, yet little research exists on how they are designed and how they are received by players. The thesis addresses that knowledge gap through four empirical studies. The first study looks at player responses to a videogame character that can be given commands through a natural language interface. The second study is a historical analysis of voice-operated games that examines the technological and cultural factors that have shaped their form and popularity. The third study develops a pattern language for voice game design based on a survey of 471 published videogames with voice interaction features. The fourth study compares player responses to videogames that feature speech-based voice interaction and non-verbal voice interaction, and applies the theoretical perspective of frame analysis to interpret their reactions.

Through these studies, the thesis makes two main contributions to the human-computer interaction and games studies literature. First, it identifies five genres of voice gameplay that are based upon fundamentally different types of vocal activities, and details the design patterns and design goals that are distinctive to each genre. Second, it presents an empirically grounded theoretical model of gameplay that accounts for players' feelings of engagement, social presence, frustration and embarrassment during voice gameplay. Overall, the thesis demonstrates that the fictional framing a videogame presents is a crucial factor in determining how players will experience its voice interaction features.

Declaration

This is to certify that:

i. the thesis comprises only my original work towards the PhD, except where indicated in the Preface

ii. due acknowledgement has been made in the text to all other material used

iii. the thesis is fewer than 100,000 words in length, exclusive of tables, figures, bibliographies and appendices.

Fraser Allison

January 2020

Preface

Peer-reviewed publications included in the thesis

This dissertation includes four published research articles, in accordance with The University of Melbourne guidelines for a *Thesis with Publication*.[1]

Chapter 4 presents the following peer-reviewed journal article:

> Allison, F., Luger, E., & Hofmann, K. (2018). How Players Speak to an Intelligent Game Character Using Natural Language Messages. *Transactions of the Digital Games Research Association*, 4(2), 1–50. https://doi.org/10.26503/todigra.v4i2.88

I led all stages of the research and writing process. The research design and data collection was conducted at Microsoft Research Cambridge, UK, during a 12-week research internship for which I took leave from my PhD studies. My supervisors for that research internship, Dr Ewa Luger and Dr Katja Hofmann, contributed feedback on the study design and written drafts.

Chapter 5 presents the following peer-reviewed journal article:

> Allison, F., Carter, M., & Gibbs, M. (2017). Word Play: A History of Voice Interaction in Digital Games. *Games and Culture*, 1555412017746305. https://doi.org/10.1177/1555412017746305

I led all stages of the research and writing process. My thesis supervisors, Dr Marcus Carter and Assoc. Prof. Martin Gibbs, contributed feedback on the study design and written drafts.

Chapter 6 presents the following peer-reviewed conference paper:

> Allison, F., Carter, M., Gibbs, M., & Smith, W. (2018). Design Patterns for Voice Interaction in Games. *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, 5–17. https://doi.org/10.1145/3242671.3242712

I led all stages of the research and writing process. My thesis supervisors, Dr Marcus Carter, Assoc. Prof. Martin Gibbs and Assoc. Prof. Wally Smith, contributed feedback on the study design and written drafts.

Chapter 7 presents the following peer-reviewed conference paper:

> Allison, F., Newn, J., Smith, W., Carter, M., & Gibbs, M. (2019). Frame Analysis of Voice Interaction Gameplay. *Proceedings of the 2019 CHI*

---

[1] The University of Melbourne. (2019). *Graduate Research Training Policy (MPF1321)*. Retrieved from https://policy.unimelb.edu.au/MPF1321

Conference on Human Factors in Computing Systems, 393:1–393:14. https://doi.org/10.1145/3290605.3300623

I led all stages of the research and writing process. My fellow PhD candidate, Joshua Newn, contributed to the technical setup, data analysis and writing of the paper. My thesis supervisors, Wally Smith, Marcus Carter and Martin Gibbs, contributed feedback on the study design and written drafts.

Peer-review publications based upon the thesis research

Two further publications drew upon the research conducted for this thesis, but are not included in the thesis itself:

> Allison, F. (2016). A History of Voice Interaction in Games. *Abstract Proceedings of the First International Joint Conference of DiGRA and FDG*. DiGRA/FDG '16. http://www.digra.org/digital-library/publications/a-history-of-voice-interaction-in-games/

> Allison, F., Luger, E., & Hofmann, K. (2017). Spontaneous Interactions with a Virtually Embodied Intelligent Assistant in Minecraft. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2337–2344. https://doi.org/10.1145/3027063.3053266

For both of these papers, I led all stages of the research and writing process. My co-authors on the second paper, Ewa Luger and Katja Hoffmann, contributed feedback on the study design and written drafts, and assisted with the data collection.

Additional peer-reviewed publications produced during candidature

The following papers were produced during my candidature, but did not draw upon or contribute directly to the research conducted for this thesis:

> Allison, F. (2015). Whose mind is the signal? Focalization in video game narratives. *Proceedings of the 2015 DiGRA International Conference*, *12*. http://www.digra.org/digital-library/publications/whose-mind-is-the-signal-focalization-in-video-game-narratives/

> Carter, M., Allison, F., Downs, J., & Gibbs, M. (2015). Player Identity Dissonance and Voice Interaction in Games. *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, 265–269. https://doi.org/10.1145/2793107.2793144

> Allison, F., Carter, M., & Gibbs, M. (2015). Good Frustrations: The Paradoxical Pleasure of Fearing Death in DayZ. *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*, 119–123. https://doi.org/10.1145/2838739.2838810

Allison, F. (2016). Are You Out of Your Mind? Focalization in Digital Games. *Transactions of the Digital Games Research Association*, *2*(3), 31-60.

Carter, M., & Allison, F. (2017). Fear, loss and meaningful play: Permadeath in DayZ. *Journal of Gaming & Virtual Worlds*, *9*(2), 143–158. https://doi.org/10.1386/jgvw.9.2.143_1

Newn, J., Velloso, E., Allison, F., Abdelrahman, Y., & Vetere, F. (2017). Evaluating Real-Time Gaze Representations to Infer Intentions in Competitive Turn-Based Strategy Games. *Proceedings of the 2017 Annual Symposium on Computer-Human Interaction in Play*, 541–552. https://doi.org/10.1145/3116595.3116624

Carter, M., & Allison, F. (2019). Guilt in DayZ. In K. Jørgensen & F. Karlsen (Eds.), *Transgression in Games and Play*. MIT Press.

Newn, J., Allison, F., Velloso, E., & Vetere, F. (2018). Looks Can Be Deceiving: Using Gaze Visualisation to Predict and Mislead Opponents in Strategic Gameplay. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3173574.3173835

Newn, J., Singh, R., Allison, F., Madumal, P., Velloso, E., & Vetere, F. (2019). Designing Interactions with Intention-Aware Gaze-Enabled Artificial Agents. In D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, & P. Zaphiris (Eds.), *Human-Computer Interaction – INTERACT 2019* (pp. 255–281). Springer International Publishing. https://doi.org/10.1007/978-3-030-29384-0_17

Acknowledgement of funding

This thesis was financially supported by an Australian Government Research Training Program Scholarship, top-up scholarships from the Microsoft Research Centre for SocialNUI and the Melbourne Networked Society Institute (both at The University of Melbourne) and a paid research internship at Microsoft Research Cambridge, UK. Thank you to everyone who facilitated this support.

Acknowledgements

I am enormously grateful to an enormous number of people. This thesis would not exist without them.

To my supervisors. Martin Gibbs, whose wisdom, patience and ability spot a bad idea were indispensable. Marcus Carter, who always let me know which questions I *should* be asking, both in research and in academic life. Wally Smith, who improved my thinking and writing every time I thought I had already got it right.

To everyone at the Interaction Design Lab at The University of Melbourne. Frank Vetere and Vassilis Kostakos, for fostering a culture of high standards and mutual respect. Nicole Barbee and Kat Franks, for being there to help. Eduardo Velloso, Jenny Waycott, Greg Wadley, Ryan Kelly, Zaher Joukhadar, Steven Baker, John Downs, Bernd Ploderer and Niels Wouters, for the guidance of good example.

To my PhD cohort. Jane Mavoa, Sarah Webber, Dana McKay, Ronny Andrade, Romina Carrasco, Fernando Estrada, Deepti Aggarwal, Kayla Heffernan, Hasan Shahid Ferdous, Kagonya Awori, Behnaz Rostami Yeganeh, Patrick Pang, Zhanna Sarsenbayeva, Ahed Aladwan, Unni Krishnan, Lucy Sparrow, David Cumming, Madeleine Antonellos, Brian McKitrick, Toni McLennan, Niels van Berkel, Ikee Dimaguila, Abdulaziz Murad, Danula Hettiachchi, Xin Zhang, Nick Smith, Ebrahim Babaei, Priyankar Bhattacharjee, Yousef Kowsar and Martin Reinoso, who provided aid and amity in many forms.

Special thanks to Joshua Newn, whose assistance with the research was immensely helpful, and to Melissa Rogerson, who worked tirelessly to cultivate an inclusive and supportive local community of games researchers.

To Katja Hofmann and Ewa Luger at Microsoft Research Cambridge, from whom I learned a great deal in a short time.

To Robbie Fordyce, Daniel Golding, Tom Apperley and Brendan Keogh, whose advice was indispensable.

To my sister, Jess, for being there and lending a hand.

To my mum, Wendy, for the long talks, for the moral support and above all for helping to organise my wedding while I was distracted by the PhD.

To my wife, Ashlee Riordan, whose critical thinking, personal integrity and dedication to doing right by other people are unwavering.

For Rose Riordan Allison. You'll understand this when you're older.

Table of Contents

List of Tables

List of Third Party Copyright Material

Citation information for third party copyright material	Location of item in thesis	Permission granted Y/N
Allison, F., Luger, E., & Hofmann, K. (2018). How Players Speak to an Intelligent Game Character Using Natural Language Messages. *Transactions of the Digital Games Research Association*, 4(2), 1–50. https://doi.org/10.26503/todigra.v4i2.88	Section 4.5	Y[2]
Allison, F., Carter, M., & Gibbs, M. (2017). Word Play: A History of Voice Interaction in Digital Games. *Games and Culture*, 1555412017746305. https://doi.org/10.1177/1555412017746305	Section 5.5	Y[3]
Allison, F., Carter, M., Gibbs, M., & Smith, W. (2018). Design Patterns for Voice Interaction in Games. *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, 5–17. https://doi.org/10.1145/3242671.3242712	Section 6.6	Y[4]
Allison, F., Newn, J., Smith, W., Carter, M., & Gibbs, M. (2019). Frame Analysis of Voice Interaction Gameplay. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 393:1–393:14. https://doi.org/10.1145/3290605.3300623	Section 7.5	Y[4]
Image from: Berger, C. (1916). *Sound-operated circuit-controller* (Patent No. US1209636 A).	Section 5.5, Figure 1	N
Image from: RDI Video Systems. (n.d.). *HALCYON*. RDI Video Systems [Pamphlet].	Section 5.5, Figure 2	N
Image from: Vivarium. (1999). *Seaman*. Sega [Videogame].	Section 5.5, Figure 3	N
Image from: Harmonix. (2003). *Karaoke Revolution*. Konami [Videogame].	Section 5.5, Figure 4	N
Image from: Chrysostomou, V. (2016). *Plan Be* [Videogame].	Section 5.5, Figure 5	N
Image from: Marcelionis, A. (2013). *The Howler* [Videogame].	Section 7.5, Figure 1	N
Image from: Looking West Apps. (2013). *ATC Voice*. Looking West Apps [Videogame].	Section 7.5, Figure 1	N
Image from: Ubisoft Shanghai. (2008). *Tom Clancy's EndWar*. Ubisoft [Videogame].	Section 7.5, Figure 1	N

The fair dealing provisions of Australian copyright law allow limited use of some copyright material without requiring permission from the copyright owner.

[2] http://todigra.org/index.php/todigra/about/submissions#copyrightNotice
[3] https://au.sagepub.com/en-gb/oce/journal-author-archiving-policies-and-re-use
[4] https://www.acm.org/publications/policies/copyright-policy#permanent%20rights

1 Introduction

Voice-controlled technology has long been a signifier of the future. For decades, science fiction stories have been replete with talking computer systems, whose ability to speak intelligently to humans is a sign of their technological sophistication. Real-world voice interaction systems are often compared to these fictional counterparts, but the user experiences they create typically bear little relation. Users frequently report that they find voice interaction embarrassing, or that it makes the computer seem stupid rather than intelligent (Budiu & Laubheimer, 2018b, 2018a; Jentsch et al., 2019; Luger & Sellen, 2016; Murad & Munteanu, 2019; Rico & Brewster, 2010). This is only partly due to technical shortcomings. As this thesis will show, voice interaction faces issues of human social psychology that science fiction has largely overlooked.

A scene from the 2013 film *Her* (Jonze, 2013) stands out as an unusually apt depiction of voice interaction. The film is set in a near future in which voice interfaces have become the standard format for interacting with digital technology. In the scene, the film's protagonist, Theodore Twombly, is playing a holographic videogame by using hand gestures to control the movements of a virtual astronaut on a strange world. Suddenly, the astronaut is knocked backwards by a small, cute alien, which looks up at the astronaut suspiciously. Twombly zooms in to a first-person perspective, so that he and the alien are looking directly at one another, and asks it politely for assistance. The alien responds with a barrage of obscene insults in a childlike voice. Twombly is briefly nonplussed, but figures out that he can win the alien's cooperation by entertaining it with an equally foul-mouthed response. Then, the voice of Twombly's computer announces that he has a new email. Twombly and the computer discuss the email, while the alien interrupts with insulting remarks about their conversation.

Despite the futuristic technology, this scene captures several characteristics that resonate with the present-day reality of voice interaction. First, it depicts voice interaction as a meaningful social experience. The alien represents a gameplay challenge, but the nature of that challenge is social: to move the game forward, Twombly must understand what the alien is communicating and respect its point of view. Second, it shows voice interaction collapsing the boundary between the virtual world and the everyday social world. The distance between Twombly and his avatar literally disappears, as he shifts from controlling it from a third-person perspective to embodying it and speaking as its voice from a first-person perspective. Additionally, the game and the email exchange bleed into each other because they are conducted through sound; there is nothing to stop the alien from hearing the computer speak or *vice versa*. Finally, the scene portrays voice interaction as absurd and playful rather than serious and sophisticated. Even though Twombly's videogame is a technological marvel, the sight of a grown man cajoling a recalcitrant virtual character is inherently comical, perhaps even demeaning. This depiction of voice interaction as a social, liminal and playful activity reflects the characteristics of voice interaction that this thesis will explore.

In a time when voice interfaces, virtual characters and virtual worlds are all becoming more common and more intertwined, the thesis sets out to understand the intersection of these technologies by exploring their use in game design. It surveys four decades of videogame history to discover the trends that have shaped voice game design, and the recognisable patterns that have emerged in the field across that time. It develops a theory of voice gameplay as a complex social event with paradoxical effects, which can simultaneously heighten a player's feeling of involvement in the virtual world and increase their susceptibility to frustration, embarrassment and confusion. Ultimately, the thesis identifies and articulates the source of the tensions that run through voice gameplay, and provides a map of the design patterns and approaches that can be used to resolve them.

1.1 Background

In the past decade, voice user interfaces (VUIs) have become a prominent feature of many off-the-shelf consumer technologies. Voice assistants, such as Apple's Siri and Amazon's Alexa, have taken an increasingly central place in computer operating systems and smart home devices. These voice assistants are presented not as sets of menus and icons but as virtual characters, with humanlike personalities and names (or at least human job titles, as in the case of Google Assistant). Prior to the rise of VUIs, talking virtual characters appeared most commonly in videogames. Voice assistant design has been directly inspired by videogame character design, explicitly so in the case of Microsoft's Cortana, which bears the name and voice of the artificial intelligence (AI) assistant from the *Halo* game series (Warren, 2014). This persona implies that the experience of talking to Cortana, the voice assistant, is analogous to the experience of talking to Cortana, the futuristic AI game character.

Voice interaction is an increasingly common feature of videogames. This is not an entirely new development: voice-operated videogames (*voice games*) have existed since the 1970s, and include some of the most novel, extensive and popular examples of voice interaction with virtual characters. However, as Chapter 5 shows, their popularity has historically been constrained by the limited availability of hardware and software platforms that support voice input for games. The rise of VUIs over the past decade has eroded these constraints.

Between general-purpose computing devices like the smartphone and dedicated game consoles like the Xbox Kinect, it is now normal for a consumer in a Western country to own multiple platforms that are equipped for voice interaction gameplay.

However, the proliferation of voice games appears to be hampered by a sense that voice interaction is often more frustrating or awkward than fun. Early in my candidature, I co-authored a study that analysed online reviews and responses to voice games using grounded theory method. In this study, we found positive comments about the effectiveness and immersiveness of voice commands in some voice games, mixed in with numerous complaints that voice interaction felt unnatural, uncomfortable or embarrassing in others (Carter et al., 2015). Little other research has been conducted to explain or address these kinds of player experience issues in voice games, as Chapter 2 will show.

The paucity of research on voice interaction extends beyond games. Shortly before I began my candidature, multiple scholars had publicly lamented the neglect of voice interaction research in human-computer interaction (HCI) in general (Aylett et al., 2014; Munteanu et al., 2013). Although the rate of VUI studies has increased in the intervening years, a recent review found that the literature is still "somewhat fragmented" and lacking a "critical mass" (Clark et al., 2019, p. 16) in all but a few subtopics. HCI researchers have focused on developing and evaluating novel voice interfaces, and testing how users respond to different kinds of speech from a computer. These prior studies lay groundwork for a general understanding of voice interaction, but leave out many of the complexities of videogame play, such as the imaginative dimension of the fictional gameworld that may be populated by characters with diverse relationships to the player.

The field of game studies provides more extensive and multi-layered accounts of the complexities of games and gameplay. In particular, game studies literature emphasises the importance of understanding videogames not only as computer systems with interfaces, but also as *gameworlds*—imagined living places populated with characters, cultures and environments, into which players project some sense of themselves as an avatar or in-game persona (Bayliss, 2007; Jørgensen, 2009)—and as cultural products, situated within various socio-historical contexts and discourses about game design and the social conventions of gameplay. These are crucial factors in understanding how players experience a particular game, as this thesis will attest. However, only a handful of game studies researchers have examined voice interaction (see Chapter 2), and as a result there are fundamental questions about voice interaction in games that have gone largely unexplored.

1.2 Defining VOICE gameplay

So far, this chapter has described a distinctive type of voice interaction that has been growing in prominence, both in videogames and in other consumer technology. There is a confusing array of possible terms that can be used to refer to various elements of this type of interaction. *Automatic speech recognition* and *natural language understanding* describe the back-end speech processing systems; *speech interface* and *voice user interface* describe elements of the system that the user sees and hears; *voice command device* describes the hardware that the user operates; *conversational agent* and *voice assistant* describe the virtual character that speaks to the user; *speech command* and *voice interaction* describe the activity that the user engages in; and so on. Each of these terms captures one aspect of the phenomenon, but none are adequate as an identifier for the whole. Terms that describe this form of voice interaction through reference to games—such as *voice-controlled videogame* and *voice interaction gameplay*—obscure the fact that it increasingly exists outside of games. A more specific concept is needed to clarify what the thesis is about.

In this thesis, I refer to the central phenomenon of concern as *voice-operated interaction with characters and environments* (VOICE). The term VOICE describes any activity in which a user uses their voice to interact with a simulated virtual characters or environments. This definition is intentionally broad, but incorporates several crucial distinctions that I will now discuss, both to clarify the terminology and to indicate the boundaries of what the thesis will address.

The first half of VOICE—"*voice-operated interaction*"—restricts its scope to the act of operating a computer system with one's voice. That is to say, it excludes vocal activities that may be performed with a computer system, but do not operate the system itself directly, such as player-to-player voice chat, a common feature of online games (Wadley et al., 2015). Within this focus on operating the computer system, the concern is with any kind of voice behaviour. The term "voice-operated interaction" is intentionally agnostic about the form that the vocal action takes, encompassing any kind of vocalisation: speaking, humming, whistling, commanding, questioning, exhaling. It even allows for voice interaction based on unintended vocalisations.[5] This is in contrast to alternative terms for voice input, which typically restrict the meaning to a certain type of vocalisation.

[5] As an illustrative example, the stealth game *Alien: Isolation* uses unintended vocalisations in its "noise detection" mode. When this mode is enabled, in-game enemies respond to any sound that the game's microphone detects as though it came from the player-character. This forces the player to play quietly, as a cough or laugh can cause their character to be killed in the game.

For instance, “voice command” specifies the use of spoken words in the form of a direct instruction.

The second half of VOICE—“*with characters and environments*”—distinguishes it from the general notion of voice interaction with computers. As noted previously in this chapter, virtual characters are a central feature of many voice user interfaces. Prior HCI research has suggested that this is because humans are “voice activated” (Nass & Brave, 2005), in the sense that they tend to perceive machines that speak as humanlike, and respond to them in similar ways that they respond to people. Videogames, of course, have always been saturated with talking characters, from the simple enemy aliens of *Space Invaders* to the complex romantic companions of *Mass Effect*. Based on the research conducted for this thesis, I posit that these types of characters are central to how users make sense of voice interaction, which is why the concept of VOICE must be defined in relation to virtual characters, rather than treating them as epiphenomena. Additionally, a central finding of this thesis is that voice interaction does not only give the talking character a greater sense of life—it also gives the player a greater sense of being a participant in the gameworld. This extends to voice interactions that do not involve a virtual interlocutor, such as when a player can use their voice to cast spells or blow out candles in the gameworld. To recognise this, the term VOICE incorporates both virtual characters and virtual environments.

To summarise, the term VOICE encapsulates both the literal activity involved in voice interaction and the imaginary framing of that activity as an interaction between the player and a fictive living world populated by characters. Throughout this thesis, wherever the terms “voice games”, “voice game design” or “voice gameplay” are used, they refer to this conception of VOICE.

1.3 Aim and scope

The aim of this thesis is to provide a new understanding of VOICE in the context of videogame design. It presents four studies, two of which examine the player experience of games that use voice or natural language input, and two of which examine the history and design or previously published games. Each of these studies addresses a gap in the literature on voice interaction in games, and contributes to the growing body of research on the wider topic of voice interaction design. The aim of the thesis is therefore summarised in its overarching research question:

> **Main RQ** How does voice-operated interaction with characters and environments (VOICE) influence the player experience of digital games?

Over the course of this thesis, I set out to understand voice games as an area of interaction design for which I could find little prior research. I sought to develop an account of voice game design that encompasses all of its major variations—a map of the territory in its entirety, rather than a more detailed depiction of a single one of its enclaves. Consequently, the scope of the thesis is in some respects quite broad and open-ended: each study takes an exploratory approach, and uses inductive analysis methods to allow for unexpected insights to emerge. In fact, the initial scope of the research included text-based interaction as well as voice-based interaction, which is reflected in the design of Study 1. However, from Study 2 onwards, the thesis is firmly grounded in the concept of VOICE in games. The threads of the research are woven together in Chapter 8, which presents a unified discussion of voice game design and gameplay that draws upon all four of the studies herein.

With the exception of Study 1, which concerns text-based interaction, the thesis is solely concerned with the phenomenon of VOICE in videogames, as a question of game design and player experience. It refers to the underlying technologies that enable speech processing only where they are directly relevant to the player experience. It does not address the adjacent topics of voice communication or sound design in videogames, except insofar as they relate to voice game design and gameplay.[6] It also does not focus on "audiogames": digital games that provide auditory feedback but not visual feedback, often designed for players with visual impairments (Friberg & Gärdenfors, 2004; Kirke, 2018). The number of audiogames that use voice interaction has skyrocketed in the years since I began the thesis. In part, this can be attributed to the proliferation of smart speaker systems in this time period. These voice-operated audiogames are outside the main scope of the thesis, although a selection of popular voice-operated audiogames are examined in Chapter 6.

In summary, this thesis introduces the concept of voice-operated interaction with characters and environments (VOICE), and presents the most comprehensive body of research to date on VOICE in videogames. The aim is to explore how players experience VOICE in videogames, and how these experiences are influenced by the variety of ways in which videogames utilise VOICE in their design. The following section provides an overview of the thesis that explains its major contributions.

[6] Although I do not discuss them at length in this thesis, the doctoral dissertations of Jørgensen (2007) and Wadley (2011) were useful to me as introductions to the topics of videogame sound design and voice communication in online games, respectively.

1.4 Overview of the thesis

This section provides an outline of each chapter of the thesis. Chapters 1–3 introduce the research, review prior literature on the topic and explain the research methodology. Chapters 4–7 report the results of four studies, each of which addresses one or more ancillary research questions related to the overarching topic. Each of the results chapters consists primarily of a verbatim copy of a peer-reviewed academic article, for which I am the primary author, in accordance with The University of Melbourne guidelines for a *Thesis with Publication*.[7] Supporting documentation for these articles is included in Appendix A. Chapter 8 presents a discussion that draws upon all four studies to synthesise the major contributions of the thesis.

Chapter 2—Literature Review

Chapter 2 reviews prior work from HCI and game studies. First, it defines several key concepts that relate to the topic of VOICE. Second, it summarises the technical history of the development of speech technology. Third, it reviews the literature on voice user interfaces, focusing on three topics: embodied conversational agents, computers as social actors and voice assistants. Fourth, it discusses the literature that relates to videogames and voice interaction, beginning with a review of efforts to develop voice games for accessibility, language learning, speech therapy and non-verbal voice interaction. It examines the small number of studies on the player experience of voice games, and identifies a research gap in relation to this topic. Finally, it presents the research questions that the thesis aims to address.

Chapter 3—Methodology

Chapter 3 explains the methodological considerations that informed the research design. It first situates the thesis within the fields of HCI and game studies, the epistemological paradigm of constructionism and the theoretical perspective of interpretivism. It then presents the array of qualitative methods used in the thesis, and discusses their methodological considerations. The chapter concludes with a summary of the four studies in turn, which describes how the methods used in each case addressed the research questions.

Chapter 4—Natural Language Gameplay (Study 1)

Chapter 4 comprises a participant observation and interview study with players aged 11–15 who interacted with a videogame character through natural language text. It includes an article that was published in *Transactions of the Digital Games*

[7] The University of Melbourne. (2019). *Graduate Research Training Policy (MPF1321)*. Retrieved from https://policy.unimelb.edu.au/MPF1321

Research Association (Allison, Luger, et al., 2018). The study addresses Research Question 1:

> **RQ1** How do players make sense of and interact with a game character that responds to natural language text?

The chapter explains the research method, which involved the use of a Wizard of Oz prototype to simulate a videogame character that responded appropriately to participants' natural language messages via a text chat channel in the game. Based on a content analysis of the message logs and a thematic analysis of the post-game interviews, the study makes three main contributions to the HCI and game studies literature:

- A detailed account of relatively spontaneous player approaches to natural language interaction in a game, indicating some of the things that players might do with and expect from natural language gameplay when they have not been given detailed instructions.
- A thematic analysis of difficulties and concerns that arose from attempting to interact with a game character through natural language, based on both gameplay observations and post-game interviews.
- A hypothesis about two key variables that differentiated player reactions to a "talking" game character, namely the degree to which they engaged with the character and the degree to which they anthropomorphised the character.

Study 1 is distinctive in this thesis, as it concerns text-based gameplay rather than voice gameplay. At the time when the study was conducted, I had defined the scope of the thesis in terms of "conversational interaction" in games, encompassing both text and voice interaction. Chapter 4 includes an introductory discussion that situates Study 1 in relation to the rest of the thesis and reviews its contribution towards the overall research question.

Chapter 5—Voice Game History (Study 2)

Chapter 5 presents a history of VOICE in videogames, from experimental research projects in the 1970s to popular commercial entertainment products in the 2010s. It includes an article that was published in *Games and Culture* (Allison et al., 2017). The study addresses Research Question 2:

> **RQ2** What is the historical lineage of voice game design and what are the major influences that have shaped it?

The chapter introduces the study as a work of media archaeology (Parikka, 2012). It describes the research method, which involved an extensive review of primary and secondary sources to construct a database of 247 videogames published between 1973 and 2016, accounting for the vast majority of commercially published videogames from that time period that featured any form of voice interaction. Based on an analysis of these sources, the study puts forward three main findings:

- There have been seven distinct phases of voice game design, distinguished by their time period, design focus and geographic centre of activity.
- Voice game design has been highly reactive to changes in the hardware and software platforms that became popular for videogames in each time period.
- There have been systematic differences in the way that voice games from Japan and the West present the player-character relationship.

Chapter 6—Voice Game Design Patterns (Study 3)

Chapter 6 presents the development of a pattern language for voice game design. It includes an article that was published in the *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Allison, Carter, et al., 2018). The study addresses Research Question 3:

> **RQ3** What design patterns are currently implemented in voice games?

This chapter builds upon the work in the previous chapter by surveying the design of 471 videogames that afford voice input. This represents a comprehensive and systematic approach to compiling a complete census of published voice games, including those from academic, commercial and hobbyist developers. Based on a range of primary and secondary sources, I recorded a description of each game's voice interaction mechanisms, as well as information about its development, format and presentation. Through a process of iterative, inductive coding, I developed a set of 25 design patterns that represent the main repeated variations in voice game design. The chapter also contains a further analysis that identifies five core genres of voice gameplay, each of which represents a different kind of voice-based activity.

Chapter 7—Voice Gameplay (Study 4)

Chapter 7 presents an investigation into the player experience of voice gameplay, which uses Erving Goffman's frame analysis (Goffman, 1974) as an interpretive framework. It includes an article that was published in the *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Allison et al., 2019). The study addresses Research Question 4 and Research Question 5:

> **RQ4** What is the distinctive nature of voice gameplay?

> **RQ5** How do players experience verbal and non-verbal voice gameplay?

This chapter describes a participant observation and interview study, in which participants played a selection of videogames alternately using voice and traditional manual controls. The games represented different forms of voice gameplay, including both verbal (speech-recognition-based) and non-verbal (non-speech-recognition-based) voice interaction. The study empirically validates an existing theoretical model of gameplay events (Conway & Trevillian, 2015), and develops it with an additional level of detail based on player comments. Finally, the chapter presents seven qualitative themes that describe the player experiences of verbal and non-verbal voice gameplay, based on a thematic analysis of the participant interviews informed by the theoretical perspective of frame analysis (Goffman, 1974).

Chapter 8—Discussion

The final chapter draws together findings from all four studies and discusses how they have contributed towards addressing each of the research questions. The two main contributions of the thesis are a theoretical model of voice gameplay based in frame analysis and a typology of voice game genres grounded in design patterns. Due to the constraints of a *Thesis with Publication* structure, the previous chapters were limited in their ability to discuss these contributions in the context of the thesis as a whole; consequently, Chapter 8 provides a holistic discussion of both the model and the typology that draws upon the findings of all four studies. Following this discussion, the chapter considers the implications of the thesis for smart speaker games, voice assistants and virtual/augmented reality experiences. Next, it acknowledges the limitations of the research design, and reflects upon the value of including historical analysis in HCI research. The thesis ends with a discussion of opportunities to extend this research into the domains of virtual and augmented reality systems and smart speakers.

2 Literature Review

The preceding chapter introduced the topic of voice-operated interaction with characters and environments (VOICE) and described how an idealised version of VOICE has long been presented as a desirable goal for the future of human-computer interaction, particularly in the context of videogames. This chapter reviews prior work from the research fields of human-computer interaction (HCI) and game studies on several topics relating to voice game design, to set the foundation for the research in the following chapters. First, it defines several of the key terms and concepts on which the thesis is based (Section 2.1). Second, it presents a brief account of the history of speech technology, to explain how its current capabilities differ from those of past decades (Section 2.2). Third, it provides a synthesis of the most relevant research on voice user interfaces, drawing upon research related to embodied conversational agents, computers as social actors and voice assistants (Section 2.3). Fourth, it summarises the research that has been conducted on voice game design and gameplay, and evaluates prior work on the role of game characters as social actors that shape the player experience (Section 2.4). Finally, it identifies gaps in the literature that pertain to the history, design and player experience of voice games, and presents the research questions that have been formulated to address these gaps (Section 2.5).

As a *Thesis with Publication*,[8] this thesis contains a full peer-reviewed published article in Chapters 4, 5, 6 and 7. Each of these articles contains a review of prior work that relates to its specific area of that study. This chapter presents a broader overview of the literature that provides the foundation and context for the thesis as a whole.

2.1 Definitions and scope

Voice user interface (VUI) denotes an interface through which a user may interact with a machine system by using their voice as a form of input. The term is sometimes used to refer specifically to interfaces that employ automatic speech recognition and respond to spoken language (e.g. Cohen et al., 2004, pp. 5–6). However, in this thesis "VUI" includes any interface that responds to any aspect of a user's voice, including *verbal voice input* (spoken language) and *non-verbal voice input* (Igarashi & Hughes, 2001), such as pitch and volume.

Voice interaction refers to the act of a user operating a system through a VUI. Following Hornbæk and Oulasvirta's definition of *interaction* (2017, p. 5049), it includes the user's responses to the system, the system's responses to the user,

[8] The University of Melbourne. (2019). *Graduate Research Training Policy (MPF1321)*. Retrieved from https://policy.unimelb.edu.au/MPF1321

and the ways in which the two entities determine each other's behaviour over time. As with VUI, "voice interaction" refers to either verbal or non-verbal voice input, whereas "speech interaction" refers specifically to verbal voice input. Voice interaction is distinct from *voice communication* and *voice chat*, which in the context of this thesis refer to players in a videogame or virtual world conversing with each other through a voice-over-internet-protocol system, rather than using their voice to control the game (Wadley et al., 2015; Wadley & Gibbs, 2010). Voice communication and voice chat are out of scope for this thesis.

The particular type of voice interaction with which this thesis is concerned is *voice-operated interaction with characters and environments* (VOICE), defined in Chapter 1. That is to say, the research does not look at the types of voice interaction that a user engages in when dictating a manuscript, when commanding a television to "pause" and "play", or when being passively monitored by a system that records their affect. It looks specifically at voice interaction in which the player's voice affects a virtual character or a gameworld.

Gameworld denotes the imaginary universe that a videogame depicts through visual and audio representations.[9] Jørgensen defines gameworlds as "world representations designed with a particular gameplay in mind and characterized by game-system information that enables meaningful player interaction" (2013, p. 3). A gameworld is simultaneously an imaginary world and an interface to a system governed by the logic of its game mechanics, as Jørgensen notes: "gameworlds are presented in part as similar to other world systems, such as the physical and fictional worlds, and in part as a system for human–computer interaction that does not have any obligation to fictional coherence" (2013, p. 4). Gameworlds are the ideal context for studying VOICE, as they are the most common and widely familiar examples of complex virtual environments populated by interactive virtual characters.

Game sound design merits a brief mention, as a topic that is closely related to voice game design but is outside the scope of this thesis. Sound design is recognised as an important part of human-computer interaction, with different sounds having widely varying effects on a user's recognition, understanding and emotional response in an interaction (Dingler et al., 2008; Fernström, 2003; Franinović & Serafin, 2013). Videogame sounds have a substantial impact on the player's experience, and particularly the player's emotional reaction to events and characters in the gameworld (Collins, 2013, 2011; Jørgensen, 2007). The way an avatar makes sounds in response to the player's actions is a particularly important element of sound design, as it can strengthen the player's feeling of connection to their avatar: "players have a direct, embodied interaction with the sounds that they evoke and hear in games, and coupled with physical or kinesonic-congruent

[9] Usually, although there are videogames with audio-only interfaces (Friberg & Gärdenfors, 2004; Röber & Masuch, 2005), tactile interfaces (Parisi, 2008, 2013) and other such variations.

action, these sounds (and thus the game character) can become an extension of the self" (Collins, 2013, p. 44). Sound design may be of heightened importance in voice gameplay, since voice games raise the potential for sound output to conflict, overlap or be incongruent with sound input. However, the sound design of voice games has not been explored at length in any prior literature, to the best of my knowledge, and due to time and space limitations this thesis includes only limited discussion of this aspect of voice game design.

2.2 A brief history of speech technology

Building a machine to recognise human speech is a pursuit that predates digital computers. As early as 1916, children's toys with names like Radio Rex used a mechanical form of proto-speech-recognition to respond to sounds in the human vocal range (Cohen et al., 2004, p. 15; Gernsback, 1916, p. 104; 'Toys that obey your voice', 1916, p. 718). In 1952, Bell Labs created a system that could recognise the sound of ten digits, when spoken at a set pace by a voice for which the machine had been pre-configured (Davis et al., 1952). Over the next two decades, researchers in Japan, UK, USA and the Soviet Union developed new techniques to grow the vocabulary of speech recognition systems, still using the approach of matching individual words against stored sound patterns (Moskvitch, 2017).

Progress received a boost in 1971, when the US Defence Advanced Research Projects Agency (DARPA) funded a five-year program of speech recognition research across multiple labs (Lea & Shoup, 1979; Rabiner & Juang, 2008). Two of the most successful projects to emerge from this venture were developed at Carnegie Mellon University: Hearsay and Harpy. One version of Hearsay could recognise standard Chess instructions such as "Bishop to Queen Three" and update an on-screen Chess board accordingly (Reddy et al., 1973); it is the earliest documented example of a voice-operated digital game that I have found in my research. Harpy was the first speech recognition system to successfully use a language model to recognise multiple words in a sentence (Lea & Shoup, 1979; Moskvitch, 2017).

A particularly important contribution for speech recognition was the development of hidden Markov models, a statistical technique to infer the probability of hidden information based on observable information (Baum et al., 1970). In 1988, Lee combined a hidden Markov model approach with elements of Harpy to create Sphinx, "the world's first accurate large-vocabulary speaker-independent continuous speech recognition system" (K.-F. Lee, 1988, p. 375). Hidden Markov models remain at the core of speech recognition technology today, and updated versions of Sphinx continue to be among the most widely used speech recognition systems, including in some videogames (see Chapter 5).

The next big leap came in the 2010s. A combination of increased computing power, better optimised computing infrastructure, greatly increased datasets and breakthroughs in machine learning techniques powered a rapid improvement in the error rates and fluency of speech recognition systems, including in noisy environments (Huang et al., 2014; J. Li et al., 2014). Much of this recent progress has been driven by large software companies, such as Apple, Microsoft, Google and Amazon, seeking to put forward so-called "intelligent personal assistants" with which their users can engage in conversation. Whereas previous waves of progress in speech recognition had introduced the experience of speaking to an automated call centre menu or dictating words to a computer, this new wave of progress introduced the experience of speaking to virtual characters as a regular occurence in everyday life. It is this experience—using one's voice to interact with a virtual character or virtual environment—that is the focus of this thesis, albeit in the specific domain of videogames.

2.3 Voice user interface research

While research on speech recognition technology has been extensive since the 1950s, several researchers have noted that the user side of speech interaction has received comparatively little investigation (Aylett et al., 2014; Clark et al., 2019; Munteanu et al., 2013; Murad et al., 2019). Murad et al. (2019) argue that this has contributed to a lack of design guidelines that are specific to speech interaction, leading to a situation in which voice user interfaces are typically designed according to guidelines that were intended for graphical user interfaces. Non-verbal voice interaction, it must be noted, has received even less attention than speech interaction (Poláček, 2014, p. 17).

A recent review of the HCI literature found a total of 99 peer-reviewed research papers, all published since 1990, that focused on spoken interaction and included some kind of empirical evaluation with users (Clark et al., 2019). Of these, 22 concerned synthesised speech as an output from the system, rather than user speech as an input to the system, and thus do not involve "voice interaction" as defined in Section 2.1. The authors noted that a great deal of the design research focused on narrow application areas and specific interface components, whereas more generalisable design guidelines were absent: "there is still a need to establish design considerations and robust heuristics for developing user centred speech interactions" (Clark et al., 2019, p. 17). In addition, they noted that few studies have looked at speech interaction in situations that involve more than one user—a scenario that this thesis examines in Chapter 7.

2.3.1 Embodied conversational agents

Much of the research on voice interaction centres on the role of conversational agents, and particularly embodied conversational agents. A conversational agentis a computer system that converses with the user through language, while an embodied conversational agent is such a system that features a visual depiction of a character to represents the agent that is speaking. Justine Cassell, who conducted some of the most significant early research on this topic, defines an embodied conversational agent as: “an interface in which the system is represented as a person, information is conveyed to human users by multiple modalities such as voice and hand gestures, and the internal representation is modality independent and both propositional and nonpropositional” (2001, p. 67). This definition encompasses social robots as well as characters that are animated on a screen. A videogame character that converses with a player and expresses itself with both bodily and vocal actions—such as Pikachu in *Hey You, Pikachu!*—is an example of an embodied conversational agent.

As Cassell describes, the visible body of an embodied conversational agent gives it a greater ability to complement “propositional” information (the content of the conversation itself) with “interactional” information: signalling understanding, confusion, wish to speak, readiness to listen and other such states through the use of bodily cues such as gaze behaviour and gestures (Cassell, 2000, pp. 74–75; Cassell et al., 2001, pp. 57–58). Embodied conversational agents can express a sense of personality through non-verbal actions, such as demonstrating extroversion by making larger gestures and signalling greater readiness to take turns in conversation. Sproull et al. (1996) found that users judged the personality of an embodied conversational agent with an animated face differently from that of an equivalent conversational agent with no animated face. Users in that study also sought to present themselves in a more positive light and reported feeling less relaxed and confident when interacting with the embodied conversational agent—a set of behaviours that suggests a heightened perception of being under social scrutiny. Chen et al. (2010) found that users preferred an extroverted embodied conversational agent to an introverted one. However, it is unclear whether this finding represents a general preference for extroversion in embodied conversational agents or an effect of something else, such as cue consistency. Isbister and Nass (2000) found that users preferred embodied conversational agents that exhibited consistent personality cues across their verbal and non-verbal behaviour.

Isbister and Doyle (2002) identified a tendency for embodied conversational agent studies to lack a clear indication of the specific design problem they are trying to solve, and therefore failing to make clear advances on well-defined problems. To address this, they presented a taxonomy of research related to embodied conversational agents that consists of five areas of enquiry, each with a different overall aim. Research on “appearance and behavioural believability” aims to understand how embodied conversational agents can present an illusion of life to the user. Research on “social interface techniques” aims to understand how embodied conversational agents should engage in social roles, including scenarios that involve multiple embodied conversational agents interacting with a user. Research on "application domains" aims to understand how embodied conversational agents can be of use to specific user groups and for specific contexts of use. Research on "agency and computational issues" aims to understand how to improve the underlying technical architecture that drives embodied conversational agents. Finally, research on "production" aims to understand how embodied conversational agents can be produced in a way that is highly consistent and lacking in disruptions. As a great deal of the research in this thesis concerns embodied conversational agents in the form of videogame characters, the thesis itself can be categorised as an example of embodied conversational agent research on the *application domain* of videogames. Believability and social interface techniques make up a secondary focus of the research, as these two aspects are often central to understanding the success or failure of embodied conversational agent interactions in videogames. Technical production and agency and computational issues are not the focus of the thesis.

2.3.2 Computer voices as social actors

Nass et al. (1994) proposed the idea that "computers are social actors" (CASA). By this, they meant that humans respond to other humans and to computers in similar ways, such that a generalised finding from social science about human behaviour or attitudes towards other humans could often be replicated if the other human was replaced by a computer that had characteristics associated with humans. This effect was dubbed the "media equation" (Reeves & Nass, 1996). Nass et al. (1994) speficied a procedure for testing the media equation, which involved providing a computer system with four "human" characteristics: language output, responses based on multiple prior inputs, responding in a human-sounding voice, and performing a role that is traditionally performed by a human (Nass et al., 1994, p. 72). Note that three of the four characteristics that were intended to make a computer function as a humanlike social agent are related to language and speech. A range of experiments have shown that users do indeed follow some of the norms of inter-human social behaviour when interacting with a computer, and particularly a conversational agent. Nass and Moon (2000) attribute these effects not to users being fooled into believing that computers are humanlike, but to users acting "mindlessly" according to familiar behavioural scripts—an interpretation that has been supported by at least one subsequent experiment (D. Johnson & Gardner, 2009).

Early studies in the CASA paradigm showed that users follow certain norms of politeness when speaking to computer agents. Nass et al. (1999) found that users provided a more positive evaluation of a conversational agent when interviewed directly by the agent than when filling out a separate questionnaire—an apparent example of social desirability bias (Fisher, 1993), as though the users were attempting to ingratiate themselves with the computer interviewer. The same effect was replicated by Hoffman et al. (2009) with an embodied conversational agent. Nass et al. (1994) found that users were more swayed by a computer agent's praise of another computer agent's performance on a tutoring task when the two agents spoke in different voices, indicating a perception that the praise was more reliable because it came from a different "person". This result was unaffected by whether the voices came from the same physical computer or two separate physical computers, indicating that voice overrode physical embodiment as an indicator of the computer agent's "identity". This is consistent with the theory, put forward by Nass and Brave (2005), that humans are especially inclined to respond to voices in technology as though they came from a human being.

Some studies have challenged the extent of the politeness effect. Goldstein et al. (2002) found that users gave lower evaluation scores for a handheld computing device when it was present during their evaluation than when it was absent, rather than higher scores as the CASA paradigm would have predicted, and used this finding to challenge the generalisability of the CASA politeness effect. However, the devices in their study were not provided with conversational agent features, such as language understanding or speech output, as Nass et al. (1994) prescribed, and this could account for the discrepancy. More recently, Lang et al. (2013) found that users did not provide more positive evaluations of an embodied conversational agent when responding to the agent directly. The authors speculated that perceptions of computers may have changed in the years that internet usage has become ubiquitous, such that users are less likely to perceive a computer program as an isolated social entity. Once again, however, Lang et al. (2013) did not provide their agent with voice output, and so an alternative interpretation of their results is that it is voice output that is the strongest driver of the CASA effect, as Nass and Brave (2005) assert. This interpretation is supported by Qvarfordt et al. (2003), who found that users' perception of a computer system as being humanlike increased as it provided a greater volume of spoken feedback.

Conversely, a number of studies have shown that users demonstrate rudeness and disinhibition towards conversational agents. In a multi-user interaction with an embodied conversational agent, Rehm (2008) found that users partly followed human-human behavioural norms when interacting with the agent, but also engaged in typically anti-social behaviour such as verbally insulting it and staring at it. De Angeli and Brahnam (2008) studied textual dialogues with conversational agents, and found that a great deal of the dialogue was abusive, pornographic or simply rude. When the conversational agents were embodied with different gender presentations, the female-presenting agent was subject to a higher proportion of swear words, sexual comments and negative stereotyping (Brahnam & De Angeli, 2012). Silvervarg et al. (2012) likewise found that a female-presenting embodied conversational agent received more verbal abuse than a male-presenting agent and an androgynous agent. These results are not inconsistent with the notion of computers as social actors. Rather, as De Angeli and Brahnam suggest (2008), they imply that users sometimes engage with conversational agents as social actors that are not protected by the ordinary boundaries of appropriate social behaviour; users draw some sense of social entertainment or novelty from abusing conversational agents, precisely because the users have the impression of the agents as social actors while they are simultaneously aware that the agents are not human beings whose feelings can be hurt.

The presence of gender stereotypes is one of the most common findings in the literature on CASA, although the direction and influence of these stereotypes is not always consistent. Nass et al. (1997) tested a conversational agent as a tutor and an evaluator of information with a human user, and found that when the agent was given a female voice as opposed to a male voice, both male and female users rated it lower in performance at either role, less friendly as an evaluator, more informative about the stereotypically feminine topic of love and relationships, and less informative about the stereotypically masculine topic of computers. Lee et al. (2000) found that a female-voiced conversational agent was perceived to be less socially attractive and trustworthy than a male-voiced conversational agent. Conversely, Linek et al. (2010) found that users preferred and learned more from a female-voiced conversational agent in an instructional task. Zanbaka et al. (2006) found that male users were more persuaded by a female-presenting conversational agent and vice versa. Consistent with the "mindlessness" explanation of CASA effects, Lee (2008) found that users attributed greater competence and conformed more to the instructions of a male-presenting conversational agent than a female-presenting conversational agent, but only when distracted by other tasks during the interaction. Interestingly, when the conversational agent was given a synthesised male or female voice, users rated its apparent gender as more salient, but responded to it in a less gender-stereotypical fashion—perhaps because speech output pulled the users' attention to the focal task more effectively than text output, and thus reduced the "mindlessness" of their responses. In response to the accumulated literature on gender-stereotyped responses to computers, several scholars have argued that the use of gendered voices in computer systems needs to be approached with much greater care and concern for the wider societal implications—in particular, the possibility that the proliferation of female-sounding voice assistants may reinforce gender stereotypes of women as polite and subservient—than has previously been the case (Brahnam & Weaver, 2015; Guzman, 2016; Phan, 2017; Woods, 2018).

Research from outside the field of HCI also supports the conclusion that people make various snap judgements about a speaker based on characteristics of their voice, although once again the effects are not always as simple as a universal preference for one type of voice over another. The gender effect is supported by studies such as that of O'Connor and Barclay (2017), which have found that people rate voices as more or less trustworthy according to a combination of factors including the gender and pitch of the voice and the context in which the trust would be granted (e.g. economic trust versus mating-related trust). McAleer et al. (2014) found that a large group of respondents gave consistent impressions of a speaker's personality based on recordings of the speaker saying "hello" that lasted for less than a second. Ryan and Sebastian (1980) found that middle-class Anglo-American students rated Spanish-accented English speakers less favourably than American-accented English speakers on a range of judgements when the speaker was introduced as coming from a lower-class background, but not when they were introduced as coming from a middle-class background. In a study that involved users playing a trust-based economic game with a conversational agent, Torre et al. (2018) found that conversational agents that acted in a generous fashion were trusted more if they had a Standard Southern British English accent than if they had a Liverpool or Birmingham accent, whereas conversational agents that acted in a less generous fashion were trusted more if they had one of the latter accents. This suggests that user ratings of an agent's personality are partly dependent on congruency between the agent's demonstrated behaviour and the expected behaviours that users associate with the sound of its voice. A meta-analysis of research on the effect of accents on interpersonal evaluations supported the general hypothesis that standard accents (those spoken by a majority of a population) are generally preferred to non-standard accents (those spoken by a minority of a population) in almost all instances (Fuertes et al., 2012). Taken together, the literature presents clear evidence that voice output is a strong driver of user perceptions that a computer system has a "character" or "personality", in combination with other features such as the system's actions and the context of use.

2.3.3 Voice assistants

In the past decade, the focus of HCI research on voice interaction has turned to commercial voice assistants, such as Amazon's Alexa, and the devices in which they are situated, such as the Amazon Echo smart speaker. Market research suggests that voice assistants have become highly popular despite a widespread sense that they have poor usability (Adobe, 2019; Budiu & Laubheimer, 2018b; Whitenton & Budiu, 2018). Studies on the use of voice assistants have shown that both usage and satisfaction tend to decline over time, as users become frustrated and unwilling to spend time learning the boundaries of what the assistant can and cannot do (Lopatovska et al., 2018; Luger & Sellen, 2016). This fits with a study by Knijnenburg and Willemsen (2016), which demonstrated that users tend to form more anthropomorphic mental models[10] of conversational agents that exhibit more humanlike characteristics. The authors argue that the failure of voice assistants to provide an enjoyable user experience lies in their overly humanlike appearance, which causes users to assume they have a degree of humanlike capabilities—an overestimation that leads inevitably to disappointment. Multiple HCI researchers have suggested that the "human metaphor" for voice assistants should be reconsidered in light of its tendency to

[10] For a discussion of mental models in human-computer interaction, see Norman (1983).

attract unrealistic attributions of intelligence or other human characteristics (Cowan et al., 2017; Culley & Madhavan, 2013; Knijnenburg & Willemsen, 2016; Luger & Sellen, 2016).

Perhaps because of their apparent human-ness, voice assistants have shown a notable tendency to be integrated into domestic social life. Smart speakers are most often placed in communal areas such as the home kitchen or living room, and their usage is interwoven with social activities such as meal times and discussions between family members (Lopatovska et al., 2018; Porcheron et al., 2018; Purington et al., 2017; Sciuto et al., 2018). A study of online reviews found that users who described using Alexa for more social interactions, including those with a higher number of members in the household, were more likely to talk about Alexa in personified terms (Purington et al., 2017). Collectively, these studies support the principle that "computers are social actors" (Nass et al., 1994) in the case of voice assistants, and that it is therefore important to study VUIs not only as interfaces but as characters.

2.4 Games research

The study of games has coalesced in the past two decades, from a pursuit that had been scattered across many academic fields and disciplines in a loosely connected fashion to a distinct field of research in its own right. In the early 2000s, major debates in the newly minted field of game studies centred on the boundary-work (Gieryn, 1983) of distinguishing "the game", and particularly "the videogame", as a unique phenomenon that merits dedicated study due to its inherent differences from other forms of media (Copier, 2003; Deterding, 2017). Game scholars were largely successful in this pursuit, establishing dedicated game studies institutions such as scholarly associations, conferences and journals, as well as helping to legitimise the study of games in related fields such as HCI and communication studies (Deterding, 2017, pp. 530–531). This success was no doubt aided by a broader societal shift, in which videogames and tabletop games became increasingly popular, socially accepted and culturally prominent across the world (Brand et al., 2019, p. 10; T. Cross, 2017; Duggan, 2015; Jolin, 2016; Keogh, 2016).

Multiple scholars have sought to compile a picture of the topics that games research has focused on over its history. Using a network analysis of keyword co-occurences in publication from core games research venues, Melcer et al. (2015) found a separation between technical games research (mostly related to artificial intelligence methods and topics) and non-technical games research (including human-computer interaction and related topics such as player experience). These findings were supported by Martin (2018), who conducted a wider network analysis of keyword co-occurences and author co-citations, and also analysed how the literature has changed over time. Martin found that the focus of games research has shifted from a heavy concentration on technology and media effects before the turn of the millennium to a wider range of topics in the new millennium, including games as cultural products, online gaming, game-based learning and medical applications of games, as well as a burst of interest in artificial intelligence in games. Coavoux et al. (2017) conducted a content analysis on three core game studies publication venues,[11] and noted a divide between humanities and social science approaches to the study of games. The major limitations that Coavoux et al. noted in the literature was an overrepresentation of a small number of videogames (notably *World of Warcraft*) and game genres (notably online gaming, first-person shooters and role-playing games), with a long tail of uncited games and “forgotten genres” that were rarely if ever studied despite being widely popular. It is to avoid such arbitrary narrowness of focus that the studies in Chapter 6 and Chapter 7 have been designed to include as many relevant videogames as possible in their analysis.

Games research has long been a subcategory within the field of HCI, which received a boost in prominence in 2014 with the founding of the ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play (CHI PLAY) conference series. Based on an analysis of HCI literature in the 10 years prior to this date, Carter et al. (2014) described four major research paradigms in which HCI games research had been conducted: operative, epistemological, ontological and practice. Operative research uses games as instruments to achieve some objective that is outside the normal purpose of game-playing, such as games that are designed to change the player’s habits or attitudes. Epistemological research uses games as research tools to generate insights into some other topic, such as using videogames as test-beds for novel interaction techniques. Studies in both of these paradigms, the authors argue, have tended to frame gameplay as an insufficiently valuable or worthwhile activity when it is not instrumentalised for an external purpose. In contrast, the ontological and practice paradigms of games research involve studies that seek to understand games and the interactions that emerge from gameplay (respectively) for their own sake.

Caroux et al. (2015) conducted a systematic review of HCI research to map the concepts that have been studied in relation to player-videogame interaction. This review was limited to studies that have presented empirical evidence about player-videogame interaction in entertainment situations. They described six main conceptual themes within this research: player engagement, player enjoyment, input and output techniques, game contents, multiplayer gaming, and global (i.e. holistic) approaches to studying player-videogame interaction. In relation to input techniques, Caroux et al. found empirical studies on motion-based interaction, tactile interaction, tangible interface interaction, gaze-based

[11] The three publication venues analysed in the study were the journals *Games & Culture* and *Game Studies* and the proceedings of the Digital Games Research Association (DiGRA) conference series.

interaction and brain-computer interaction. They did not find any empirical studies on voice interaction in videogames that met their criteria for inclusion.

There have been a significant number of past studies on voice interaction in videogames, as the following sections will describe. However, nearly all of these studies have focused on the development of a specific novel voice-controlled game or game interface. Most reflect either the operative paradigm of HCI games research (Carter et al., 2014, pp. 32–33), in that their stated intention is to develop a voice game that provides some kind of extrinsic benefit to the player, or the epistemological paradigm of HCI games research (*ibid.*, p. 33), in that their stated intention is to use a game as a tool to generate knowledge about another topic. Few papers have reflected the ontological paradigm (*ibid.*, p. 33), in that they have sought to understand voice game design in and of itself, or the practice paradigm (*ibid.*, p. 33), in that they have sought to understand player experiences and interactions with voice games. Those that have done so are discussed in Section 2.4.5.

2.4.1 Voice interfaces for accessibility in videogames

Grammenos et al. (2007) conducted a program of research to design three universally accessible videogames as case studies for how to improve the accessibility of videogames in general. The games were an online version of *Chess* (Grammenos et al., 2005) and two multiplayer variants of *Space Invaders* (Grammenos et al., 2009, 2006). Similarly, Norte and Lobo (2008) created a voice-controlled version of the number puzzle game *Sudoku*. Each of these games was designed to be inherently controllable through multiple interaction modalities, including voice commands. In addition, multiple HCI researchers have developed systems that allow voice-based control of existing videogames, using either speech input (Derboven et al., 2014; Harada et al., 2011; Mustaquim, 2013; Sporka et al., 2006) or non-verbal voice input (Harada et al., 2011; Igarashi & Hughes, 2001; Sporka et al., 2006). Collectively, these papers represent a valuable contribution towards the design of more accessible videogames. However, the investigation of player responses to any of these voice game prototypes is limited to tests of usability, and so this body of research does not contribute to an understanding of player experiences of voice gameplay.

2.4.2 Voice games for language learning

Voice games have an obvious appeal as educational tools that can provide structure and motivation to support self-directed language learning. This was recognised early by DARPA, which commissioned the development of the *Tactical Language and Culture Training System*, an educational simulation videogame designed to teach US Armed Forces personnel how to communicate in a foreign language (W. L. Johnson, 2007; W. L. Johnson & Beal, 2005; W. L. Johnson & Valente, 2009; W. L. Johnson & Wu, 2008). A report found that trainees showed significant improvements in their knowledge of Iraqi language and culture after 40 hours of training with the system, indicating that the approach was at least somewhat successful (W. L. Johnson & Valente, 2009, p. 82). Voice games have since been developed for learning Mandarin Chinese (McGraw, Yoshimoto, et al., 2009; McGraw & Seneff, 2008), Portuguese (Silva et al., 2011), French and German (Roy, 2016), as well as games that can be adapted to multiple languages (Anderson et al., 2008; Jewell, 2011). A study on one of these educational games found that its players indicated high levels of engagement, enjoyment and motivation to play again. However, it noted that there were inherent problems with getting a speech recognition system that was trained on native speaker pronunciations to recognise the pronunciation of a non-native speaker of the language, even if the pronunciation was recognisably correct to a human ear (Anderson et al., 2008, pp. 615–616). Overall, this literature suggests that voice games are a promising avenue for language learning, in part because voice gameplay has the potential to be entertaining for the players, but it offers few generalisable insights into the design of voice games that are not intended for language learning or other education purposes.

2.4.3 Therapeutic voice games

A category that has emerged relatively recently, but already received a substantial amount of attention, is voice games that are designed to support players in conducting therapeutic exercises. Such voice games have been developed to facilitate speech rehabilitation (Cler et al., 2017; Loaiza et al., 2013; Lv et al., 2017, 2015), speech therapy (Duval et al., 2018; Krause et al., 2013; Lan et al., 2014; Navarro-Newball et al., 2014; Rubin & Kurniawan, 2013; Zajc et al., 2018) and emotion regulation (Fernández-Aranda et al., 2012; Kostoulas et al., 2012; Sonne & Jensen, 2016). Of these, the study by Navarro-Newball et al. (2014) provided the most extensive user evaluation, based on observations of gameplay as well as questionnaires and interviews with the players, the parents and the therapist. It determined that the voice game used in the study—which presented a cartoon animal scenario, and required players to practice pronouncing phonemes correctly to proceed—was received very positively, receiving the highest possible scores for enjoyment and engagement, despite some usability challenges. These studies point to voice games as a viable format for the provision of therapeutic voice exercises, but provide little insight into the player experience of these games, nor general findings about the design of voice games in a broader sense.

2.4.4 Non-verbal voice input for videogames

While the majority of the research literature on voice games has concentrated on the use of spoken language, there is a smaller body of work that has looked at non-verbal voice interaction in games. Igarashi and Hughes (2001) proposed three interaction techniques for non-verbal voice control of games: control by continuous voice, in which the player causes an action to continue by producing a consistent vocal sound until they wish the action to stop; rate-based parameter control by pitch, in which an action is modified by the pitch of the player's voice as they produce a consistent vocal sound; and discrete control by tonguing, in which a player produces discrete sounds to count out a number that they wish to select. The authors acknowledged that the techniques require "an unnatural way of using the voice" (Igarashi & Hughes, 2001, p. 156) and may be tiring for long-term usage. However, the paper did not address the social acceptability of making unusual vocal sounds to control a computer, which subsequent HCI research has indicated can present a barrier to usage of voice interfaces (Al Hashimi, 2007; Rico & Brewster, 2010).

More recent research has demonstrated the possibility for games to be controlled by the player's breath, rather than their speech (Sra et al., 2018; Tennent et al., 2011). Sra et al. (2018) found that players gave a breath-controlled virtual reality (VR) game relatively high ratings for novelty, accessibility and usability compared to an equivalent VR game controlled with a motion trackpad. Players described the experience of breath-based interaction as "intuitive" and "natural" and stated that it heightened their sense of presence in the VR gameworld. However, players also reported that breath-based interaction was more physically demanding than controller-based interaction, and that they had greater difficulty remembering breath inputs than controller-based inputs (Sra et al., 2018, pp. 8–9). These findings closely echo the results of the player experience study presented in Chapter 7.

2.4.5 Voice gameplay

The prior research on voice games described in previous sections has included limited exploration of player experiences of voice gameplay. In the studies that have reported on player responses, there has been a general theme of high levels of enjoyment, engagement or immersion being reported during voice gameplay, despite usability challenges that are greater than those for traditional manual controls. In all of those studies, however, players were evaluating a novel voice game that was created or adapted by the researchers. As a result, the findings may be influenced by demand characteristics,[12] which may have reduced the likelihood that players would report negative experiences with the voice games. In addition, the fact that the studies were all on voice games developed for research

[12] "Demand characteristics" are aspects of a study that may cause participants to change their responses, intentionally or unintentionally, to match what they believe the researcher expects of them, thus causing the findings to differ from what would have been observed if the participants did not have a expectations about the researcher's expectations. The importance of demand characteristics in research was established by Martin Orne (1962). In a user study on a novel computer system, anything that causes the participant to think that the researcher has designed the system and would like it to succeed can be a demand characteristic, which may influence the participant to give the system a more positive evaluation.

or training makes it uncertain how closely the results would apply to voice games that have been developed for entertainment.

The most widely popular category of voice game to date has been singing-based games, as Chapter 5 describes. These games are designed to be played at social events involving multiple players and spectators, and so a proper understanding of their player experiences requires an investigation of how they are played in that context. Accordingly, Fletcher and Light (2011) conducted two ethnographies of social groups that played the karaoke game *SingStar*. They provided a detailed account of how karaoke gameplay was physically conducted by the players, with widely varying configurations of different numbers of players filling different roles and appropriating the game's microphones in a variety of ways. Fletcher and Light emphasised the role of *SingStar* as an activity around which social activities and relationships were structured, characterising it as a "glue technology that assists in crafting and strengthening social linkages amongst players" (Fletcher & Light, 2011, p. 1).

Hämäläinen et al. (2004) conducted informal testing of two games controlled by non-verbal voice input, including one controlled by singing. They noted that their results supported the view that voice interaction was unlikely to succeed as a replacement for a traditional controller, but had potential to be popular in games that were designed around vocal action as their core activity, rather than as a control mechanism: "the interface makes gaming a spectator sport—you can show off your musical skills much the same way as when singing karaoke or performing on stage in a concert" (Hämäläinen et al., 2004, p. 370). This echoes the findings of Fletcher and Light (2011) described above, and suggests that a playful social spectacle may be an important part of the appeal of voice interaction—a possibility that has received little attention in the wider HCI literature on voice user interfaces.

A further question is whether the satisfactions of social voice gameplay can only be provided by playing in the presence of other people, or whether non-player characters can provide some of the same kinds of engagement in voice gameplay. Fraser et al. (2018) tested one approach to making non-player characters more responsive to the player in voice gameplay. They developed a prototype game that could process natural-language speech input from the player, analyse the sentiment of the player's words and the tone of their speech, and cause a non-player character (to whom the player was speaking in the game) to respond in a manner that reflected the emotion that the player had expressed. Participants in the study rated this emotionally responsive voice game more positively than an alternative version of the game that did not use the emotion detection features, and also spend more time playing the emotionally responsive version of the game, which the authors interpret to mean that they were more engaged with it. This provides some evidence that non-player characters can provide a socially engaging single-player experience in a voice game. However, the ability to generalise from this study is limited, as the comparison was only between a videogame that was designed for emotionally responsive dialogue and an alternative version of the same game with the emotionally responsive dialogue (the core game mechanic around which the game was designed) turned off. A broader understanding of the player experience of voice gameplay requires a broader study of fully realised voice games.

In 2015, at the beginning of my PhD candidature, I co-authored a grounded theory method study that looked at online reviews and responses to commercially published voice games (Carter et al., 2015). We noted many positive comments about the enjoyability and immersiveness of voice commands, but also a great deal of frustration with voice commands that were slower or less reliable than the traditional controller equivalent. One of the key contributions of this research that had not previously been noted was the theme of "identity dissonance": numerous players, in a variety of games, reported that voice commands felt uncomfortable or unnatural when they were phrased as being directed at the player-character, rather than phrased as something that the player-character might have said. We argued that this was due to a dissonance between the normal game controls, which evoked a sense that the player-character was the player's own embodied persona in the gameworld, and the voice controls, which evoked a sense that the player-character was a separate identity to whom the player was speaking. Voice commands that aligned the player's own identity with the player-character identity were much more positively received, and less subject to complains about feeling "unnatural". The findings of this study provided the inspiration for the topic of my PhD research: initially on conversational interaction with game characters (the focus of Chapter 4), and then on VOICE in games as I developed a clearer understanding of the role of voice in driving the "identity dissonance" effect.

2.4.6 Game characters as social actors

This chapter has reviewed a range of studies that indicate that users respond to conversational agents, to a certain extent, as though they are human. This does not mean that users are deluded into believing that these agents are human or have humanlike consciousness, as HCI researchers have been at pains to point out (D. Johnson & Gardner, 2009; Nass & Moon, 2000). Rather, users appear to respond to these agents *as if* they were humanlike, in a similar manner to how film viewers and novel readers respond to fictional characters as if they were real people. This is further complicated in videogames by the fact that there are multiple overlapping types of characters and identities that players may respond to—and which may come into conflict with each other, as the phenomenon of "identity dissonance" (Carter et al., 2015) suggests. Consequently, this section reviews prior research on the different types of characters and identities in gameplay, based on studies of non-voice-based games.

Games researchers have defined a number of identities that a person may take on when interacting with a videogame. Many studies have focused on two identities in particular: that of the *player* or *user* as the physical person operating the game system, and that of the *character* or *avatar* as their controllable virtual figure in the gameworld (e.g. Banks, 2015; Bessière et al., 2007; Hart, 2017; Klimmt et al., 2009; Yee & Bailenson, 2007). These studies typically conceptualise the character (or avatar) as an externalised projection of the player's sense of self, or as a means by which to temporarily inhabit the role of an alternative self. However, this presupposes a form of gameplay in which the player controls a singular avatar that represents their character in the gameworld.

Multiple aspects of this supposition are variable across different types of games. As Jørgensen (2009) has pointed out, the number of figures that a player controls in a gameworld may be more than one; or it may be zero; or it may be ambiguous, as in a real-time strategy game in which the player issues commands from a disembodied overhead perspective to units that follow those commands but are otherwise autonomous. In addition, as Bayliss (2007) has argued, the notions of *avatar* and *character* are distinct from one another and not always synchronised: a player character such as Lara Croft has a defined identity that stands apart from the identity of the player and the identity of the avatar; one imagines that the character of Lara Croft could drink at a bar or play cricket, even though the avatar of Lara Croft is not programmed with these abilities. This distinction was empirically supported by a series of interviews conducted with players of an online game by Carter et al. (2012), who found that players made a distinction between their *character* and *avatar* identities along the lines that Bayliss (2007) suggested, and also distinguished between the identities of themselves as the *user* and themselves as a *player*. The interviewees tied their identity as a player to their reputation within that game's community, and felt that it was separate from the "offline" identity that they had outside the context of the game community (Carter et al., 2012).

Within the conceptual category of game characters, there is an additional distinction to be made between *player characters* and *non-player characters*. Player characters represent the player's identity role (or roles) in the gameworld, and their actions and decisions are usually controlled to some extent by the player. Non-player chracters represent other people or beings in the gameworld, and their actions and decisions are usually controlled primarily by the game's artificial intelligence programming or a predetermined script (although some non-player characters also respond to instructions from the player). Player responses to non-player characters have been studied at length in the game studies literature, with much of the research focused on the question of how to design non-player characters to be believable or engaging (e.g. Gillies, 2018; Lankoski et al., 2011; Lankoski & Björk, 2007; Rivera et al., 2012). However, this thesis is concerned with the narrower question of how players form a sense of social relationship between themselves (as the player or as the player character) and non-player characters, in the manner of the "computers as social actors" paradigm (Nass et al., 1994).

Some insights into the multifaceted nature of the relationship between a player and a player character or non-player character have been provided in a survey study by Bopp et al. (2019). The survey asked players to explain their thoughts and emotional attachments towards a game character for whom they had felt especially strong appreciation, desire, respect, devotion or care. The authors developed seven themes to describe the types of emotional attachments that players reported: "cool and capable", "respected nemesis", "admired paragon", "crush", "concern for one's protégé", "sympathetic alter ego" and "trusted close friend". Although these themes were developed to describe types of emotional attachments, each theme in fact describes a type of social relationship by which the character is positioned in relation to the player.[13] Players mentioned both player characters and non-player characters in their responses, and the authors found that each theme had a clear tendency to be applied to one or the other type of character. For example, player characters were more likely to elicit the sense of being an admired paragon, whereas non-player characters were more likely to elicit the sense of being a trusted friend. This study shows that players experience games as complex social situations in which there are multiple ways for the player, player character and non-player character identities to relate to one another.

Much of the research on identity in games does not address the role of non-player characters, and uses the term "character" to mean "player character" without clarification. This is perhaps attributable to the fact that a large proportion of these studies (e.g. Banks, 2015; Bessière et al., 2007; Carter et al., 2012; Gilbert et al., 2014; Yee & Bailenson, 2007) were conducted in the context of massively multiplayer online role-playing games (MMORPGs) such as *World of Warcraft* and online virtual worlds such as *Second Life*, in which constructing a distinctive player character identity is a central part of gameplay, and in which it is typical for the majority of a player's non-violent interactions to be with the avatars of other players. However, the distinction between player character identities and non-player character identities is critical for this thesis, as the relationships between these identities will emerge as important drivers of the player experience of voice gameplay in the following chapters.

2.5 Research gaps and questions

The research summarised in this chapter supports the argument that humans are "voice activated", as Nass and Brave (2005) have asserted: that we respond to

[13] The social relationship is apparent in the names of all seven themes except for "cool and capable". This theme describes characters that participants found exciting or amusing, like a funny or skilful friend, or a sense of empowerment in their own player character identity (Bopp et al., 2019, p. 317).

virtual agents that speak in some of the same ways that we respond to humans, consciously or unconsciously following some of the same behavioural scripts and expectations that we would normally follow in person-to-person social interactions. The literature review shows that HCI researchers have explored the potential of voice interaction to evoke a sense of social presence through embodied conversational agents, voice assistants and game characters. It also shows that voice interaction has been used in videogames to work towards a number of practical purposes, such as language learning and speech therapy.

Despite this body of work on voice interaction and voice game design, the existing literature contains few studies that provide a wider understanding of what voice game design is, outside of research labs, or how it is experienced by players. Only a handful of studies have investigated voice gameplay as it is practiced, rather than seeking to develop new forms of voice games that are oriented towards some extrinsic purpose beyond the inherent pleasures, meanings and experiences of gameplay. There is a gap in the literature on the topic of how voice games are designed, how players respond to voice games, and even what types of voice games exist beyond the serious games developed for research purposes.

To address that gap, this thesis sets out to answer the following research question:

> **Main RQ** How does voice-operated interaction with characters and environments (VOICE) influence the player experience of digital games?

The research question is deliberately broad and open-ended. Rather than attempting to pinpoint one narrow facet of voice game design, this thesis aims to shine a light on this under-examined area of game design in its entirety. The research question reflects the fact that the thesis is concerned with both how voice games have been designed in the past—as an indicator of the breadth of ways in which voice games may be used in the future—and how players experience different forms of voice gameplay.

The overall research question is broken down into the following more specific research questions, which are addressed in separate studies:

> **RQ1** How do players make sense of and interact with a game character that responds to natural language text?

The first research question addresses the topic of how players respond to dialogue with a game character. Unlike the following questions, RQ1 concerns text-based interaction rather than voice interaction. This reflects the initial focus of the research, which was intended to cover "conversational interaction" with game characters across both voice and text, before it was narrowed to focus on voice gameplay in the subsequent studies. RQ1 is addressed primarily in Chapter 4, through a participant observation study of text-based interaction with a game character in *Minecraft*.

RQ2 What is the historical lineage of voice game design and what are the major influences that have shaped it?

The second research question aims to alleviate the lack of information about what voice games are, outside of research labs. RQ2 is addressed primarily in Chapter 5, through the construction of a history of voice games from their earliest beginnings to 2016, which accounts for how these games have developed both as technological systems and as cultural products.

RQ3 What design patterns are currently implemented in voice games?

The third research question also addresses the lack of information about voice games, but this time with a focus on the specific variations that have appeared in the design of voice games. RQ3 is addressed primarily in Chapter 6, through the development of a pattern language for voice game design and a typology of five fundamental genres of voice games.

RQ4 What is the distinctive nature of voice gameplay?

The fourth research question is once again deliberately broad. This thesis seeks to understand how players experience voice gameplay, and specifically what is different or noteworthy about voice gameplay compared to other forms of gameplay, without presupposing what that might be. RQ4 is addressed primarily in Chapter 7, through a participant observation study and a frame analysis approach to interpreting player accounts of voice gameplay.

RQ5 How do players experience verbal and non-verbal voice gameplay?

The fifth research question is a response to the observation that non-verbal forms of voice input have been used successfully in both academic and commercial voice games, and yet the player experience of non-verbal voice gameplay has received even less study in prior literature than speech-based voice gameplay. RQ5 is addressed primarily in Chapter 7, through a participant observation study in which players engage in both verbal and non-verbal voice gameplay and discuss both experiences in relation to each other.

3 Methodology

The literature review in the last chapter highlighted that voice game design and gameplay are not well understood. It concluded with the five research questions that guide the thesis. This chapter explains how each of these research questions was addressed. It situates the thesis first within the fields of human-computer interaction and game studies (Section 3.1), and then within the constructionist-interpretivist research paradigm (Section 3.2). It details the range of research methods used to study gameplay in Chapters 4 and 7 (Section 3.3) and to study voice game design in Chapters 5 and 6 (Section 3.4). It concludes with an overview of the research design that explains how each method was chosen to address one or more of the research questions (Section 3.5).

3.1 Disciplinary orientation

This thesis is situated at the intersection of human-computer interaction (HCI) and game studies. Both of these fields concern the study of interactive systems and their users; for game studies and for a subset of HCI, the interactive systems in question are digital or analog games. The fields are also similar in that they both employ a wide variety of research methods and theoretical perspectives drawn from other areas of scholarship. HCI has been described as "polycentric" (Liu et al., 2014, p. 3561) and an inter-discipline (Blackwell, 2015, p. 505) due to its lack of standardised research perspectives and themes, just as game studies has been described as "polyphonic" (Mäyrä, 2014) and multidisciplinary (Deterding, 2017, p. 537; Mäyrä, 2008, p. 6) for much the same reason.

3.1.1 Game studies

The multidisciplinarity of game studies stems partly from its youth. Although individual scholars have studied games in a piecemeal fashion for many decades,[14] game studies did not emerge as a truly cohesive field until after videogames reached a critical mass of popularity in the 1990s. Aarseth declared that "2001 can be seen as the Year One of Computer Game Studies[15] as an emerging, viable, international, academic field" (Aarseth, 2001). He noted that game studies was greatly influenced by the range of disciplinary backgrounds of its members: "we all enter this field from *somewhere else*, from anthropology, sociology,

[14] Early scholars of play and games include Huizinga (1949 [1938]), Caillois (2001 [1958]), Suits (1978) and Sutton-Smith (1997). Sudnow (1979) is noteworthy as one of the first individuals to study the experience of playing digital games.
[15] Although Aarseth used the term "Computer Game Studies", the field has come to be known simply as "Game Studies", and includes a rich tradition of research on non-digital games.

narratology, semiotics, film studies, etc, and the political and ideological baggage we bring from our old field inevitably determines and motivates our approaches" (Aarseth, 2001, emphasis in original). That remains broadly true today. However, Deterding (2017) has observed a countervailing trend, in which an increasing volume of games-related research is conducted under the banner of other disciplines—including HCI—which has caused those viewpoints to become less well represented within mainstream game studies journals and conferences. Both trends are reflected in this thesis, which has drawn upon multiple disciplines and methodologies for its research design, and published its findings in a combination of game studies and HCI venues.

3.1.2 Human-computer interaction

Human-computer interaction has been constituted as a recognisable field for considerably longer, and undergone more clearly defined shifts over time. Its roots stretch back before the advent of the digital computer, to the fields of Industrial Psychology and Human Factors in the early 20th Century (Grudin, 2012, pp. 4–7), and some of the earliest examples of what we would today call HCI research were conducted in the 1950s (Grudin, 2012, p. 8). The historical development of HCI is commonly divided into three waves (Bødker, 2015, 2006) or paradigms (Harrison et al., 2007, 2011). Each wave reflects a different set of goals and concerns, and is therefore typified by a different mix of research approaches.

The first wave of HCI has its roots in the engineering discipline, and particularly the field of human factors, beginning in the late 1950s (Grudin, 2012, pp. 8–10). Research in this paradigm treats human-computer interaction as an ergonomic design problem to be solved. Its goal is to optimise users' performance on computer tasks by resolving any problems that would impair their productivity. It typically achieves this through systematic testing and the establishment of fixed guidelines (Bødker, 2006, p. 1), under a broadly positivist/postpositivist research paradigm (Duarte & Baranauskas, 2016, p. 2). How the user understands the interaction is considered only insofar as it affects output; first-wave HCI "tends to take a pragmatic approach to meaning, ignoring it unless it causes a problem" (Harrison et al., 2007, p. 7).

The second wave, by contrast, is centrally concerned with what and how the user thinks. Influenced by cognitive science, it depicts computer use as a continuous communication loop between two information processors: the computer system and the user's mind. Typically, second wave HCI research aims to understand the manner in which information flows between computers and people, with the end goal of improving the efficiency and accuracy of this exchange (Harrison et al., 2007, p. 10). The second wave emerged in the 1980s, at a time when computer usage was becoming progressively more widely practiced and also more discretionary (Grudin, 2012, p. 16). Accordingly, second-wave HCI is characterised by a greater focus on understanding the context of interaction, and on considering interaction scenarios that involve more than one person or computer system (Bødker, 2006, p. 1). It makes greater use of proactive, naturalistic and participant-oriented research methods, such as participatory design and contextual inquiry, but remains primarily concerned with computing activities as a form of purposeful work (Bødker, 2006, p. 1; Duarte & Baranauskas, 2016, p. 3).

The third wave arose in the first decade of the 2000s, partly in response to the growing presence of computer systems in non-work contexts, such as homes, classrooms and mobile applications (Bødker, 2006, p. 1; Harrison et al., 2007, pp. 5–6). It represents a shift away from the study of computing as productive information processing work, and towards the study of computing as meaningful participation in culture. As Bødker describes it: "the focus of the third wave, to some extent, seems to be defined in terms of what the second wave is not: non-work, non-purposeful, non-rational, etc" (2006, pp. 1–2). The third wave is concerned with embodiment (Dourish, 2004), emotion (Harrison et al., 2007, p. 6), meaning-making (Harrison et al., 2011, p. 388) and experience (Bødker, 2006, p. 5; McCarthy & Wright, 2007) as central elements of human-computer interaction, rather than as marginalia. To understand these phenomena, the third wave moves the focus of HCI from testing and evaluation to more exploratory, discover-oriented research (Bødker, 2006, p. 2). It makes greater use of ethnographic and ethnomethodological approaches, which "tend to eschew *a priori* categories of interest in favor of discovering what emerges from interaction" (Harrison et al., 2011, p. 387). Third-wave HCI exhibits the concern for lived experience and socially constructed meaning that are the hallmarks of a constructionist-interpretivist research paradigm. However, as Duarte and Baranauskas point out, the third wave frequently "inherits" (2016) certain features of the positivist/postpositivist paradigm in which HCI first developed, namely a common use of quasi-experimental methods as opposed to more naturalistic inquiry, and an emotionally neutral and objective rhetorical style.

This thesis sits firmly within the third wave of HCI research. It is ultimately concerned with understanding the ways in which players experience voice gameplay as a meaningful event, rather than with the discovery of universal laws about the operation or efficiency of voice interaction in games. This is reflected in the research questions, which address how players "make sense of" and "experience" voice gameplay. Accordingly, I have approached this research from a constructionist-interpretivist perspective, as the next section will describe.

3.2 Epistemology and research approach

This section explains the epistemology (constructionism) and theoretical perspective (interpretivism) that have guided this thesis. It describes their consequences for the research methodology, and ends with a summary of the overall research design.

Constructionism is an epistemology, or theory of how knowledge is formed (Crotty, 1998, p. 3). It holds that people's understanding and experience of the world is a creation of their interactions and their beliefs (Neuman, 2014, p. 104). Human beings have no direct access to objective reality, as "all knowledge, and therefore all meaningful reality as such, is contingent upon human practices, being constructed in and out of interaction between human beings and their world, and developed and transmitted within an essentially social context" (Crotty, 1998, p. 42). In the constructionist viewpoint, a research finding is not an objective fact recorded by a neutral observer, but an account of an idea that the researcher has actively developed. The researcher constructs their own understanding, and they are influenced in this process by the concepts and cultural-social processes to which they have previously been exposed.

Interpretivism is the theoretical perspective or philosophical stance adopted by practitioners of interpretivist social research (Crotty, 1998, pp. 2–8; Neuman, 2014, p. 103). Interpretivist research is concerned with "meaningful social action" (Neuman, 2014, p. 104), which is to say, how people construct meaning within a situation and how they experience life in a particular setting. The meaning of a human action is dependent on the meaning system—such as the culture or language—in which it occurs, and through which it is interpreted by other social actors (Neuman, 2014, p. 104). There is no neutral perspective from which to explain an action in terms of its meaning; it can only be understood from one or more subjective viewpoints. Therefore, interpretivist research generally requires the researcher to develop an understanding of a situation as it appears to its participants, "from within". To this end, research in this tradition typically employs qualitative methods of data collection, and inductive forms of analysis in which theories are derived in a bottom-up fashion from empirical observation (Neuman, 2014, p. 70).

The studies described in this thesis use three different methodological frames and a variety of research methods, all within the overarching paradigm of constructionist-interpretivist research. Study 1 and Study 4 are player experience studies in the traditional mould of lab-based HCI research, with the details of their research design engineered to produce rich qualitative insights. Study 2 is a historical analysis of voice game design, conducted under the aegis of media archaeology. Study 3 is a work of design pattern analysis. The methodologies and methods used in each study are shown in Table 3.1.

Table 3.1 Research approach (see Crotty, 1998)

Epistemology	Constructionism			
Theoretical perspective	Interpretivism			
Main research question	How does voice-operated interaction with characters and environments (VOICE) influence the player experience of digital games?			
Study	**Natural Language Gameplay (Chapter 4)**	**Voice Game History (Chapter 5)**	**Voice Game Design Patterns (Chapter 6)**	**Voice Gameplay (Chapter 7)**
Ancillary research questions	RQ1 How do players make sense of and interact with a game character that responds to natural language text?	RQ2 What is the historical lineage of voice game design and what are the major influences that have shaped it?	RQ3 What design patterns are currently implemented in voice games?	RQ4 What is the distinctive nature of voice gameplay? RQ5 How do players experience verbal and non-verbal voice gameplay?
Methodology	Observation and interview study	Media archaeology	Design pattern language	Observation and interview study
Data generation methods	Participant observation Semi-structured interviews Interaction logs Wizard of Oz prototype	Artefact review Paratext review	Artefact review Paratext review	Participant observation Semi-structured interviews
Data analysis methods	Inductive thematic analysis Qualitative content analysis (informed by conversation analysis)	Historical analysis	Design pattern analysis	Inductive thematic analysis Frame analysis

3.3 Research methods for studying gameplay

This section describes and discusses the research methods used in Study 1 (Chapter 4) and Study 4 (Chapter 7), which focused on *voice gameplay* as an activity. The methods are: participant observation, Wizard of Oz prototyping, semi-structured interviews, thematic analysis, content analysis and frame analysis. The research methods used to study *voice games* as designed artefacts are described in Section 3.4.

3.3.1 Participant observation

Observational studies are among the most common approaches to HCI (Sharp et al., 2019, p. 260) and games research (Sangin, 2018). In a typical observational study, the researcher recruits participants to engage in a task that involves a computer system. The task may take place in a controlled setting, such as a usability lab, or in a more naturalistic setting ("in the field"). The researcher may observe participants directly (in person) or indirectly (through video recordings, sensor recordings or system logs). Participants in observation studies often answer questions about their experience, either after the task is finished through a questionnaire or interview process, or during the task itself using a think-aloud process (Lewis, 1982).

Study 1 and Study 4 were *participant observation studies*. For both studies, participants were recruited to attend a session in a controlled lab setting, during which they played through multiple game scenarios. For Study 1, the sessions involved text-based natural language interaction with a game character in *Minecraft*, operated as a Wizard of Oz prototype (see Section 3.3.2). For Study 4, the sessions involved voice and non-voice gameplay with three different commercially published videogames. Data was collected through a combination of direct observation, system logs and interviews (see Section 3.3.3). The following sections will discuss the methodological considerations of each of these factors.

Observation in a controlled environment

Lab-based studies provide several benefits for observational research. They allow for controlled comparisons, by minimising the variance between sessions and conditions to isolate the cause of any observed differences. They enable the researcher "to investigate use of technology without deploying it, to compress time and thereby study infrequent phenomena, to study interaction that would otherwise be unethical or impossible to obtain data about, to control external factors, and to collect fine-grained data" (Hornbæk, 2013, p. 6). These factors made the controlled-environment observation approach well suited for Study 1 and Study 4, as the aim of these studies was to produce a detailed comparative analysis of players' experiences with multiple uncommon forms of gameplay.

As an additional benefit, the lab setting also allowed me to control the degree of privacy with which the games were played. While privacy was not a significant ethical concern, prior research has shown that participants may find voice interaction embarrassing in the presence of other people (Carter et al., 2015; Williamson, 2011). As a result, the presence of others was a factor of interest in this research. Study 4 utilised the controlled lab study design to ask participants to perform different types of voice gameplay in the presence of two people: myself as the researcher, and one other participant. This generated a number of insights into the social psychology of voice gameplay, which are detailed in Chapter 7.

The benefits of a lab study come at the cost of external validity. The artificiality of the lab environment and the potential for social desirability bias introduced by the observing researcher can mean that participants behave differently than they would during an unobserved interaction (Gergle & Tan, 2014, p. 195; Sharp et al., 2019, p. 301). Gergle and Tan (2014, p. 195) note three strategies to mitigate these issues, each of which was employed in one or both of the observational studies as discussed below.

The first strategy is to make the participants' task match as closely as possible to the reality of the activity that it is intended to recreate. To this end, both lab studies used commercially published videogames rather than custom research prototypes, and involved gameplay tasks that matched the normal goals and structure of gameplay in those games. In Study 4, the study room was furnished and decorated as a domestic lounge room, with comfortable sofas, pot plants, bookshelves, board games, a coffee table and a television. In Study 1, the study room was furnished as an office environment, but decorated with toys and props that represented objects from *Minecraft*, the videogame used in the study, to create a more relaxed feel.

The second strategy is to choose participants who are similar to those who would normally undertake the task, outside the research context. For Study 1, participants were recruited from a demographic group (11–15 year olds in England) among which *Minecraft* is almost ubiquitously played, based on a consultation with a senior developer on the game. For Study 4, there was no clear profile of a typical voice game player, and so participants with a range of ages, genders and gameplay experience levels were recruited (see Chapter 7 for further details).

The third strategy is to assess how closely the behaviours observed in the lab reflect behaviours observed in the field. Prior to Study 4, I had conducted extensive analyses of video recordings and written reviews of voice gameplay for Study 2, Study 3 and an additional grounded theory method study of voice gameplay on which I was a co-author (Carter et al., 2015). As a result of this experience, I was able to verify that the behaviour and attitudes observed in Study

4 were a good match for behaviours and attitudes observed outside the lab settting.

The knowledge gained from prior studies also meant that I began the research with preconceived expectations about what the findings might be, and these expectations may have the influenced the research outcomes. From a positivist or postpositivist perspective, such influences are sources of bias that ought to be minimised through precautions in the data collection and analysis processes. For example, the original formulation of grounded theory method (GTM) instructs researchers not to read anything that might influence their interpretation of the research object before they conduct the research: "literally to ignore the literature of theory and fact on the area under study, in order to assure that the emergence of categories will not be contaminated" (Glaser & Strauss, 1967, p. 45). From a constructionist perspective, there is no such "uncontaminated" perspective to protect, because the researcher's existing mental constructions are the only tools by which they can construct an understanding of any phenomena, and these constructions are always socially and experientially based (Guba & Lincoln, 1994, pp. 110–111). As Thornberg notes, attempting to preserve an unbiased perspective is only likely to prevent the researcher from admitting to biases that they inevitably have—not to mention the practical problem that Glaser and Strauss' dictum "makes it impossible for researchers to conduct studies in their own areas of expertise which appears odd and counter-intuitive" (Thornberg, 2012, p. 244). The appropriate way to address one's preconceptions as a constructionist researcher is to acknowledge them and to be reflexive in considering how they may have influenced the research findings, as I have sought to do in this thesis.

External factors must also be considered when evaluating the findings of the participant observation studies. Study 1 and Study 4 were conducted in relatively formal settings, under visible observation, and under instruction rather than through spontaneous play; each of these is likely to have had some effect on how the participants interacted with the games, how they interpreted the gameplay experience and how they spoke about it afterwards. The controlled lab study approach was effective in generating insights into voice interaction and conversational interaction gameplay, and particularly for allowing close comparisons between different modalities of gameplay, but it would be valuable for these insights to be further investigated and validated through more naturalistic inquiry.

Interaction logs

In addition to direct observation and video recordings, participant behaviour can be observed indirectly through *interaction logs*. Interaction logs are records of participants' specific interactions with a system, such as their key presses, mouse movements, task-flow sequences or activity durations (Sharp et al., 2019, pp. 299–300) As a data generation method, interaction logging has the benefit of being unobtrusive, allowing a large volume of information to be collected about users' behaviour without interrupting the flow of their activity.

Study 1 used interaction logs to record participants' natural-language dialogue with a game character in *Minecraft*. Messages to and from the character were sent through the game's built-in text chat console. These messages were recorded and saved as a plain text file at the end of each session, for later analysis (see Section 3.3.5).

3.3.2 Wizard of Oz method

The *Wizard of Oz method* is an approach to prototyping and evaluating an interactive system, in which the system is presented to research participants as being fully automated, but some of its workings are operated in secret by a human researcher (Kelley, 1984). The method was originally developed to study user interactions with natural language systems (Gould et al., 1982; Green & Wei-Haas, 1985; Kelley, 1983), and it has since become widely used to study user interactions with other types of systems, including embodied AI agents (Goodrich & Schultz, 2007; Maulsby et al., 1993; Riek, 2012). Originally designed for use in an iterative fashion over several cycles of prototyping and development on a single system (Kelley, 1984), the method has since become commonly used in single-cycle studies to evaluate or explore user responses to a certain type of system design (Riek, 2012, p. 129).

The Wizard of Oz approach presents several benefits. It allows researchers to rapidly prototype and test "hypothetically possible systems" (Clark et al., 2019, p. 6) that would be prohibitively expensive, time-consuming or difficult to construct in reality. It enables researchers to collect samples of human-computer dialogue (or other interaction data) that are representative of what a future system will be built for, effectively reversing the normal construct-then-test pipeline to allow the system's design to be informed by "real" user inputs from an early stage (Dahlbäck et al., 1993, p. 259). It is particularly suited to natural language applications, as it allows an application to be tested with users without the constraints of a flawed and inflexible speech recognition or language processing system, thus isolating usability issues that are not attributable to those subsystems.

Study 1 utilised the Wizard of Oz approach to simulate a virtually embodied AI-driven game character in *Minecraft*, who could respond to natural language text inputs and adapt its behaviour over time based on player instructions. To ensure the interaction was a fair representation of a plausible near-future gameplay scenario, the behavioural script for the prototype game character was based on extensive background research, which included: reading literature review papers on artificial intelligence in games (Yannakakis, 2012; Yannakakis & Togelius, 2015), natural language processing (Dahl, 2013; Picca et al., 2015; J. O. Ryan et al., 2015; Zagal et al., 2012), human-robot interaction (Goodrich & Schultz, 2007;

Kahn et al., 2008; Lallée et al., 2010; Riek, 2012) and interactive machine learning (Amershi et al., 2014; Knox et al., 2012); consulting with experts in commercial game AI (two senior game developers) and machine learning (my co-author and project supervisor, Katja Hofmann); and examining various existing examples of language-based assistant AI characters for games like *Minecraft*.

The aim of Study 1 is not to create a perfect representation of what natural-language-based gameplay interactions will look like in the near future. Rather, as Dahlbäck et al. (1993) suggest, it was to use the Wizard of Oz approach to collect information about how humans respond to such a system, so that these reactions can be taken into account for future developments in natural language gameplay (and, in the context of this thesis, to provide insights into language-based gameplay more generally that connect with the research on voice gameplay).

3.3.3 Semi-structured interviews

The *interview* is a core method of data generation in HCI (Sharp et al., 2019, p. 260) and games research (Bromley, 2018; Cote & Raz, 2015). It has been defined simply as "conversation with a purpose" (Bingham & Moore, 1959, p. 3). Interviews can be conducted in a multitude of ways: in person or remotely; verbally or through text dialogue; with individuals or with groups; in a single session or across multiple sessions; structured, semi-structured or unstructured (Fontana & Frey, 2005, p. 361). Unstructured and semi-structured interviews are a powerful tool for gaining insight into topics that research participants know a great deal about (Charmaz, 2014, p. 29), and particularly for understanding how their social worlds look from the inside (Miller & Glassner, 2010, p. 137). For games research, interviews "offer one of the only ways to understand players' motivations, [and] their understanding of mechanics or features in the game" (Bromley, 2018, p. 163). In comparison to other methods of generating player data, such as surveys and gameplay logs, interviewing allows an unparalleled level of depth and detail about each individual player's thoughts, at the cost of being time-intensive and labour-intensive per participant (Cote & Raz, 2015, p. 95).

In-depth, semi-structured, face-to-face interviews were the primary method of data generation used in Study 1 and Study 4. This choice was informed by the research questions for these studies as well as the overall epistemological stance. As the research questions for these two studies were broad and exploratory—asking how players respond to, make sense of and experience different forms of gameplay—it was important to allow for insights to emerge that the researcher did not anticipate in advance. Unstructured or semi-structured interviewing is suitable for this, as its lack of rigid boundaries enables the researcher "to understand the complex behavior of members of society without imposing any a priori categorization that may limit the field of inquiry" (Fontana & Frey, 1994, p. 366). This accords with the constructionist epistemology of the research, as it positions the interviewer as being in conversation with the interviewee, collaboratively constructing an account of their experience, whereas a positivist approach would aim to record the interviewee's words from a neutral perspective (Miller & Glassner, 2010, p. 131).

That is not to completely discount my influence as an interviewer, whose preconceptions and aims are reflected in the questions that I chose to ask. For both Study 1 and Study 4, I used a loose interview structure that began with broad, non-specific prompts for the participant to describe their experience of the gameplay, followed by progressively more specific questions that focused in on the research questions if the participant did not address them more spontaneously. However, I adapted my questions to the flow of each session, by adding questions about specific events that had occurred or remarks that were made during gameplay, and by probing for further explanations when a participant made an observation that was outside the scope of the planned interview. More specific details about the interview approach are provided in Chapter 4 and Chapter 7.

In Study 1, I encouraged participants to vocalise their thoughts during the gameplay sessions, in accordance with the "think aloud" protocol that is commonly used in HCI research (Sharp et al., 2019, p. 524). I did not give this instruction in Study 4, as the participants needed to use their voices to interact with the game system; requiring them to alternate between voice commands and verbal commentary would have added a confounding complication to the research. By conducting semi-structured interviews immediately after gameplay, and prompting the participants with questions about specific events that had occurred during the gameplay, I was able to gain a relatively fresh account of their experiences without the use of the think-aloud protocol.

Reflections on my role as an interviewer

My own capabilities and tendencies as an interviewer will have inevitably influenced the outcomes of the research. Prior to my graduate studies, I worked for several years as a market researcher, in the course of which I conducted numerous interviews and focus groups. From this experience, I developed skills in building rapport with interviewees, balancing a structured approach with following new threads in the conversation, and taking shorthand notes while remaining visibly attentive to what the interviewee is saying. These skills helped to ensure the interviews were productive.

However, Study 1 presented new challenges for me as an interviewer. I had not previously interviewed middle-school-aged children, as I did in that study. While all interviews have an imbalanced social dynamic, due to the inherent asymmetry in the roles of the interviewer and the interviewee, the effect was heightened in Study 1 by the fact that I was an adult interviewing children and young teenagers. I found it relatively difficult to build rapport, particularly in the short time-frame of the lab study. I also noticed that I needed to adjust my communication style (question wording, speed, tone of voice, body language, physical proximity, use of props and so on) substantially between participants, tailoring my interviewing approach to what I sensed the individual participant was most comfortable with. This may have limited the richness and depth of insight I was able to gain in this study—although many of these young participants did provide interesting comments, as detailed in Chapter 4.

3.3.4 Thematic analysis

A central challenge for qualitative research is how to synthesise diverse data to produce findings and theories, in the absence of the kinds of mathematical methods that underpin quantitative analysis. Qualitative methods allow a researcher to explore a topic in depth by examining the complexities of the research data in a manner that goes beyond measurement and testing (Creswell, 2014, p. 4); but it is neither practical nor meaningful to simply describe such data in its totality. The researcher must form interpretations about what the data means, and codify those interpretations into findings that will be meaningful to other people. Qualitative research in general, and qualitative research under a constructionist-interpretivist paradigm in particular, acknowledges that the researcher acts as the primary instrument of analysis by constructing their own interpretation of the data (Creswell, 2014, p. 185). This means that a qualitative analysis is always in some sense idiosyncratic to the researcher's perspective, but by following a structured analysis process, the researcher can ensure that they approach the data methodically, and limit the natural human tendency to assume that datapoints fit their preconceptions. A structured process also allows the researcher to expose their decision-making framework so that readers can evaluate the validity of their claims (Braun & Clarke, 2006, p. 93; Creswell & Miller, 2000, p. 128).

The core analysis method used in Study 1 and Study 4 was *thematic analysis*. Thematic analysis is a method for identifying patterns of meaning in a dataset, variations of which have been employed in qualitative research for decades. The specific form of thematic analysis used in these studies was *reflexive thematic analysis* (Braun & Clarke, 2019), codified by Braun and Clarke (2006) as a series of six steps or phases, outlined in Table 3.2. The researcher proceeds through these six steps sequentially, but in a recursive process, such that they often loop back to an earlier step to iterate upon their previous work (Braun & Clarke, 2006, p. 86).

Table 3.2 Phases of thematic analysis (from Braun & Clarke, 2006)

Phase	Description of the process
1 Familiarising yourself with your data	Transcribing data (if necessary), reading and re-reading the data, noting down initial ideas.
2 Generating initial codes	Coding interesting features of the data in a systematic fashion across the entire data set, collating data relevant to each code.
3 Searching for themes	Collating codes into potential themes, gathering all data relevant to each potential theme.
4 Reviewing themes	Checking if the themes work in relation to the coded extracts (Level 1) and the entire data set (Level 2), generating a thematic 'map' of the analysis.
5 Defining and naming themes	Ongoing analysis to refine the specifics of each theme, and the overall story the analysis tells, generating clear definitions and names for each theme.
6 Producing the report	The final opportunity for analysis. Selection of vivid, compelling extract examples, final analysis of selected extracts, relating back of the analysis to the research question and literature, producing a scholarly report of the analysis.

Braun and Clarke (2006, pp. 83–85) note three main variations within thematic analysis: inductive or deductive code development, semantic or latent code content, and essentialist/realist or constructionist epistemology. Study 1 and Study 4 used primarily inductive code development, a bottom-up approach in which codes are created in response to the data rather than according to a pre-set coding scheme. This reflects the research aims, which in both cases were to understand *how* players would respond to an unusual form of gameplay. Inductive code development allowed for the development of a new, holistic understanding of the subject grounded in the data, rather than only collecting answers to specific questions based on a pre-existing theory. The content of my codes was primarily semantic, in that it described the patterns that I saw directly in the data, and reserved theorising about the implications of those patterns for a later stage of analysis.

The epistemology requires some further discussion. My overall epistemology in this research is constructionist, as explained in Section 3.2. However, the themes that I have developed more closely resemble the types of themes that Braun and Clarke describe as characteristic of an essentialist/realist epistemology, in that they report "experiences, meanings and the reality of participants" (2006, p. 85) and do not examine "the ways in which events, realities, meanings, experiences and so on are the effects of a range of discourses operating within society" (2006, p. 81). The discrepancy, as I understand it, comes from assumptions about the object of study. Braun and Clarke write about performing thematic analysis in the discipline of psychology. In psychology, where the object of study is internal psychological processes, a "social constructionist approach" is by definition concerned with the way in which human experiences are socially produced, and this defines what the researcher can meaningfully examine. The studies in Chapters 4 and 7 do concern players' experiences, but only insofar as those experiences are externally expressed in the participants' words and actions. That is to say, for the purposes of these studies I am not concerned with how the meanings that participants ascribed to their experiences arose; but I do approach the research from the constructionist epistemological position, by which I acknowledge that both the participants' descriptions of their experiences and the concepts that I use to describe them *are* socially constructed phenomena, inescapably contingent on meanings that have emerged from social interactions (Crotty, 1998, p. 42).

On inter-rater reliability

Inter-rater reliability is a statistical measure of the level of consistency with which two or more people coded the same dataset (McDonald et al., 2019, p. 2). A metric with roots in positivist research, it is intended to demonstrate whether the findings of an analysis are replicable across coders—a type of reliability that Krippendorff terms "reproducibility" (2004, p. 215). Although there were two coders for the thematic analysis in Study 4, we do not report an inter-rater reliability score for this study. This is in accordance with the guidelines set out by McDonald et al. (2019), who state that inter-rater reliability is not suitable for inductive analyses, in which the objective is to develop emergent themes or theories, nor for analyses in which one coder has particular expertise in the subject matter. Both of these conditions are applicable in the case of Study 4. Chapter 7 explains the process that was followed to reach a consensus on the research themes, which alternated between phases of independent coding and inter-coder discussion.

3.3.5 Content analysis

Content analysis is a research method that involves categorising media texts and other forms of qualitative data, and then systematically analysing the frequency and distribution of categories within the dataset (Sharp et al., 2019, p. 332). It was originally intended to be a strictly quantitative technique, with roots in a method for reducing the contents of newspapers to countable variables that was known as "quantitative newspaper analysis" (Krippendorff, 2004, pp. 3–17). Modern variations on this quantitative approach continue to abound in the digital age, with online media providing effectively endless sources of data about human communication behaviours to be analysed (McMillan, 2000).

Content analysis now appears in qualitative as well as quantitative forms. The distinction between the two is not entirely clear-cut; Krippendorff argues that content analysis is always partly qualitative, since “all reading of texts is qualitative, even when certain characteristics of a text are later converted into numbers” (Krippendorff, 2004, p. 16). In general terms, however, “qualitative content analysis” describes an approach that involves a greater focus on close interpretation than frequency analyses of counted variables (Mayring, 2004).

Study 1 used qualitative content analysis to provide an overview of research participants’ dialogue with a game character through a text interface. The variables included surface-level attributes of the text, such as the length and grammatical structure of each messages, as well as implicit attributes, such as the intended speech acts that each message performed. The content analysis provided a basic description of what the participants had done during the study, which was supplemented by a thematic analysis of the interviews to ascertain what the participants had been thinking and how their attitudes had informed their actions.

3.3.6 Frame analysis

Frame analysis (Goffman, 1974) is a dramaturgical approach to understanding everyday social situations, in which people’s interaction is understood through a broad metaphor of theatre. Goffman argued that just as actors on the stage reproduce a script, so social situations are enacted and experienced as instantiations of more general patterns. Goffman referred to these shared patterns for understanding social situations as “frames”. A frame could be anything that provides an answer to the question “What is it that is going on here?” A major innovation in Goffman’s approach, relative to other conceptions of social patterns, is the observation that situations can be multiply framed. On seeing a chess game, for example, we may answer the question “What is it that is going on here” at different levels, by responding that people in the situation are: moving pieces of wood; playing chess; striving to prove their intellectual superiority; or that one person is distracting the other while their house is being burgled. In Goffman’s terms, an “original” or “primary” behaviour of moving pieces of wood is transformed by the frame of chess, then transformed again by the frame of a battle of intellect, and finally transformed by the frame of a diversion for a home burglary. Frames in which all participants are mutually aware of the framing are described as “keys”, and the process of creating them as “keying”, by analogy to musical keys (Goffman, 1974, pp. 43–45). Frames in which some participants are not aware of the framing, such as the deceptive home burglary frame above, are known as “fabrications” (Goffman, 1974, p. 83).

Among social science methods, frame analysis is unusually well-suited to the study of gameplay. It is a research approach explicitly designed to illuminate the layers of role-playing in which people engage, and gameplay is a situation in which people take on particularly well-defined and distinguished roles. Goffman himself referred to gameplay extensively in his writings (Goffman, 1961, 1974), and a considerable number of scholars have continued to do so in subsequent years.

Fine (1983) adapted frame analysis to the study of role-playing games, identifying how Dungeons & Dragons players alternate between a fantasy frame and a primary frame of real life. Fine identified two reframings in this situation: the original social behaviour of a person was reframed into that of a player, which was reframed again into the activity of a game character. An important observation was the way people moved rapidly between frames, sometimes marked by subtle shifts in spoken style to signify the role-playing frame ("What be the date?" "It's the end of Langala.") or the primary framework of the untransformed social situation ("No, in this world." "August tenth.") (Fine, 1983, p. 202).

In more recent years, researchers have further adapted both Goffman and Fine's work to the study of digital games. Deterding (2009) offered frame analysis as a theory to explain the so-called "magic circle" (Salen & Zimmerman, 2004, p. 95) boundary that differentiates the rules governing gameplay from the rules governing ordinary social behaviour. Linderoth (2012) applied frame analysis to role-players in an online game, and found that the technologically mediated nature of online gameplay created tensions between the untransformed frame of the game and the fantasy frame of the imaginative role-playing, which required effort on the part of the players to overcome. Building upon all of the above, Conway and Trevillian (2015) proposed an ontological model that separates the event of gameplay into three frames: the Social World, Operative World and Character World.

Study 4 employed frame analysis as an interpretive framework for understanding player responses to voice gameplay. The research provides an empirical validation for the model proposed by Conway and Trevillian (2015), and builds upon that model to form the core of one of the main contributions of this thesis (see Chapter 8). Frame analysis did not provide a "method" for the research *per se*. Goffman was infamously vague about the details of his research approach; so much so that an author of several books about the sociologist commented that "Many would question whether Goffman had a method in any of its usual senses" (Smith, 2006, p. 11). Consequently, Study 4 used thematic analysis as the core method for organising and analysing the data, and frame analysis as the theoretical lens through which to interpret the initial findings.

3.4 Research methods for studying games as designed artefacts

This section describes and discusses the research methods used in Study 2 (Chapter 5) and Study 3 (Chapter 6), which focused on *voice games* as designed artefacts. It begins with a discussion of design research and the use of paratexts, and continues with an overview of historical analysis, media archaeology and design pattern analysis. The research methods used to study *voice gameplay* as an activity are described in Section 3.3.

3.4.1 The study of artefacts as designed

Study 2 and Study 3 involved direct analyses of games as designed artefacts. In both studies, the core method of data generation was to compile a database of information on previously published voice games by examining the games themselves and secondary information about the games. The database included commercially published games for entertainment, serious games for education and training, research prototypes described in academic papers, small games by independent hobbyist programmers, microgames produced in a weekend at game jams, fan-made modifications ("mods") for existing games, and several voice control software systems marketed for use with digital games. Detailed information about each game was added to the dataset according to the needs of the two studies.

This type of research has been labelled in various ways. In HCI, it has been described as the study of "the artefact as designed" (Dix et al., 2003, p. 79) or "archaeologically-inspired artefact analysis" (Dix et al., 2003, p. 79). In the field of Design Research, it has been described as "design phenomenology" (N. Cross, 1999, p. 6). Under either title, its purpose is the same: to interpret the knowledge that is implicitly encoded in the end product of a design process, "in the forms and materials and finishes which embody design attributes" (N. Cross, 1999, p. 6), as opposed to the knowledge that can be drawn from designers, users and design processes.

Initially, I conceived of a small study of voice games-as-designed as a preliminary step to be undertaken before the "real" research on voice gameplay. However, once I began the process, I discovered that there was a great deal to learn from past voice games, which reflected important lessons about the history of the format and the variety of ways in which voice input has been used in digital games. I also found that these topics had not been the subject of any substantial reporting, either in the academic literature or in the wider world of professional and hobbyist games media. To study voice gameplay without first establishing a thorough understanding of voice games would be to build a house without a foundation; any conclusion drawn from a particular configuration of gameplay could turn out to be either irrelevant to real-world game design, or already anticipated and solved by designers of existing games. Furthermore, the process of documenting, narrativising and codifying voice games was an inherent contribution to knowledge in this design space.

Paratexts

The term *paratext* was coined by the literary theorist Genette (1997), originally to describe the parts of a printed work that are not the main text, such as a book cover, table of contents and appendix (Genette & Maclean, 1991). It was introduced to game studies by Consalvo (2007, 2017), who applied the concept to the secondary industries that had sprung up around digital games: “Gaming magazines, strategy guides, mod chip makers, the International Game Exchange, Even Balance and other companies” (2007, p. 9). The term has since been used in game studies primarily to refer to secondary media that are about specific games, such as online forum posts (Iacovides et al., 2013), patch notes (Paul, 2010) and gameplay videos (Carter, 2015).

To put together the database on voice interaction game design, I drew upon a large variety of paratexts about voice games. The primary paratexts were an assortment of scholarly papers and books, game manuals, game-related promotional materials, gameplay videos, game reviews, wikis, strategy guides and online forum posts. Details about the processes used to identify and select these sources are included in the published papers in Chapter 5 and Chapter 6.

These paratexts were used to compile a detailed description of the presentation and voice gameplay functionality of games that were impractical or impossible to play directly. Sources that provided unfiltered information about the game’s design, such as lightly edited “Let’s Play” gameplay videos (see Glas, 2015) and game manuals, were prioritised over sources that provided second-hand accounts, such as game reviews and forum posts. Wherever possible, information was triangulated from multiple sources to ensure that the details recorded about each game were accurate.

The majority of the sources consulted were in English. However, I made efforts to search for sources in other languages—mostly through Metacritic, a review aggregator website that collects game reviews in several languages. I used Google Translate to interpret these sources. The games themselves were primarily available in English, often alongside one or more versions in other languages. Only a few dozen games were available solely in a language other than English, primarily either Japanese or a different European language.

3.4.2 Historical analysis

Study 2 used the information compiled about voice games to conduct a history of voice game design, under the aegis of media archaeology. The process that was used to search and select games for analysis in Study 2 is described in the published article (see Chapter 5).

Media archaeology

Media archaeology is somewhat difficult to introduce. Scholars in the field acknowledge that “there is no general agreement about either the principles or the terminology of media archaeology” (Huhtamo & Parikka, 2011, p. 3). It has been described as a “travelling discipline” (Parikka, 2012, p. 15), as it has not found a home within any one parent discipline—and this is reflected in a deliberate avoidance of prescriptive methodologies. The common thread that holds media archaeology together is a focus on the overlooked histories of current media cultures. Media archaeology seeks to understand the influences and alternative pathways that have influenced present media, but have been disregarded in the standard narrative. As part of this, media archaeologies often investigate "failed" and "dead" media formats as case studies, putting these mostly forgotten cases on equal footing with more successful media as a basis for understanding present culture through the lens of the past (Parikka, 2012, pp. 12–14).

While media archaeology often involves a narrow but in-depth analysis of a small number of case studies, Study 2 presents a history of voice games that prioritises encompassing the entire breadth of the format, which precludes it from exploring in depth aspects such as the media-cultural contexts in which the games were developed. According to the typology of game histories proposed by Suominen (2016), it is a genealogical and pathological history: genealogical in that it maps the lineages and categories of voice interaction game that have appeared at different times; and pathological in that it puts forward an explanation of why voice interaction games developed in the ways they did, with reference to material preconditions and conceptual antecedents. It focuses on identifying trends and links between games based on their chronology, genre, geographic area of production and technological platform.

Primary and secondary historical sources

The quality of historical research depends to a large extent on the quality of its sources. One of the most important distinctions to consider is between *primary sources* and *secondary sources*. Primary sources are "original unfiltered evidence" (Benjamin, 2006, p. 27): first-hand accounts or recordings from contemporary witnesses. Reliable primary sources are the *sine qua non* of historical research, as they are the closest a scholar can get to a direct observation of a past situation or object. As Marwick bluntly states: "historical work is generally esteemed serious and scholarly to the extent that it is properly based on the primary sources" (Marwick, 1989, p. 199). Secondary sources are those that report on a situation or object indirectly. They have an important role in historical research as a guide to the topic, a map of the historical territory to steer the historian's search, and a source of alternative perspectives to triangulate an understanding of the subject (Benjamin, 2006, p. 28; Marwick, 2001, p. 26), but do not carry the same weight as primary sources.

The difference between primary and secondary sources is not always clear-cut, particularly in histories of media. Marwick (2001, p. 156) draws a broad distinction based on contemporaneity, in which primary sources are those created within the time period that is being studied. He encapsulates this in a simple maxim: "Rule number one: look at the date!" (2001, p. 27, emphasis in original). Benjamin (2006, p. 26) draws a finer distinction, illustrated with the example of a news report: a written report from a direct eyewitness to an event is a primary source for that event, but a report that draws its information from the accounts of other people is a secondary source; however, the latter would be a primary source if the object of study is the journalistic coverage of the event, or the individual journalist. Benjamin's definition and Marwick's definition appear to be in conflict. However, the apparent disagreement can be resolved by taking into account their implied objects of study. Marwick writes of historians studying a time period rather than a specific event or object; anything from that time period is therefore a direct record of some part of the subject. In contrast, Benjamin refers to the study of a specific event. This shows that, although it is simple enough to define primary sources as those that provide direct evidence, determining what constitutes direct evidence requires some consideration.

Study 2 and Study 3 used five main categories of primary sources to understand the design of each game: the game itself, video recordings of the game, the game manual, game reviews and academic papers. I prioritised examining the game itself over other sources, but as the work encompassed more than 400 games by the completion of Study 3 (many of which required specific platforms to operate), I was able to play only around one in ten of the games directly. For most of the games in the dataset, I was able to find multiple video recordings by different players on YouTube, often in the form of "Let's Play" videos (see Glas, 2015), which provided direct evidence of how the game operated. The publisher's website often provided written manuals for the game that detailed its functionality, including its voice interaction affordances. The publisher's website often provided a written manual for the game that detailed its functionality, including voice interaction affordances. Game reviews provided direct, first-hand accounts of most of the games in both studies; reviews from professional game journalists and critics were prioritised first, followed by user reviews on websites such as Steam and Metacritic. Finally, for the games that were produced as part of an academic research project, the main source was usually a peer-reviewed, published article by the researchers who had developed the game and conducted some form of evaluation. Academic books and papers also provided first-hand accounts of a small number of commercial voice games, including details of their design and functionality. These primary sources were supplemented by a variety of websites that aggregated information about games, ranging from large wikis such as Wikipedia and Giant Bomb to game discussion forums such as GameFAQs and NeoGAF. Those secondary sources were particularly useful for the initial identification of new voice games.

3.4.3 Design pattern analysis

For Study 3, the research aim was to understand major design variations across the category of voice games, rather than lineages and historical phases that connected games across time. The core approach was similar to Study 2, in that it involved the creation of a large database of published games that featured VOICE aspects. The number of games in the database was nearly doubled, from 247 in Study 2 to 471 in Study 3. This reflected several factors: the inclusion of more academic research prototypes and other types of serious games; a large spike in the number of voice games published in the year after the end of data collection for Study 2; and the addition of 37 games that I had overlooked in the previous study (mostly small web-based games by independent game makers, plus several games published only in Japanese that were not well known outside Japan, and a handful of English-language commercial games). The process used to search and select games for Study 3 is described in the published article (see Chapter 6). The full list of voice games analysed is included in Appendix B.

For each game in the database, the following information was collected and recorded for Study 3:

- The year in which it was first published
- Whether or not it was developed as part of an academic research project.
- The type of platform for which it was designed (e.g. "mobile" or "PC")
- The primary videogame genre to which it belongs (e.g. "sports" or "RPG")
- How central voice input is to its overall design (e.g. "central feature" or "optional alternative control scheme")
- Whether it uses verbal or non-verbal voice input
- The general type of speech act that its voice input enables (e.g. "command" or "singing")
- Whether it frames voice inputs as coming from the player (i.e. from outside the diegetic gameworld) or a player-character (i.e. a character that exists within the diegetic gameworld)
- The type of recipient to whom the player speaks (e.g. "a non-player character" or "the game system itself")
- A summary description of the voice-operated gameplay mechanisms

Based on these features, I conducted an iterative process of open coding (functionally equivalent to the inductive thematic analysis procedure described in Section 3.3.4) to develop a *pattern language* that describes the major variations that observed within this population of games.

A pattern language is a framework for identifying, cataloguing and describing common variations in a design field. The concept was first advanced by Alexander and colleagues (1977) in architecture and urban planning, and has since been used in various fields including software engineering (Gamma et al., 1995), interaction design (Borchers, 2001), human-computer interaction (Pan & Stolterman, 2013) and game design (Björk & Holopainen, 2004). A pattern language is a "vocabulary" of individual design patterns. Each design pattern is a formalised explanation of a reusable approach to a design challenge, typically illustrated with one or more real-world exemplars. Other elements of a design pattern vary, but they typically include an idiomatic title, an explanation of the circumstances in which the pattern may be used, a description of what the pattern entails in practice, and an explanation of how the pattern overlaps with or connects to other patterns from the same pattern language.

The purpose of a pattern language is to provide designers and researchers with a set of reusable solutions to design problems. There is some debate over whether these “solutions” should be prescriptive rules, based on normative value judgements about what designers should be trying to achieve, or descriptive showcases, which simply communicate *one possible approach* to a design problem without implying that it is correct or preferable to other approaches. The original pattern language of Alexander et al. (1977) was certainly prescriptive, and the merits of a prescriptive approach to game design patterns has been championed by Dormans (2013) and McGee (2007). However, researchers in game studies have commonly declined to put forward normative pattern languages, and instead presented pattern languages as records of common design approaches. The precedent for this was set by Björk et al., who were among the first to apply the pattern language approach to game design, and who argue that framing design patterns as problem-solving tools makes them inappropriate for use as analytical tools: “not all aspects of design can or should be seen as solving problems, especially in a creative activity such as game design which requires not only engineering skills but also art and design competences” (2003, p. 183). Accordingly, Björk et al. adapted and redefined Alexander et al.’s original pattern template for use in game studies.

The design patterns presented in Chapter 6 follow a slightly simplified version of the template developed by Bjork et al. (2003). This fits with the constructionist epistemology under which I have approached this research. For a constructionist researcher, any argument about how well a design solution works is bound to the perspective from which the judgement was made, and so it is not possible to justify a universal judgement that a certain design solution is generally preferable to others.

3.5 Summary of the research design

The preceding sections of this chapter have described the methodological foundations of this thesis and the research methods that it has employed. This section concludes the chapter with a synopsis of the four studies, outlining the research questions and research methods that are specific to each one.

3.5.1 Study 1: Natural Language Gameplay

The first study was an exploratory investigation of how players engage in conversational interactions with a game character, and what sort of attitudes and expectations this might elicit. The study was designed to answer Research Question 1:

> **RQ1** How do players make sense of and interact with a game character that responds to natural language text?

A Wizard of Oz prototype of a plausible near-future artificially intelligent game character was designed based on a review of recent literature in game AI, natural language processing, human-robot interaction and interactive machine learning. The details of this process and the study method are described in Chapter 4. Participants aged 11–15 (n=18) were recruited to play a game scenario that involved conducting collaborative tasks with the character in the sandbox videogame, *Minecraft*. A qualitative content analysis was performed on logs of the in-game messages to construct a detailed account of how the players used natural language to interact with the game character. Semi-structured interviews were also conducted with the participants. A thematic analysis of the interviews and gameplay recordings produced a set of themes that explain the attitudes, expectations and concerns that informed the players' actions.

3.5.2 Study 2: Voice Game History

The second study was a historical analysis of voice games as a media format, conducted under the rubric of media archaeology. The study was designed to answer Research Question 2:

> **RQ2** What is the historical lineage of voice game design and what are the major influences that have shaped it?

An extensive array of primary and secondary sources (see Section 3.4.2) were analysed to construct a database of information about 247 voice-operated videogames, published between 1973 and 2016. Chapter 5 presents a narrative history of voice games based upon this database, which describes seven distinct historical phases of voice game design distinguished by a combination of time period, design focus and geographic centres of activity. It

identifies an ebb-and-flow pattern to the production of voice games, indicating that voice game design has been highly reactive to changes in the hardware and software platforms that predominated in each time period. Finally, it identifies systematic differences in the ways that videogames from Japan and videogames from Western countries have presented voice gameplay.

3.5.3 Study 3: Voice Game Design Patterns

The third study involved the creation of a pattern language for voice game design. The study was designed to answer Research Question 3:

> **RQ3** What design patterns are currently implemented in voice games?

For this study, the database of voice games used in Study 2 was substantially updated and extended, nearly doubling the number of videogames included to 471. New information was encoded about every game in the database, to describe its format, development, presentation and mechanisms. As in the previous study, this information was drawn from a variety of primary and secondary sources. A process of inductive open coding on this database produced a pattern language consisting of 25 design patterns for voice games. The patterns range from specific game mechanisms to high-level structures of voice input. Furthermore, an analysis of the co-occurrence of these design patterns within games identified five genres of voice gameplay, each based on a different type of voice interaction. The complete pattern language and genre analysis are presented in Chapter 6.

3.5.4 Study 4: Voice Gameplay

The fourth and final study was a comparative investigation of the player experience of different voice games, using frame analysis as an interpretive framework. The study was designed to answer both Research Question 4 and Research Question 5:

> **RQ4** What is the distinctive nature of voice gameplay?

> **RQ5** How do players experience verbal and non-verbal voice gameplay?

Adult participants (n=24) were recruited in pairs to play three three commercially published voice games. The set of games was chosen to offer both verbal and non-verbal voice gameplay, and to reflect a range of aesthetic game genres. Participants played each game using both voice input and a traditional manual control scheme, and were interviewed in pairs after playing each game. The interview recordings were analysed by two researchers using inductive thematic analysis, with frame analysis as an interpretive theoretical framework. Chapter 7 presents the two main outcomes of this research: a set of seven themes that characterise the player experience of voice gameplay, and a general model of gameplay based in frame analysis that accounts for the experiences that participants described.

4 Natural Language Gameplay

This chapter presents Study 1 in the form of a peer-reviewed journal article:

> Allison, F., Luger, E., & Hofmann, K. (2018). How Players Speak to an Intelligent Game Character Using Natural Language Messages. *Transactions of the Digital Games Research Association*, 4(2), 1–50. https://doi.org/10.26503/todigra.v4i2.88

The article is included verbatim in its published format, following a discussion of how it contributes to the thesis. I led all stages of the research and writing process. Data collection was conducted during a 12-week research internship at Microsoft Research Cambridge, UK. My supervisors for that research internship, Ewa Luger and Katja Hofmann, contributed to the study design and writing of the paper.

4.1 Introduction

The preceding chapters introduced and contextualised the topic of this thesis: voice-operated interaction with characters and environments (VOICE) in games. They also note that I began the research by exploring the broader topic of "conversational interaction" in games, encompassing both text and voice interaction, before refining the focus to voice gameplay. This chapter presents the results of Study 1, which investigated player interactions with a game character using natural language text dialog. The study addresses Research Question 1:

> **RQ1** How do players make sense of and interact with a game character that responds to natural language text?

To answer this question, I devised and conducted a lab-based study in which participants interacted with a Wizard of Oz prototype (Kelley, 1984, 1983) of a game character that could respond appropriately to natural language messages, sent via a text dialogue interface. The study makes three main contributions to the HCI and game studies literature:

- A detailed account of relatively spontaneous player approaches to natural language interaction in a game, indicating some of the things that players might do with and expect from natural language gameplay when they have not been given detailed instructions.

- A thematic analysis of difficulties and concerns that arose from attempting to interact with a game character through natural language, based on both gameplay observations and post-game interviews.

- A hypothesis about two key variables that differentiated player reactions to a "talking" game character, namely the degree to which they engaged with the character and the degree to which they anthropomorphised the character.

Study 1 is distinct from Studies 2–4 in that it involves text-based gameplay, rather than voice gameplay. I chose to study text-based gameplay in the first study on the basis that it would allow me to develop an understanding of language-based gameplay in isolation from sound-related factors, such as speech recognition errors, player embarrassment at needing to speak aloud, and the pitch, tone, accent and gender of the game character's voice. The results bear this out, to an extent: some of the player experience themes documented in this chapter echo themes that were later observed in voice gameplay (Chapter 7), as Section 4.4 describes. However, many of the detailed findings are particular to text-based interaction, such as the disruptive effect of spelling errors and the way in which punctuation is used to indicate intent.

This sections that follow set the context for Study 1. They describe its research method, summarise its main findings, and discuss how its results have contribute to the aims of the thesis overall. The chapter concludes with a report on the study and its outcomes in the form of a published peer-reviewed journal article.

4.2 Research method

Participants aged between 11 and 15 (n=18) were recruited, all of whom had prior experience with the videogame *Minecraft*. This age group was chosen to represent a demographic in which playing *Minecraft* was especially common, following consultation with one of the game's developers. Each player was asked to perform a series of tasks in the game with the assistance of an artificial intelligence (AI) agent, which was represented as a game character with a robotic appearance. During one of the tasks, players could write to the AI agent using natural language messages through a text interface. In response, the agent would send a simple text reply and update its current behaviour according to what the player wrote. Players were informed that the agent would attempt to provide them with assistance, and would also learn to adapt its behaviour based on their feedback. After each task scenario, players were interviewed about their experience.

The "AI" agent was in fact a Wizard of Oz prototype (Kelley, 1984, 1983): a system designed to represent a collection of plausible near-future AI behaviours, but secretly controlled by a hidden human researcher. The agent's behavioural script was formulated based on a review of recent research literature in videogame AI, natural language processing and interactive machine learning, as well as two interviews with expert game developers. Players were given intentionally minimal instructions for how to interact with the agent. The aim of this approach was to elicit relatively spontaneous responses to an AI character that reacts intelligently to natural language.

A content analysis was conducted on the logs of the natural language messages. The coding scheme was informed by both conversation analysis (Sacks et al., 1974) and the theory of speech acts (Searle, 1969): it recorded each message's syntax, function, subject and direct object, as well as whether it included an indicator of amounts or locations, or involved conversational breakdown or repair. This coding was used to create a detailed description of the types of messages that players sent. This was supplemented by a thematic analysis of the interviews, which developed the themes described in the next section.

4.3 Summary of findings

The content analysis showed that players' use of the text dialogue varied considerably, in terms of message frequency, message length and specific wording. The majority of messages were in imperative syntax, with a substantial minority in interrogative or declarative syntax. Most messages were intended as commands, while a small proportion were intended as cancellations, queries, explanations, acknowledgements, self-corrections or tests. The paper describes a number of detailed findings about the grammatical construction of the various types of messages, as well as their contexts of use.

One such finding concerned the messages with interrogative syntax. Some of these messages were intended as genuine questions, but a majority were intended as polite commands. This distinction is easy for humans to make, due to the cooperative principle (Grice, 1975), but potentially difficult for computers. The content analysis found that genuine questions were far more likely to end with a question mark than polite commands, providing a potential textual clue to the player's intention. Players also stated in the interviews that they had attempted to use politeness as a form of positive feedback to reinforce certain behaviour by the AI agent.

Players encountered several types of recurring difficulties with the text dialogue. They found communicating amounts, locations and object boundaries to be challenging, both because it was difficult to formulate sufficiently precise commands, and because they did not trust the AI agent to interpret ambiguous instructions according to their wishes. Conversely, players were often unsure how to interpret or investigate the meaning of responses from the agent that they found ambiguous. Despite these difficulties, nearly all players stated that they found the conversational interaction enjoyable to use, particularly because it created an impression that the agent character was alive and understanding.

Players exhibited varying expectations about the agent's capabilities, reflecting differences in their mental model (see Norman, 1983) of the nature of its intelligence and how it processed the game environment. In the interviews, players cited a range of external cues that informed these expectations. Common cues were: analogy to human intelligence, prior experience with AI characters in games, prior experience with other humans in games, knowledge of programming languages and game cheat codes, and depictions of AI in science fiction films. Notably, players often did not test their mental models carefully, but instead operated on the assumption that their expectations were accurate until they encountered evidence that something was not working as intended. This reinforces the importance of clear early cues to set player expectations.

Players' attitudes towards the AI agent varied on two primary axes. The first axis was their level of engagement, expressed through factors such as whether they spent a great deal of time interacting directly with the agent during the game and asking questions about it during the interview, or largely ignored it during the game and displayed little interest in it during the interviews. The second axis was the degree to which they anthropomorphised the agent, by expressing empathy towards it and reacting to it as though it had human-like thoughts and feelings. These two variables were at least partially independent: some players appeared to be relatively high or low on both axes, while others appeared to be relatively high on one axis and relatively low on the other.

4.4 Contributions to the thesis

This study addresses Research Question 1: "How do players make sense of and interact with a game character that responds to natural language text?" The results provide a detailed account of players' communication behaviours, in terms of both the actions themselves and the intentions behind those actions. The study indicates the categories of "speech acts" that players see natural language interaction as affording, as well as some of the difficulties that players encounter in attempting to express their wishes.

One of the conclusions of the study is that players' expectations of a language-understanding game character are strongly influenced by contextual cues. Some of the cues identified in this study relate specifically to *Minecraft*: the normal behaviour of its AI characters, the social practices of its online player community, and the format of the cheat codes it accepts. Other cues relate to the wider world: human intelligence, human body language, knowledge of programming languages, depictions of AI in films. Because these factors influence the players' responses, it follows that player responses to conversational interaction are likely to vary under different conditions. This limits how the results of this study can be generalised to the overall research question of the thesis, since voice interaction may elicit a different set of associations for the player than text interaction. However, the larger finding of this study remains relevant: that players interpret conversational interaction with a game character through a variety of lenses, not all of which are prompted by the game itself, and that these lenses cause the player to experience gameplay in ways that cannot be predicted based on the mechanics of the dialogue system in isolation.

This core finding informed the design of the studies that followed. Both Study 2 (Chapter 5) and Study 3 (Chapter 6) are based upon a wide-ranging analysis of voice game design, aiming to understand not only how videogames use voice interaction in a mechanical sense, but also how videogames frame the meaning of voice input by depicting its effect on the gameworld. For Study 4 (Chapter 7), the research aim was to understand how players understand and interpret voice gameplay, using interviews and frame analysis to develop a theoretical model of their game experience. Two of the findings of Study 4 strongly echo the findings of Study 1: that spoken interaction elicits a heightened impression of game characters being alive and aware of the player, and that players often struggle to connect their voice commands to specific places and actions in the gameworld. Collectively, the results outlined in this section indicate that text interaction cannot be treated as a direct equivalent of voice interaction in games, but that these two forms of language-based gameplay carry some similar usability challenges and a shared potential to imbue game characters with a sense of social presence.

1.

How Players Speak to an Intelligent Game Character Using Natural Language Messages

Fraser Allison, Ewa Luger, & Katja Hofmann

Transactions of the Digital Games Research Association

November 2018, Vol 4 No 2, pp 1-49

ISSN 2328-9422

***All authors were at Microsoft Research Cambridge, UK, during data collection stage.**

ABSTRACT

AI-driven characters that learn directly from human input are rare in digital games, but recent advances in several fields of machine learning suggests that they may soon be much more feasible to create. This study explores the design space for interacting with such a character through natural language text dialogue. We conducted an observational study with 18 high school students, who played Minecraft alongside a Wizard of Oz prototype of a companion AI character that learned from their actions and inputs. In this paper, we report on an analysis of the 186 natural language messages that players sent to the character, and review key variations in syntax, function and writing style. We find that players' behaviour and language was differentiated by the extent to which they expressed an anthropomorphic view of the AI character and the level of interest that they showed in interacting with it.

Keywords

Natural language, AI, human-agent interaction, Wizard of Oz, Minecraft

INTRODUCTION

In recent years, advances in machine learning have driven rapid improvements in artificial intelligence (AI) techniques, both within and outside of videogames. Game worlds make effective testing grounds for learning AI due to their ability to simulate real-world challenges in incremental, constrained stages in a controllable and measurable environment ("Why AI researchers like video games" 2017). As a result, games have been the focus of a great deal of AI research involving machine learning (Yannakakis & Togelius 2015; 2017), from digitised board games (Silver et al. 2016) to real-time arcade games (Mnih et al. 2015; Shaker et al. 2013) and three-dimensional gameworlds that more closely approximate physical space (Johnson et al. 2016).

Despite this, game developers have been slow to take up machine learning techniques for in-game character AI.Games that do feature AI characters that learn from player inputs have been either notable for their novelty, such as *Black & White* and *Forza Motorsport*, or confined to academic projects, such as *NERO:*

NeuroEvolving Robotic Operatives (Stanley et al. 2005). Both the tools and the imagination needed to change this are surfacing. On the tools side, several companies including Unity (Juliani 2017) and Microsoft (Johnson et al. 2016) have released open-source platforms that facilitate the training of machine learning agents in game environments, and these platforms can be adapted for imitation learning (in which agents learn from actions demonstrated by a player or another agent) and reinforcement learning (in which agents learn from a reward signal provided by the environment, such as a game score). On the imagination side, the idea of AI characters that learn from players is becoming more prevalent in games. For example, the 2017 stealth-action game *Echo* is built around the design conceit that enemy characters learn and replicate the actions taken by the player (Robertson 2017). And in 2014, *Middle-Earth: Shadow of Mordor* featured a system in which individual enemy characters were permanently transformed and adapted by their encounters with the player (Taljonick 2014). However, these conceptual examples do not appear to be driven by machine learning technology in any real sense. Shaping the user experience for a game character that actively learns through interaction with a player remains challenging, as there are concerns that the interaction may be inconsistent, difficult for players to understand, or simply not fun (Muñoz-Avila et al. 2013; Yannakakis & Togelius 2015).

The present study is positioned as a pilot study designed to support the long-term aim of developing AIs that can learn to make sense of complex game environments (Johnson et al. 2016), especially in multi-agent settings that include AI and human players. Its purpose is to take a speculative look into the near future, and consider what the user interface considerations would be for an AI character that dynamically changes its behaviour based on what it learns from the player. As this character would have the potential to be less predictable and more adaptable than a character with static AI, an ideal interaction modality should allow for a wider variation of player inputs than a traditional gamepad. It should also be expressive enough to convey detailed state changes in the AI back to the player, including ones that may not be obvious in the character's actions. Despite this increased scope for complexity, it should remain intuitive enough for the player to understand and formulate new input combinations without needing to spend a great deal of time learning how to do so. A natural language dialogue system is a common proposal for an interface that meets these criteria, as evidenced by the near ubiquity of spoken interaction among science fiction AIs, from the androids in *Westworld* to Holly the ship's computer in *Red Dwarf*.

Natural language interaction presents an intriguing mode for interacting with relatively independent and teachable AI agents. In principle, it could give players great flexibility to direct and interact with an AI character in an "off-script" fashion, without the need to learn and navigate an extensive graphical or physical user interface. And language understanding systems have improved rapidly in recent years, driven by the success of neural network models of machine learning.[1] Machine reading comprehension has approached and even exceeded human standard in some constrained scenarios (Eckersley et al. 2017), although open domain language understanding by computers remains far below human level. However, natural language interaction also introduces design challenges. For example, the conversational mode of dictating action to a computer is a departure from the widely accepted interaction paradigm of "direct manipulation" (Shneiderman 1982), in which the player's point of control is represented as a clear and direct link between the physical controls in their hands and a singular locus of manipulation in the computer system or gameworld (Bayliss 2007). This unfamiliarity and the unbounded nature of natural language can make it difficult to immediately formulate the "right" thing to say, and can require a process of learning how to adapt one's phrasing to the system

1. Many of the recent improvements in natural language processing have come from the use of neural network models, particularly recurrent or recursive neural networks (RNN), and specifically variations of long short-term memory (LSTM) and gated recurrent unit (GRU) RNN models. For a full explanation of the use of neural network models in natural language processing, see Goldberg (2016).

(Luger and Sellen 2016, 5289). On the developer's side, the openness of natural language can make it difficult to anticipate what kind of syntax and concepts players will use.

To contribute towards addressing these challenges, we conducted a study of players interacting with an autonomous, learning-capable game character using natural language. This study employed the Wizard of Oz research method, which involves presenting participants with a convincing replica of an automated system in which some of the functions are secretly operated by a human (Kelley 1983). We designed a conceptual prototype of a plausible near-future AI agent that learns from player actions and uses natural language text messages to communicate, which we named help_bot. We invited participants to perform tasks in the videogame, Minecraft, with the assistance of help_bot, and provided intentionally minimal instructions for how to do so, so that we could observe the ways that players spontaneously attempted to engage and speak with an agent of this type. We conducted a content analysis on the natural language messages that players sent to help_bot, to study what type of syntax they used, what kind of commands they gave and how the use of language as opposed to traditional game controls created opportunities and problems for the interaction.

Given the rapid progress in machine learning-based natural language processing, it is our belief that these techniques have the potential to dramatically change in-game interaction using natural language. However, it is extremely unlikely that the available technology will immediately jump to human-level language understanding; much more partial and constrained language interactions appear likely in the near term. Therefore, potential usage scenarios and player perceptions and behaviours need to be thoroughly understood, to allow game developers to craft language-based interactions that suit players' expectations and desires within the constraints of the available technology. We see the key contribution of this work in mapping this space of player perceptions and behaviours, paving the way towards the required understanding and development of novel designs.

PRIOR WORK

In the past decade, there has been a steady increase of research interest in AI applications in digital games (Yannakakis & Togelius 2017, 19-20). Much of this work has focused on training game-playing agents using reinforcement learning (RL), in which an agent is taught to associate combinations of actions and environmental conditions with a reward signal (such as the character's health or a score counter), and learns through trial and error to maximise the reward signal through its choice of actions (Sutton and Barto 1998). Famously, Google DeepMind used a combination of RL and supervised learning from expert human moves to train a Go-playing program (Silver et al. 2016) that beat one of the world's best human Go players, Lee Sedol. RL research is also being conducted to train agents in real-time digital games, from older Atari 2600 games (Bellemare et al. 2012) to more recent games with complex spatial environments such as Doom (Kempka et al. 2016), Starcraft (Farooq et al. 2016) and Minecraft (Johnson et al. 2016). A related branch of research has looked at developing agents that learn from player actions, either from pre-recorded play data or through direct interactions with players. The goal of this work can be to learn higher-level performance strategies, create more convincingly human-like game characters or adapt to individual players' preferences and playstyles (Bakkes et al. 2012).

Studies in interactive machine learning look at scenarios in which a human actively provides feedback to a learning agent to update its behaviour. Researchers in this area have consistently found that users exhibit strong preferences for teaching styles that do not always align with the learning model of the agent (Amershi et al. 2014). Whereas RL-based agents are often designed to learn from explicit feedback on their recent actions, human teachers give relatively little explicit feedback, and instead focus on communicating the desired behaviour conceptually through demonstrations and positive prompts (Amershi et al. 2014; Kaochar et al. 2011; Knox et al. 2012). When required to give repetitive and simplistic input, users often experience impatience and frustration, and a resulting decline in their performance as teachers (Cakmak et al. 2010; Guillory and Bilmes 2011). In a study by Fischer et al. (2013), human users were better at adapting their teaching behaviour for a learning robot when the robot's feedback mimicked the human's social behaviour (in the form of gaze), which indicates that their mental models of how the robot was learning and attending to things were influenced by their knowledge of human learning and attention. Similarly, a study by Koenig et al. concluded that human users' failures to adapt their teaching behaviour effectively based on feedback from a robot learner resulted from a "tendency to map a human-like model onto the capabilities of the robot" (2010, 1111). A review of interactive machine learning studies by Amershi et al. (2014) concluded that a wide range of interactions are possible for human teaching of agents, but studying the human users of these systems will be critical to ensuring their success.

A common barrier to studying user behaviour with both intelligent agents and natural language interfaces is the difficulty in implementing such systems to a high level of reliability. When the research question is not how users respond to the current state of the art but how they would respond to a hypothetical version of the technology, implementing the technology can be prohibitively difficult or expensive. To circumvent this, researchers in human-computer interaction often implement Wizard of Oz prototypes instead. In the Wizard of Oz method, an interface is presented to the user as being fully automated, but is operated out of sight by a human facilitator without the user's knowledge (Maulsby et al. 1993). This approach, first developed for studying user responses to natural language interfaces (Kelley 1983), is also commonly used for studying user interactions with intelligent agent systems (Goodrich and Schultz 2008; Riek 2012). Bernotat et al. (2012)

used a Wizard of Oz design to test how people responded to a futuristic “smart home” without specific instructions, and found that most users defaulted to speech control, demonstrating that language-based interaction is associated with intelligent systems in the public imagination. In another Wizard of Oz study, Xu et al. determined that users could recognise an unsignalled change in the behavioural pattern of an agent, and adapt their own behaviour to suit. These studies demonstrate that the Wizard of Oz approach is well suited to an exploration of how users interact with intelligent agents, particularly in language-based interactions.

A great deal of research has been conducted on natural language interfaces, but for the purposes of this study we are primarily interested in studies of user behaviour in natural language interactions with embodied virtual characters. Most prominent in this field is the work of Cassell, who formulated the concept and early prototypes of the “embodied conversational agent” (2000). Cassell focused on the role of non-verbal behaviours in sustaining the experience of human-like conversation, and the ways in which these factors make conversation fundamentally multimodal. Mateas and Stern (2005) incorporated expressive and affective embodied conversational agents into an interactive drama game, Façade, which was built around natural language interactions; Sali et al. (2010) compared this version of Façade with alternative versions wherein the player selected dialogue responses from a menu rather than typing their own, and found that although the natural language interface generated frustrating errors and reduced players’ feeling of control, it was still the most preferred modality as it provided the greatest sense of presence and engagement. More recently, Lessard has produced several natural language interaction games designed around “conversational puzzles” (2016, 6), making conversation itself more of a game mechanic than a pseudo-social interaction. Lessard concludes that the natural language interaction in his games is easy for players to understand and to start playing with, but that the highly scripted nature of current game dialogue systems restricts the ability of games to take advantage of more emergent gameplay possibilities that would theoretically be possible with natural language.

WIZARD OF OZ CHARACTER DESIGN

Following a review of literature on AI research in games and other fields of application, we extrapolated a set of abilities that we thought represented a reasonable approximation of what an autonomous agent in a game-world such as Minecraft could be made capable of within a few years' time. We named this hypothetical agent help_bot, and defined its abilities in a manner that we could represent through a human-controlled character.

Help_bot could "see" the same visual input as a player, and use this vision to understand and navigate unfamiliar terrain. It could recognise simple objects by sight within the game-world, including objects that were defined items in that game (such as the block types in Minecraft) as well as geometric shapes and patterns formed from the arrangement of objects in the game. It could add new objects to its recognised list through being given labelled examples (such as learning to associate a new shape with the label "pyramid"). It could learn and imitate behavioural patterns by watching the actions of a player-controlled avatar, and update its behaviour based on positive and negative feedback from the player, as well as behavioural prompts such as being hit or being given a particular tool or material. It had a limited ability to infer a higher-level goal from a player's actions, such as predicting what larger shape a player might be constructing from the initial placement of a few blocks.

Notably, help_bot was designed to be a "friendly" or companion character, in contrast to the majority of AI-controlled characters in digital games who take an "enemy" or oppositional role. Help_bot's behaviour followed a simple loop. Its starting state was to follow the player's avatar from a short distance and observe what they did. Periodically, it would categorise the player's current action (e.g. building with bricks) and infer a short-term goal of that action (e.g. building a straight wall out of bricks). Help_bot would then attempt to assist in that task by continuing the action, such as by adding more bricks onto the wall to extend it in the same direction. At irregular intervals, or when prompted by direct interaction by the player, help_bot would reassess what the player was doing and either continue its current action or choose a new action accordingly.

The player could override this behaviour by sending help_bot messages through Minecraft's built-in chat channel. Help_bot understood natural language input through this channel, within constraints. It could distinguish between commands, questions, statements and acknowledgements and choose an appropriate response. It looked for verbs in a message that matched an action in its behavioural repertoire, such as "build", "follow" and "attack", and it looked for a grammatical subject and object to determine what the verb referred to. In this way, long or fragmented sentences could have their meaning inferred from key elements without fully understanding every word, but more nuanced or obscure meanings would not be understood. Messages that were understood prompted standard responses from help_bot: <ok> for commands, <done> after the command was completed, and <yes> or <no> in response to questions. Messages that were not fully understood prompted a request for clarification: <show me where>, <show me how> or <?> (see *Responses to prompts for additional information*).

In accordance with the Wizard of Oz research protocol, help_bot was secretly controlled during the study by a researcher in a separate room. As described above, the Wizard of Oz approach is often used for research on intelligent agent and natural language system prototypes (Goodrich and Schultz 2008; Riek 2012). Indeed, the method was first developed for studying the responses of "computer-naïve, first-time users" (Kelley 1983, 193) to a natural language application. One of the considerations of this method is that the deception should not be too obvious, which goes hand-in-hand with ensuring that the prototype is not unrealistically high-performing. In our case, the researcher controlling help_bot was instructed not to make its behaviour appear too intelligent or natural. They controlled its movements entirely through a keyboard, rather than a mouse and keyboard, to reduce its fluidity of movement. Occasionally, the researcher made help_bot make deliberate "errors" by choosing actions that were plausible but against the player's apparent intentions, such as building over an open space that the player had created. This was to reinforce the impression that help_bot was computer-controlled, and to allow us to observe how players attempted to correct unwanted behaviour. To further support the impression of being an AI-controlled character, help_bot was given a robotic appearance.

METHODOLOGY

We recruited students from two high schools in the United Kingdom to participate in an observational user study. Excluding two participants who dropped out, we had 18 participants (11 female, 7 male) aged between 11 and 15 complete the study. As a rough indicator of sufficiency, this is equal to the mean sample size for in-person user studies presented at the ACM Conference on Human Factors in Computing Systems in 2014 (Caine 2016, 986). Parental consent was obtained for all participants, and parents were given the option to be nearby and observe the study. All participants were required to have played Minecraft before, and their level of experience varied from a few hours to over a hundred hours of play.

The study was conducted across two weeks, with each participant in a separate session. Each session lasted approximately 90 minutes. After greeting the participant and their parent and explaining the study, the facilitator showed the participant to a private room with a computer running Minecraft and a video camera. Participants were asked whether they would like to opt out of having their image appear in any publications about the study, which two did. The facilitator then interviewed the participant briefly about their experience with Minecraft: how often they played, which game modes and activities they preferred, and whether they played in single-player or multiplayer mode. Each participant was set three building tasks to complete in sequence, in a pre-saved Minecraft gameworld created for the purpose.

The first task was to build a model boat, without assistance. This was a warm-up task, which gave the player a chance to get used to the game controls if they needed to, and to become accustomed to the study environment. It also allowed the researchers to observe how the player behaved in “normal” solo Minecraft play, to understand their habits and strategies so that they could be contrasted with how they played in the subsequent tasks. Five minutes was allowed for this task.

After the first task, the facilitator explained that they would be introducing an AI assistant character named “help_bot” into the world for the following tasks. Help_bot was described as an experimental prototype developed by the researchers, which could learn how to act in Minecraft by watching and interacting with players. Players were told that help_bot would try to assist them in their next building task, and that they could teach it or show it what to do if they wanted.

In either the second or third task, the player was also told that help_bot could understand messages that were sent through Minecraft’s chat function. If the text interaction was introduced in the second task, for the third task the player was told that the text interaction was disabled. The order of the language and non-language interaction conditions was rotated, ensuring a balanced allocation of age, gender and previous Minecraft experience for participants.

The facilitator deliberately avoided giving specific instructions or examples of how to interact with help_bot. Before the non-language-input condition, the player was told that help_bot would learn from what it saw them do; that it would try to help them with whatever they were doing; and that they could teach it things or show it what to do. Before the language-input condition, the player was again told that help_bot would learn from what it saw them do, and also what they wrote in the chat channel; that it could understand normal sentences; and that they could try to tell it what to do, teach it things, or give it feedback on what it had done. The instructions did not specify that players were required to interact with help_bot, and once the task had begun the facilitator did not direct the player further except to answer questions. Players who ignored help_bot or lost interest in their building task and moved on to other activities were allowed to do so.

Both the second and third task lasted 15 minutes (although players were allowed to go overtime by up to five minutes to finish what they wanted to do in the game). The instruction for the second task was to build a house, and the instruction for the third task was to add to it with a construction of their choice. After the first eight participants this task was changed, as most participants were familiar with the task of building a house from previous Minecraft experience, and so created it too quickly and often with little planning required. For the remaining participants, the instruction was to build a maze.

After each task, the facilitator conducted a short semi-structured interview with the player, prompting them to explain what they had been thinking at various moments. Players were also asked what they thought of help_bot; what strategies they used when they wanted help_bot to change its behaviour; how well they felt help_bot understood what they wanted, and what made them think so; how playing with help_bot compared to playing with another human; and what features they would change or add. After the final interview, the Wizard of Oz research approach was explained and participants were informed that help_bot had been a human-controlled character. Prior to this debriefing, no participant indicated a suspicion that help_bot may not be computer-controlled.

Content analysis of message logs

Across the 18 sessions of the study, players sent a total of 186 messages through the chat channel. These messages, along with the responses from help_bot, were saved in a log file. We conducted a content analysis on these messages to study the language players had used. These messages were analysed and coded by the first author in an iterative open coding process. Each message was coded according to seven main categories:

- Message syntax (e.g. interrogative)
- Message function (e.g. query)
- Message subject (e.g. “I”, “you”)
- Message direct object (e.g. “me”, “this”)
- Amount indicator (e.g. “a”, “some”)
- Location indicator (e.g. “here”, “back”)
- Repair process (e.g. reformulation)

Our analysis approach was informed by conversation analysis (Sacks et al. 1974), as the high-level goals of this study are to some extent aligned with the goals of conversation analysis. We focus on understanding the structures and patterns that can be discerned among pairs or longer sequences of “utterances” (to use the conversation analysis term), and how the structure of messages relates to the ways in which they are employed to elicit specific actions, rather than pure exchanges of information. We also look at players’ strategies for repairing failures of communication through their messages (Schegloff et al. 1977). However, the full method of conversation analysis is not suitable for this context, as it is substantially concerned with the mutual organisation of dialogue, ordinarily in the form of verbal speech. In this study, by contrast, the organisation of dialogue was one-sided, with the player initiating and directing nearly all of the conversation under the expectation that help_bot would act as a passive responder.

Accordingly, we draw on conversation analysis conceptually in the definition of codes, but pragmatically take a content analysis approach that is more tailored to the log transcript data available.

In analysing the messages as action-oriented inputs, we consider each one as a “speech act” (Searle 1969), intended by the player to serve a functional purpose. We note that this purpose can be different to the literal meaning of the sentence, and so a complete analysis must determine the true function of a message from contextual information. We draw on both the in-game context and players’ interview comments to infer the function of each message, and we identify indirect speech acts (Searle 1975) in which the literal and functional meaning of the message disagree.

We do not intend through this analysis to lay down firm or fixed rules of conversational procedure for natural language-based interaction. However, as Button and Sharrock (1995) argue, we believe that examining the function and form of players’ spontaneous natural language messages will provide useful guidance for natural language interactions with game characters, by pointing out what naïve players might want or expect such a character to be able to understand and respond to appropriately.

RESULTS

The focus of this paper is on the natural language interactions between players and help_bot. However, these textual inputs were highly multimodal with the players’ actions using the traditional game controls, so some discussion of these actions is also required. To capture this, we use the following notation when describing interactions from the study: text inputs in angle brackets, verbal comments in quotation marks, and physical or in-game actions in italics. For example:

Player:<come back and give be the wood>

Help_bot:<?>

Player: “Question mark. So do I have to do every command in one line? Because I did two commands there.”

Player reads their previous message again.

Player: “Oh, because I wrote ‘give be the wood’.”

The intended meaning of a text command was often dependent on the virtual space in which the player’s avatar and help_bot were standing. Some commands directly referred to a visible object, as in <kill the zombie>, while others carried an implicit expectation that they would be carried out in the nearest relevant location, as in <can you get some wood please>. A few commands were paired explicitly with the player’s actions, as in <copy me>, or with help_bot’s actions, as in the following exchange:

Player:<break blocks>

Help_bot mines blocks directly in front of it, leaving other blocks above.

Help_bot:<done>

Player:<there is some at the top>

Help_bot:<?>

Player:<look up>

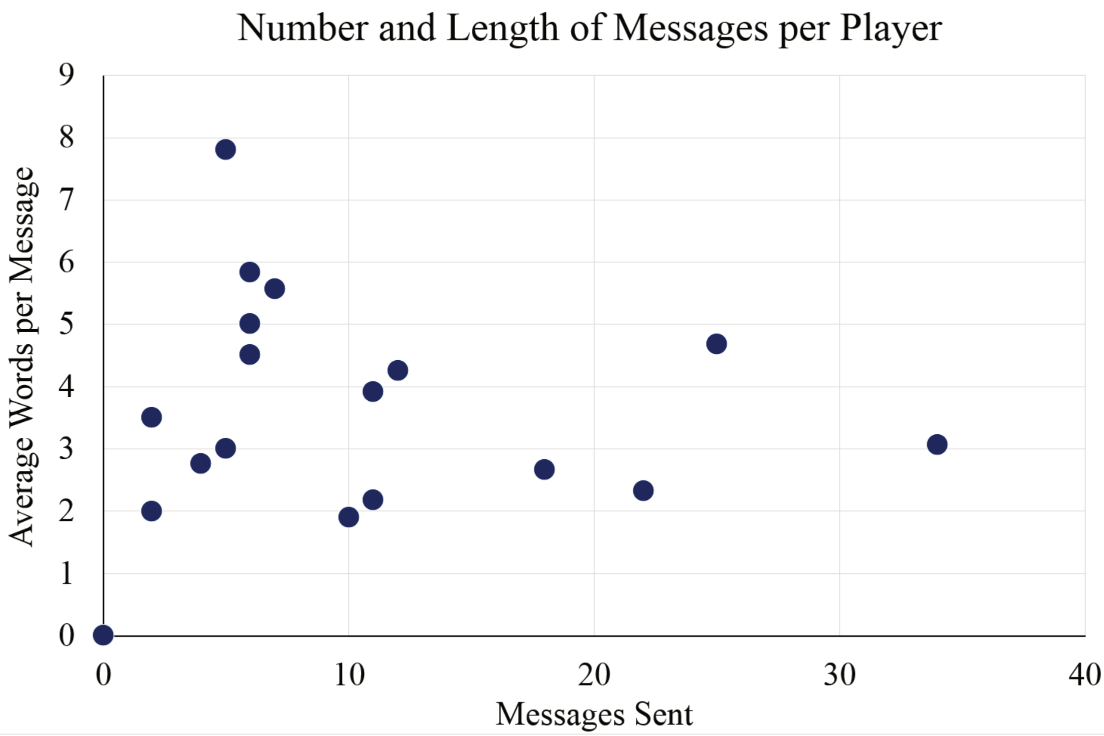

Figure 1: Scatter plot of players in the study by the number of messages they sent (horizontal) and the average length of those messages in words (vertical).

The extent to which players engaged with help_bot through the natural language system varied considerably (see figure 1). Several players sent less than five messages in total, including one who sent none at all. Conversely, several players sent help_bot more than one message per minute across the 15-minute task, with 34 being the highest.A few players wrote full sentences of up to 11 words (see figure 2), including polite phrasings and compound sentences, but the majority wrote primarily in terse phrases that consisted of no more than three words. There was no strong correlation between the length of a player's messages and the number of messages they sent.

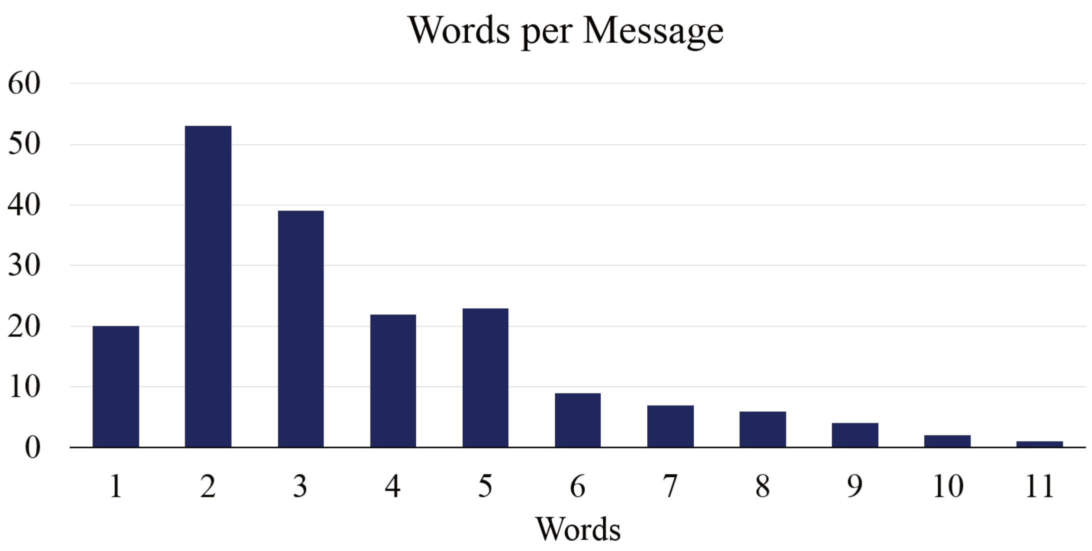

Figure 2:Distribution of messages by the number of words per message.

In the post-task interviews, some players said that they had kept their messages short to be sure that they would be understood. Although they had been told that help_bot could understand normal writing, players pre-emptively avoided longer or more complex sentences on the assumption that they would not be understood, as one player explained:

> "It's quite hard to get the wording correctly to get it to do stuff. Because like when I told it to bring in the material I was going to say 'collect this and give it to me' – I wasn't sure if it would understand that. So you have to be quite simple."

At the same time, however, players were concerned that their messages could be lacking in necessary detail, and that help_bot might make incorrect assumptions due to a lack of specificity in their instructions:

> "[I needed] to be quite simple and not overcomplicate it, so it would understand. [. . .] Probably if I were to say 'collect wood', it might have got *any* wood, so you have to be quite specific with that, what type of wood."

As a result of these conflicting tensions, players often hesitated over the wording of their messages and expressed uncertainty about how to engage help_bot in more complex tasks. A few players suggested that a customisable menu of commands might be useful for defining some more complicated requests. We observed that the moments of uncertainty often came after a player had begun to type a request, but was unsure how to finish it; a contextual autocompletion function could provide timely assistance.

The exact wording used in messages was diverse. Of the 186 messages sent throughout the study, 128 were unique in their wording. A further 11 messages were sent twice by the same player. Only 12 messages were sent by multiple players, including four that were sent by three different players: <thank you>, <follow me>, <come back> and <build a house>. The latter was influenced by the construction task we assigned the first eight players, which was to build a house. Only two messages longer than three words were used by multiple players: <bring me oak wood> and <give me the wood>.

Message structure and purpose

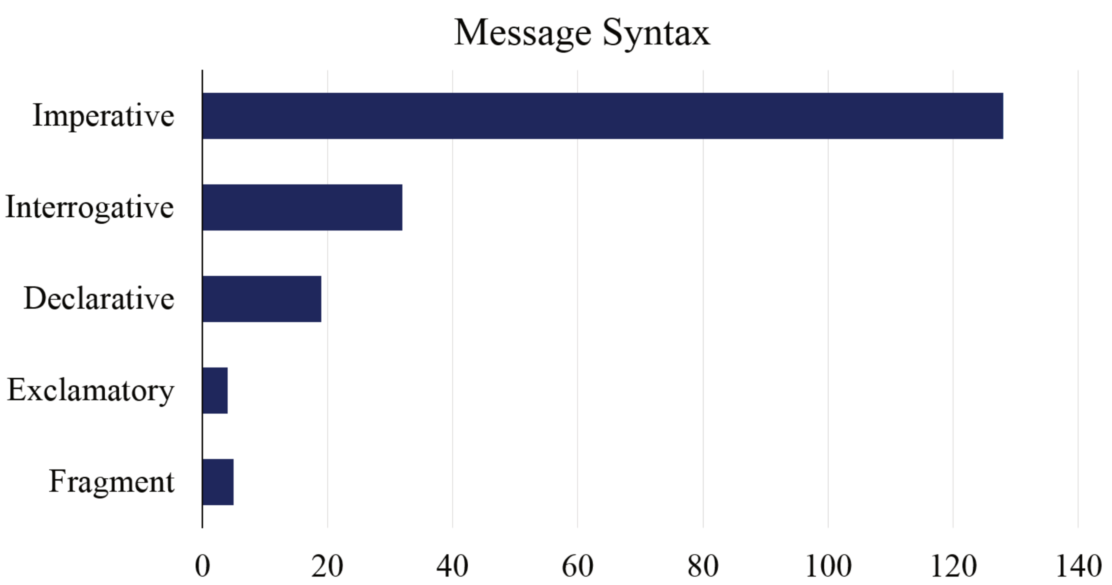

Figure 3:Distribution of messages by sentence type.

Most messages were structured as *imperatives* (see figure 3). These were usually short phrases such as <get stone> and <come here>. Some imperatives were longer, including a few compound sentences containing multiple commands, such as <come to me and give me the oak wood>. Every imperative message was either directions to undertake an activity (*command*) or directions to cancel the current activity (*stop*). We show the relative proportion of message functions in figure 4.

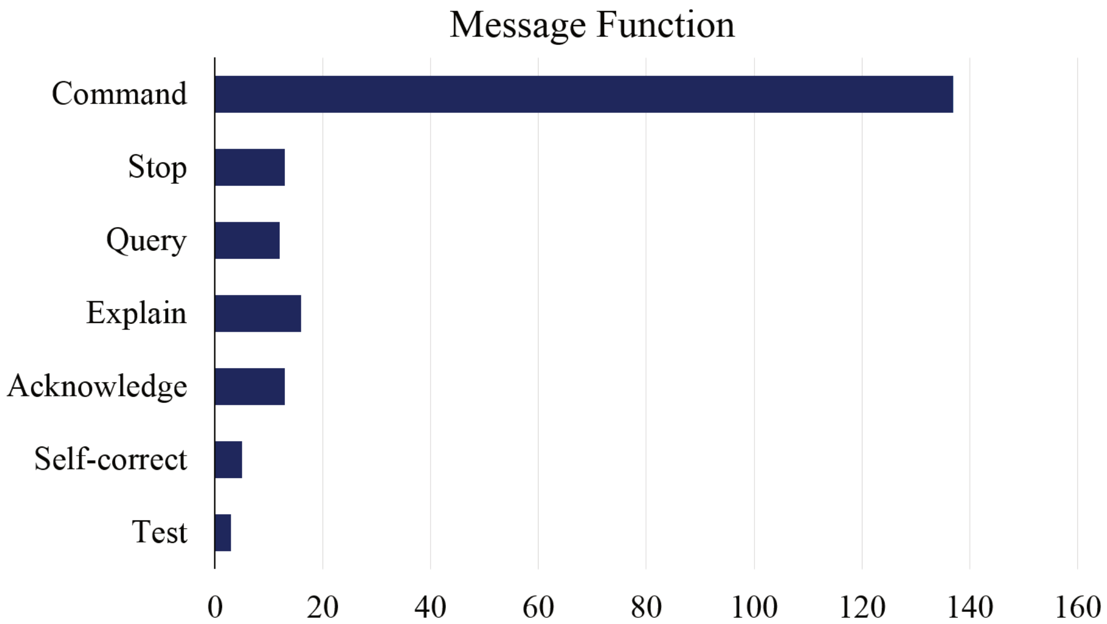

Figure 4:Distribution of messages by intended purpose, inferred from game context and player comments.

A large majority of messages omitted a grammatical subject (see figure 5). This was because most messages were imperatives, and standard English grammars omits the subject in an imperative phrase – the implied subject is the receiver of the message. One message specified its subject by naming help_bot (<help bot come here>). Six messages lacked an explicit subject, because the intended subject was the same as the previous message; we label these as “Antecedent (implied)”.

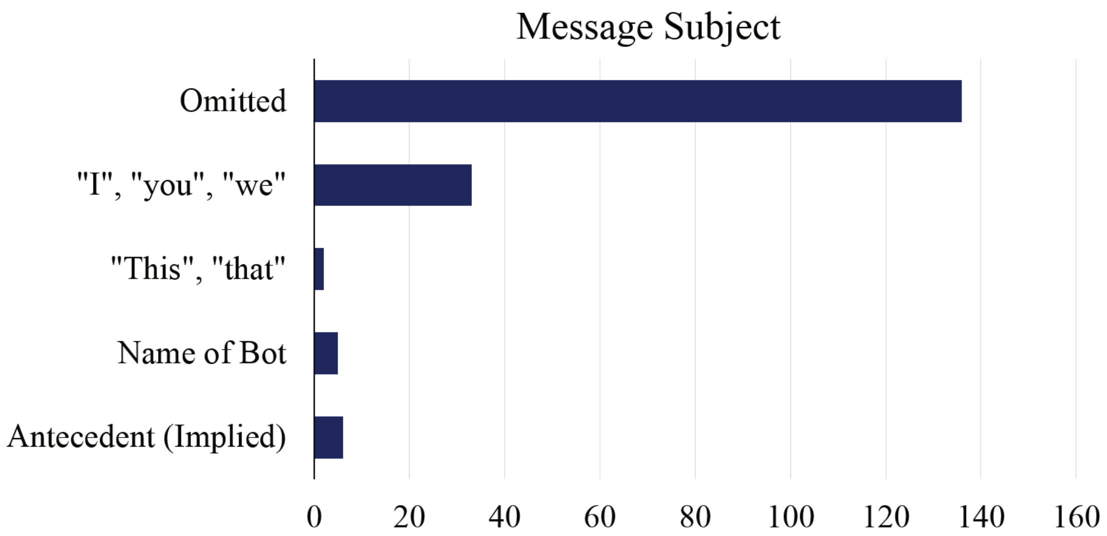

Figure 5:Distribution of messages by intended purpose, inferred from game context and player comments.

The direct object of most messages was indefinite rather than specific, as in <get stone> or <get some stone> as opposed to <get the stone> or <get that stone>. Where the direct object referred to “me” or “you”, this was usually a learning command, as in <copy me>, or an acknowledgement, as in <thank you>. The frequency of each type of direct object is shown in figure 6.

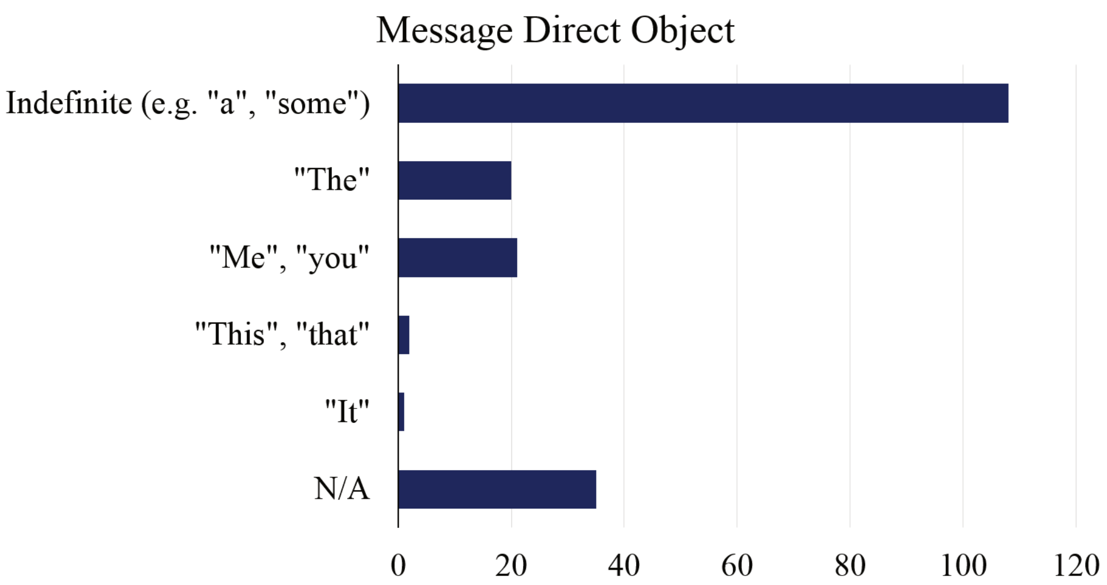

Figure 6:Distribution of messages by intended purpose, inferred from game context and player comments.

Most command imperatives were conceptually simple, requesting actions that had only one step (as in <come to me>) or two steps (as in <bring me oak wood>, requiring help_bot to collect oak and bring it to the player). Players also tested more conceptually complex requests, such as <build a house>. In response to these, help_bot would prompt the player to <show me how>. At this point, players typically simplified or abandoned the request, although some proceeded to demonstrate or explain the task (see *Responses to prompts for additional information*).

Stop imperatives, such as <stop> and <don't mine that>, were typically used when help_bot had completed a task to the player's satisfaction, or when it had made a categorical error. By this, we mean an action that was different in kind from what the player intended, such as when this player tried to ask help_bot to collect fruit:

> *Player:*<get food>
>
> *Help_bot:*<ok>
>
> *Player:* "Oh, it's about to go and get food! Yay! I really hoped that would work."
>
> *Help_bot moves towards a cluster of sheep and cows.*
>
> *Player:*"Don't kill everything please. I'm going to follow it and make sure it doesn't kill everything. Oh – I'm not going to watch, because I feel like it's going to kill everything. . . Maybe I should tell it don't kill everything."
>
> *Player:* <don't kill everything>
>
> *Help_bot:*<?>
>
> *Player:* "You don't. . ." *[nervous laugh]* "OK then, I'm just not going to watch and pretend that help_bot isn't slaughtering animals behind me."
>
> *Player:* <stop>

Players did not use stop commands when help_bot made smaller-scale mistakes, such as placing blocks in the wrong location or digging a hole too deeply. In these cases, players often verbalised their frustration, but in the game they simply reversed help_bot's actions using their own avatar. In the post-task interview, players said they wanted help_bot to interpret either a <stop> command or a reversal of its recent actions by the player as implicit negative feedback, so that it would be less likely to take those actions in the future. That is, to update its behavioural algorithm. However, there were concerns that textual feedback may be too ambiguous, so help_bot might unlearn the wrong behaviour.

Interrogative or questioning phrasing was the second-most common message structure. Only two-fifths of these messages were punctuated with a question mark, as would be grammatically expected. We inferred that less than half of the interrogatives were truly intended as questions (e.g. "do you have any wood?"), and the remainder were indirect commands (e.g. "can you get me some coal please"). Interestingly, the use of a question mark was a strong indicator of a genuine question: three-quarters of the queries, but only one-fifth of the interrogative commands ended with a question mark, as figure 7 shows. One participant surmised that a question mark might be required for a message to be understood as a question:

Player:<do you have any wood?>

Help_bot:<yes>

Player: <how much>

Player: "I probably should have done a question mark."

Help_bot:<?>

Player: <how much wood do you have?>

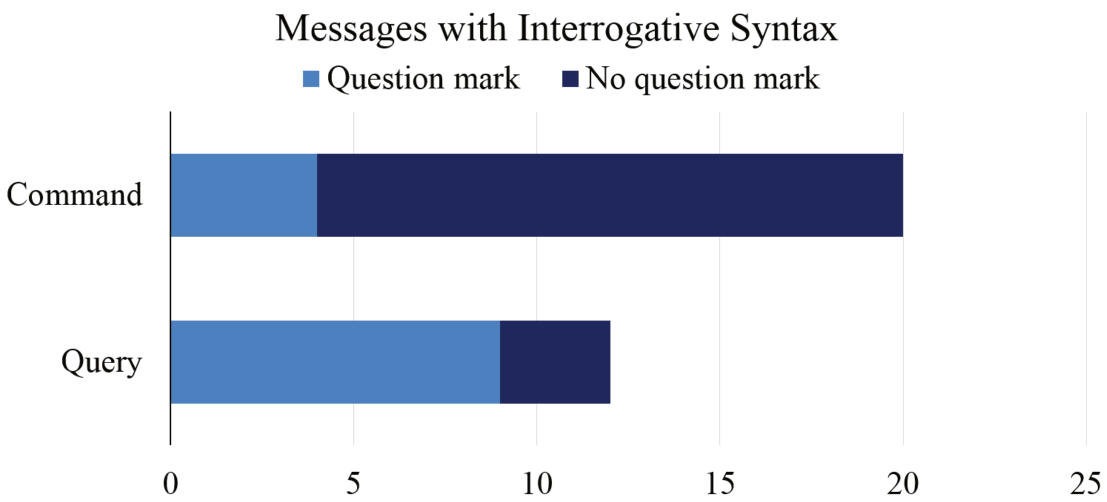

Figure 7:Breakdown of messages with interrogative syntax. Most of those intended as queries ended with a question mark, whereas most of those intended as commands did not.

Interrogative queries were used to learn about help_bot itself. Questions referred to either the contents of its inventory (<do you have any wood?>), the types of actions it was capable of (<what can you do?>), and its status (<are you lost?>). Players used these questions to gain information about help_bot that was not available through other means.

The third-most common message type was *declarative*, or straightforward statements of fact. The primary uses of declarative messages were to identify an object (<this is a shelter>) or to acknowledge help_bot's actions (<thank you> or <well done>). In a few cases, declarative statements were paired with learning commands to teach help_bot a behaviour that it could later use. For example, one participant instructed help_bot to <watch> as they built a simple hut shape, typed <this is a shelter>, then typed <build a shelter>. Other participants described more elaborate versions of this teaching approach as a way of automating repetitive work:

"One of my friends, every world he makes he always has this thing where he has [. . .] a boat with a pig in it and it always just spins around eternally. He could name it 'pigspin' or something like that, and every world he goes in he could give [help_bot] the things it needs and go ‘build pigspin’.”

Two players used declarative messages as indirect stop commands for help_bot (<you don’t need to make any more planks> and <that’s fine>).

In the post-task interviews, players mentioned that they thanked or complimented help_bot as a form of positive feedback, with an idea that this might reinforce its learning of the recent behaviour:

“It’s about being polite, and also saying ‘thanks, you did the *right thing*’.”

Four messages were *exclamatory* in format, of which three were greetings (<hello> and <hi>) and one was a celebratory statement (<yay we finished the maze!>). We classified the purpose of all four as acknowledgements to help_bot, with the latter also being an explanatory message intended to reinforce the idea that what the player and help_bot had just built was a <maze>.

Finally, five messages were incomplete fragments. Two of these were self-corrections by the player (<bring me the wod> followed by <wood>), intended to update the meaning of the preceding message. One was similarly an addition to a prior message, adding <two blocks high> after <build wall>. The remaining two messages were nonsensical (<jeff> and <s>), and at least one of these was deliberately so. The player explained in the post-task interview that they had entered a nonsensical message to test help_bot’s responses:

“I was trying to see – because I [previously] did ‘thank you’ and it said ‘ok’, so I wasn’t sure if that was a response to ‘thank you’ or if it was just its generic response if it doesn’t understand something. And I did some random stuff and [learned that] the generic response if it doesn’t understand something is [a] question mark.”

Indeterminacy of amounts, places and boundaries

Specificity was an issue for many participants. Players were forced to confront the fact that their everyday phrasing contains a great deal of ambiguity, which is resolved by human conversation partners through context and common sense. The Minecraft setting facilitated references to objects in the world, as it contains only a limited set of clearly labelled and categorised objects. However, players were uncertain how to specify locations and amounts in a way that help_bot would understand. Players also expressed uncertainty about help_bot's understanding of boundaries.

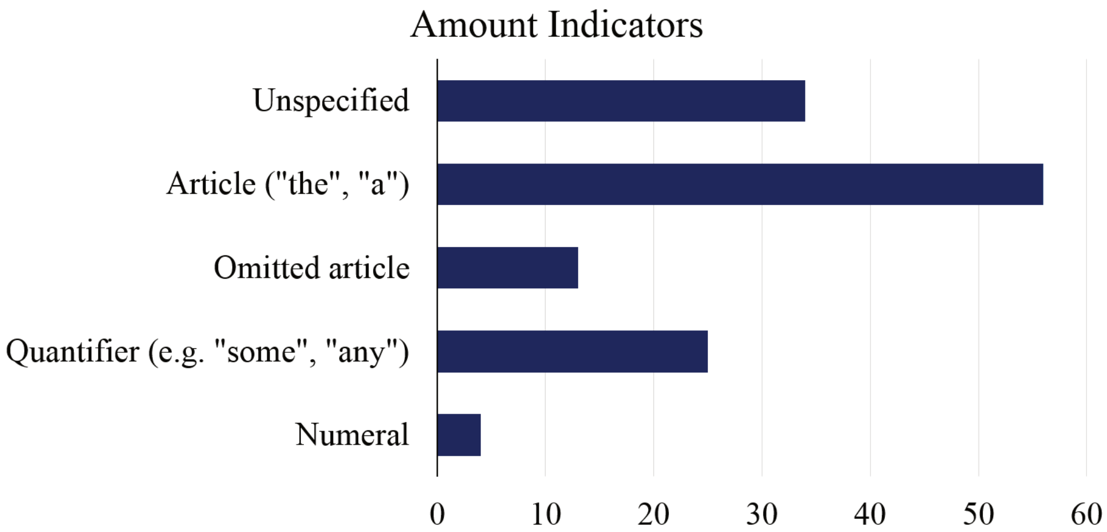

Figure 8:Breakdown of messages by how they indicated an amount for their direct object.

Amounts were rarely specified with a numeral; only one player did so (see figure 8). Most references to multiples of an object left the amount unmentioned, as in <collect bricks>. The rest of the time a linguistic quantifier was used, as in <get some wood> and <can you get me more inc sacks please>. Players would then leave help_bot to collect the objects until they felt it had enough, and stop or call it back at that point. Single objects were less difficult, as players included the relevant article ("a" or "the"), depending on whether the required object was generic (as in <build a box>) or specific (as in <bring me the coal ore>). In some cases, the article was implied but omitted, as in <craft sign>, which could increase the difficulty for a natural language system to parse correctly.

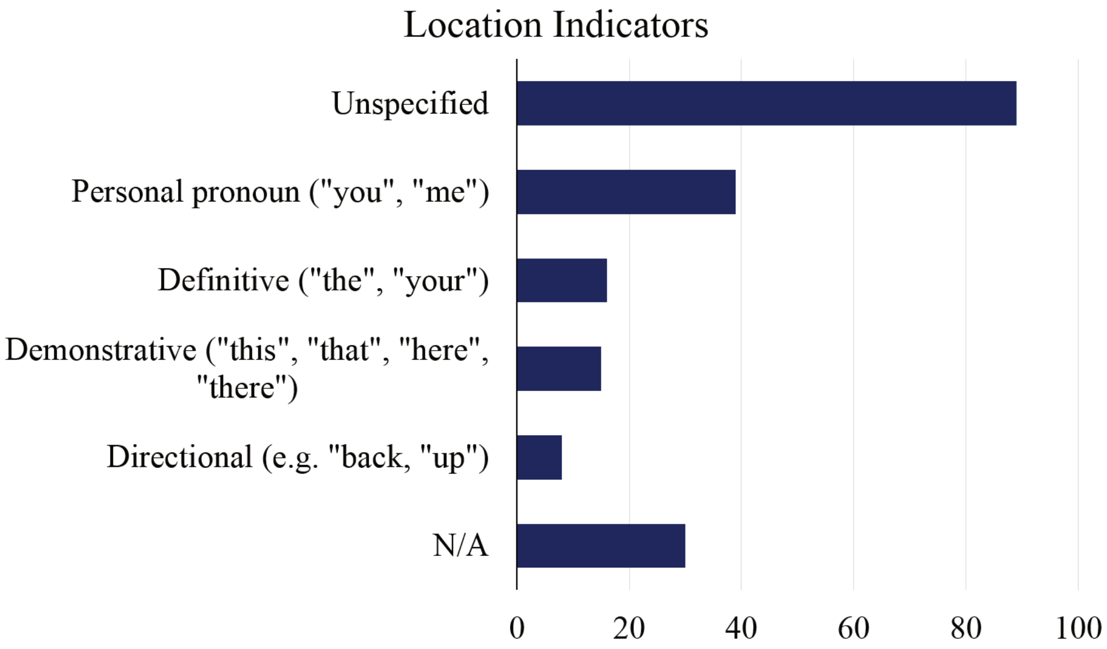

Figure 9:Breakdown of messages by how they indicated a location for the action of their message.

Locating the object of the message introduced further difficulties. In most cases, no location was specified, but there was an implicit deixis to many of these messages (see figure 9). For a typical fetch request, such as <get wood>, the player wanted help_bot to take the item from a specific location (usually the nearest source, unless that source was something the player had built) and bring it back to the player. Similarly, a typical build request such as <build a box> contained an unstated assumption that help_bot would place the building on a suitable flat piece of ground near to the player, but not close enough to interfere with their current activity.

“It built it quite close to where I built it. Like when I did the three-by-three square, it built right next to it. With this it didn’t really matter, but I think if you could tell it where to build that would be quite good. . . Maybe like coordinates or something, I don’t know.”

Some players attempted to be more specific about the location for their requests by using demonstrative terms such as “this” and "here" as linguistic pointers. These were paired with the player's avatar moving to the relevant location, or even tapping on a specific spot with the avatar's hand, as in this exchange:

> *Player:* <get me some spruce wood>
>
> *Help_bot:* <ok>
>
> *Help_bot starts mining a house that the player had built out of spruce blocks.*
>
> *Player:* "So it is mining what I've done earlier, which is really weird."
>
> *Player:* <don't mine that>
>
> *Help_bot stops mining:* <ok>
>
> *Player looks at a tree:* <mine this>
>
> *Help_bot starts mining the tree.*
>
> *Player:* "At least it knows what 'this' is."

The need to demonstrate the location with the player's avatar meant that proximal locations were much easier to point out than distal locations. Many players expressed frustration at their inability to designate specific places at a distance.

As mentioned previously, players expressed some concern about help_bot's judgement regarding the boundaries of locations and objects. There were times when a player might request more of a resource that they were currently working on, and worried that help_bot might mine that resource straight out of their built structure, as it was the closest source.

> "Does it know when to stop? And if I had a house full of oak wood, and I said 'go get some oak wood', would it take the oak wood from my house?"

Responses to prompts for additional information

We defined three prompt messages for help_bot to use in response to messages that were outside the parameters we had set for its understanding. These indicated varying levels of comprehension, and elicited different types of responses from the players.

The first prompt was used when the player entered a request that was comprehensible to help_bot, but for which there was no clear or default location. This included requests for a resource type that was not currently visible, and requests for building a shape that did not specify where the building should go. In the pilot phase, the prompt text was <where?> or <from where?>, but players expressed confusion about how to answer this. One player asked if they needed to enter coordinates on Minecraft's grid reference, which was not a feasible solution. For the remainder of the study, we changed help_bot's prompt to <show me where>, and observed that players showed less uncertainty about this prompt, and consistently responded by physically moving their avatar to the desired location. We inferred from this that a query is easier for the player to respond to if it suggests a form of response, compared with one that is less prescriptive about how it should be answered.

The second prompt was <show me how>, used when the player entered a message that was recognisable in form as a building, collection or other action request, but which required knowledge that help_bot was not initially given. A typical case was a request to construct a complex object, since help_bot recognised only the names of geometric shapes unless taught other configurations by the player. Players most often responded to this prompt by giving up the request, but in several cases responded by demonstrating the behaviour they wanted and asking help_bot to copy it, as in this example:

Player: <build a tree>

Help_bot: <show me how>

Player: <no build a maze>

Help_bot: <show me how>

Player: <just watch>

Help_bot: <ok>

Player constructs a maze out of blocks.

Player: <now build a maze>

In these demonstrations for help_bot, players consistently marked either the start or the end of the action sequence with an explicit message, but not both. Help_bot was expected to recognise the completed structure as a Gestalt "object", and thereby understand where the relevant actions had begun or ended.

In contrast, a few players used more detailed messages to explain what they wanted, rather than demonstrating and expecting help_bot to replicate their actions:

Player:<help build a house>

Help_bot:<show me how>

Player:<put some cobble on top of each other>

Help_bot:<ok>

Help_bot begins stacking blocks of cobble. Player places more blocks alongside it to form a box shape, and help_bot follows this shape.

Help_bot:<done>

Player:<put some cobble on the top to make a roof>

Help_bot:<ok>

When messages were not recognisable as a command or a request, help_bot's fallback prompt was <?>. This was used in response to spelling mistakes, sentence fragments and verbs that were outside help_bot's abilities (such as <die> and <write hello>). When players received this response, most were quick to assume that help_bot was unable to perform the action they had requested, and simplified or abandoned their request. However, in many cases they were mistaken: the action was within the abilities we had defined for help_bot, but the message formulation was not. Spelling was particularly notable, as players sometimes did not notice their own spelling mistakes, which left them with the false impression that help_bot could not understand the message they *intended* to write, rather than the message that they actually wrote. From this, we infer a need for feedback messages that are specific about the source of the lack of understanding. For example, rather than saying <?> or <I didn't understand that>, a message might say <I don't recognise the word 'whool'>.

Feedback on natural language input

Despite the challenges of natural language input, most players commented that they enjoyed it more than interaction via the traditional game controls or menus. A few players noted that menu-based interactions, such as those that exist for some friendly characters in Minecraft, would be preferable to natural language text as they provide more structure and clarity about the character's abilities. However, most considered the natural language input style an overall benefit as it opened up the possibility space for what they could potentially do with the character.

Players used the dialogue to consider how help_bot "thought", in some cases actively probing it with questions about its abilities or variations on a text prompt to test how it responded. The text responses provided a relatively clear channel to understand what was happening inside help_bot's "brain", and gave a sense that it was updating its behaviour.

> "With the one that I couldn't speak to, if I had tried to get wood or something, it didn't have the same feel that it was learning. So the text one, it felt like it was learning because it was saying ok, yes, I know how to do this."

Players also stated that natural language gave help_bot a greater sense of being alive and engaged with the player, compared with the non-speaking version, which made it more enjoyable to interact with:

> "Typing feels more interactive, like you're talking to a real person. Pressing stuff doesn't feel like that, like you're just talking to a computer."

Two players commented that they would like to have the option to talk to help_bot through speech rather than text messages, to further extend the feeling of a living character. However, they were uncertain about the ability of speech recognition technology to work well enough to support this.

> "It would be cool if you could do voice commands. But then again, you'd have to have the most to-your-country accent, otherwise it wouldn't understand it."

Expectations and cues for understanding the agent

The way in which players interacted with help_bot was influenced by their prior knowledge in several domains. Most obviously, players' experiences with Minecraft guided many of their initial attempts to understand and engage the character. Players compared help_bot to Minecraft's villager and wolf NPCs (non-player characters), for example in surmising that help_bot would attack any enemy NPC that the player attacked, because this was the behaviour for a tamed wolf. Players also drew on their experiences with other humans in the multiplayer game; two players tried to engage help_bot by repeatedly crouching their avatar in front of it, which is a social custom in online Minecraft equivalent to waving "hello". Finally, players compared the text inputs in our study to the console commands in Minecraft. These commands allow players to edit the state of the game by entering a text string beginning with a forward slash – for example, </time set 6000> to change the in-world clock to midday. Several players initially started their messages to help_bot with a forward slash, until the facilitator pointed out that it was unnecessary. One player asked whether command strings in the same format could be used to manage help_bot's learning, such as </train help_bot X> to learn a current action and </set help_bot X> to repeat that action at a later time. These influences show that the context in which an AI character is deployed will influence the way users understand it and expect it to behave.

Expectations were also drawn from sources beyond Minecraft. Players who had programming experience compared the natural language inputs in our study to the programming language that they had used, and this guided their thinking about what might be possible. Real-world social cues were also applied: players expressed discomfort when help_bot followed their avatar too persistently, stood too close to it, or stared at it for too long. And unsurprisingly, several players referred to film depictions of robots and AI characters, such as *The Terminator*:

> "At one point I was like, 'ok ok ok ok, that's enough!' It kind of reminded me of in a film where there's like robots and they go out of control. That's why I was afraid to start digging the ground to get a flat bit, because I was afraid it would just start levelling the whole world."

Players' preconceptions about help_bot were, in some cases, well suited to the protocols we had designed for it, and in other cases beyond its abilities. What we found notable was that these expectations were sometimes assumed to be true, if only unconsciously, without having been tested. For example, one player travelled far away from help_bot and was confused when it was unable to find its way to them, as their experience with other NPCs had taught them to expect friendly characters in Minecraft to teleport near the player if they strayed too far away. This raises the importance of providing the right contextual cues for players to form the right mental model of how an AI character works, as incorrect expectations may otherwise be set and not tested.

Variation in engagement and anthropomorphisation

We observed substantial variations between players in the ways that they engaged with, reacted to and spoke about help_bot. Putting these differences together, we hypothesise that they reflect two main dimensions on which attitudes towards help_bot varied. The first was level of engagement, or the extent to which players were interested in interacting with help_bot. The second was anthropomorphisation, or the extent to which players acted as though help_bot had human-like thoughts and feelings.

Differences in the level of engagement were apparent in the time each player spent interacting with help_bot during the tasks. As figure 1 shows, more than half of the players sent no more than seven messages to help_bot throughout the task, or less than one every two minutes, whereas several players sent more than twice this many messages. This variation carried over to other behaviours as well, including the amount of time the player spent watching help_bot, and the amount of interest they expressed verbally during and after the tasks. Highly engaged players spent more time experimenting with help_bot to determine its capabilities and asking the facilitator questions about it, and in some cases largely abandoned the construction task we had set in favour of playing with help_bot. We observed that several of the players who showed the greatest interest in help_bot also held higher expectations that it was capable of complex behaviour, although whether there was a causal relationship is unclear; it could simply be that these players thought and spoke more about the possibilities.

There are several behaviours wrapped up in what we are calling "anthropomorphisation", each of which represents an attitude that the AI character has human qualities. Players varied in the language they used in text messages, from those who entered only terse verb-noun commands such as <kill sheep> to those who greeted help_bot with a <hello>, framed their commands as polite requests such as <can you bring me some birch wood please>, and thanked help_bot for completing tasks. Players showed varying levels of empathy for help_bot, from those who casually hit it with an axe when it was in the way, to those who expressed concern about its wellbeing. One such player avoided clicking on help_bot, concerned that they would hit it by accident, and expressed guilt at making their avatar eat food in front of help_bot:

> "I feel kind of bad eating it – can I give this to you?"
>
> *Player gives some of the food to help_bot and demonstrates eating it with their avatar.*
>
> "Did they eat it? I don't see it, I assume they ate it. Now I feel slightly less bad."

When talking about instructing help_bot, some players described it in terms of a brainless instrument that could be programmed to perform repetitive actions, whereas others gave it tasks that required more independent, sophisticated and arguably human-like judgement. As an example of the latter, one player repeatedly set up pits for help_bot to fall into, explaining that they were trying to teach it to avoid the situation by looking out for and filling in any pits that it encountered in the future. Players also expressed an expectation that help_bot would prioritise tasks in a common-sense fashion, so that when a hostile creature attacked, for example, they were surprised if help_bot did not automatically come to their assistance. As such cases happened only rarely during the test, it was not certain whether this expectation was higher among participants who had higher expectations for other aspects of help_bot's judgement.

The behaviours that indicated low or high anthropomorphisation appeared to cluster together in individual participants. A player who expressed empathy for help_bot was also frequently one who gave it higher-level instructions with more room for autonomy, and one who described it more as a character with a mind than as a plain instrument. That is not to say that players who showed higher anthropomorphisation believed help_bot had human intelligence or emotions, but they appeared more inclined to act as though it did.

	Low anthropomorphism	**High anthropomorphism**
High engagement	Treated agent as an instrument to be programmed	Treated agent as a character capable of judgement
Low engagement	Inattentive to agent	Polite to but uninterested in agent

Table 1: 2×2 model of player attitudes to the AI agent (help_bot), showing how variations in engagement and anthropomorphism resulted in different behaviours.

Notably, the level of engagement and the level of anthropomorphisation were at least partially independent of each other. Some players showed relatively little interest in help_bot, but addressed it courteously in the few messages they did send. Other players spent considerable time testing out help_bot's abilities and talked about it with enthusiasm, but as an interesting tool that they could program rather than as a character. There were players who liked help_bot as a potential sidekick character that could exercise independent judgement, and those who largely ignored it and said little to suggest that it had an inner life. The variations are summarised in table 1. This study is too small-scale and unstructured to draw firm conclusions about these variables, but we put them forward as a possibility to investigate in later research.

DISCUSSION

Our analysis shows that players' messages to help_bot primarily used simple syntax and direct commands, but that there was substantial complexity and variation in the details of wording and the way in which text messages were paired with in-game actions. Few messages were repeated between different players, and players invested considerable thought into their choice of words due to the difficulty of communicating with an AI that does not have the common ground of knowledge and judgement shared by most humans. Indirect speech acts (Searle 1975) were common, particularly in the form of commands with interrogative phrasing, which highlights the need for a natural language interface to either distinguish between direct and indirect commands, or remind players to use direct syntax. A contextual autocompletion function would seem suitable, to provide guidance to players as they are formulating the phrasing to translate their intention into words, which was often a moment of hesitation.

The findings also show that messages rarely contained all the information needed to interpret them correctly within the words themselves. Contextual information was also required. Much of this missing information could be inferred quite simply in the scenario we studied, but would become ambiguous in other contexts. For example, it could be assumed in our study scenario that help_bot was the subject of imperative messages as there was no other conversation partner in the game, but in a multiplayer or multi-agent game situation the subject would need to be stated by the player or inferred by the agent. Other messages required reference to the dynamic game state to be accurately interpreted, as in the use of deictic words such as "there", "that" or "away" matched with the player's avatar's location and gaze direction. (This is suggestive of one of the earliest multimodal interface models, *Put-That-There*, which combines speech, gesture and gaze to determine the user's input (Bolt 1980).) Players were aware of the inherent indeterminacy of their language, and expressed concerns about help_bot's ability to make judgements that would seem sensible for a human, such as distinguishing between the “natural” and built environment (that is, between what the game generated and what the player constructed). Thus, the challenge for natural language agents in games is not only to make correct judgements about the player’s intentions, but also to communicate the results of those judgements to the player.

The need to communicate the agent’s internal decision-making is even greater in the context of interactive machine learning tasks, wherein the player is directly teaching the agent to learn new behaviour or change its existing behaviour. Help_bot’s messages reassured players that it was learning, and went some way towards clarifying what it was learning and what it was not, despite consisting of only a few words. This appears to be linked to the sentiment that help_bot felt more like a living, thinking person when it talked: these signals conveyed the sense that it had a mind, rather than just a behavioural algorithm. Accordingly, players’ strategies for teaching help_bot focused on communicating concepts through demonstration and example, rather than training help_bot through frequent feedback on its actions. (Note that players were told before each task that help_bot could learn from their feedback.) This is consistent with previous studies (Amershi et al. 2014; Kaochar et al. 2011), and poses a technical difficulty as many interactive machine learning approaches rely on user feedback to iteratively adapt the agent’s behaviour (for example, Knox 2013). Some players did perceive themselves to be giving help_bot feedback, although this came in the form of implicit signals, such as acknowledging messages like <thank you> and contradictory actions like undoing help_bot’s work.

Users may be encouraged to take a more direct teaching role by giving them tools to make their feedback more precise, and by providing clear feedback on what has been learned. Participants in our study were hesitant to engage in teaching, partly due to a perceived ambiguity about precisely what help_bot was learning from their actions. In addition, the fact that our help_bot was designed to continuously observe and adapt its behaviour to the player's actions made it hard for our participants to recognise behaviour that was permanently learned, as opposed to momentary imitation. Participants generally preferred to be able to control when and where help_bot was taking in information for its own learning.A suggestion to reduce the ambiguity around help_bot's learning was to give players the ability to toggle it between a learning mode and a non-learning mode.

The way in which a subset of players anthropomorphised help_bot is consistent with past studies of conversational interaction (Luger and Sellen 2016) and empathic agents (Paiva et al. 2017), and representative of a wider effect in human-computer interaction: the tendency for people to respond to computers as though they are human, which Reeves and Nass have dubbed the "Media Equation" (1996). Nass and Moon argue that users "mindlessly" (2000, 82) apply social rules and expectations to computers, and Nass and Brave (2005) suggest that this effect is particularly strong for interactions involving speech. The strength of this effect has been challenged (Shechtman & Horowitz 2003; Lang et al. 2013), but there is some evidence that users apply more social behaviour to computer characters as their appearance becomes more human-like (Gong 2008). This implies that natural language interaction is a modality that will elicit more social reactions, as players in our study ascribed a greater sense of humanity and intelligence to help_bot when it used text, despite the text being limited to only a few formulaic phrases.

These variations in players' attitudes and expectations towards help_bot are important because they show that players have different mental models (Norman 1983) of how the agent works and what it is capable of. Our findings suggest that there is no universal starting point or blank slate in how players will perceive an agent. Expectations about an agent's degree and form of "intelligence", its adaptability and its responsiveness to different inputs are influenced by both the presentation and context of the agent, and background knowledge drawn from science fiction, previous game experiences, and real-world human social customs.

To facilitate players having a smooth experience with an agent like help_bot, designers will need to evaluate the context of the game genre and the appearance of the character to anticipate what kinds of expectations players may have, and consider how both implicit cues and explicit messages may serve to guide players to adopt the right mental model for working with their character.

Limitations and future directions

Our concern in this study was not only to evaluate what players said to the agent, but to observe players' actions in the context of gameplay, and to examine their reasoning in the post-task interviews. By nature, this was a limited study of a relatively small group of participants of one age bracket in one geographical location in a single language. A broader-based study would be needed to determine how representative our measurements are of natural language interactions with game characters by other groups of users, or to make statistical comparisons of message patterns under different conditions. One intriguing question, which we plan to address in a future study, is whether players use language differently when (knowingly) communicating with an AI compared to another human player, and in what respects their language is different or similar.

Our study has not been designed to elicit comparative preferences between natural language messages and conventional game dialogue systems such as branching conversation trees. As we have reported, our participants did compare the natural language system to other methods of interacting with game characters, but unsurprisingly these comparisons centred on Minecraft's own system, which is mostly wordless and oriented towards trading goods. In the most directly relevant comparative study, Sali et al. (2010) found that players preferred typing natural language messages to game characters over choosing messages from a menu, even though they encountered a higher number of errors, and frustration with the natural language mode. This is consistent with our findings. However, both studies looked at only a single play session, and involved participants for whom natural language was a novel way of interacting with game characters. It remains to be seen whether a preference for natural language would continue over a longer time period.

CONCLUSION

Overall, our study shows that there are substantial commonalities in the syntax and concepts that players use in natural language interactions with a game character with learning AI, but also significant variations in the specific wording, behaviour and expectations for the character, which are driven by players' prior knowledge and contextual cues. While natural language interaction offers the promise of a flexible, engaging and intuitive way to interact with and teach AI agents in games, much work will be required to realise this prospect. Our study did not specifically set out to measure the extent of anthropomorphism in players' mental models of the character, but we have identified this as an area for further investigation to determine how it can be used either as a design resource to shape the interaction, or as a pitfall to be avoided. In future work we will look more systematically at the effects of anthropomorphism on language-based interaction with an AI character.

ACKNOWLEDGEMENTS

This work was supported by Microsoft Research and the Australian Government Research Training Program.

BIBLIOGRAPHY

Amershi, Saleema, Maya Cakmak, W Bradley Knox, and Todd Kulesza. 2014. "Power to the People: The Role of Humans in Interactive Machine Learning." *AI Magazine*35 (4).

Bakkes, Sander C. J., Pieter H. M. Spronck, and Giel van Lankveld. 2012. “Player Behavioural Modelling for Video Games.” *Entertainment Computing* 3(3): 71–79. https://doi.org/10.1016/j.entcom.2011.12.001.

Bayliss, Peter. 2007. “Notes Toward a Sense of Embodied Gameplay.” In *DiGRA ’07 – Proceedings of the 2007 DiGRA International Conference: Situated Play*, The University of Tokyo, Japan. http://www.digra.org/digital-library/publications/notes-toward-a-sense-of-embodied-gameplay/.

Bellemare, Marc G., Yavar Naddaf, Joel Veness, and Michael Bowling. 2012. “The Arcade Learning Environment: An Evaluation Platform for General Agents.” *ArXiv:1207.4708 [Cs]*, July. https://doi.org/10.1613/jair.3912.

Bernotat, Jasmin, Birte Schiffhauer, Friederike Eyssel, Patrick Holthaus, Christian Leichsenring, Viktor Richter, Marian Pohling, et al. 2016. “Welcome to the Future – How Naïve Users Intuitively Address an Intelligent Robotics Apartment.” In *International Conference on Social Robotics 2016*, edited by Arvin Agah, John-John Cabibihan, Ayanna M. Howard, Miguel A. Salichs, and Hongsheng He, 982–92. Cham, Switzerland: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-47437-3_96.

Bolt, Richard A. 1980. “‘Put-That-There’: Voice and Gesture at the Graphics Interface.” In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, 262–270. SIGGRAPH ’80. New York, NY: ACM. https://doi.org/10.1145/800250.807503.

Button, Graham, and Wes Sharrock. 1995. “On Simulacrums of Conversation: Toward a Clarification of the Relevance of Conversation Analysis for Human-Computer Interaction.” *Cambridge Series on Human Computer Interaction*, 107–25.

Caine, Kelly. 2016. “Local Standards for Sample Size at CHI.” In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 981–992. CHI ’16. New York, NY: ACM. https://doi.org/10.1145/2858036.2858498.

Cakmak, Maya, Crystal Chao, and Andrea L Thomaz. 2010. “Designing Interactions for Robot Active Learners.” *IEEE Transactions on Autonomous Mental Development* 2(2): 108–18.

Cassell, Justine, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill. 2000. *Embodied Conversational Agents*. Cambridge, MA: MIT Press.

Eckersley, Peter, Yomna Nasser, Yann Bayle, Owain Evans, Gennie Gebhart, and Dustin Schwenk. 2017. “AI Progress Measurement Project.” Electronic Frontier Foundation. 12 June 2017. https://www.eff.org/ai/metrics.

Farooq, Sehar Shahzad, In-Suk Oh, Man-Jae Kim, and Kyung Joong Kim. 2016. “StarCraft AI Competition Report.” *AI Magazine* 37(2): 102–7.

Fischer, Kerstin, Katrin S Lohan, Chrystopher Nehaniv, and Hagen Lehmann. 2013. “Effects of Different Kinds of Robot Feedback.” In *Proceedings of the 5th International Conference on Social Robotics*, 260–69. New York, NY: Springer. https://doi.org/10.1007/978-3-319-02675-6_26.

Goldberg, Yoav. 2016. “A Primer on Neural Network Models for Natural Language Processing.” *Journal of Artificial Intelligence Research* 57: 345–420.

Gong, Li. 2008. “How Social is Social Responses to Computers? The Function of the Degree of Anthropomorphism in Computer Representations.” *Computers in Human Behavior*24 (4): 1494–1509. https://doi.org/10.1016/j.chb.2007.05.007.

Goodrich, Michael A., and Alan C. Schultz. 2008. "Human–Robot Interaction: A Survey." *Foundations and Trends in Human–Computer Interaction* 1(3): 203–75. https://doi.org/10.1561/1100000005.

Guillory, Andrew, and Jeff A. Bilmes. 2011. "Simultaneous Learning and Covering with Adversarial Noise." In *Twenty-Eighth International Conference on Machine Learning (ICML 2011)*, 11:369–76.

Johnson, Matthew, Katja Hofmann, Tim Hutton, and David Bignell. 2016. "The Malmo Platform for Artificial Intelligence Experimentation." In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 4246–47.

Juliani, Arthur. 2017. "Introducing: Unity Machine Learning Agents." *Unity Blog*(blog). 2017. https://blogs.unity3d.com/2017/09/19/introducing-unity-machine-learning-agents/ (accessed Feb. 2018).

Kaochar, Tasneem, Raquel Torres Peralta, Clayton T. Morrison, Ian R. Fasel, Thomas J. Walsh, and Paul R. Cohen. 2011. "Towards Understanding How Humans Teach Robots." In *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*, 347–352. UMAP'11. Berlin, Heidelberg: Springer-Verlag.

Kelley, J. F. 1983. "An Empirical Methodology for Writing User-Friendly Natural Language Computer Applications." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 193–196. CHI '83. New York, NY: ACM. https://doi.org/10.1145/800045.801609.

Kempka, Michał, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. 2016. "ViZDoom: A Doom-Based AI Research Platform for Visual Reinforcement Learning." *ArXiv:1605.02097*.

Knox, W. Bradley, Brian D. Glass, Bradley C. Love, W. Todd Maddox, and Peter Stone. 2012. “How Humans Teach Agents.” *International Journal of Social Robotics*, 4(4): 409–21. https://doi.org/10.1007/s12369-012-0163-x.

Knox, W Bradley, Peter Stone, and Cynthia Breazeal. 2013. “Training a Robot via Human Feedback: A Case Study.” In*International Conference on Social Robotics*, 8239:460–470. Berlin, Heidelberg: Springer.

Koenig, Nathan, Leila Takayama, and Maja Matarić. 2010. “Communication and Knowledge Sharing in Human–Robot Interaction and Learning from Demonstration.” *Neural Networks*, Social Cognition: From Babies to Robots, 23 (8–9): 1104–12. https://doi.org/10.1016/j.neunet.2010.06.005.

Lang, Helmut, Melina Klepsch, Florian Nothdurft, Tina Seufert, and Wolfgang Minker. 2013. “Are Computers Still Social Actors?” In *CHI ’13 Extended Abstracts on Human Factors in Computing Systems*, 859–864. CHI ’13. New York, NY: ACM. https://doi.org/10.1145/2468356.2468510.

Lessard, Jonathan. 2016. “Designing Natural-Language Game Conversations.” In *Proceedings of the First International Joint Conference of DiGRA and FDG*, Dundee, Scotland. http://www.digra.org/digital-library/publications/designing-natural-language-game-conversations/.

Luger, Ewa, and Abigail Sellen. 2016. “‘Like Having a Really Bad PA’: The Gulf Between User Expectation and Experience of Conversational Agents.” In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297. CHI ’16. New York, NY: ACM. https://doi.org/10.1145/2858036.2858288.

Mateas, Michael, and Andrew Stern. 2010. “Structuring Content in the Façade Interactive Drama Architecture.” In *Proceedings of the First AAAI Conference on Artificial Intelligence and Interactive*

Digital Entertainment, 93–98. AIIDE’05. Marina del Rey, CA: AAAI Press.

Maulsby, David, Saul Greenberg, and Richard Mander. 1993. “Prototyping an Intelligent Agent through Wizard of Oz.” In *Proceedings of the INTERACT ’93 and CHI ’93 Conference on Human Factors in Computing Systems*, 277–284. CHI ’93. New York, NY: ACM. https://doi.org/10.1145/169059.169215.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. 2015. “Human-Level Control through Deep Reinforcement Learning.” *Nature*, 518 (7540): 529–33. https://doi.org/10.1038/nature14236.

Muñoz-Avila, Hector, Christian Bauckhage, Michal Bida, Clare Bates Congdon, and Graham Kendall. 2013. “Learning and Game AI.” In *Artificial and Computational Intelligence in Games*, edited by Simon M. Lucas, Michael Mateas, Mike Preuss, Pieter Spronck, and Julian Togelius, 6:33–43. Dagstuhl Follow-Ups. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Nass, Clifford, and Scott Brave. 2005. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. Cambridge, MA: MIT Press.

Nass, Clifford, and Youngme Moon. 2000. “Machines and Mindlessness: Social Responses to Computers.” *Journal of Social Issues*, 56 (1): 81.

Norman, Donald A. 1983. “Some Observations on Mental Models.” In *Mental Models*, edited by Dedre Gentner and Albert L. Stevens, 7–14. Mahwah, NJ: Lawrence Erlbaum Associates.

Paiva, Ana, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. “Empathy in Virtual Agents and Robots: A Survey.” *ACM Transactions on Interactive Intelligent Systems*, 7(3): 11:1–11:40. https://doi.org/10.1145/2912150.

Reeves, Byron, and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. CSLI Publications and Cambridge University Press.

Riek, Laurel. 2012. "Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines." *Journal of Human-Robot Interaction,* 1(1): 119–36. https://doi.org/10.5898/JHRI.1.1.Riek.

Robertson, John. 2017. "Echo: Ex-Hitman Devs Bring Machine Learning to Stealth Games." Ars Technica. 15 August 2017. https://arstechnica.com/gaming/2017/08/echo-hitman-preview/ (accessed Feb. 2018).

Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. "A Simplest Systematics for the Organization of Turn-Taking for Conversation." *Language,* 50(4): 696–735. https://doi.org/10.2307/412243.

Sali, Serdar, Noah Wardrip-Fruin, Steven Dow, Michael Mateas, Sri Kurniawan, Aaron A. Reed, and Ronald Liu. 2010. "Playing with Words: From Intuition to Evaluation of Game Dialogue Interfaces." In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, 179–186. FDG '10. New York, NY: ACM. https://doi.org/10.1145/1822348.1822372.

Schegloff, Emanuel A., Gail Jefferson, and Harvey Sacks. 1977. "The Preference for Self-Correction in the Organization of Repair in Conversation." *Language,* 53 (2): 361–82. https://doi.org/10.2307/413107.

Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, UK: Cambridge University Press.

Searle, John R. 1975. "Indirect Speech Acts." In *Syntax and Semantics, Volume 3: Speech Acts*, edited by Peter Cole and Jerry L. Morgan, 59–82. New York, NY: Academic Press.

Shaker, Noor, Julian Togelius, Georgios N. Yannakakis, Likith Poovanna, Vinay S. Ethiraj, Stefan J. Johansson, Robert G. Reynolds, Leonard K Heether, Tom Schumann, and Marcus Gallagher. 2013. "The Turing Test Track of the 2012 Mario AI Championship: Entries and Evaluation." In *IEEE Conference on Computational Intelligence in Games (CIG)*, 1–8. IEEE. https://doi.org/10.1109/CIG.2013.6633634.

Shechtman, Nicole, and Leonard M. Horowitz. 2003. "Media Inequality in Conversation: How People Behave Differently When Interacting with Computers and People." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 281–288. CHI '03. New York, NY: ACM. https://doi.org/10.1145/642611.642661.

Shneiderman, Ben. 1982. "The Future of Interactive Systems and the Emergence of Direct Manipulation." *Behaviour & Information Technology*, 1 (3): 237–56. https://doi.org/10.1080/01449298208914450.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature*, 529 (7587): 484–89. https://doi.org/10.1038/nature16961.

Stanley, Kenneth O., Bobby D. Bryant, and Risto Miikkulainen. 2005. "Evolving Neural Network Agents in the NERO Video Game." *Proceedings of the IEEE*, 182–89.

Sutton, Richard S., and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*. 1st ed. Cambridge, MA: MIT Press.

Taljonick, Ryan. 2014. "Shadow of Mordor's Nemesis System Is Amazing–Here's How It Works." GamesRadar. 29 August 2014. http://www.gamesradar.com/shadow-mordor-nemesis-system-amazing-how-works/ (accessed Feb. 2018).

The Economist. 2017. “Why AI Researchers like Video Games.” 13 May 2017, sec. Science and Technology. https://www.economist.com/news/science-and-technology/21721890-games-help-them-understand-reality-why-ai-researchers-video-games (accessed Feb. 2018).

Yannakakis, Georgios N, and Julian Togelius. 2015. “A Panorama of Artificial and Computational Intelligence in Games.” *IEEE Transactions on Computational Intelligence and AI in Games,* 7 (4): 317–335.

Yannakakis, Georgios N., and Julian Togelius. 2017. *Artificial Intelligence and Games*. Springer.

5 Voice Game History

This chapter presents Study 2 in the form of a peer-reviewed journal article:

> Allison, F., Carter, M., & Gibbs, M. (2017). Word Play: A History of Voice Interaction in Digital Games. *Games and Culture*, 1555412017746305. https://doi.org/10.1177/1555412017746305

The article is included verbatim in its published format, following a discussion of how it contributes to the thesis. I led all stages of the research and writing process. My thesis supervisors, Marcus Carter and Martin Gibbs, contributed to the study design and writing of the article.

5.1 Introduction

The previous chapter reported on a study of players interacting with a game character using natural language through a text interface. This study allowed me to explore some of the issues inherent in conversational interaction gameplay, in isolation from the particular complexities of voice interaction—such as the (un)reliability of speech recognition systems and the potential for self-consciousness about being overheard during gameplay. However, some of the findings that emerged from that study were specific to text input: the way that unnoticed misspellings could cause a breakdown in communication, for instance, or the way that the act of typing sometimes distracted players from visually observing the game character. The remaining three studies in the thesis look directly at voice-operated interaction with characters and environments (VOICE) in games.

This chapter presents a complete history of voice games as a subcategory of digital game design. The results address Research Question 2:

> **RQ2** What is the historical lineage of voice game design and what are the major influences that have shaped it?

This history of voice game design was composed through an analysis of 247 voice games, published between 1973 and 2016. Based on this research, the chapter puts forward three main arguments:

- There have been seven distinct phases of voice game design, distinguished by their time period, design focus and geographic centre of activity.
- Voice game design has been highly reactive to changes in the hardware and software platforms that became popular for digital games in each time period.

- There are consistent and systematic differences in voice game design between Japan and the West, particularly with regards to the way that voice games have presented the player-character relationship.

The methodology, findings and contributions of this study are summarised in the sections below, and discussed at greater length in the published article.

5.2 Research method

For this study, a media archaeological approach was used to compose a history of the use of voice interaction in videogames. A variety of primary and secondary sources (detailed in the published article) were used to compile a descriptive database of information about commercially published voice games. Any videogame that has allowed players to use their voice to affect the game state was included in the database, with the exception of games that feature only an audio interface and no screen-based visual interface, such as smart speaker games. The final list comprised 247 games published between 1973 and 2016, which accounts for the vast majority of commercially published videogames from that time period that featured any form of voice interaction.[16] This database was examined for evidence of design trends and lineages between games of different time periods. The methods used to identify voice games and to collect and analyse the data about each game are described in more detail in Chapter 3 and in the published article below.

5.3 Summary of findings

The production and popularity of voice games have ebbed and flowed repeatedly over the past four decades. There are seven distinct phases in which various types of voice games have achieved relative prominence, interspersed by fallow periods in which few voice games (at all or in particular genres) were produced. These phases are distinguished by a combination of their time period, design focus and geographic centre of activity, and characterised as follows:

1. *Research Games:* Starting in around 1973, various academic and industrial research labs have created voice-controlled videogames as prototypes for speech recognition systems.

[16] Data collection for Study 3 (Chapter 6) identified an additional 37 games that were not included in the analysis for Study 2. Most were small, independently created Flash games published through online portals such as Newgrounds.com, but a handful were commercially published titles. Subsequent analysis of these games has not changed any of the findings or conclusions of the historical study, but it does make the claim to have "identified and classified all games that feature some form of voice interaction and have received a public release", which was made in the published article, strictly inaccurate.

2. *False Starts:* In the mid–1980s, a series of videogame consoles made extensive use of speech recognition, but were commercially unsuccessful. A basic microphone in the Nintendo Famicom led to the inclusion of simple voice gameplay features in many popular videogames in Japan.
3. *Digital Companions and the DS in Japan:* In the late 1990s and early 2000s, a series of Japanese videogame consoles (most notably, the Nintendo DS) afforded voice input. Many games made use of this feature by presenting a virtual "pet" or "friend" for the player to converse with.
4. *Command and Control in the West:* In the mid-2000s, a succession of military strategy videogames incorporated voice command gameplay. These games, developed in North America and Europe, offered a distinctly more authoritarian style of voice gameplay compared to the voice games that had been produced in Japan.
5. *The Karaoke Boom:* Between the early 2000s to the early 2010s, karaoke was one of the most popular videogame genres worldwide.
6. *The Kinect Effect:* Microsoft's Kinect console, which combined a speech recognition system with a motion-sensing camera, led to the inclusion of voice gameplay features in a wide range of high-profile videogames for several years in the early 2010s.
7. *Add-Ons, Mods, and Open Platforms:* From the mid-2010s onwards, voice gameplay began to appear in an increasing number of small-scale, low-budget, independently produced videogames and videogame mods.

The overall impression of these phases is that voice games have gradually improved in both their sophistication and their cultural centrality over time. However, the historical analysis also demonstrates that voice game design has been heavily constrained and shaped by the platforms that happened to dominate the commercial videogame market at different points in time. The article describes a pattern of "platform reactivity", in which voice games repeatedly proliferate on game consoles that afford voice input, but become scarce again when future game consoles exclude voice input features. This highlights the importance of studying game platforms as a basis for understanding design trends in games themselves, as exemplified in the disciplines of platform studies (Bogost & Montfort, 2009) and media archaeology of games (Parikka & Suominen, 2006).

The final argument of the published article is a counterpoint to the notion of platform reactivity. Despite dramatic changes in the underlying technology of voice interfaces, the basic genres of voice games have remained surprisingly consistent over time. Games from the 1980s present a similar range of voice-based activities to games from the 2010s, when viewed in terms of what the game fiction indicates that the player is doing with their voice. Across national boundaries, however, game genres have shown consistent differences. Voice games developed in Japan have often involved relatively intimate and egalitarian conversation between the player and a game character, whereas voice games developed in North America and Europe have more often involved authoritative command-based interactions with game characters that serve as strategic pawns rather than fleshed-out characters. The research approach used in this study cannot determine whether these stem from broader cultural variations. However, prior empirical studies have found differences in North American and East Asian players' responses to game characters, particularly with regards to perceptions of social presence (Luo et al., 2019; Yoon & Vargas, 2016), which is consistent with the hypothesis that cultural attitudes may be driving the regional variations in game design. Regardless of their origin, these inter-country differences show that the design of voice affordances in videogames has been influenced by factors other than the capabilities of the underlying voice technology.

5.4 Contributions to the thesis

This chapter primarily answers Research Question 2: "What is the historical lineage of voice game design and what are the major influences that have shaped it?" The media archaeology study traces both commonalities and changes in voice games, across the entire period of time in which voice interaction has been a feature of digital game design. The details of the seven historical phases provide a genealogical map of the lineages behind modern voice game design, and the broad historical trends observed in the study—platform reactivity and geographic variance—illustrate some of the factors that have influenced it.

This chapter also provides some initial insights towards Research Question 3: "What design patterns are currently implemented in voice games?" It describes various recognisable forms of voice gameplay, many of which have remained surprisingly consistent over time despite dramatic changes in the underlying voice technologies. The following study, presented in Chapter 6, expands upon this with a detailed analysis of the specific recurring design patterns that have been evident in voice game design over the course of its history. The findings of these two studies form the basis for one of the major contributions of this thesis: the five genres of voice gameplay discussed in Chapter 8.

As a final note, subsequent evidence collected for the thesis has supported the conclusions of Study 2. The published article points to the beginnings of a boom in small, independently produced voice games that use voice interaction as a central feature rather than a optional mode; it concludes with a prediction that this boom would continue, based on the historical trends identified in the study. That prediction was borne out in Study 3 (Chapter 6), which shows that the number of voice games released in 2017 was more than double the total of any prior year, even without considering the rapid growth in smart-speaker audiogames, most of which fit the profile described in the section entitled "Add-Ons, Mods, and Open Platforms".

Article

Games and Culture
1-23

Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1555412017746305
journals.sagepub.com/home/gac
SAGE

Word Play: A History of Voice Interaction in Digital Games

Fraser Allison[1], Marcus Carter[2] and Martin Gibbs[1]

Abstract

The use of voice interaction in digital games has a long and varied history of experimentation but has never achieved sustained, widespread success. In this article, we review the history of voice interaction in digital games from a media archaeology perspective. Through detailed examination of publicly available information, we have identified and classified all games that feature some form of voice interaction and have received a public release. Our analysis shows that the use of voice interaction in digital games has followed a tidal pattern: rising and falling in seven distinct phases in response to new platforms and enabling technologies. We note characteristic differences in the way Japanese and Western game developers have used voice interaction to create different types of relationships between players and in-game characters. Finally, we discuss the implications for game design and scholarship in light of the increasing ubiquity of voice interaction systems.

Keywords

voice interaction, speech recognition, media history, game design, media archaeology

[1] Interaction Design Lab, School of Computing and Information Systems, The University of Melbourne, Parkville, Victoria, Australia

[2] Department of Media and Communications, The University of Sydney, Camperdown, New South Wales, Australia

Corresponding Author:

Fraser Allison, Interaction Design Lab, School of Computing and Information Systems, The University of Melbourne, Parkville, Victoria 3010, Australia.

Email: fraser.allison@unimelb.edu.au

A person speaks casually to a computer opponent over a game of chess. The computer answers in a synthesized but seemingly intelligent voice; they exchange witticisms and debate recent events. Such scenes have long been predicted by both computer scientists and science fiction writers. For decades, we have imagined that a world of playful, talkative, genuinely intelligent artificial intelligence (AI) is close on the horizon—but like a mirage, it has seemed to melt away at our approach, as each step forward has revealed unanticipated challenges and technical setbacks (Huang, Baker, & Reddy, 2014; Munteanu et al., 2013). This pattern of "false dawns" (Aylett, Kristensson, Whittaker, & Vazquez-Alvarez, 2014, p. 754) and disillusionment has led to a period of neglect for the study of voice interaction (Munteanu et al., 2013).

Three parallel developments have given voice interfaces renewed relevance, particularly in the context of game design. The first is a series of rapid improvements in speech recognition, driven by increased computational power, cloud computing services, larger data sets, and improved machine learning techniques (Huang et al., 2014). The second is a growing presence of conversational AI characters in technology systems outside of digital games, sometimes modeled directly on game characters, such as Microsoft's Cortana (Warren, 2014). The third is a rapid proliferation of microphones and speech processing systems in consumer gaming devices from all-purpose devices such as smartphones and home computers to more specialist gaming equipment such as the Microsoft Kinect. These developments point to a growing opportunity for game designers to build novel gameplay experiences around voice interaction[1] and for voice interface designers in other fields to learn how games have made spoken interaction with virtual characters enjoyable and engaging for users (Luger & Sellen, 2016, p. 1). The history of digital games provides guidelines for both of these perspectives.

In this article, we review the history of voice interaction in digital games from its origins in computer science research in the 1970s to the present day. The patterns of experimentation, popularity, and commercial failure seen in this time period offer clear indications about how certain interaction styles are received by users, in a context where a positive user experience is the overriding consideration. We identify seven distinct phases, chronologically overlapping but distinguishable by their content, design focus, and geographic centers of production. We find that the development of voice interaction games has been highly reactive to changes in the technological platforms that have enabled digital gaming and that hardware has both shaped and constrained the design space of voice interaction games. We also identify regional differences in voice interaction game design and compare the characteristic style of games designed in Japan to those designed in North America and Europe. Finally, we argue that the current phase of voice interaction game design is qualitatively different to those that have come before, with a proliferation of small and independent developers creating low-budget games that have a tighter focus on the affordances of the voice interface, suggesting that voice interaction game design may be in a signal period of experimentation and development of the form.

Data and Method

Histories of speech recognition have been written by Huang, Baker, and Reddy (2014), Rabiner and Juang (2008), Baker et al. (2009), and Furui (2005). In each of these, the focus is firmly on describing developments in the underlying technologies that enable speech recognition systems. However, as several scholars have noted, there has been a comparative lack of research on the interface design and user applications of voice interfaces (Aylett et al., 2014; Luger & Sellen, 2016, p. 1; Munteanu et al., 2013). Rather than rehash the contents of existing histories by explaining in detail the technology that enables voice interaction in games, we focus in this article on reporting the applications for which voice interfaces have been used in games from a player's perspective. This is intended to complement the existing literature that provides a more technically oriented perspective. However, where relevant we refer to aspects of the technology that are informative for understanding the allowances and limitations of a particular voice interaction system.

To reconstruct the history of voice interaction in digital games, we set out to archive and categorize every digital game that has received a public release and features voice input as a mode of interaction. To achieve this, we conducted searches on related key words within selected publications in game studies and human–computer interaction, before broadening our search to general academic databases, Google Scholar, and the wider web. We searched for key words on the game review aggregator site Metacritic, the digital distribution platform Steam, game-related wikis such as Giant Bomb, discussion forums such as Reddit, and enthusiast online game catalogues such as the Video Game Console Library. Additional information about each game was compiled from voice interaction research papers, platform studies texts, historical books, magazine articles, news reports, game reviews, published interviews, sales data, marketing materials, publisher websites, and the games themselves. Wherever possible, findings were checked with multiple sources. We used the automated translation service Google Translate to review websites in languages other than English; however, this remains a study primarily informed by English-language sources.

In our analysis of these materials, we have sought to classify design patterns and trace connections between games that employ similar styles of voice interaction. According to Suominen's (2016) categorization of game histories, our approach is primarily genealogical, as we map out the lineages and subcategories that have emerged among voice interaction games, and to some extent pathological, as we seek to explain these developments by reference to their conceptual antecedents and material preconditions. In a loose sense, we are engaged in media archaeology (Apperley & Parikka, 2015; Parikka & Suominen, 2006); although within the constraints of this article, we can account only briefly for the broader media-cultural contexts in which voice interaction games have developed.

Antecedents to Digital Voice Interaction Games

The idea of conversing with inorganic objects is far older than the digital age. Around the third century before common era (BEC), the Greek sophist Callistratus wrote of a statue of an African king that was able to speak:

> There was in Ethiopia an image of Memnon, the son of Tithonus, made of marble; however, stone though it was, it did not abide within its proper limits nor endure the silence imposed on it by nature, but stone though it was it had the power of speech. (Fairbanks, 1931, p. 407)

Speech recognition technology itself predates the digital computer by three decades in the form of children's toys. The first such device was a grumpy-looking toy bulldog (Figure 1), variously known as Radio Rex or the "Wireless Pup," that sprang out of its kennel when its name was called (Cohen, Giangola, & Balogh, 2004, p. 15; Gernsback, 1916, p. 104; "Toys That Obey Your Voice," 1916, p. 718). First patented in 1916 and refined in subsequent years (Berger, 1916, 1917, 1918), Rex worked on a mechanical form of protospeech recognition: the lever that pushed Rex forward was held back by an electromagnet, whose power connection was tuned to disconnect when vibrated at around 500 Hz. This matches the frequency of the vowel sound in "Rex"—ironically calibrated to the vocal pitch of an average adult male rather than a child.

With the advent of computers, machines that could think and speak entered the realm of plausibility, and stories of humans matching wits and words against AI quickly became a staple of science fiction. In the 1960s, both *2001: A Space Odyssey* and *Star Trek* showed audiences a futuristic spaceship AI that addressed its crew by voice and played chess against them, only a few years before voice-controlled computer chess would become a reality.

Research Games

In 1973, Reddy, Erman, and Neely at Carnegie Mellon University (CMU) published a description of *Voice-Chess*, a digital chess game built on the Hearsay-I speech recognition system. Like the AIs of *Star Trek* and *2001: A Space Odyssey*, *Voice-Chess* could recognize standard chess instructions such as "Bishop to Queen Three" and update the on-screen board accordingly. This is perhaps the earliest example of a prototype speech recognition system that used a game as its use case. The turn-taking structure and well-defined vocabulary of chess moves provided a template for utterances that were constrained and standardized enough for Hearsay-I to process.

The development of Hearsay-I was funded by the U.S. Advanced Research Projects Agency Speech Understanding Research program (Lea & Shoup, 1979; Rabiner & Juang, 2008). Several other pioneering systems were developed under this program, including Hearsay-II and Harpy at CMU and Hear What I Mean at Bolt Beranek and Newman. Lee (1988) at CMU would later combine elements of Harpy

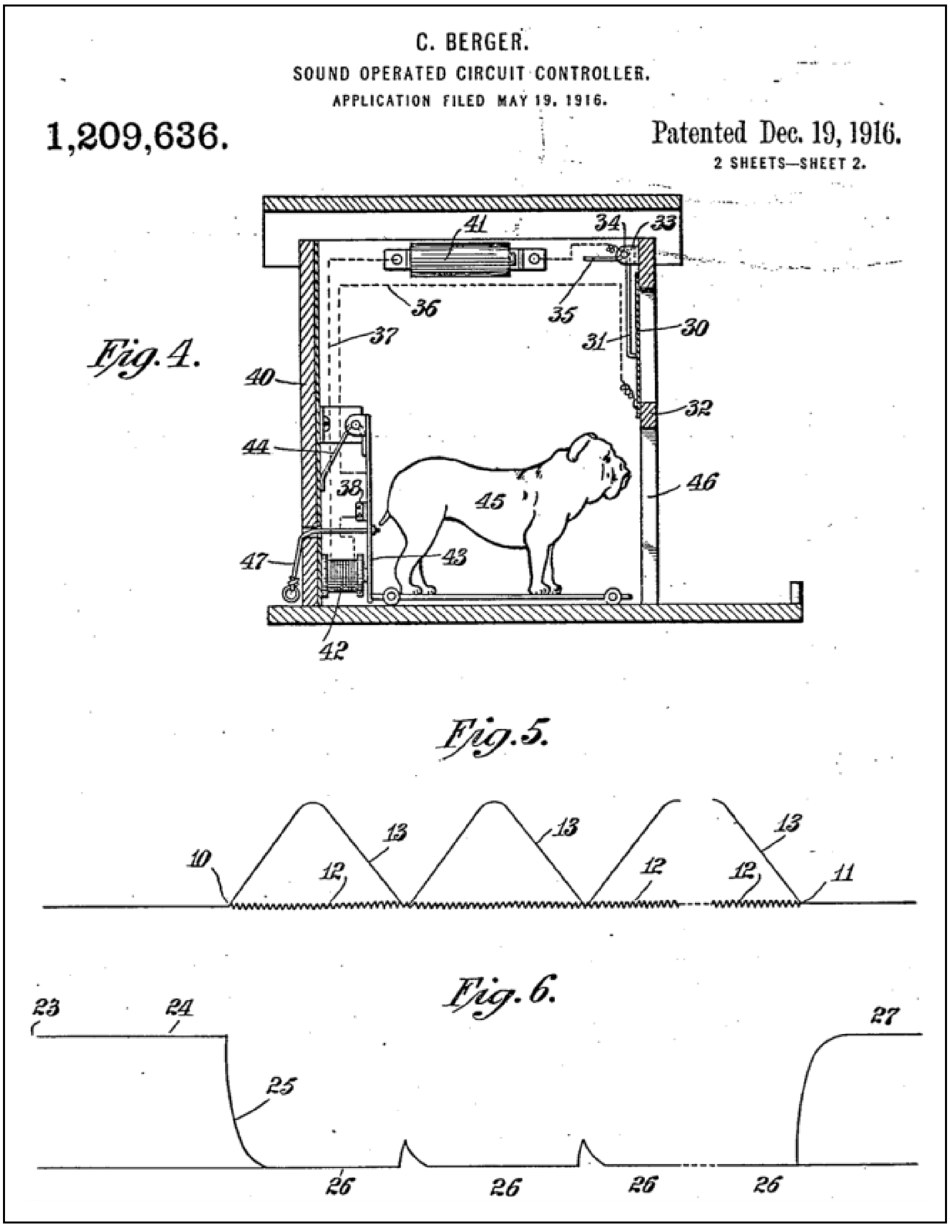

Figure 1. Patent illustrations for Radio Rex (Berger, notes, p. 2).

with improved statistical methods to create the first Sphinx system, whose successors continue to be among the mostly widely used tools for speech recognition today, including in digital games such as *In Verbis Virtus*.

As speech recognition technology has improved, researchers have continued to use digital games as prototypes or research objects, particularly when exploring novel interaction methods (e.g., Dow, Mehta, Harmon, MacIntyre, & Mateas, 2007; Nanjo, Mikami, Kunimatsu, Kawano, & Nishiura, 2009). One notable branch of human–computer interaction research has explored voice input as an alternative control modality for users with motor impairments (Harada, Wobbrock, & Landay, 2011; Mustaquim, 2013), sometimes using nonverbal voice actions such as volume

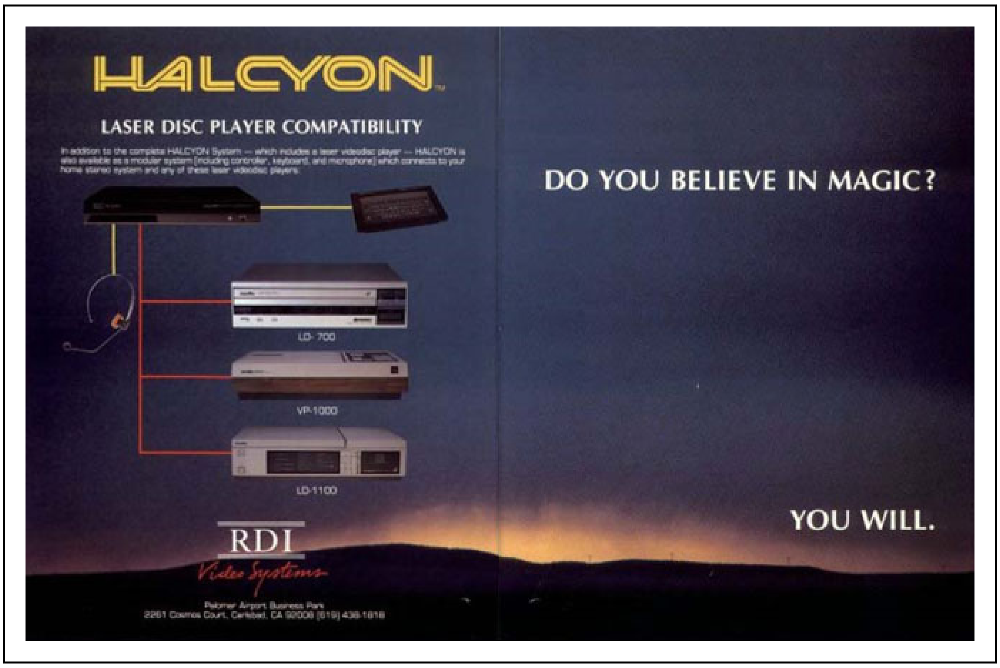

Figure 2. Flyer for the Halcyon (Kinder & Hallock, n.d.).

or pitch rather than words (Harada et al., 2011; Igarashi & Hughes, 2001; Sporka, Kurniawan, Mahmud, & Slavík, 2006).

False Starts

By the early 1980s, speech recognition technology was being added to mass-market products. The Milton Bradley Company was first to bring voice interaction to the commercial digital games market. In 1983, it released the "MBX" module for Texas Instruments' TI-99/4A computer, which added speech recognition and speech synthesis capabilities (Herman, 1997, p. 282; Mace, 1983, p. 27). Milton Bradley published seven games that took advantage of the speech recognition feature, including *Championship Baseball*, in which players could tell their baseball team where to throw the ball. Seeing this, Atari commissioned Milton Bradley to develop a "Voice Commander" module for the Atari 2600 console (Herman, 1997, p. 92).

An even more ambitious project was the "Halcyon" game console (Figure 2), created by Rick Dyer's RDI Video Systems, developer of the popular arcade game *Dragon's Lair*. Unlike the Milton Bradley add-ons, Halcyon was a whole console dedicated to voice input and designed to be controllable entirely through spoken commands. Marketing for the Halcyon proclaimed it to be "the first and ONLY home entertainment system featuring speech synthesis, voice recognition, and artificial intelligence" (Kinder & Hallock, n.d.), and in interviews, Dyer described the system as being similar to the AI HAL 9000 from *2001: A Space Odyssey* ("Lasers and Computers," 1985). The console contained a dedicated speech recognition subsystem, which could recognize a player saying their own name, and had the capacity to learn more than 1,000 words. Commands aside from "yes" and "no" had to be pretrained by the player before the system could recognize them, and even then the error rate was frustratingly high (Dark Watcher & 98PaceCar, n.d.; "Lasers and Computers," 1985), but nevertheless this technology was highly advanced for a consumer product at the time.

All three platforms fell victim to bad timing. In 1982, the markets for both home computers and video game consoles had experienced a glut, which by 1983 had spilled over into a costly price war. A flood of cheap game hardware and poor-quality games caused a sharp drop in profits for video game companies and a decline in consumer interest, which led to the industrywide North American video game crash of 1983 (Wolf, 2008, p. 105). Texas Instruments withdrew from the home computer business, dooming the MBX module a month after it began production (Mace, 1983, pp. 22–27). Atari shelved the voice commander indefinitely, prompting Milton Bradley to bring a lawsuit against it for breach of contract (Herman, 1997, p. 92). Halcyon was launched in 1984 with only two games and a price tag of more than US$2,000; sales were nonexistent, and RDI Video Systems went bankrupt almost immediately (Wolf, 2008, p. 101; Wolf, 2012, pp. 353–354).

At the same time, however, Japan was beginning its rise to dominance in video game production. In 1983, Nintendo released the Family Computer (Famicom) console to unprecedented success (Altice, 2015, p. 200). The Famicom included voice interaction almost as an afterthought: It came with two controllers, one of which had a built-in microphone that could amplify the player's voice through the television speaker. The audio processing system was basic, capable only of registering sound as a binary on-or-off signal (Altice, 2015, p. 25). For the first 2 years of the Famicom's release, no games made use of its microphone; but from 1985, a number of developers began to add voice interaction elements to their games usually in the form of an optional or secret ability rather than a core game mechanic. Given the extreme limitations of the hardware, Famicom games show a surprisingly broad range of voice actions: from shouting to kill enemies (*The Legend of Zelda*), to blowing air to spin a roulette wheel (*Kamen Rider Club*), to "bargaining" with a merchant (*Kid Icarus*).

In 1987, the first dedicated karaoke video game was published on the Famicom: *Karaoke Studio*. The cartridge for *Karaoke Studio* included a physical attachment for a stage-style microphone. The player sang into this microphone while music played and song lyrics appeared on screen, accompanied by an animated video. The player's singing was amplified through the television speakers and given a score at the end of the song; although this was based on little more than timing, and as with later karaoke games, there was no element of speech recognition ("Karaoke Studio," n.d.). It would be 10 years until the next karaoke video game was released, but since that time the genre has continued to follow the basic form laid down by *Karaoke Studio*.

Despite its novelty, voice interaction was never seen as a big part of the Famicom's success. Later versions of the console, including the Nintendo Entertainment

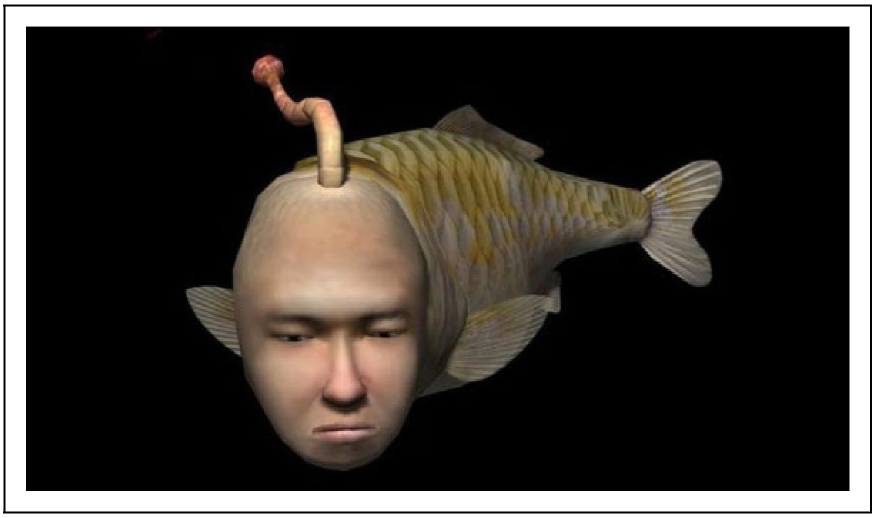

Figure 3. The main character in *Seaman*, a fish with a human face.

System (NES) that was sold outside Japan, removed the microphone and the flow of games on the system that used the feature gradually dried up (Altice, 2015, p. 26). It would be a number of years before Nintendo, or any other company, reintroduced voice interaction technology to a games console.

Digital Companions and the DS in Japan

The success of the Famicom/NES helped to revive the video game industry and moved its center of gravity from the United States to Japan. Nintendo, Sega, and Sony dominated the game console market throughout the 1990s, and they defined the capabilities of their respective game consoles in terms of computing power. The early and late stages of the decade were commonly referred to as the "16-bit era" and "32-bit era" in games, reflecting the popular focus on CPU size at the time. None of these consoles featured a built-in microphone, and consequently, no voice interaction games were released for those platforms.

In 1998, Nintendo again brought voice interaction back to console gaming with the voice recognition unit, an add-on to the Nintendo 64 console that paired a microphone with a speech recognition pod that could encode the player's voice inputs (Provo, 2000). However, only two games made use of the unit—1998s *Hey You, Pikachu*! and 1999s *Densha De Go*! *64*—and only the former was sold outside Japan. Although Nintendo quickly abandoned the voice recognition unit, *Hey You, Pikachu*! earns a significant place in this history as the first "virtual pet" game to use voice interaction.[2] The game allows players to befriend and converse with the title character, a friendly Pokémon, while controlling a relatively nondescript child avatar. This design pattern was repeated in several games by Japanese developers over the following few years: *Seaman* on the Sega Dreamcast (Figure 3), *Lifeline* on the Sony PlayStation 2, and *N.U.D.E.@* on the Microsoft Xbox. Like *Hey You, Pikachu*!, these games use a first-person perspective but place the narrative focus on a second-person “friend” or “pet” like Pikachu. In all four games, the player’s main task is to converse with the virtual companion to teach them or guide them through the gameplay scenarios, as the relationship between the characters develops over time. This relationship-oriented voice interaction has remained a characteristic of Japanese game design and a rarity in Western game design for the next decade.

Yet another configuration of voice interaction game design appeared on the PlayStation 2 in 2001. *Yoake no Mariko* positions the player as an actor in various B-grade films and presents them with an on-screen script including stage directions for each line’s delivery. The player is scored on how well they perform their lines, leading to an experience somewhere between karaoke and role-play. Although not an especially popular or influential game, *Yoake no Mariko* is notable as an example of the use of voice input to create a distinct new game genre.

The fact that this crop of voice interaction games emerged within the space of a few years after the long winter of the 1990s was not a coincidence, but neither was it driven by a sudden surge in consumer demand for voice interaction. Rather, it was a response to the hardware. This was the era in which consoles began to allow online multiplayer gaming, beginning with the Sega Dreamcast. Online voice chat was a fast and socially engaging way for players to communicate without taking hands off the controller (Wadley, Carter, & Gibbs, 2015, p. 360), and so each of the major game platforms received a microphone peripheral that players could purchase, often bundled with an online game. Once microphones had become common, voice interaction became a much more cost-effective prospect for game developers to explore, as consumers no longer needed to pay for a dedicated microphone for the sake of just one or two games. Conversely, a voice interaction game that came bundled with a microphone, such as *Seaman*, was a more attractive prospect to consumers, as the microphone could be used for online voice chat after the game itself was done with.

Meanwhile, Nintendo continued to experiment with voice interaction systems, separate from any online voice chat applications. Several games on the Nintendo GameCube were bundled with a microphone to allow voice control, including the bizarre *Odama*, in which the player commands the movement of feudal Japanese soldiers to avoid being crushed by an enormous mystical pinball. This culminated in the system that brought voice interaction gaming to its largest audience yet: the Nintendo DS.

The Nintendo DS was the successor to Nintendo’s Game Boy series, and it extended the handheld console format in a number of ways. Most prominent were its two LCD screens including a touch screen and stylus. It was also the first game console since the Famicom to include a built-in microphone. Game developers responded to this unorthodox set of features with new and unusual game mechanics. *Brain Age*, one of the console’s flagship titles, tests the player’s mental agility with rapid-fire color naming. *Nintendogs* lets players teach tricks to a virtual dog by calling commands like “sit” or “roll over.” The *Phoenix Wright* series, about the trials of a cartoon lawyer, allows players to interrupt court proceedings with a shout of “Objection!” *The Legend of Zelda: Phantom Hourglass* asks the player to extinguish torches and spin windmills by blowing into the microphone. Each of these games was critically and commercially successful, and the DS itself sold more than 150 million units worldwide, rivalling the PlayStation 2 as the highest selling game console of all time (Lynch, 2013). This brought voice interaction gaming to a much wider audience than had ever experienced it before, in a variety of styles. Speech recognition was not yet a seamless technology, however, and the novelty of voice inputs was offset by the console's tendency to misunderstand or fail to respond to the player's commands.

Command and Control in the West

In North America and Europe, voice interaction gaming took a long time to recover from the crash of the 1983. When it did return, it took on a distinctly different character than the style of voice interaction games that were being made in Japan.

The first commercial video game on the PC to feature built-in voice interaction was *Command: Aces of the Deep* in 1995. In this World War II submarine simulator, the player takes on the role of a German U-boat captain and can speak orders to their crew through a microphone. Although *Aces of the Deep* was well received, the voice recognition was criticized for being unreliable ("150 Best Games of All Time," 1996, p. 74; Stevens, 1999). It proved to be the last game in the *Aces* series, and for several years, it remained the only major PC game to include voice interaction. Unlike developers for the original Famicom or later platforms, PC game developers at the time could not rely on their consumers to have a microphone that could take advantage of any voice interaction features.

In retrospect, however, *Command: Aces of the Deep* clearly foreshadows the style of voice interaction that was to become characteristic of Western game design in the decade to come. Firstly, it puts the player in the role of a military authority figure in a combat situation, who barks orders at a team of underlings. Secondly, it presents a fundamentally top-down communication structure, in which the player's subordinates are semiautonomous agents of the player's strategic will, rather than fully fledged characters with inner lives and thoughts of their own to discuss. Thirdly, it affords voice control as an option that can easily be replaced with mouse clicks and hot keys rather than a central and indispensable part of the game's design. This impersonal command-and-control style is indicative of the common patterns of voice interaction that came to predominate in games made by North American and European developers. Games made in Japan, by contrast, often employed a more conversational style that placed the virtual character on a more even level with the player and involved peaceful rather than combat-based scenarios.

The new wave of voice interaction games began to arise in North America after headset microphones were released for the Sony PlayStation 2 and Microsoft Xbox consoles. These headsets were primarily designed to allow voice chat between players in online games but were immediately used for voice interaction. The first game to use the PlayStation 2 headset was *SOCOM: U.S. Navy Seals*, a tactical third-person perspective shooter, which included an option for the player to give verbal orders to the squad members under their command. This model was repeated in four more games in the *SOCOM* series, along with a string of other squad-based shooters: *Rainbow Six 3* and *Rainbow Six Vegas*, *SWAT: Global Strike Team*, *Greg Hastings Tournament Paintball Max'd*, and, on the PC, *Unreal Tournament 2004*. In each case, the player was positioned as the leader of a combat team who could issue instructions to their semiautonomous and largely characterless teammates.

The tactical combat command trend culminated in *Tom Clancy's EndWar*, the first game to attempt real-time strategy on a console using voice commands. Unlike the squad games that came before it, voice command was central to the design of *EndWar*. Although it could be controlled using the game pad, the game encouraged voice input as the primary method of control. *EndWar* featured an unusually flexible speech recognition system that could understand multipart command strings in a Who-What-Where pattern such as "unit two, attack hostile one" or "unit four and all gunships, create group and move to target." Reviewers praised the "superbly implemented voice-recognition system" (Reed, 2008) for providing an elegant and reliable solution to the ergonomic constraints of the console format. Unfortunately, the nonvoice aspects of the game were criticized for lacking strategic depth and narrative interest. *EndWar* attracted a small cult following but ultimately entered the annals of gaming as a curio rather than a way forward.

Following the failure of *EndWar* to find a large audience, voice interaction petered out in most game genres. However, one category bucked the trend and by itself took voice interaction gaming to its greatest ever heights of popularity: karaoke.

The Karaoke Boom

Karaoke has been far and away the most successful genre of voice interaction gaming in its history. The first karaoke video games appeared on the Famicom, first in the form of a minigame within the action-adventure title *Takeshi No Chōsenjō* and later as a fully developed game in *Karaoke Studio*, described above. The genre fell dormant in the 1990s, but with the revival of voice interaction gaming at the turn of the millennium, karaoke began a 10-year run as one of the most popular game genres of the era. Karaoke was a centerpiece of the "party game" style at the same time as video games began to shake off their children's toy image and achieve popularity with older demographics. Karaoke games became a feature of adult social events as a social bonding activity, often accompanied by alcohol (Fletcher & Light, 2011).

Once again, the revival began in Japan before it reached the rest of the world. The short-lived *Dream Audition* series, beginning on the PlayStation 2 in 2000, followed the basic template laid down in *Karaoke Studio*: Players were given a stage-style microphone and a list of well-known pop songs and were scored based on their ability to hit the right notes as they sang along to an instrumental version of the song. Lyrics appeared on screen in time to the music, although as the system tracked only

Figure 4. Piano roll-style pitch indicator in *Karaoke Revolution*.

vocal pitch and not phonemes, the player was free to sing whatever words (or sounds) they chose. In 2003, *Karaoke Revolution* took up the format and added a piano roll-style pitch indicator for the vocals (Figure 4). It depicted the player as a character singing to a crowd and framed the scoring as the singer's ability to entertain the crowd; the player could be cut off mid-song if the "crowd meter" dropped too low due to out-of-tune singing. This format became the new standard that subsequent karaoke games would imitate.

The following year, *SingStar* was released in Europe and Oceania, following the template laid down by *Karaoke Revolution*. They were joined in 2007 by *Rock Band*, and shortly afterward by *Guitar Hero World Tour*, which expanded the karaoke format to include plastic guitars and drums that could be played alongside the vocals. The *Lips*, *Boogie*, and *Sing It*! series also joined the karaoke stable at this time. Many of these series released multiple games per year across multiple gaming platforms, and every major game console featured at least one karaoke series.

In 2011, near the height of the karaoke boom, Fletcher and Light conducted two ethnographies of *SingStar* player groups. They describe *SingStar* as a "glue technology" (Fletcher & Light, 2011, p. 1) that facilitated social bonding and intergenerational connections within the groups. They also documented a variety of collaboration and co-play practices, wherein the microphone was shared between varying numbers of players, with secondary singers often acting as a social support to the primary player holding the microphone.

The karaoke craze lasted through two generations of consoles. From the launch of *Karaoke Revolution* in 2003 to the end of 2012, more than 120 different karaoke games were released (counting multiplatform releases as a single game). More than 70 of these were published in just 3 years between 2008 and 2010. But as the market became oversaturated, consumer interest in the genre finally waned. In 2011, the rate of new releases began to wind down, and two attempts to revive the genre with new games in 2015 met with little success (David, 2016; Yin-Poole, 2016).

The Kinect Effect

By 2010, when the karaoke craze was at its peak, other forms of voice interaction gaming had fallen into a lull. The Nintendo DS was 6 years old and soon to be superseded and had always been constrained by the limited sensitivity of its microphone. Voice-controlled strategy games had failed to gain traction after the limited success of *EndWar*, and the *SOCOM* series that pioneered voice control in the tactical shooter genre had stopped including it as a feature.

The situation was reversed in late 2010 when Microsoft released the Kinect, a multipurpose sensor bar for the Xbox 360. The Kinect was primarily marketed as a motion sensing camera, which could track players' movements in three-dimensional space. However, it also included a microphone array and a speech recognition system. This provided a ready-made facility for game developers to add voice interaction to their games without needing in-house expertise in speech recognition. It also quickly established a substantial consumer market for voice interaction games, as within its first 2 months the Kinect became the fastest-selling consumer electronics device in history (Peckham, 2011).

This resulted in a new wave of voice interaction games, although it took some time to gather momentum as most early Kinect games focused on motion control rather than voice interaction. The launch title *Kinectimals*, a virtual pet game following the *Nintendogs* template, allows players to teach their animal tricks using voice commands such as "sit down" and "roll over." The following year, limited voice controls were included in *Dance Central 2*, *Kinect Sports: Season 2*, and *Halo: Combat Evolved Anniversary*. In 2012, the sci-fi squad-based shooters *Binary Domain* and *Mass Effect 3* both allowed players to give verbal orders to their teammates in battle and to converse with characters by reading out scripted lines of dialogue. The crest of the wave came in 2013, with the launch of the Xbox One console. For the first 6 months of its release, every Xbox One came bundled with a Kinect as a mandatory add-on (Turner, 2014), further expanding the potential user base for voice interaction. There followed a run of games that took advantage of voice input, in a diverse range of genres: action games such as *Ryse: Son of Rome*, family-friendly fare such as *Zoo Tycoon*, sports games such as *FIFA 14*, racing games such as *Forza Motorsport 5*, and role-playing games such as *Dragon Age: Inquisition*. More than 30 major releases to date have used the Kinect's voice recognition. The voice affordances vary considerably between these games, although it is notable that all of them make voice input an optional feature rather than a core element of the design—perhaps reflecting the fact that many Xbox One owners do not have a Kinect. In most of these cases, voice commands are simply an alternative means of activating actions that are already afforded by the primary control scheme; they form a redundant control layer rather than a novel implementation of voice-specific game mechanics.

Not all the Kinect games use spoken commands in this way, however. Some focus on creating the illusion that the characters in the game world are simply able to hear the sound of the player. The stealth horror game *Alien: Isolation* includes an optional “noise detection” mode, in which any moderately loud noise detected by the Kinect microphone alerts the game’s fearsome antagonist (the titular alien) to the location of the player–character. Similarly, the stealth action game *Splinter Cell: Blacklist* allows the player to call out to in-game enemies, prompting them to investigate.[3] This feature was well received by players and reviewers alike for supporting a sense of connection with the game world (Carter, Allison, Downs, & Gibbs, 2015, p. 267).

Unlike in earlier eras, player feedback on the voice interaction features of Kinect games did not focus on the accuracy of speech recognition, which was generally considered to be acceptably reliable for many accents. Instead, voice interaction was critically appraised largely according to two considerations: its efficiency compared to regular controls and its appropriateness within the context of the game’s imaginary and overall design (Carter et al., 2015, p. 266). Voice commands that allowed players to role-play momentarily as their character were praised, whereas voice commands that called attention to the player’s position outside the game world were criticized as feeling “unnatural” and “embarrassing”—an effect that Carter, Allison, Downs, and Gibbs (2015) termed “identity dissonance” (p. 268).

Add-Ons, Mods, and Open Platforms

Most of the games described so far have been the products of well-established development and publishing studios. One reason for this is that the early game platforms on which voice interaction was enabled were consoles, which had higher barriers to access and distribution than home computers. However, the greater openness and flexibility of the home computer as a platform has meant that voice interaction gaming has been available to those willing to seek out extra software or game mods.

The predecessors to today’s more open approach to voice interaction games emerged in the 1980s, when speech recognition systems became available for home PCs that included voice control schemes for virtual flight simulators (Schoen, 1985). By the end of the 1990s, such voice control software was marketed specifically for PC games including Sontage Interactive’s Game Commander and Microsoft Sidewinder Game Voice. These systems worked by converting spoken commands into user-specified keyboard or mouse inputs. Thus, the voice control was a separate overlay rather than an integrated part of the game itself, although a few games came with preconfigured profiles for voice control software, including *Star Trek: Bridge Commander.*

Over time, as free speech recognition software became more accessible, some players began to take it upon themselves to add voice interaction to PC games by

Figure 5. Overhead map view in *Plan Be*.

creating and distributing unofficial game modifications ("mods") that leveraged software from open-source speech recognition systems such as CMU Sphinx. A notable example of player-led development of voice interaction is the role-playing game *The Elder Scrolls V: Skyrim* by Bethesda Game Studios. Magical incantations are a central action in Skyrim's imaginary and game mechanics, but when it was first released in 2011, the game had no voice interaction features. Modders quickly responded by creating mods that enabled players to perform their incantations vocally; one such mod had been downloaded by more than 49,000 different users as of June 2017.[4] In 2012, Bethesda responded by adding similar voice command features in an official update to the Xbox 360 version of the game.

In the 4 years since voice commands were brought to *Skyrim*, there has been a proliferation of smaller scale, smaller budget voice interaction games produced by independent developers. These have appeared for both PCs (*Bot Colony*, *There Came an Echo*, *In Verbis Virtus*, *Plan Be*) and mobile phones (*The Howler*, *Ah*! *Bird*, *Mayday*! *Deep Space*), the latter having the advantage that microphones are a standard feature of the platform. Notably, all of these small-scale independently developed games have made voice interaction a central focus of their player experience; several are unplayable without a microphone. As a result, their design engages more closely with the affordances of voice than most larger studio productions, for which voice input is merely a bonus. For example, both *Plan Be* and *Mayday*! *Deep Space* are stealth games that provide the player with a "radio" connection to an in-game character, whom they must guide through a dangerous facility with verbal directions while looking at an abstracted overhead map (Figure 5). Through the voice of the character "on the ground," these two games evoke a richer sensory environment than they depict visually, creating a heightened tension and allowing the player's imagination to fill in the scene (Shimomura, 2015). Reviews attributed a heightened experience of social involvement to the spoken interaction format:

> This is the rarely explored power of voice: speaking to a character with your own actual voice connects you to them, forges a bond. (Shimomura, 2015)

Discussion

Voice interaction has had a place in digital games, for research and entertainment, since 1973. But its path toward acceptance and popularity has not been smooth. Hardware and software platforms have acted as constraints on the design space of voice interaction, leading to a tidal pattern in which voice interaction games have repeatedly risen on a wave of novelty with a new game platform, only to fall away as its limitations are discovered. We have shown that the development of voice interaction game design has been responsive to the platforms on which digital games are played, but that this influence is not deterministic, as developers in different countries have taken distinctive approaches to their use of the technology.

Platform Reactivity

Although voice interaction gaming had something of a false start in the 1980s, achieving only limited success with the severely limited voice input system on the Famicom, it is notable how many different genres of voice interaction games were present in those early years. From karaoke singing to bargaining with merchants to blowing air and directing a sports team, the games of that era foreshadowed many of the genres that would emerge on more capable platforms. Nevertheless, the technology could not support the ambitions of game makers or the demands of players. The long winter that followed for voice interaction games lasted until game platforms once again found a reason to include microphones, with the advent of online gaming. That latent consumer interest in voice interaction remained throughout this period is suggested by the stratospheric success of the Nintendo DS, which was the first platform to foreground voice interaction as a central game mechanic in its most popular titles. We do not mean to imply that voice interaction was the primary reason for the success of the DS, but its prominence in the design and marketing of bestselling games such as *Brain Age* and *Nintendogs* suggests that it was a contributing factor. On other platforms around the same time, the digital companion games and tactical combat games that adopted voice interaction attracted a smaller audience, reflecting the lower market penetration of headsets and other microphones on these consoles.

What this history has clearly shown is that each boom in voice interaction has proceeded from a new platform development, beginning with the Famicom microphone and continuing most recently with smartphone gaming. Voice interaction game design has been both driven and constrained by the affordances of the technology platforms on which digital games have been built at each point in time. While examples of voice interaction exist in almost every era of digital games, the great majority have closely followed the release of a platform that afforded new voice input or speech processing capabilities. This highlights the importance of studying platforms as context for understanding the games that are played upon them, as exemplified in the interlinked fields of platform studies (Bogost & Montfort, 2009) and media archaeology (Apperley & Parikka, 2015; Parikka & Suominen, 2006). The patterns we have observed in this history illustrate the influence dynamic that Bogost and Montfort (2009) ascribe to platforms, wherein the technology is not deterministic but does "encourage and discourage different sorts of expressive new media work" (p. 5).

Karaoke games may be considered as an exception to the platform-driven rule, as games in this genre have often been sold with the microphones that enable their use rather than taking advantage of preexisting microphones that come with the console on which they are played. This can be explained by the extra functions that microphones play in karaoke games as a social prop and a cultural referent (Fletcher & Light, 2011, p. 11) as well as a controller. Whereas in most digital games, the imagined action is situated in the virtual space and oriented toward virtual characters, in karaoke games, the imagined action is situated in the room and (often) oriented toward a real audience. The karaoke genre has traditionally used large, stage-style microphones rather than discreet headsets or controller microphones; these provide both a clear visual cue for the audience to see the player in the context of a stage performance rather than an ordinary living-room scene and a focal point away from the screen around which multiple singers can orient themselves (Fletcher & Light, 2011, p. 10). Thus, the voice input hardware is an even more essential feature of karaoke games than many other voice interaction games, for which the physical format of the microphone is less relevant.

Voice Interaction Game Cultures

To say that technology and platforms have driven the development of voice interaction games is not to say that they have fully determined how these game experiences have been designed. As this history has described, voice interaction games produced in different regions show signs of being the product of different ways of thinking about how players might communicate with virtual characters. This shows that cultural, individual, and economic factors still play a large role in what kinds of experience we see as possible and desirable for voice interaction games.

The clearest stylistic differences are those that emerged between the voice interaction games made in Japan and those made in North America and Europe. This is particularly apparent in the way that each of these regions typically approached the player's relationship with the virtual characters in the games. Japanese developers have tended to create conversational interactions in which the player engages in two-way communication with a singular character who is presented as an autonomous being. This style is exemplified by *Hey You, Pikachu*! and *Seaman* and might be called "digital companion" or "virtual pet" games. While the virtual character is clearly responsive to the player's inputs rather than fully independent, they are presented by the game as being in a reciprocal relationship with the player, in which relationship building is at least partly the goal. In contrast, Western developers have typically created militaristic strategy games in which the player uses their voice to issue commands to multiple units, in an authoritative, top-down communication style that suits the hierarchical structure of the player–character relationship; the inner thoughts of the virtual characters are of minimal relevance. *Tom Clancy's EndWar* and *Command: Aces of the Deep* are archetypal examples. While not every voice interaction game from a Japanese, North American, or European developer fits these profiles, there is a strong tendency toward these styles of voice interaction among most of the games from each region.

Interestingly, the stylistic differences we have noted alternately support and contradict the argument made by Nass and Brave (2005) that all speech interaction automatically evokes an imagined social relationship with the technology. Nass and Brave report on a number of experiments that suggest people's responses to computers accord with their responses to other humans in equivalent interactions, and that this effect is particularly pronounced for voice interactions. This effect might be expected to be even more pronounced in digital games, as virtual characters are such a central feature of the medium. Indeed, the Japanese style of two-way, relatively equal communication humanizes (or anthropomorphizes) the virtual characters and suggests a relationship that very much treats the game character as a true conversation partner with a mind of its own. Conversely, the Western style of top-down control implies a much less humanized view of the virtual characters and a greater sense that the player is simply moving tokens on a game board.

The peculiarities of Western and Japanese game design have long been identified by contrast with one another, although it would be reductive to suggest that there is one monolithic culture or approach to game design on either side (Consalvo, 2016, pp. 4–5). The differences we have observed may be due to broad cultural factors or they may be the result of individual preferences; the number of people making voice interaction games has always been small enough that a single developer could have a substantial influence on the style of games produced in their region. What these differences clearly demonstrate is that the content and style of voice interaction games have not simply been determined by the affordances of the platforms on which they are built, although these platforms have certainly shaped what is possible.

The tidal pattern of voice interaction gaming that this history has documented may now be coming to an end, as microphones and voice interfaces become an increasingly standard component of gaming devices. There appears to be little prospect of a return to a situation like the winter of voice interaction games in the

1990s, when common gaming platforms simply did not feature voice interfaces due to hardware constraints. At the same time, the barriers to developing and publishing a voice interaction game have been lowered by open-source speech processing systems and digital distribution channels, which has enabled the proliferation of smaller scale games by independent developers. Without the platform-driven cycle of novelty to rely on, designers will need to understand how past attempts at voice interaction in games have succeeded or failed in engaging users. The history outlined in this article has shown that video games have a long record of experimentation and creativity with the format, much of it let down by the limitations of the technology of the day. Recognizing what kinds of interactions are desirable and how different configurations of the player–character relationship create meaning for players will benefit both game designers and developers of voice interfaces for other contexts.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Microsoft Research Centre for Social Natural User Interfaces and the Australian Government Research Training Program.

Notes

1. Note that we use the term voice interaction exclusively for the use of voice to interact with technology, as distinct from voice communication or voice chat, which refer to online conversation between humans.
2. Apple Town Monogatari, a life simulation game for the Famicom, is arguably an earlier case, although its voice interaction is limited to two hidden "Easter eggs": By shouting at the right moment, the player can make their virtual friend fall off a ladder or cause a chirping bird to appear.
3. A similar trick had been included 10 years earlier in Manhunt, another stealth action game.

References

150 Best Games of All Time. (1996, November). *Computer Gaming World*, *148*, 63–80.

Altice, N. (2015). *I am error: The Nintendo family computer/entertainment system platform*. Cambridge, MA: MIT Press.

Apperley, T., & Parikka, J. (2015). Platform studies' epistemic threshold. *Games and Culture*. doi:10.1177/1555412015616509

Aylett, M. P., Kristensson, P. O., Whittaker, S., & Vazquez-Alvarez, Y. (2014). None of a CHInd: Relationship counselling for HCI and speech technology. In *CHI '14 extended abstracts on human factors in computing systems* (pp. 749–760). New York, NY: ACM. doi:10.1145/2559206.2578868

Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., ... O'Shaughnessy, D. (2009). Developments and directions in speech recognition and understanding, Part 1 [DSP Education]. *IEEE Signal Processing Magazine*, *26*, 75–80. doi:10.1109/MSP.2009.932166

Berger, C. (1916, December 19). *Sound-operated circuit-controller*. Retrieved August 15, 2017, from https://worldwide.espacenet.com/publicationDetails/biblio?II=0&ND=3&adjacent=true&locale=en_EP&FT=D&date=19161219&CC=US&NR=1209636A&KC=A

Berger, C. (1917, October 16). *Automatically-operating toy or the like*. Retrieved August 15, 2017, from https://worldwide.espacenet.com/publicationDetails/biblio?II=0&ND=3&adjacent=true&locale=en_EP&FT=D&date=19171016&CC=US&NR=1243380A&KC=A

Berger, C. (1918, September 24). *Sound-operated toy or instrument*. Retrieved August 15, 2017, from http://worldwide.espacenet.com/publicationDetails/biblio;jsessionid=cmLm1-lKKSLAdeqyB-BeTcCz.espacenet_levelx_prod_2FT=D&date=19180924&DB=&locale=&CC=US&NR=1279831A&KC=A&ND=1

Bogost, I., & Montfort, N. (2009). Platform studies: Frequently questioned answers. In *Digital Arts and Culture Conference*. Retrieved November 22, 2017, from http://escholarship.org/uc/item/01r0k9br

Carter, M., Allison, F., Downs, J., & Gibbs, M. (2015). Player identity dissonance and voice interaction in Games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (pp. 265–269). New York, NY: ACM. doi:10.1145/2793107.2793144

Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). *Voice user interface design*. Boston, MA: Addison-Wesley Professional.

Consalvo, M. (2016). *Atari to Zelda: Japan's videogames in global contexts*. Cambridge, MA: MIT Press.

Dark Watcher & 98PaceCar. (n.d.). *RDI Halcyon*. Retrieved November 29, 2016, from http://www.videogameconsolelibrary.com/pg80-rdi.htm

David, E. (2016, February 12). *Was the world not ready for Rock Band 4 and Guitar Hero Live?* Retrieved November 29, 2016, from http://siliconangle.com/blog/2016/02/12/was-the-world-not-ready-for-rock-band-4-and-guitar-hero-live/

Dow, S., Mehta, M., Harmon, E., MacIntyre, B., & Mateas, M. (2007). Presence and engagement in an interactive drama. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1475–1484). New York, NY: ACM. doi:10.1145/1240624.1240847

Fairbanks, A. (Ed.). (1931). *Philostratus, imagines. Callistratus, descriptions*. London, England: William Heinemann.

Fletcher, G., & Light, B. (2011). *Interpreting digital gaming practices: SingStar as a technology of work*. Presented at the European Conference on Information Systems, Helsinki, Finland. Retrieved August 15, 2017, from http://usir.salford.ac.uk/17245/

Furui, S. (2005). 50 Years of progress in speech and speaker recognition research. *ECTI Transactions on Computer and Information Technology*, *1*, 64–74.

Gernsback, H. (1916, June). An electric pup that heeds your call. *The Electrical Experimenter*, *4*, 104.

Harada, S., Wobbrock, J. O., & Landay, J. A. (2011). Voice games: Investigation into the use of non-speech voice input for making computer games more accessible. In P. Campos, N. Nunes, N. Graham, J. Jorge, & P. Palanque (Eds.), *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction—Volume Part I* (pp. 11–29). Berlin, Heidelberg: Springer-Verlag.

Herman, L. (1997). *Phoenix: The fall & rise of videogames* (2nd ed.). Union, NJ: Rolenta Press.

Huang, X., Baker, J., & Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM*, *57*, 94–103. doi:10.1145/2500887

Igarashi, T., & Hughes, J. F. (2001). Voice as sound: Using non-verbal voice input for interactive control. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology* (pp. 155–156). New York, NY: ACM. doi:10.1145/502348.502372

Karaoke Studio. (n.d.). Retrieved June 7, 2017, from http://famicomworld.com/system/other/karaoke-studio/

Kinder, J., & Hallock, D. (n.d.). *Halcyon interactive laserdisc system*. Retrieved December 13, 2016, from http://www.dragons-lair-project.com/community/related/homesystems/halcyon/

Lasers and Computers. (1985, January 14). *Computer Chronicles*. Boston, MA: PBS.

Lea, W. A., & Shoup, J. E. (1979). *Review of the ARPA SUR project and survey of current technology in speech understanding* (No. N00014-77-NaN-570) (p. 162). Los Angeles, CA: Office of Naval Research, Speech Communications Research Lab.

Lee, K.-F. (1988). Word recognition in large vocabularies: On large-vocabulary speaker-independent continuous speech recognition. *Speech Communication*, *7*, 375–379. doi:10.1016/0167-6393(88)90053-2

Luger, E., & Sellen, A. (2016). 'Like having a really bad PA': The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5286–5297). New York, NY: ACM. doi:10.1145/2858036.2858288

Lynch, K. (2013, November 18). *PlayStation 4: Can Sony's new console live up to its predecessor's record-breaking legacy?* Retrieved November 29, 2016, from http://www.guinnessworldrecords.com/news/2013/11/playstation-4-can-sony%E2%80%99s-new-console-live-up-to-its-predecessor%E2%80%99s-record-breaking-legacy-47090

Mace, S. (1983, November 21). TI retires from home-computer market. *InfoWorld*, *5*, 22–27.

Munteanu, C., Jones, M., Oviatt, S., Brewster, S., Penn, G., Whittaker, S., … Nanavati, A. (2013). We need to talk: HCI and the delicate topic of spoken language interaction. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (pp. 2459–2464). New York, NY: ACM. doi:10.1145/2468356.2468803

Mustaquim, M. M. (2013). Automatic speech recognition: An approach for designing inclusive games. *Multimedia Tools and Applications*, *66*, 131–146. doi:10.1007/s11042-011-0918-7

Nanjo, H., Mikami, H., Kunimatsu, S., Kawano, H., & Nishiura, T. (2009). A fundamental study of novel speech interface for computer games. In *ISCE: 2009 IEEE 13th International Symposium on Consumer Electronics, Vols 1 and 2* (pp. 291–293). New York, NY: IEEE.

Nass, C., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. Cambridge, MA: MIT Press.

Parikka, J., & Suominen, J. (2006). Victorian snakes? Towards a cultural history of mobile games and the experience of movement. *Game Studies*, *6*. Retrieved August 15, 2017, from http://gamestudies.org/0601/articles/parikka_suominen

Peckham, M. (2011, March 9). *Kinect breaks Guinness record, sells 10 million systems, tops iPhone and iPad*. Retrieved November 29, 2016, from http://www.pcworld.com/article/221738/kinect_breaks_guinness_record_sells_10_millions_systems.html

Provo, F. (2000, November 3). *Hey You, Pikachu*! *Review*. Retrieved May 18, 2016, from http://www.gamespot.com/reviews/hey-you-pikachu-review/1900-2650135/

Rabiner, L., & Juang, B.-H. (2008). Historical perspective of the field of ASR/NLU. In J. Benesty, M. M. Sondhi, & Y. Huang (Eds.), *Springer handbook of speech processing* (pp. 521–538). Berlin, Heidelberg: Springer.

Reddy, D. R., Erman, L., & Neely, R. (1973). A model and a system for machine recognition of speech. *IEEE Transactions on Audio and Electroacoustics*, *21*, 229–238. doi:10.1109/TAU.1973.1162456

Reed, K. (2008, November 6). *Tom Clancy's Endwar Review*. Retrieved November 29, 2016, from http://www.eurogamer.net/articles/tom-clancys-endwar-review

Schoen, J. (1985, March 5). When you talk, Your PC listens. *PC Mag*, *4*, 122–132.

Shimomura, D. (2015, January 21). *Mayday*! *Deep Space relies on the panic in your mind*. Retrieved June 8, 2017, from https://killscreen.com/articles/mayday-deep-space-relies-panic-your-mind/

Sporka, A. J., Kurniawan, S. H., Mahmud, M., & Slavík, P. (2006). Non-speech input and speech recognition for real-time control of computer games. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 213–220). New York, NY: ACM. doi:10.1145/1168987.1169023

Stevens, N. (1999, October). *Command aces of the deep*. Retrieved November 29, 2016, from http://www.subsim.com/ssr/commandaces.html

Suominen, J. (2016). How to present the history of digital games: Enthusiast, emancipatory, genealogical, and pathological approaches. *Games and Culture*. doi:10.1177/1555412016653341

"ThuuMic—Use your mic for dragon shouts" by DeadlyAzuril and PsychoHampster, retrieved June 8, 2017, from http://www.nexusmods.com/skyrim/mods/5626/

Toys That Obey Your Voice (1916, December). *Popular Science Monthly*, *89*, 718–719. New York, NY: Modern Publishing Company.

Turner, A. (2014, May 14). *Microsoft backflips on Kinect for Xbox One*. Retrieved May 20, 2016, from http://www.smh.com.au/digital-life/computers/gadgets-on-the-go/microsoft-backflips-on-kinect-for-xbox-one-20140514-zrcc7.html

Wadley, G., Carter, M., & Gibbs, M. (2015). Voice in virtual worlds: The design, use, and influence of voice chat in online play. *Human–Computer Interaction*, *30*, 336–365. doi:10.1080/07370024.2014.987346

Warren, T. (2014, April 2). The story of Cortana, Microsoft's Siri killer. *The Verge*. Retrieved August 15, 2017, from http://www.theverge.com/2014/4/2/5570866/cortana-windows-phone-8-1-digital-assistant

Wolf, M. J. P. (2008). *The video game explosion: A history from PONG to playstation and beyond*. Santa Barbara, CA: ABC-CLIO.

Wolf, M. J. P. (2012). *Encyclopedia of video games: The culture, technology and art of gaming* (Vol. 2). Santa Barbara, CA: Greenwood.

Yin-Poole, W. (2016, February 12). *Guitar Hero Live failed to do the business, too*. Retrieved November 29, 2016, from http://www.eurogamer.net/articles/2016-02-12-guitar-hero-live-failed-to-do-the-business-too

Author Biographies

Fraser Allison is a PhD candidate at the Microsoft Research Centre for Social Natural User Interfaces at the University of Melbourne. His doctoral research examines the use of conversational interfaces to interact with virtual characters.

Marcus Carter is a lecturer in digital cultures at the University of Sydney. His website is http://marcuscarter.com/

Martin Gibbs is an associate professor in the School of Computing and Information Systems at the University of Melbourne. His current teaching and research interests lie at the intersection of science, technology studies, and human–computer interaction and are focused on the sociable use of interactive technologies with a particular interest in the study of digital and analogue games.

6 Voice Game Design Patterns

This chapter presents Study 3 in the form of a peer-reviewed conference paper:

> Allison, F., Carter, M., Gibbs, M., & Smith, W. (2018). Design Patterns for Voice Interaction in Games. *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, 5–17. https://doi.org/10.1145/3242671.3242712

The paper is included verbatim in its published format, following a discussion of how it contributes to the thesis. I led all stages of the research and writing process. My thesis supervisors, Marcus Carter, Martin Gibbs and Wally Smith, contributed to the study design and writing of the paper.

6.1 Introduction

The previous chapter identified a trend in which different types of voice games have risen to prominence in different eras, such as the conversational digital companion era of 1998-2007 and the global karaoke boom of 2003-2012. In spite of these fluctuations, however, the historical analysis showed the basic forms of voice gameplay to be surprisingly consistent over time: the earliest and crudest voice games present the player with much the same types of activities as the newest and most sophisticated—at least in terms of the imagined action—despite large differences in the games' speech processing capabilities. This suggests that there are resilient underlying patterns in game design for voice interaction, which have remained relevant over several decades of technological change and are likely to continue to remain relevant into the future.

This chapter presents the results of a study to identify and describe these patterns of voice game design. The study primarily addresses Research Question 3:

> **RQ3** What design patterns are currently implemented in voice games?

Study 3 addresses this research question through an extensive analysis of published voice games. The variations in voice game design are codified according to the pattern language framework first developed by Alexander et al. (1977) and adapted to game design by Björk and Holopainen (2004). The study identifies 25 recurring design patterns in voice games, and notes that these patterns have tended to co-occur within particular games in ways that represent five distinct voice gameplay genres. Based on these findings, this chapter presents a pattern language that can be used to recognise, conceptualise and communicate common elements in voice game design.

6.2 Research method

This study set out to to identify and catalogue every published videogame that has afforded voice interaction, whether it was released as a commercial product, a free piece of software, an educational tool or an academic prototype. To achieve this, the database of voice games used in Study 2 (Chapter 5) was expanded through an extensive process of desktop research. As in Study 2, any screen-based videogame that has allowed players to use their voice to affect the game state was included in the database, with the exception of most audiogames. However, a sample of 22 smart-speaker audiogames was tested to determine whether the results for that type of game would be similar. The final database included 471 games, with a summary description of the voice affordances and interface characteristics of each one. An inductive analysis process was used to identify patterns in this data, and these patterns were translated into a game design pattern language following the format codified by Björk and Holopainen (2004). The research method is described at greater length in Chapter 3 and in the published article below. The list of games analysed in this study is included in Appendix B.

6.3 Summary of findings

The study produced a vocabulary of 25 design patterns for voice game design. These range from specific actions that can be performed with voice (such as SHOUT TO SHOOT) to higher-level patterns that describe how voice interaction is structured (such as CHOICE OF OPTIONS) or how the game situates the player within the imaginary gameworld (such as SPEAK AS A CHARACTER). The published paper describe each pattern with a one- or two-sentence definition, an illustrative example from an existing game, a one-paragraph discussion of its gameplay consequences, and a list of other design patterns that it overlaps or conflicts with.

The design patterns are classified into six categories. The first category, *Diegetic Framing*, contains the patterns that describe how the game fiction frames the player's voice input. The second category, *Dialogue Structure*, contains the patterns that describe how the dialogue between players and NPCs is configured. The four remaining categories describe specific voice-based actions and game mechanics. The headings for these four categories (*Selection*, *Navigation*, *Control* and *Performance*) indicate the type of action that the pattern involves, as represented in the fictional space of the gameworld.

An additional analysis was conducted to determine which design patterns have tended to appear together within the same games. Two such clusters of patterns were identified, one based on "command" and another based on "branching dialogue". The paper also notes that a different set of patterns were found to appear almost always in isolation, and that these isolated patterns are notable for representing forms of direct action rather than mediated action. The next section will argue that these clusters of patterns constitute five distinct genres of voice gameplay, each of which reflects a different type of voice activity.

Two final analyses look at the design patterns that have been used in two specific sets of voice games: academic research prototypes and smart speaker games. The academic voice games have been highly focused on two particular design patterns (PRONUNCIATION and SPEAK AS THE PLAYER) that rarely appear in non-academic games, indicating that the research literature is not representative of the wider world of voice game design. Similarly, the sample of smart speaker used a narrow subset of design patterns, oriented around the CHOICE OF OPTIONS and QUESTION AND ANSWER structures, reflecting the constraints of their audio-only format.

6.4 Voice gameplay genres

The published paper contains an analysis of how voice game design patterns have often co-occurred within games. It notes that certain patterns have tended to appear in combination, while other patterns have nearly always appeared in isolation. This section summarises that analysis and extend it one step further to argue that the co-occurrence of design patterns reflects the existence of five implicit genres of voice gameplay, most of which have not been formally identified in prior literature. It begins by reviewing the distinction between communication-based voice gameplay and direct-action-based voice gameplay, and then break these categories down further into the five genres.

As the paper describes, certain sets of design patterns have usually appeared together in the same games, while other patterns have usually not appeared in combination with any others. The first group reflects communication-based gameplay, which is centrally concerned with conveying information between the player and a character. Voice commands and conversational dialogue are both forms of communication-based gameplay. Communication-based gameplay typically requires a combination of voice affordances to reflect the inherent complexity of speech, hence the need for a whole cluster of patterns in these games. The second group reflects direct-action-based gameplay, in which the player uses their voice to carry out an action for themselves. Example actions include blowing out a virtual candle, matching a certain frequency pattern or correctly pronouncing a specific word. Because these actions do not have the modular complexity of speech, they are inherently simpler, and each one can therefore be encapsulated in a single design pattern.

Within the overarching categories of communication-based and direct-action-based voice gameplay, I have identified five core genres of voice games. The classification of each genre is based upon the fundamental type of voice activity involved, defined by voice game design patterns as shown in Table 6.1. This follows Björk et al.'s (2003, p. 190) argument that game design patterns may be used to define the genre of the games that they appear in. Conversely, the genre classifications do not reflect any consideration of the fictional setting or aesthetic style of the videogames that they describe, following Apperley's proposal that scholarly definitions of game genres should be based on the "nonrepresentational, specifically interactive, characteristics of video games" (2006, p. 7). The purpose of this genre typology is to identify how voice game design varies beyond the set of activities that are most commonly associated with it or studied as constituents of it.

The five genres are Voice Practice, Voice Performance, Voice Command, Voice Conversation and Voice Incantation (see Table 6.1). Voice Practice has appeared exclusively in serious games (Abt, 1970). This genre of videogames is designed to guide the player through a set of vocal exercises, such as foreign language pronunciation exercises or speech rehabilitation exercises, for the purpose of self-improvement rather than play for its own sake. All of the games in the study that used the PRONUNCIATION pattern are examples of this genre, as are most of the games that used the PITCH CONTROL pattern. Voice Performance games are designed to create social fun by prompting players to make a spectacle of themselves through some kind of performance, whether it is a display of skill or self-expression (as in the KARAOKE pattern) or a display of comical absurdity (as is often the case with the VOLUME CONTROL and SHOUT TO SHOOT patterns). Voice Command and Voice Conversation are perhaps the most widely recognised types of voice gameplay, as they reflect the same forms of voice interaction that are used in common VUI such as Siri and Alexa. Both genres employ a set of patterns in combination to sustain the illusion of the player engaging in dialogue with virtual characters, as the published paper describes. Finally, the Voice Incantation genre allows the player's voice to affect the game environment directly, rather than through communication. The archetypal design pattern for this genre is SPELLCASTING, in which the player utters words of power to create a magical effect within the game. Also included is BREATH PHYSICS, in which the player can blow air into the microphone to create a physical wind effect in the game. All five of these genres are discussed at greater length in Chapter 8.

Table 6.1 Genres of voice gameplay and associated design patterns

Voice gameplay genre	Associated design patterns
Voice Practice	SPEAK AS THE PLAYER PRONUNCIATION PITCH CONTROL
Voice Performance	SPEAK AS THE PLAYER or SPEAK AS A CHARACTER KARAOKE VOLUME CONTROL SHOUT TO SHOOT
Voice Command	SPEAK AS A CHARACTER SITUATED COMMANDER FLOATING COMMANDER WHO-WHAT-WHERE COMMANDS SELECT BY NAME SELECT BY TYPE NAME AN ACTION NAME A BEHAVIOUR WAYPOINTS RELATIVE DIRECTIONS ABSOLUTE DIRECTIONS
Voice Conversation	SPEAK AS A CHARACTER CHOICE OF OPTIONS QUESTION AND ANSWER SCRIPTED CONVERSATION UNSCRIPTED CONVERSATION
Voice Incantation	SPEAK AS A CHARACTER SPELLCASTING BREATH PHYSICS OVERHEARD NOISE

6.5 Contributions to the thesis

The results of this study answer Research Question 3: "What design patterns are currently implemented in voice games?" The pattern language encodes a catalogue of the major variations in voice game design from 1973 to 2018, across both serious games and games for leisure. It also serves as a vocabulary for designers and researchers to apply to voice games, a common language that can be used to describe the format based on clear points of reference.

The study provides a clear account of how variable the design of voice game mechanics has been. Whereas non-game voice interfaces primarily use speech-based input (voice commands and conversational dialogue), this study reveals a range of non-verbal voice mechanics that have been used to great effect in games, such as BREATH PHYSICS, PITCH CONTROL and OVERHEARD NOISE. This prompted the addition to the thesis of Research Question 5—"How do players experience verbal and non-verbal voice gameplay?"—and informed the decision to include both verbal and non-verbal forms of voice gameplay in Study 4 (Chapter 7).

Finally, this chapter identifies the five core genres of voice gameplay, defined in the previous section. Chapter 8 provides an expanded discussion of these five genres, with an explanation of what each one involves and how it can best be understood based on insights from all four of the studies in the thesis.

Session: Paper Presentation CHI PLAY 2018, October 28–31, 2018, Melbourne, VIC, Australia

Design Patterns for Voice Interaction in Games

Fraser Allison[1], **Marcus Carter**[2], **Martin Gibbs**[1], **Wally Smith**[1]
[1]Interaction Design Lab, School of Computing and Information Systems, The University of Melbourne, Parkville, VIC, Australia
[2]Department of Media and Communications, The University of Sydney, Camperdown, NSW, Australia
[1]{fraser.allison,martin.gibbs,wsmith}@unimelb.edu.au, [2]marcus.carter@sydney.edu.au

ABSTRACT
Voice interaction is increasingly common in digital games, but it remains a notoriously difficult modality to design a satisfying experience for. This is partly due to limitations of speech recognition technology, and partly due to the inherent awkwardness we feel when performing some voice actions. We present a pattern language for voice interaction elements in games, to help game makers explore and describe common approaches to this design challenge. We define 25 design patterns, based on a survey of 449 videogames and 22 audiogames that use the player's voice as an input to affect the game state. The patterns express how games frame and structure voice input, and how voice input is used for selection, navigation, control and performance actions. Finally, we argue that academic research has been overly concentrated on a single one of these design patterns, due to an instrumental research focus and a lack of interest in the fictive dimension of videogames.

CCS CONCEPTS
• Applied computing~Computer games
• Computing methodologies~Speech recognition

Author Keywords
Design patterns; pattern language; interaction design; game design; voice interaction; voice control; speech recognition.

INTRODUCTION
Many visions of future technology, and future videogames, have us talking to virtual characters. In the film *Her* (2013), the main character falls in love with a Siri-like talking virtual assistant and spends his downtime playing a videogame with an alien character who communicates through foul-mouthed natural language conversation. In *2001: A Space Odyssey* (1968), an astronaut plays voice-command chess against his spaceship's talking computer. These scenarios are characterised by the fluency and ease of communication between humans and computer agents, which supports engaging and enjoyable gameplay. But this remains a hallmark of fiction, as the current reality of voice interaction gameplay is not nearly as gratifying. Feedback on voice interaction games often criticises the experience as feeling "unnatural" or "forced" [13:268], highlighting that game designers have found it challenging to design a satisfying player experience of voice interaction.

CHI PLAY '18, October 28–31, 2018, Melbourne, VIC, Australia

ACM 978-1-4503-5624-4/18/10…$15.00
https://doi.org/10.1145/3242671.3242712

To help designers work in this space, and to crystallise some of the common approaches that have been taken in the past, we propose a pattern language for voice interaction game design. A pattern language is a collection of design patterns, each one of which describes the core of a common solution to a recurring design problem [1]. It is a practical vocabulary for classifying a design space, which formalises a body of knowledge that is otherwise implicit in the commonalities between finished works. Originating in the field of architecture, design patterns gained widespread popularity in computer science as a way of sharing reliable, repeatable approaches to common design goals [31]. Björk et al. [9] introduced pattern languages to game studies as a tool to help understand the form of games and to identify design choices that can affect the player experience in predictable ways.

In this paper, we present a pattern language for the use of voice interaction in games, based on a comprehensive survey of academic and commercial games that have used any form of voice interaction from 1973 to 2018. It catalogues the major game mechanics, dialogue structures and diegetic framing devices that have been employed, to give designers and scholars a library of options for addressing the challenges of designing in this notoriously difficult modality. We identify clusters of patterns that are often used together, and patterns that have so far resisted being used in combination with any other. We find that human-computer interaction (HCI) research to date has focused on only a narrow selection of design patterns, concentrated around pronunciation exercises, and paid little attention to the fictive and experiential elements that provide players with a sense of presence in a gameworld.

RELATED WORK

Voice Interaction in Digital Games
Research on voice control of digital games has been undertaken since at least the 1970s, more than a decade before voice control first appeared in commercial videogames [2]. Academic interest in the topic has predominantly come in the form of instrumental research

[14], using game design to create an engaging format for language learning practice [39,49] and speech rehabilitation exercises [54,69]. Alongside this, a substantial body of research has explored the implementation of voice input as an alternative game control scheme, to enable access for players with motor impairments or other disabilities that prevent them from using physical controls [33,68]. Several research prototypes have explored options for controlling games through non-speech qualities of voice, such as the volume [76] and pitch [36,59,68] of vocal input; these approaches lend themselves to relatively simple arcade-style game mechanics such as one-dimensional movement.

Three-dimensional gameworlds introduce a particular challenge for voice interaction. Wadley and Ducheneaut [79] found that players struggle to communicate spatial deixis (referential concepts such as “there” and “on your left”) when speaking to other human players in a virtual world. This difficulty is exacerbated with AI characters, which require a situation model of the game space before they can interpret commonplace spatially deictic phrases such as “come over here and pick this up” [32:140]. Several of the patterns that we identify (notably *Waypoints*, *Absolute Directions*, *Relative Directions*, *Select by Name* and *Select by Type*) are aimed at resolving such referential ambiguity without requiring complex situation models.

Studies on the player experience of voice interaction games have tended to emphasise its social character. Fletcher and Light describe the karaoke game as a “glue technology that assists in crafting and strengthening social linkages amongst players” [29:1]. Conversely, Dow et al. highlight the “uncomfortable” and “awkward” [26:9] feeling that some players experienced when talking to a computer. A study by Carter et al. found that players’ feelings of embarrassment and awkwardness were most pronounced in games that required them to say things that were inconsistent with their in-game persona as the player-character – an effect they termed “identity dissonance” [13]. All together, these studies suggest that social framing is an important element in voice interactions in games. Consequently, we have captured the social framing of voice actions in our analysis of game design patterns.

In a previous study [2], we traced the history of voice interaction games from the 1970s to 2016, and noted characteristic differences between Japanese and Western voice interaction games. We identified the beginnings of a boom in small-scale games that focus on voice interaction as their central modality, and predicted that this boom would continue thanks to the increasing availability of microphone-equipped game platforms and effective speech recognition systems. The current study shows that this has eventuated, with 106 new voice interaction videogames released in 2017 alone, more than twice as many as any previous year. That excludes audio-only games, such as those on Amazon Alexa, of which there have been hundreds released since 2015. However, most of these recent games employ only a small set of simple and repetitive game mechanics. This context makes it even more timely to consider a pattern language that can illuminate the wider design space for voice interaction gameplay.

Design Patterns

Design patterns are a way of identifying and describing common solutions to design problems. The term was first used by Christopher Alexander in relation to architecture and urban planning [1], and has since been applied to interaction design fields such as human-robot interaction [41] and game design [8]. A design pattern consists of a formalised explanation of a repeatable approach to a specific design scenario. The elements of a pattern vary from Alexander’s original formulation, but generally include an idiomatic title, an explanation of the problem or goal addressed, a description of the solution, one or more illustrative examples, and a note on connections with other patterns. Importantly, design patterns are not meant to be detailed blueprints, and do not include specific rules for how they should be implemented. Rather, Alexander conceptualises patterns as abstractions of a design solution:

> Each pattern describes a problem which occurs over and over again in our environment, and then describes the core of the solution to that problem, in such a way that you can use the solution a million times over, without ever doing it the same way twice. [1:x]

Design patterns were introduced to the games research literature by Björk et al. [9]. They avoided framing patterns as solutions to problems, as this “creates a risk of viewing patterns as a method for only removing unwanted effects of a design” [9:185], rather than as a tool for more creative design work. Instead, they identify their value as tools for inspiring game ideas, understanding how existing games are put together, and helping to make choices about how new game will work and what elements they should include [8]. Of course, patterns like *Power Ups* and *High Score Lists* have been known in game design for decades, but collecting and formalising these patterns creates a sort of reference library in which the concepts can be detailed, annotated and cross-referenced for the benefit of designers who may be less familiar with their use. It also establishes a common language for discussing voice interaction, with clear points of reference to ensure mutual understanding [28].

Björk et al. [9] define a five-part format for game design patterns, comprising a short, specific and idiomatic *name*; a concise *description*, with examples; a discussion of *consequences* of applying the pattern; common choices involved in *using the pattern*; and a list of *relations* between patterns. Variations on this format have been used to collect patterns in game dialogue [11], character design [44,45,50,62], level design [35,52,67] and psychological factors in game design [24,46,51,80]. Pattern languages have often been used to map relatively under-documented design spaces, including mobile games [22], pervasive games [10], AI-based games [17,71], public games [7], disaster response games [70] and games controlled by the player's gaze [75].

Most pertinent for the current study is Alves and Roque's pattern language for sound design in games [3,4], which covers a wide variety of game sound patterns such as overheard conversations, contextual music and sound effects that signal player failure or success. While their collection is primarily concerned with sound output, it includes one pattern (*Sound Input*) that covers player-produced sound, including both sounds introduced to the game through a microphone and sounds made by the game in response to player button presses. This can be seen as an umbrella category for our study, which only concerns design choices related to sounds made by the player, and considers sound output only insofar as it relates to the uses of sound input.

Despite the enthusiasm for pattern languages, some cautions have been sounded about how they are applied. Erickson warns against the "common misconception" [28:361] that a pattern language is a prescriptive tool that describes universal templates for designers to apply directly from the page. The prescriptive approach is articulated by Dormans, who argues that "a design pattern library should depart from a clearly formulated notion of quality in games" [25:4] and present its patterns as prescribed solutions to common problems, on the pragmatic basis that this would be more attractive to game developers than a pattern language that does not pre-judge the quality of a design choice. Erickson counters this view with the argument that the purpose of a pattern language is generative: it provides a common meta-language to bootstrap the creation of a project-specific vocabulary in which all participants in a design process can effectively participate. Rather than being rigid prescriptions, "both the language and individual patterns are malleable and are used to generate site-specific pattern languages for particular projects" [28:361]. This is echoed in an interview study with HCI researchers [58], which found that overly formalised or prescriptive pattern languages tend to be less usable as they fail to account for the dynamic and contextual nature of design and technology. The study identified the extensive effort involved in collecting patterns and developing them into a coherent language as a barrier to their wider use, but once formulated, a pattern language was considered to be a particularly useful tool for capturing interactions, documenting conventions, sharing knowledge and encouraging design thinking.

A case study by Segerståhl and Jokela [65] noted shortcomings in the comprehensiveness, naming and organisation of some existing pattern languages which reduced their usability to designers. To improve the organisation of pattern languages, they advise that patterns should be grouped into task-related areas, such as **Searching** and **Navigation**, and that pattern names should be illustrative of the solution they represent (as in *Bread Crumbs*) rather than generic or ambiguous (as in *Main Navigation*). They emphasise the importance of examples to provide an illustration of the pattern in practice. Finally, they recommend integrating pattern collections into larger libraries with a consistent format. We have followed these suggestions as much as possible in the formation of the pattern language in this study.

NOTE ON TERMINOLOGY

Discussions about voice interaction and game characters share a common problem: their key terms are often used interchangeably, as they appear superficially similar, despite referring to distinct concepts [5,15]. To avoid ambiguity, in this section we define how we use these terms in this paper.

Voice interaction encompasses any way that voice can be used as an input modality for a computer system to process and render a response. Notably, it excludes player-to-player communication, as in online voice chat. *Voice control* refers to the intentional use of voice to directly control a system, which may be verbal or non-verbal, whereas *voice command* is specifically the use of spoken words as direct instructions to the system. A drone that moves in response to whistling would therefore be an example of voice control, but not voice command. Voice command relies on *speech recognition*, the recognition of spoken words by a computer system.

Virtual character encompasses any kind of character with an identity in a gameworld, whether controlled by a player or by AI, and whether or not it appears on screen. *Player character* is the specific virtual character that represents the player's identity in the gameworld, often with its own separately defined identity in the game fiction. It may appear in the form of an *avatar*, a virtual body over which the player exerts consistent control in the gameworld and serves as their locus of manipulation, "a tool that extends the player's ability to realise affordances within the game-world" [5:2]. All game characters that are not representative of or controlled by the player are referred to as *non-player characters* (NPCs); these are usually proximally controlled by an AI system rather than the player's inputs. Finally, *unit* is a generic term for AI-driven entities that includes those that are not strictly "characters", such as a group of NPCs that functions as one unified company, or a tank that is interchangeable with all other tanks. The term is associated with strategy games, in which characters often operate more like tokens on a game board than fully-realised characters.

Diegetic elements are part of the fiction of the gameworld; they are "real" from the perspective of the characters in the game. *Non-diegetic* elements are external to the gameworld, such as game menus or loading screens.

METHOD FOR COLLECTING DESIGN PATTERNS

To begin our design pattern collection, we first set out to identify every screen-based digital game that has used voice interaction up to early 2018. We compiled this list by conducting keyword searches on terms related to voice interaction and speech recognition in the online game libraries Steam, Giant Bomb, itch.io, Newgrounds, Kongregate and the Video Game Console Library, the review aggregator site Metacritic, the discussion website Reddit, and general web searches through Google Search. Next, we applied our search terms to Google Scholar and the ACM Digital Library to identify academic works that have described a unique voice interaction game or game control system. We looked at the prior works cited in these papers and, through their Google Scholar listings, works that had cited them since their publication to find more examples. We repeated this process with new keywords and venues, informed by our initial results, until our list converged on a set of 40 academic research games and 409 commercial and independent games published between 1973 and 2018.[1] English was by far the most common language these games used, but a substantial number were alternatively or exclusively available in other languages, primarily Japanese and a variety of European languages. Where our searches turned up non-English sources that described these games, we reviewed them using Google Translate. However, this study relies in the main on English-language sources.

For each game in our list, the first author recorded the origin, platform and game genre, how central voice interaction was to the gameplay, whether verbal or non-verbal voice input was recognised, what types of speech acts were possible, whether voice actions were diegetically framed as coming from the player or their avatar, and whether voice actions were directed towards the game system or objects within the gameworld. The last two categories were prompted by studies of voice interaction [13] and voice communication [78] that have found that voice can disrupt players' sense of presence and identity when it conflicts with the diegetic framing of the gameworld. The first author then provided a descriptive summary of each game's voice interaction mechanics, and identified potential patterns through an open-coding process. Initial codes described the specific combination of a voice input and a response in the game (such as *Blow air to move sailing ship*). As shared themes emerged across the data set, the codes were iteratively revised and abstracted to describe higher-level patterns (such as *Breath physics*). We applied additional codes to describe how voice actions were framed by the game's presentation (see **Patterns for Diegetic Framing**). We ceased coding once we arrived at a stable list of patterns that encapsulated all the main variations we observed in our list of games.

After this process was complete, we conducted a smaller survey on 22 audio-only games on Amazon Alexa, to test whether different design patterns were apparent. The patterns we found were consistent with those in the larger survey. Our findings for audiogames are discussed at the end of the paper.

[1] A complete list of the games surveyed for this study has been submitted as a separate spreadsheet.

VOICE INTERACTION GAME DESIGN PATTERNS

Our pattern language comprises 25 patterns for voice interaction (see table 1). These patterns are descriptions of emergent features of this design landscape. They are not intended to be exhaustive or mutually exclusive; this is a language, not a logical model for the categorisation of voice interaction games. Like a language, it resists neat categorisations and ontological consistency. However, to aid understanding we have arranged the pattern into six categories, which we will now explain.

1 Diegetic framing	2 Dialogue structure	3 Selection
Speak as a character	Choice of options	Select by name
Speak as the player	Question and answer	Select by type
Situated commander	Who-what-where commands	
Floating commander	Scripted conversation	
	Unscripted conversation	

4 Navigation	5 Control	6 Performance
Waypoints	Name an action	Pronunciation
Absolute directions	Name a behaviour	Karaoke
Relative directions	Shout to shoot	Breath physics
	Volume control	Exclamation
	Pitch control	Spellcasting
		Overheard noise

Table 1. Game design patterns for voice interaction.

The first category, **Diegetic Framing**, consists of the patterns for how the game fiction frames the player's voice input. The patterns in this category are relatively abstract, as they describe the collective implications of multiple game elements, from narrative text to game mechanics. They address the role-playing dimension that has been identified as an important factor in players' experience of voice interaction [2,13].

The second category, **Dialogue Structure**, contains the patterns that describe how dialogue is configured between a player and an NPC. This category looks one level above the individual utterances, to the arrangement of utterances into a particular form of dialogue, such as conversation.

The remaining patterns are categorised into four broad task areas, in accordance with Segerståhl and Jokela's [65] guidelines for usable pattern languages. The categories are **Selection**, **Navigation**, **Control** and **Performance**. Patterns therein describe individual utterances and the responses they engender from the game.

We define each pattern with an illustrative *Name*, a short *Description*, an *Example* that exemplifies the pattern, a brief discussion of *Consequences* for the player experience, and a compact list of *Relations* to other patterns, such as conflicts and parent-child relationships. This is a simplified version of Björk et al.'s [9] model, foregoing their section on *Using the Pattern* for brevity, to allow for a more comprehensive list of the design patterns we have observed.

We distinguish patterns by how they are presented to the player, rather than by their underlying mechanism in the game system. To illustrate this, consider a game that asks the player to extinguish a candle by blowing into the microphone in one scene, and to interrupt a wedding by shouting an objection in another scene. The trigger for both actions may be identical in the game code, with either one being activated by any sufficiently loud sound. But in our pattern language, the former would be described as *Breath Physics* and the latter as *Exclamation* in accordance with their diegetic framing.

1. Patterns for Diegetic Framing

Patterns in this category describe how the game fiction frames and positions the player's voice actions in relation to the game's imaginary. These patterns do not describe specific voice actions, but how the configuration of voice actions casts the player in a speaking role. The patterns are *Speak as a Character*, *Speak as the Player*, *Situated Commander* and *Floating Commander*.

1.1 SPEAK AS A CHARACTER

The player's utterances correspond to the utterances of a character in the gameworld. The game responds to the player's voice as though it is the diegetic voice of the player-character.

Examples: *Guitar Hero: World Tour* [55] represents players as a rock band on stage, performing the song that the players are performing. The virtual crowd cheers or heckles the virtual singer based on how well the player sings.

Consequences: Maintains consistency between the player's in-game persona and the persona implied by their voice actions. Player commentary suggests that this supports an experience of immersion and flow [13].

Relations: Instantiated by *Situated Commander* and *Floating Commander*. Instantiated by patterns that represent embodied diegetic actions, namely *Scripted Conversation*, *Unscripted Conversation*, *Spellcasting*, *Breath Physics*, *Overheard Noise* and often *Exclamation*. Conflicts with *Speak as the Player*.

1.2 SPEAK AS THE PLAYER

The player's utterances do not correspond to the utterances of a character in the gameworld. The game treats the player's voice actions as non-diegetic inputs.

Example: In *Tomb Raider: Definitive Edition* [20], the player can say the name of a weapon to make the player-character hold that weapon. This would not make sense as a vocalisation by the player-character, as she would be giving verbal instructions to herself.

Consequences: Carter et al. [13] noted that players sometimes experienced "identity dissonance" when speaking to (rather than as) the player-character, which suggests that *Speak as the Player* can disrupt the player's sense of presence in the gameworld if not applied carefully.

Relations: Conflicts with *Speak as a Character* and patterns that instantiate it: *Situated Commander*, *Floating Commander*, *Scripted Conversation*, *Unscripted Conversation*, *Spellcasting*, *Breath Physics*, *Overheard Noise* and often *Exclamation*.

1.3 SITUATED COMMANDER

The player gives voice commands to NPCs in the role of an on-screen avatar who they control directly.

Example: This pattern is common in tactical shooter games, such as *Rainbow Six: Vegas* [73], in which the player controls the leader of a military squad and can give the other squad members instructions such as "check your fire" and "regroup on me".

Consequences: Provides the experience of an embodied persona as the imagined source of the player's voice, but also requires the player to divide their attention between managing the avatar's actions and using voice. This can be challenging, especially when the avatar is under threat or time pressure, as speech formulation draws on some of the same cognitive resources that are needed for strategic problem solving, to a greater extent than hand-eye coordination does [66].

Relations: Instantiates *Speak as a Character*. Can be instantiated by *Who-What-Where Commands*. Conflicts with *Speak as the Player* and *Floating Commander*.

1.4 FLOATING COMMANDER

The player gives voice commands to NPCs from a free-floating perspective above the gameworld, in the role of an unseen player-character.

Example: In *Tom Clancy's EndWar* [74], the player directs a battalion of infantry, tank and helicopter squads from above the battlefield using *Who-What-Where Commands*.

Consequences: Allows the player to dedicate their full attention to issuing commands without the distraction of managing an avatar. Combined with the wide field of view granted by the bird's-eye perspective, this makes *Floating Commander* well suited to strategy games and games with complex *Who-What-Where Commands*.

Relations: Instantiates *Speak as a Character*. Can be instantiated by *Who-What-Where Commands*. Conflicts with *Speak as the Player* and *Situated Commander*.

2. Patterns for Dialogue Structure

Patterns in this category describe a systematic arrangement of speech between the player and the game. They are concerned with how utterances are configured into a certain form of dialogue. The patterns are *Choice of Options*, *Question and Answer*, *Who-What-Where Commands*, *Scripted Conversation* and *Unscripted Conversation*.

2.1 CHOICE OF OPTIONS

The game presents a list of options that the player can select from by saying an associated phrase. The phrase may be the full text of the option or a corresponding keyword.

Example: In *Thayer's Quest* [61], each scene ends with a numbered list of locations. The player selects a destination by saying the number next to the place they want to visit.

Consequences: Reduces the possible inputs, which allows the speech recognition task within the game system to be simplified to improve reliability. Also simplifies the task

for the player, minimising uncertainty and the need for creative cognition to formulate an answer.
Relations: Instantiated by *Scripted Conversation* and often *Question and Answer*.

2.2 QUESTION AND ANSWER
A series of questions and responses between the player and a game character (or the game interface). Either party may be the questioner.
Example: In the racing game *Forza Horizon 2* [60], a voice-activated virtual assistant named ANNA provides verbal directions on request. If the player asks "ANNA, what should I do?", ANNA may say "You haven't attempted a Bucket List Challenge for a while, would you like me to direct you to the nearest one?" If the player responds "Yes", ANNA gives turn-by-turn directions to the event.
Consequences: Replicates the familiar dynamic of a virtual assistant such as Siri or Alexa. Provides a relatively clear and constrained task for the game system. When the game character acts as the questioner, their questions can steer the player into giving certain kinds of responses that the game can recognise.
Relations: Often instantiates *Choice of Options*. Can instantiate *Scripted Conversation* or *Unscripted Conversation*.

2.3 WHO-WHAT-WHERE COMMANDS
The game accepts voice commands in the format of an imperative phrase that directs a subject (who) to take an action (what) in a place (where), in that order. The "who" or "what" element is sometimes unstated when the target subject or place is implicit in the rest of the command phrase or due to the context.
Example: *There Came an Echo* [38] uses an extensive vocabulary of *Who-What-Where Commands*, such as "Miranda move to Alpha Six" and "Everybody focus fire on Target Two on my mark". As these examples show, the commands can incorporate *Select by Name* or *Select by Type* to denote their subject or object, *Name an Action* or *Name a Behaviour* to denote their action, and *Waypoints* to denote their location. The second example also shows how *There Came an Echo* adds a "when" element to the pattern: the command is inactive until the player next says "mark".
Consequences: Allows the player to convert a small vocabulary of keywords into an extensive and highly configurable set of instructions, in a format that mimics everyday speech. This makes it well suited to games that focus on coordinating the movement and actions of multiple characters around a space, notably strategy games. This pattern creates the narrative implication that the player-character is in a position of authority over other characters, and seeing characters obey verbal orders supports a player experience of authority as well.
Relations: Can instantiate *Situated Commander* or *Floating Commander*. Can be instantiated by some combination of *Select by Name*, *Select by Type*, *Name an Action*, *Name a Behaviour*, *Waypoints*, *Relative Directions* and *Absolute Directions.*

2.4 SCRIPTED CONVERSATION
An in-character conversation in which the player participates by reading pre-scripted lines of dialogue that appear on screen, effectively "performing" the voice of the player-character.
Example: During conversation scenes in *Mass Effect 3* [6], each time the player-character has a turn to talk, the game shows a list of several options for what they can say. The player can choose an utterance from this list by reading it aloud. The player-character then gives a longer version of this utterance in its own voice, and the NPC responds.
Consequences: Allows the player to participate in a fully expressive conversation without the need for a sophisticated language understanding system, as the game system only needs to match their speech to a small number of pre-set utterances. Pre-scripted responses remove the task of formulating a response in a phrase the game will accept, which can interfere with a player's sense of presence [64].
Relations: Instantiates *Speak as a Character* and *Choice of Options*. Can be instantiated by *Question and Answer*. Conflicts with *Unscripted Conversation*.

2.5 UNSCRIPTED CONVERSATION
An in-character conversation in which the player participates by formulating and speaking utterances in the role of the player-character.
Example: Gameplay in *Seaman* [77] centres on *Unscripted Conversation* with a talking fish. The fish asks questions like "What do you do for a living?" and the player is given no specifications for how to answer. The game system listens for relevant keywords to match the player's answer with an appropriate response from the fish.
Consequences: This is the voice interaction format that most closely approximates everyday speech, and the one that is often imagined in depictions of future technology. Formulating an appropriate natural-language utterance on the spot can be more enjoyable than choosing from a menu of options, but it is also a more difficult task, and can leave the player with a lower sense of control or involvement in the story [64].
Relations: Instantiates *Speak as a Character*. Can be instantiated by *Question and Answer*. Conflicts with *Unscripted Conversation*.

3. Patterns for Selection
Patterns in this category describe methods of distinguishing between objects, including units and characters, with voice input. The patterns are *Select by Name* and *Select by Type*.

3.1 SELECT BY NAME
The player selects a specific object in the gameworld by saying its unique name.
Example: In *There Came an Echo*, the voice command "Miranda attack Target Two" uses *Select by Name* to identify both its subject (the friendly character Miranda) and its object (the enemy character labelled Target Two).

Consequences: One of the major limitations of voice interaction is its lack of a precise locus of manipulation, like a mouse pointer, which makes it difficult to quickly and precisely select an object on screen [79]. Unique names overcome this by providing an unambiguous and stable link between a game object and a specific utterance.
Relations: Instantiates *Who-What-Where Commands* in combination with *Absolute Directions*, *Relative Directions*, *Name an Action* or *Name a Behaviour*. Instantiated by *Waypoints*.

3.2 SELECT BY TYPE
The player selects objects of a certain type by saying a category to which they belong. May allow multiple categories to narrow down the selection.
Example: *Tom Clancy's EndWar* uses this pattern to select every unit in a category, with commands such as "calling all tanks, retreat". *I'm Hiding* [40] uses it to identify individual objects, through a combination of three words that collectively describe only one object on the screen, such as "big green crayon".
Consequences: Like *Select by Name*, this pattern helps to connect player speech to objects in the gameworld, with the added flexibility of enabling reference to multiple objects at once, and without the need to give every selectable object a unique name. However, it is more ambiguous than *Select by Name* and requires more thinking, as the player needs to appraise the qualities of the object and match them to a valid category label; for example, "green crayon" rather than "emerald marker".
Relations: Instantiates *Who-What-Where Commands* in combination with *Absolute Directions*, *Relative Directions*, *Name an Action* or *Name a Behaviour*.

4. Patterns for Navigation
Patterns in this category describe methods for delineating spaces and directions through voice input. The patterns are *Waypoints*, *Absolute Directions* and *Relative Directions*.

4.1 WAYPOINTS
The player can select a specific location in the gameworld by saying its unique name or set of coordinates.
Example: *ATC Voice* [47] uses the names of real-world air traffic control radio beacons as waypoints. The command "Learjet eight two golf fly direct blue mesa VOR" would tell a Learjet plane with the identification number 82G to fly towards a radio beacon named Blue Mesa.
Consequences: Allows the player to quickly and precisely indicate a location through voice input. Also reduces the decision space by having a limited set of locations that afford selection. This constricts the player's options and simplifies their decision-making process for navigation, which could be useful or frustrating depending on the task.
Relations: Instantiates *Select by Name*. Instantiates *Who-What-Where Commands* in combination with *Select by Name*, *Select by Type*, *Name an Action* or *Name a Behaviour*.

4.2 ABSOLUTE DIRECTIONS
The player can command a unit to move in a direction relative to the orientation of the gameworld.

Example: *ATC Voice* uses compass headings and altitudes to direct planes in flight. The command "Turn left heading zero nine zero, climb maintain one one thousand" tells a plane to turn directly East (90 degrees from magnetic North) and ascend to an altitude of 11000 feet.
Consequences: One of the rarest patterns for voice interaction. We suggest this is because it conflicts with players' tendency to focus on the orientation of their character, making *Relative Directions* more salient to the player's perspective in many cases.
Relations: Instantiates *Who-What-Where Commands* in combination with *Select by Name*, *Select by Type*, *Name an Action* or *Name a Behaviour*.

4.3 RELATIVE DIRECTIONS
The player can command a unit to move in a direction relative to the unit's own orientation.
Example: In *Northern Lights* [21], the player controls a dog sled driver, shouting "left", "right", "forward" and "stop" to tell the sled dogs where to go.
Consequences: Approximately replaces the role of the analogue stick or directional buttons in traditional game controls, but without the capacity for continuous variable input that these controls and *Volume Control* share.
Relations: Instantiates *Who-What-Where Commands* in combination with *Select by Name*, *Select by Type*, *Name an Action* or *Name a Behaviour*.

5. Patterns for Control
Patterns in this category describe ways of using voice input to control the actions of game characters. Patterns in this category imply some separation between the actions of the player and the actions of the characters in the game. The patterns are *Name an Action*, *Name a Behaviour*, *Shout to Shoot*, *Volume Control* and *Pitch Control*.

5.1 NAME AN ACTION
The player says the name of a discrete action to make a character perform that action.
Example: In *Nintendogs* [56], the player can tell a puppy to "lie down" or "roll over" on command.
Consequences: The simplicity and directness of mapping the verb for an action to the action makes it an intuitive combination. Indeed, some game designers use the term "verbs" to describe player actions even when not referring to spoken actions [18]. Its popularity bears out this impression, as it is one of the most common patterns in our dataset, applied to most scenarios for voice interaction gameplay.
Relations: Instantiates *Who-What-Where Commands* in combination with *Select by Name*, *Select by Type*, *Waypoints*, *Absolute Directions* or *Relative Directions*. Can be instantiated by *Spellcasting*.

5.2 NAME A BEHAVIOUR
The player says the name of a mode of behaviour to make a character adopt that behaviour.
Example: In the soccer game *FIFA 14* [27], the player can call out tactics such as "ultra-defensive" and "slow it down" to make their team shift behavioural priorities.
Consequences: Expands the voice command repertoire to encompass macro-instructions. In so doing, it distances the player's input from the game action and creates a more mediated sense of control, by placing a greater degree of decision-making for individual actions on the AI systems that control NPCs.
Relations: Instantiates *Who-What-Where Commands* in combination with *Select by Name* or *Select by Type*.

5.3 SHOUT TO SHOOT
When the player shouts, their avatar shoots.
Example: In the sidescrolling shoot-em-up *Pah!* [43], the player controls a spaceship flying horizontally at a constant speed. The ship fires a laser each time the player shouts "pah!" (or any word) above a certain volume threshold. Vocalisations below that threshold move the ship up and down on the screen, using the *Volume Control* pattern.
Consequences: Like *Volume Control*, this pattern is technically and conceptually simple but physically fatiguing for the player, especially if rapid shooting is required. It has almost exclusively been used in simple shoot-em-up games on mobile platforms, reflecting both its limited depth and its ability to create an amusing spectacle for an audience.
Relations: Usually conflicts with *Speak as a Character* and instantiates *Speak as the Player*, due to the dissociation between the acts of shouting and shooting.

5.4 VOLUME CONTROL
The volume of the player's voice modulates the speed or direction of a unit's movement.
Example: In the side-scrolling platforming game *Yasuhati* [30], the avatar walks forward at a speed relative to the volume of the player's voice: the louder the sound, the faster the movement. When the volume rises above a certain level, the avatar jumps. The player must alternate between shouts to jump between platforms and quiet speech to walk across platforms.
Consequences: In a technical sense, this pattern uses the simplest form of voice interaction: converting the loudness of sound into a one-dimensional continuous variable input. For players, however, it is a difficult task to maintain precise control over their voice volume. And paradoxically, the fact that this pattern disregards speech content creates more difficulty, as it forces the player to make continual arbitrary decisions about what sounds to make. As a result, players often end up uttering nonsensical combinations of words, or meaningless gibberish. This combination of amusingly absurd speech and difficult gameplay has made *Volume Control* popular for playing in front of an audience, including on YouTube [53].
Relations: Tends to conflict with *Speak as a Character* and therefore instantiate *Speak as the Player*. Although it could in theory be combined with a speech recognition-based pattern, there are no examples of this in our dataset.

5.5 PITCH CONTROL
The frequency of the player's voice modulates the speed or direction of a unit's movement.
Example: In Sporka et al.'s [68] humming-controlled Tetris variant, a hum with a rising pitch moves the puzzle piece right, a falling pitch moves the piece left, and a constant pitch moves the puzzle piece down.
Example: *Plurality Spring* [72] combines *Pitch Control* with *Match Pitch* and *Waypoints*. The screen displays several nodes that are each labelled with a series of musical notes. Matching the note sequences moves the avatar to the node.
Consequences: Affords a one-dimensional continuous variable input, like *Volume Control*, but with greater flexibility. Players can use quieter voice inputs to minimise fatigue and social embarrassment. Supports a "vocabulary" of controls based on pitch shifts and intervals, as in the examples above.
Relations: None.

6. Patterns for Performance
Patterns in this category describe voice actions that directly perform the activity they are meant to produce. In these patterns, unlike the previous category, the player enacts the in-game action with their own voice. The patterns are *Pronunciation*, *Karaoke*, *Breath Physics*, *Exclamation*, *Spellcasting* and *Overheard Noise*.

6.1 PRONUNCIATION
The player must match the pronunciation of a word as closely as possible to succeed.
Example: Most common in serious games designed for language learning, such as *Word War* [49], or speech rehabilitation, such as *Talking to Teo* [54]. These games replicate classroom speech exercises to allow players to practice independently and get feedback on their efforts.
Consequences: Focuses the game on the act of speech recognition for its own sake, rather than as a means to a broader end.
Relations: None. While intelligible pronunciation is always needed for speech recognition to work, the *Pronunciation* pattern describes cases where it is the primary goal.

6.2 KARAOKE
The player sings along to a song and is scored on the accuracy of their pitch. The game plays the backing music for the song and shows the lyrics on-screen, next to a visual indicator of their pitch.
Example: *Karaoke Revolution Volume 2* [34] was the first *Karaoke* game to include all the main elements that are now commonly found in this pattern, including a piano-roll style pitch indicator, handheld microphone, and virtual crowds on screen that respond to how well the player is singing.
Consequences: A notably social gameplay pattern, often played in front of an audience. Fletcher and Light's description of *Karaoke* as a "glue technology that assists in

crafting and strengthening social linkages amongst players" [29:1] highlights its focus on extroverted performance and the opportunities for self-expression that it facilitates.
Relations: None.

6.3 Breath Physics
The player blows into the microphone and the gameworld responds as though it felt the gust of wind.
Example: *The Legend of Zelda: Phantom Hourglass* [57] requires the player to blow into the microphone to extinguish small flames, as though blowing out a candle.
Example: *Resonance: The Lost Score* [23] uses a variation on this pattern, in which the player's voice can smash glass when it is held at the right frequency. In this case, it is the resonance rather than the sheer wind-force of the player's breath that physically affects the gameworld.
Consequences: An easy pattern for both the player and the game developer, as blowing air takes much less effort than shouting, but registers as a loud noise due to the sustained physical displacement of the microphone diaphragm. Even small and low-quality microphones can reliably detect air blowing into them, which makes this pattern technically viable on portable game devices.
As *Breath Physics* creates the illusion of directly physically affecting the gameworld, it should increase the player's feeling of presence in the gameworld. It is also likely to support the player's sense of identification with the player-character, as Carter et al. [13] describe, by aligning the actions of the player and the player-character.
Relations: Instantiates *Speak as a Character* and conflicts with *Speak as the Player*, by conflating the player with a diegetic player-character, even if the latter is only an implied presence in the gameworld.

6.4 Exclamation
The player exclaims loudly to trigger a vocal outburst by the player character.
Example: The adventure game *Phoenix Wright: Ace Attorney* [12] includes scenes of witness testimony in a courtroom. If the player yells "hold it!", "objection!" or "take that!" during these scenes, the player-character will interrupt the witness with the same shouted phrase, and then begins a cross-examination of what the witness just said.
Consequences: Allows the player to physically enact a moment of heightened emotion and drama, creating a strong (if momentary) opportunity for role-playing and identification with the character.
Relations: Instantiates *Speak as a Character* and conflicts with *Speak as the Player*, as the player's utterance is normally the same as the character's.

6.5 Spellcasting
The player performs a verbal incantation to invoke a magical effect in the gameworld. The incantation is often in the form of words from an invented language rather than a phrase in a real-world language.
Example: *In Verbis Virtus* [37] casts the player-character as a mage who casts spells from their hand by speaking phrases in the fictional Maha'ki language. Spells include "ekto lumeh" to project a beam of light and "obee kehnu" to lift and move an object telekinetically.
Consequences: The metaphor of *Spellcasting* is a useful conjunction between the tropes of the fantasy genre, which is common in games, and the effects of speech recognition technology. The pattern supports an embodied experience with a coherent sense of player identity while performing effects that would not normally be possible through speech.
Relations: Instantiates *Speak as a Character*. Conflicts with *Speak as the Player*. Can instantiate *Name an Action* if the incantation describes the effect of the spell.

6.6 Overheard Noise
When the player makes a sound, NPCs are alerted to the presence of the player-character.
Example: In the stealth horror game *Alien: Isolation* [19], the player-character must hide from an alien that is attracted towards in-game sounds, from loud gunshots to the quiet beeping of a motion tracker device. In the optional "noise detection" mode, the alien also responds to noise in the real world that is picked up by the game system's microphone, and reacts as though the sound had emanated from the player-character's position in the gameworld.
Consequences: Reduces the perception of a separation between the physical world and the gameworld [13]. This can increase the player's feeling of presence, but it can also disrupt it if sounds from other people or outside events are accidentally picked up by the game and interfere with the experience [78].
Relations: Instantiates *Speak as a Character*. Conflicts with *Speak as the Player*.

DISCUSSION
Through this pattern collection, we have provided an index of the main configurations that have been used for voice interaction in games. Its usefulness is primarily in defining a core vocabulary for describing voice interaction games, and in offering a list of tested solutions for designers who are approaching a voice-based game mechanic for the first time. However, it can also be used by scholars as a wider lens to survey the landscape of voice interaction games, and identify emergent conventions or neglected combinations.

When we do so, one of the first things that becomes apparent is that certain patterns have tended to cluster together. Most obvious is the "command" cluster: *Who-What-Where Commands* are always instantiated through some combination of *Select by Name*, *Select by Type*, *Waypoints*, *Absolute Directions*, *Relative Directions*, *Name an Action* and *Name a Behaviour*, and almost always framed by the persona of a *Situated Commander* or *Floating Commander*. This cluster is usually found in games that focus on tactically coordinating the movement of multiple units across a contested landscape, such as strategy games and team sports games. Most of the patterns in this cluster are solutions to a similar core problem: making an unbounded possibility space manageable through speech by assigning specific words as anchors for predefined action options.

A smaller cluster may be called the "branching dialogue" cluster, consisting of *Scripted Conversation*, *Choice of Options*, *Speak as a Character* and often *Question and Answer*. These patterns co-occur in games with dialogue scenes where the player (in the persona of the player-character) chooses what to say to an NPC, often asking questions that prompt some exposition about the game's narrative arc. This cluster is sometimes found in games that also feature the "command" cluster, as in *Mass Effect 3*.

Conversely, many patterns are notable for rarely or never appearing in combination with other patterns. *Shout to Shoot*, *Volume Control*, *Pitch Control*, *Pronunciation*, *Karaoke*, *Exclamation* and *Overheard Noise* are all notable for their seeming resistance to being used in combination with any other pattern. Something that distinguishes this set is that they are all patterns of direct action rather than mediated action: the player controls or performs a task directly with their voice, rather than using their voice to exchange information with a character, as in a conversation or a command. Whether these patterns' lack of integration with other patterns is a result of technical effort or because they serve different types of player experiences is not clear, but would be an interesting point for further study.

When we narrow our focus to the voice interaction patterns found in academic papers on games, we see a reflection of what Carter et al. call the operative paradigm of HCI games research [14]. That is to say, research has focused on using games instrumentally, as mechanisms for producing secondary benefits rather than for producing meanings or positive experiences directly. As such, games research has explored only a narrow set of voice interaction patterns. *Pronunciation* is particularly common among academic research games (e.g. [16,42,48,54,63,69]), as it provides a gamified format for language learning or speech therapy exercises. Accordingly, games in academic research are much more likely to use *Speak as the Player* than commercial and independent games; even those that include a central character are more likely to ask the player to talk *to* it than to talk *as* it. The patterns employed in these games shows that they are generally less engaged with the fictive and narrative elements of the videogame format, and do relatively little to give the player a sense of presence in a meaningful gameworld.

Audiogames

To create this pattern language, we sought to review the design of every screen-based digital game that has used voice interaction, and we are confident that our dataset includes almost all such games that are identifiable today. However, it excludes audiogames: digital games that have no significant visual component. This category has grown rapidly in the past three years with the expansion of the smart speaker market, such that there are now hundreds of apps for Google Home and thousands of skills for Amazon Alexa categorised as games. Audiogames merit a separate analysis, which our approach here is poorly suited to provide, due to the large number of minimally-documented audiogames, the fundamental difference of the format, and the fact that nearly all the academic literature to date has focused on voice interaction with screen-based games.

To test whether our pattern language is transferrable to audiogames, we surveyed 22 audiogames on Amazon Alexa that had been widely reviewed. *Choice of Options* was the core interaction pattern for 17 of these games, and 11 implemented their choices in a *Question and Answer* dialogue. Four games used *Who-What-Where Commands* as the core interaction pattern. Interestingly, four games used *Absolute Directions*, which was the rarest pattern in our main dataset. We interpret this as a sign that audiogames do not convey as strong a sense of orientation in the gameworld as screen-based games, due to the lack of a visible avatar for the player to orient themselves with.

CONCLUSION

In this paper, we have introduced the idea of a pattern language for voice interaction in games. We discussed the background of pattern languages in game design, reviewed studies on best practices for design patterns for HCI, and provided an overview of challenges that have been identified for voice interaction in games – primarily spatial ambiguity, and social awkwardness due to conflicts between the identity roles of the player and the player-character. We surveyed 449 videogames and 22 audiogames to discover 25 design patterns, related to diegetic framing, dialogue structure, selection, navigation, control and performance. We noted clusters of patterns around command and conversation, and a number of isolated patterns that used voice for direct action. Finally, we observed that academic research has focused on a narrow subset of design patterns, especially *Pronunciation*.

We believe the narrow focus of academic research partly reflects a discourse around voice interaction that is overly concerned with speech recognition, and with word recognition accuracy rates in particular, as a perceived barrier to a satisfying player experience. As our pattern language has demonstrated, this does not need to be the case. Game designers have been able to create enjoyable and popular game experiences using entirely non-verbal forms of voice interaction, with *Karaoke* and *Volume Control* being the most common examples to date but far from exhausting the possibilities of this modality. And patterns like *Spellcasting* and *Waypoints* have been used successfully to constrain speech input to clear, easily distinguishable phrases that allow imperfect speech recognition systems to recognise utterances with a high degree of accuracy. We hope that cataloguing these possibilities as design patterns encourages more experimentation with the format, and reminds both game designers and researchers of options that they have to take voice interaction in less obvious directions.

ACKNOWLEDGMENTS

Fraser Allison's work was supported by an Australian Government Research Training Program Scholarship and by the Microsoft Research Centre for Social NUI.

REFERENCES

1. Christopher Alexander, Sara Ishikawa, Murray Silverstein, Max Jacobson, Ingrid Fiksdahl-King, and Shlomo Angel. 1977. *A Pattern Language: Towns, Buildings, Construction*. Oxford University Press, New York.
2. Fraser Allison, Marcus Carter, and Martin Gibbs. 2017. Word Play: A History of Voice Interaction in Digital Games. *Games and Culture*: 155541201774630.
3. Valter Alves and Licinio Roque. 2010. A Pattern Language for Sound Design in Games. In *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound* (AM '10), 12:1–12:8. https://doi.org/10.1145/1859799.1859811
4. Valter Alves and Licinio Roque. 2011. A Deck for Sound Design in Games: Enhancements Based on a Design Exercise. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology* (ACE '11), 34:1–34:8. https://doi.org/10.1145/2071423.2071465
5. Peter Bayliss. 2007. Beings in the Game-world: Characters, Avatars, and Players. In *Proceedings of the 4th Australasian Conference on Interactive Entertainment* (IE '07), 4:1–4:6. Retrieved June 1, 2015 from http://dl.acm.org/citation.cfm?id=1367956.1367960
6. BioWare. 2012. *Mass Effect 3*. Electronic Arts, Redwood City, CA.
7. Staffan Björk. 2012. *Gameplay Design Patterns for Public Games*. EXPERIMEDIA.
8. Staffan Björk and Jussi Holopainen. 2004. *Patterns in Game Design*. Charles River Media, Inc., Rockland, MA.
9. Staffan Björk, Sus Lundgren, and Jussi Holopainen. 2003. Game Design Patterns. In *Proceedings of the 2003 DiGRA International Conference: Level Up* (DiGRA '03).
10. Staffan Björk and Johan Peitz. 2007. Understanding Pervasive Games through Gameplay Design Patterns. In *DiGRA '07 - Proceedings of the 2007 DiGRA International Conference: Situated Play*.
11. Jenny Brusk and Staffan Björk. 2009. Gameplay Design Patterns for Game Dialogues. In *Proceedings of DiGRA 2009*.
12. Capcom. 2005. *Phoenix Wright: Ace Attorney*. Capcom, Osaka, Japan.
13. Marcus Carter, Fraser Allison, John Downs, and Martin Gibbs. 2015. Player Identity Dissonance and Voice Interaction in Games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (CHI PLAY '15), 265–269. https://doi.org/10.1145/2793107.2793144
14. Marcus Carter, John Downs, Bjorn Nansen, Mitchell Harrop, and Martin Gibbs. 2014. Paradigms of Games Research in HCI: A Review of 10 Years of Research at CHI. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play* (CHI PLAY '14), 27–36. https://doi.org/10.1145/2658537.2658708
15. Marcus Carter, Martin Gibbs, and Michael Arnold. 2012. Avatars, Characters, Players and Users: Multiple Identities at/in Play. In *Proceedings of the 24th Australian Computer-Human Interaction Conference* (OzCHI '12), 68–71. https://doi.org/10.1145/2414536.2414547
16. Gabriel J. Cler, Talia Mittelman, Maia N. Braden, Geralyn Harvey Woodnorth, and Cara E. Stepp. 2017. Video Game Rehabilitation of Velopharyngeal Dysfunction: A Case Series. *Journal of Speech, Language, and Hearing Research* 60, 6S: 1800–1809. https://doi.org/10.1044/2017_JSLHR-S-16-0231
17. Michael Cook, Mirjam Eladhari, Andy Nealen, Mike Treanor, Eddy Boxerman, Alex Jaffe, Paul Sottosanti, and Steve Swink. 2016. PCG-Based Game Design Patterns. *arXiv:1610.03138 [cs]*. Retrieved February 27, 2018 from http://arxiv.org/abs/1610.03138
18. Chris Crawford. 2004. *Chris Crawford on Interactive Storytelling*. New Riders, Berkeley, CA.
19. Creative Assembly. 2014. *Alien: Isolation*. Sega, Tokyo.
20. Crystal Dynamics. 2014. *Tomb Raider: Definitive Edition*. Square Enix, Tokyo.
21. Augustin Dagoret, Jean-Louis Boudrand, Lorris Giovagnoli, Nicolas Terlon, Aymeric Thevenot, and Marie Vilain. 2016. *Northern Lights*. École Nationale du Jeu et des Médias Interactifs Numériques, Angoulême, France.
22. Ola Davidsson, Johan Peitz, and Staffan Björk. 2004. *Game Design Patterns for Mobile Games*. Nokia Research Centre.
23. Demerara Games. 2017. *Resonance: The Lost Score*. Demerara Games, Natal, Brazil.
24. Claire Dormann, Jennifer R. Whitson, and Max Neuvians. 2013. Once More With Feeling: Game Design Patterns for Learning in the Affective Domain. *Games and Culture* 8, 4: 215–237. https://doi.org/10.1177/1555412013496892
25. Joris Dormans. 2013. Making Design Patterns Work. In *Proceedings of the Second Workshop on Design Patterns in Games, DPG*.
26. Steven Dow, Manish Mehta, Ellie Harmon, Blair MacIntyre, and Michael Mateas. 2007. Presence and Engagement in an Interactive Drama. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '07), 1475–1484. https://doi.org/10.1145/1240624.1240847
27. EA Canada. 2013. *FIFA 14*. EA Sports, Redwood City, CA.
28. Thomas Erickson. 2000. Lingua Francas for Design: Sacred Places and Pattern Languages. In *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (DIS '00), 357–368. https://doi.org/10.1145/347642.347794
29. Gordon Fletcher and Ben Light. 2011. Interpreting digital gaming practices: SingStar as a technology of work. Retrieved March 30, 2015 from http://usir.salford.ac.uk/17245/
30. Freedom Crow. 2017. *Yasuhati*. Freem Inc., Nishinomiya, Japan.
31. Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. 1995. *Design Patterns: Elements of Reusable Object-oriented Software*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
32. Peter Gorniak and Deb Roy. 2005. Probabilistic Grounding of Situated Speech Using Plan Recognition and Reference Resolution. In *Proceedings of the 7th International Conference on Multimodal Interfaces* (ICMI '05), 138–143. https://doi.org/10.1145/1088463.1088489
33. Susumu Harada, Jacob O. Wobbrock, and James A. Landay. 2011. Voice Games: Investigation into the Use of Non-speech Voice Input for Making Computer Games More Accessible. In *Proceedings of the 13th IFIP TC 13*

International Conference on Human-computer Interaction - Volume Part I (INTERACT'11), 11–29.

34. Harmonix. 2004. *Karaoke Revolution Volume 2*. Konami, Tokyo.
35. Kenneth Hullett and Jim Whitehead. 2010. Design Patterns in FPS Levels. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games* (FDG '10), 78–85. https://doi.org/10.1145/1822348.1822359
36. Takeo Igarashi and John F. Hughes. 2001. Voice As Sound: Using Non-verbal Voice Input for Interactive Control. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology* (UIST '01), 155–156. https://doi.org/10.1145/502348.502372
37. Indomitus Games. 2015. *In Verbis Virtus*. Indomitus Games, Borgo a Mozzano, Italy.
38. Iridium Studios. 2015. *There Came an Echo*. Iridium Studios, Los Angeles.
39. W. Lewis Johnson and Andre Valente. 2009. Tactical Language and Culture Training Systems: Using AI to Teach Foreign Languages and Cultures. *AI Magazine* 30, 2: 72. https://doi.org/10.1609/aimag.v30i2.2240
40. Joyce Hakansson Associates. 1983. *I'm Hiding*. Milton Bradley, East Longmeadow, MA.
41. Peter H. Kahn, Nathan G. Freier, Takayuki Kanda, Hiroshi Ishiguro, Jolina H. Ruckert, Rachel L. Severson, and Shaun K. Kane. 2008. Design Patterns for Sociality in Human-Robot Interaction. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction* (HRI '08), 97–104. https://doi.org/10.1145/1349822.1349836
42. Anuj Kumar, Pooja Reddy, Anuj Tewari, Rajat Agrawal, and Matthew Kam. 2012. Improving Literacy in Developing Countries Using Speech Recognition-supported Games on Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 1149–1158. https://doi.org/10.1145/2207676.2208564
43. Labgoo. 2011. *Pah!* Labgoo, Tel Aviv.
44. Petri Lankoski and Staffan Björk. 2007. Gameplay Design Patterns for Believable Non-Player Characters. In *Proceedings of the 2007 DiGRA International Conference: Situated Play* (DiGRA '07).
45. Petri Lankoski, Anja Johansson, Benny Karlsson, Staffan Björk, and Pierangelo Dell'Acqua. 2011. AI Design for Believable Characters via Gameplay Design Patterns. In *Business, Technological and Social Dimensions of Computer Games: Multidisciplinary Developments*. IGI Global, 15. https://doi.org/10.4018/978-1-60960-567-4.ch002
46. Chris Lewis, Noah Wardrip-Fruin, and Jim Whitehead. 2012. Motivational Game Design Patterns of 'Ville Games. In *Proceedings of the International Conference on the Foundations of Digital Games* (FDG '12), 172–179. https://doi.org/10.1145/2282338.2282373
47. Looking West Apps. 2013. *ATC Voice*. Looking West Apps, Denver, CO.
48. Marta Lopes, João Magalhães, and Sofia Cavaco. 2016. A Voice-controlled Serious Game for the Sustained Vowel Exercise. In *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology* (ACE '16), 32:1–32:6. https://doi.org/10.1145/3001773.3001807
49. Ian McGraw, On Yoshimoto, and Stephanie Seneff. 2009. Speech-enabled Card Games for Incidental Vocabulary Acquisition in a Foreign Language. *Speech Communication* 51, 10: 1006–1023. https://doi.org/10.1016/j.specom.2009.04.011
50. Bria Mears and Jichen Zhu. 2017. Design Patterns for Silent Player Characters in Narrative-driven Games. In *Proceedings of the 12th International Conference on the Foundations of Digital Games* (FDG '17), 59:1–59:4. https://doi.org/10.1145/3102071.3106366
51. David Milam, Lyn Bartram, and Magy Seif El-Nasr. 2012. Design Patterns of Focused Attention. In *Proceedings of the First Workshop on Design Patterns in Games* (DPG '12), 5:1–5:8. https://doi.org/10.1145/2427116.2427121
52. David Milam and Magy Seif El-Nasr. 2010. Design Patterns to Guide Player Movement in 3D Games. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games* (Sandbox '10), 37–42. https://doi.org/10.1145/1836135.1836141
53. Beckett Mufson. 2017. Watch People Play a Hilarious, Frustrating, Voice-Controlled iPhone Game. *VICE Creators*. Retrieved April 14, 2018 from https://creators.vice.com/en_au/article/vvjvk9/hilarious-frustrating-voice-controlled-iphone-game
54. Andrés Adolfo Navarro-Newball, Diego Loaiza, Claudia Oviedo, Andrés Darío Castillo, A. Portilla, Diego Linares, and Gloria Inés Álvarez. 2014. Talking to Teo: Video game supported speech therapy. *Entertainment Computing* 5, 4: 401–412. https://doi.org/10.1016/j.entcom.2014.10.005
55. Neversoft. 2008. *Guitar Hero: World Tour*. Activision, Santa Monica, CA.
56. Nintendo Entertainment Analysis & Development. 2005. *Nintendogs*. Nintendo, Kyoto, Japan.
57. Nintendo Entertainment Analysis & Development. 2007. *The Legend of Zelda: Phantom Hourglass*. Nintendo, Kyoto, Japan.
58. Yue Pan and Erik Stolterman. 2013. Pattern Language and HCI: Expectations and Experiences. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '13), 1989–1998. https://doi.org/10.1145/2468356.2468716
59. James R. Parker and John Heerema. 2008. Audio Interaction in Computer Mediated Games. *International Journal of Computer Games Technology* 2008: 1:1–1:8. https://doi.org/10.1155/2008/178923
60. Playground Games. 2014. *Forza Horizon 2*. Microsoft Studios, Redmond, WA.
61. RDI Video Systems. 1984. *Thayer's Quest*. RDI Video Systems, Carlsbad, CA.
62. Gabriel Rivera, Kenneth Hullett, and Jim Whitehead. 2012. Enemy NPC Design Patterns in Shooter Games. In *Proceedings of the First Workshop on Design Patterns in Games* (DPG '12), 6:1–6:8. https://doi.org/10.1145/2427116.2427122
63. Zak Rubin and Sri Kurniawan. 2013. Speech Adventure: Using Speech Recognition for Cleft Speech Therapy. In *Proceedings of the 6th International Conference on Pervasive Technologies Related to Assistive Environments* (PETRA '13), 35:1–35:4. https://doi.org/10.1145/2504335.2504373

64. Serdar Sali, Noah Wardrip-Fruin, Steven Dow, Michael Mateas, Sri Kurniawan, Aaron A. Reed, and Ronald Liu. 2010. Playing with Words: From Intuition to Evaluation of Game Dialogue Interfaces. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games* (FDG '10), 179–186. https://doi.org/10.1145/1822348.1822372
65. Katarina Segerståhl and Timo Jokela. 2006. Usability of Interaction Patterns. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '06), 1301–1306. https://doi.org/10.1145/1125451.1125693
66. Ben Shneiderman. 2000. The Limits of Speech Recognition. *Communications of the ACM 43*, 63–65.
67. Gillian Smith, Ryan Anderson, Brian Kopleck, Zach Lindblad, Lauren Scott, Adam Wardell, Jim Whitehead, and Michael Mateas. 2011. Situating Quests: Design Patterns for Quest and Level Design in Role-Playing Games. In *Interactive Storytelling* (Lecture Notes in Computer Science), 326–329. https://doi.org/10.1007/978-3-642-25289-1_40
68. Adam J. Sporka, Sri H. Kurniawan, Murni Mahmud, and Pavel Slavík. 2006. Non-speech Input and Speech Recognition for Real-time Control of Computer Games. 213–220. https://doi.org/10.1145/1168987.1169023
69. Chek Tien Tan, Andrew Johnston, Kirrie Ballard, Samuel Ferguson, and Dharani Perera-Schulz. 2013. sPeAK-MAN: Towards Popular Gameplay for Speech Therapy. In *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death* (IE '13), 28:1–28:4. https://doi.org/10.1145/2513002.2513022
70. Zachary O. Toups, William A. Hamilton, and Sultan A. Alharthi. 2016. Playing at Planning: Game Design Patterns from Disaster Response Practice. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play* (CHI PLAY '16), 362–375. https://doi.org/10.1145/2967934.2968089
71. Mike Treanor, Alexander Zook, Mirjam P Eladhari, Julian Togelius, Gillian Smith, Michael Cook, Tommy Thompson, Brian Magerko, John Levine, and Adam Smith. 2015. AI-based game design patterns. In *Proceedings of the 10th International Conference on the Foundations of Digital Games 2015 (FDG 2015)*, 8. Retrieved from http://www.fdg2015.org/papers/fdg2015_paper_23.pdf
72. Paul Turowski and Simon Hutchinson. 2017. *Plurality Spring*.
73. Ubisoft Montreal. 2006. *Rainbow Six: Vegas*. Ubisoft, Paris.
74. Ubisoft Shanghai. 2008. *Tom Clancy's EndWar*. Ubisoft, Paris.
75. Eduardo Velloso and Marcus Carter. 2016. The Emergence of EyePlay: A Survey of Eye Interaction in Games. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play* (CHI PLAY '16), 171–185. https://doi.org/10.1145/2967934.2968084
76. Marco Filipe Ganança Vieira, Hao Fu, Chong Hu, Nayoung Kim, and Sudhanshu Aggarwal. 2014. PowerFall: A Voice-controlled Collaborative Game. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play* (CHI PLAY '14), 395–398. https://doi.org/10.1145/2658537.2662993
77. Vivarium. 1999. *Seaman*. Sega, Tokyo.
78. Greg Wadley, Marcus Carter, and Martin Gibbs. 2015. Voice in Virtual Worlds: The Design, Use, and Influence of Voice Chat in Online Play. *Human–Computer Interaction* 30, 3–4: 336–365. https://doi.org/10.1080/07370024.2014.987346
79. Greg Wadley and Nicolas Ducheneaut. 2009. The 'Out-of-Avatar Experience': Object-focused Collaboration in Second Life. In *Proceedings of the 11th European Conference on Computer Supported Cooperative Work* (ECSCW 2009), 323–342. https://doi.org/10.1007/978-1-84882-854-4_19
80. José P. Zagal, Staffan Björk, and Chris Lewis. 2013. Dark Patterns in the Design of Games. In *Proceedings of the 8th International Conference on the Foundations of Digital Games (FDG 2013)*, 39–46.

7 Voice Gameplay Experience

This chapter presents Study 4 in the form of a peer-reviewed conference paper:

> Allison, F., Newn, J., Smith, W., Carter, M., & Gibbs, M. (2019). Frame Analysis of Voice Interaction Gameplay. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 393:1–393:14. https://doi.org/10.1145/3290605.3300623

The paper is included verbatim in its published format, following a discussion of how it contributes to the thesis. I led all stages of the research and writing process. My fellow PhD candidate, Joshua Newn, contributed to the technical setup, thematic analysis and writing of the paper. My thesis supervisors, Wally Smith, Marcus Carter and Martin Gibbs, contributed to the study design and writing of the paper.

7.1 Introduction

The previous chapter detailed the creation of a pattern language to encapsulate the major variations that have been evident in voice game design throughout its history. An analysis of these design patterns led to the finding that voice gameplay has come in five distinct genres: Voice Practice, Voice Performance, Voice Command, Voice Conversation and Voice Incantation. Of those genres, Voice Practice has received the lion's share of academic research, as it is the form of voice gameplay typically used in serious games (see Chapter 6). Voice Performance and Voice Command were the two most common genres identified in the study, but the player experience of either of these forms of voice gameplay has received little study. No prior literature has examined the experience of this kind of voice gameplay in depth, with the exception of a grounded theory study that I co-authored early in my candidature (Carter et al., 2015), and no study has provided a model for understanding how the experience of playing a game with one's voice differs from the traditional experience of playing a game with one's hands. Research Question 4 addresses this gap in the literature:

> **RQ4** What is the distinctive nature of voice gameplay?

The previous chapter also shows that voice game design has made extensive use of non-verbal voice input—that is, voice input that does not involve speech recognition. This is in contrast to non-game voice interfaces, which have rarely used non-verbal voice input outside of accessible control schemes for users with disabilities. The Voice Performance genre primarily consists of non-verbal voice gameplay. Karaoke games are a prominent example: although they present the player with a set of lyrics to sing along to, the games respond only to the pitch of the player's voice, and do not process its lexical content. The review of prior literature on non-verbal voice games found an ethnographic study (Fletcher & Light, 2011) and several usability evaluations (Hämäläinen et al., 2004; Harada et al., 2011; Sporka et al., 2006), but no comparisons of the player experience of verbal and non-verbal voice gameplay. Research Question 5 addresses this gap:

RQ5 How do players experience verbal and non-verbal voice gameplay?

This chapter presents the fourth and final study of the thesis. It provides answers to both RQ4 and RQ5 through a lab-based participant observation and interview study. Thematic analysis was used to developed seven themes that characterise the player experience of voice gameplay, and frame analysis (Goffman, 1974) was used to construct a general model of gameplay that accounts for those seven themes.

7.2 Research method

Adult participants (n=24) were recruited for a lab-based study. In pairs, the participants each played three different digital games, alternately using voice controls and manual controls for each one. The games were chosen to represent the two most common genres of voice gameplay identified in the previous chapter: Voice Command and Voice Performance. Participants were interviewed after each round of play. Two researchers (myself and a co-author on the published paper) analysed the interview transcripts using an inductive open coding process, and I developed these codes into a full thematic analysis, using frame analysis as an interpretive framework. The research methods and interpretive framework are discussed in more detail in Chapter 3 and in the published paper below.

7.3 Summary of findings

Players were somewhat divided in their responses to Voice Command (verbal) and Voice Performance (non-verbal) gameplay. The majority of players preferred the two Voice Command games, explaining that they felt more comfortable, more in control and more focused on the strategic gameplay during these games than during the Voice Command game. However, a third of the players enjoyed the Voice Performance game the most, due to its lack of constraints on voice input and its potential to facilitate an enjoyable and unusual shared social experience.

An initial analysis determined that four frames (see Goffman, 1974) appeared repeatedly in players' descriptions of their gameplay experience. These frames corresponded to a pre-existing theoretically grounded model of gameplay proposed by Conway and Trevillian (2015), although with one of its three frames subdivided into two. The paper presents a refined version of this model of gameplay, called the SFSC Model, which consists of four frames: Social World, Functional World, Strategic World and Character World. Each frame is a "transformation" (Goffman, 1974, p. 41) of the meaning of the action in the frame that precedes it. For example, the act of speaking to a videogame console (Social World frame) is transformed into the act of selecting commands from a virtual menu (Functional World frame), which is further transformed into the act of directing in-game units towards a certain objective (Strategic World frame), which is further transformed into the act of playing the role of a general commanding troops (Character World frame). Because the frames are built upon each other in sequence, a disruption in an earlier frame tends to disrupt the later frames as well. In the aforementioned scenario, if the player accidentally uses the wrong phrasing for a voice command (Functional World frame), their utterance loses its meaning as a game instruction (Strategic World frame), which interrupts their role-playing as the army general (Character World frame); but the meaning of the utterance in the Social World frame remains unaffected.

A thematic analysis of player experiences with voice and manual controls produced a rich description of players' experience with voice gameplay. The major themes were as follows:

1. Thematically appropriate voice commands increase engagement with the Character World.
2. Vocalisations that sound odd in the Social World disrupt engagement with the Character World.
3. All vocalisations are interpreted as meaningful communication.
4. Players remember voice command phrases by meaning, not always by wording.
5. Speech intended for the Social World can infringe upon the Functional World.
6. Voice commands encourage concentration on the Strategic World.
7. Voice commands feel disconnected from their effects in the Strategic World.

The first five themes describe aspects of the voice gameplay experience that resulted from interactions between two or more of the four frames of gameplay. Most of these frame interactions were disruptive to the player's sense of engagement with the gameworld. The last two themes describe ways in which the experience of voice interaction gameplay using voice commands differed from traditional manual gameplay, according to participants' responses.

7.4 Contributions to the thesis

The results of this study address Research Question 4: "What is the distinctive nature of voice gameplay?" All seven of the major themes from the qualitative interviews describe experiences that differentiate voice gameplay from non-voice gameplay. Players reported that they felt more engaged with the Character World of a game when they played it with voice controls than when they played it with manual controls (Theme 1). Some players perceived themselves to be paying greater attention to the Strategic World of the game during voice gameplay (Theme 6). However, players also reported that they felt a reduced sense of personal agency during voice gameplay, and correspondingly perceived game characters to have greater autonomy over their own actions when directed with voice commands (Theme 7). These findings have informed the two major contributions of the overall thesis, discussed in Chapter 8: namely, a general model of gameplay experiences (SFSC Model) that accounts for these issues, and a typology of voice game genres that connects the player experience themes to the design patterns that formulate each type of voice gameplay.

The study also addresses Research Question 5: "How do players experience verbal and non-verbal voice gameplay?" The responses show that non-verbal voice gameplay is a markedly different experience to verbal voice gameplay, characterised by a heightened engagement with the Social World frame (Themes 3 and 5) and a lower engagement with the Character World frame (Themes 1 and 2). Several of the key dynamics of voice command gameplay involve issues of speech formulation and speech recognition (Themes 4, 6 and 7), which are not relevant to non-verbal voice gameplay. Players described a heightened awareness of the tensions between social frames during non-verbal voice gameplay as compared to other forms of gameplay; the articulation of these frame tensions contributed substantially to the development of the SFSC Model. All of these findings are discussed at greater length in the published paper, as well as in Chapter 8.

Frame Analysis of Voice Interaction Gameplay

Fraser Allison
School of Computing and Information Systems, The University of Melbourne
Melbourne, Australia
fraser.allison@unimelb.edu.au

Joshua Newn
School of Computing and Information Systems, The University of Melbourne
Melbourne, Australia
joshua.newn@unimelb.edu.au

Wally Smith
School of Computing and Information Systems, The University of Melbourne
Melbourne, Australia
wally.smith@unimelb.edu.au

Marcus Carter
Dept. of Media and Communications, The University of Sydney
Sydney, Australia
marcus.carter@sydney.edu.au

Martin Gibbs
School of Computing and Information Systems, The University of Melbourne
Melbourne, Australia
martin.gibbs@unimelb.edu.au

ABSTRACT

Voice control is an increasingly common feature of digital games, but the experience of playing with voice control is often hampered by feelings of embarrassment and dissonance. Past research has recognised these tensions, but has not offered a general model of how they arise and how players respond to them. In this study, we use Erving Goffman's frame analysis [16], as adapted to the study of games by Conway and Trevillian [9], to understand the social experience of playing games by voice. Based on 24 interviews with participants who played voice-controlled games in a social setting, we put forward a frame analytic model of gameplay as a social event, along with seven themes that describe how voice interaction enhances or disrupts the player experience. Our results demonstrate the utility of frame analysis for understanding social dissonance in voice interaction gameplay, and point to practical considerations for designers to improve engagement with voice-controlled games.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; • **Applied computing** → *Computer games.*

KEYWORDS

Voice interaction; voice control; games; frame analysis

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05...$15.00
https://doi.org/10.1145/3290605.3300623

ACM Reference Format:
Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, and Martin Gibbs. 2019. Frame Analysis of Voice Interaction Gameplay. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3290605.3300623

1 INTRODUCTION

In our narratives about the future, voice control is consistently depicted as an easy and enjoyable way of interacting with technology. It routinely appears as a source of entertainment in science fiction stories, from the chess-playing ship's computer in *Star Trek* to the android playground of *Westworld*. Even in work-oriented narratives, such as Apple's 1987 *Knowledge Navigator* concept film that showcased a virtual talking butler as a near-future technology, voice interaction is shown as being fun and sociable. Despite these aspirations, however, current voice-controlled systems are more often characterised as troublesome [30], disappointing [23] or embarrassing [7].

Since at least the 1980s, game designers have sought to make voice interaction fun; with some success [1]. Researchers have documented an array of approaches to voice-controlled games, including some that recreate the characterful speech-driven technology of science fiction [2]. To date, however, this form of conversational interaction has proven less popular in games than simpler voice interactions that do not use speech recognition, such as karaoke games that only respond to vocal pitch [1]. When videogames have used speech recognition to emulate meaningful communication, it has been known to introduce new tensions into the player experience; a study of online responses to voice interaction games found that players were sensitive to the alignment between spoken interaction and their character identity in the gameworld, with misaligned speech creating "identity dissonance" [7].

This suggests that the imaginary or narrative dimension of a voice interaction game is an important factor in how players experience it, but it is a dimension that has been largely treated as peripheral in academic studies [2]. There is a substantial literature on the characterisation of conversational agents and social robots outside of games (e.g. [4, 11]). However, we consider voice interaction in videogames to be an ontologically different phenomenon—typically distinguished by being situated within a deeper structure of interconnected narrative, challenge and world-simulation elements—which requires study on its own terms to be properly understood.

In this paper, we present a theoretical account of the player experience of voice interaction gameplay through the lens of Erving Goffman's frame analysis [16]. We have applied frame analysis, as adapted to the study of games by Conway and Trevillian [9], to an interview-based study with 24 participants who played three different voice interaction games. We conducted a thematic analysis of the interviews and gameplay recordings to develop an understanding of how different social frames supported or conflicted with each other in the players' attempts to engage with the game. Based on this, we identify several inter-frame tensions that are fundamental to voice interaction gameplay, and propose a revised version of Conway and Trevillian's frame analytic model [9] as a theoretical framework for understanding the player experience of voice interaction. The paper, therefore, provides three interlocking contributions: a general frame analytic model of gameplay; an extensive account of factors that influence the player experience of voice interaction gameplay; and a concise model that combines the two to show which specific sources of friction can hinder players' engagement with different elements of the player experience.

2 RELATED WORK

This study builds upon a body of work that has identified social cognition as an influencer of how we use and experience voice interaction. Nass and Moon put forward evidence to suggest that "individuals behave toward and make attributions about voice systems using the same rules and heuristics they would normally apply to other humans" [26], and show social behaviours such as politeness [24], gender stereotyping [27] and personality-based responses [25] when interacting with voice interfaces. This suggests that voice interfaces alter the operative relationship between a user and a device into a pseudo-social interaction with an ersatz human. This is complicated by the fact that voice interaction takes places in a human social context. In day-to-day usage, people split their conversation between the device and each other [30].

Research on voice interaction in videogames has noted social context as an important factor in how the games are played [1, 14, 17]. An analysis of online responses to voice interaction games found that players preferred voice commands that allowed them to speak in the voice of their character; commands that did not match what the player-character might say were described by players as "uncomfortable" and "embarrassing" [7]. This suggests that alignment of player roles or identities across different frames of reference is an important factor in the experience of voice interaction games, and leads us to consider frame analysis as a perspective that can analyse these overlapping identities.

Frame Analysis

To investigate the social structure of situations involving voice-controlled games, our analysis draws on the theoretical perspective of frame analysis developed by the sociologist Erving Goffman [16]. Goffman pioneered a dramaturgical approach to understanding everyday social situations in which people's interaction is understood through a broad metaphor of theatre. Just as actors on the stage reproduce a script, so social situations are enacted and experienced as instantiations of more general patterns, or *frames*, such as weddings, football matches, business deals, and so on. A frame could be anything that provides an answer to the question "What is it that is going on here?" A major innovation in Goffman's approach is the observation that situations can be multiply framed. On seeing a chess game, for example, we may answer the question "What is it that is going on here" at different levels, responding that people in the situation are: moving pieces of wood; playing chess; or striving to prove their intellectual superiority. In Goffman's terms, a *primary* behaviour of moving pieces of wood is transformed by the frame of chess, then transformed again by the frame of a battle of intellect.

Games researchers have used frame analysis extensively to understand the multi-layered nature of gameplay situations [6, 9, 10, 13, 18, 22, 29, 35]. Fine [13] adapted Goffman's frame analysis to study role-playing games, identifying how Dungeons & Dragons players alternate between a fantasy frame and a primary frame of real life. Fine identified two reframings in this situation: the original social behaviour of a person was reframed into that of a player, which was reframed again into the activity of a game character. An important observation was the way people moved rapidly between frames, sometimes marked by subtle shifts in spoken style to signify the role-playing frame ("What be the date?" "It's the end of Langala.") or the primary framework of the untransformed social situation ("No, in this world." "August tenth.") [13]. As noted by Harrop et al., "people tend to engage in role embracement [15], rather than becoming the role, in order to allow distancing oneself from one's role, so that a failure of the character is not taken to mean a failure of the person" [18].

Conway and Trevillian [9] used frame analysis as the basis of a three-level model of "the game event", dividing player experience into a Social World, an Operative World and a Character World (abbreviated as SOC). The Social World is the primary frame of the untransformed social situation, which I experience as my everyday self. The Operative World is the

Figure 1: Selected games: The Howler (left), Air Traffic Control Voice (middle) and Tom Clancy's EndWar (right).

frame in which I experience the game as a user of an operable system with rules and affordances; "football player" is a role that exists in this frame. The Character World is the frame of the imaginary diegetic space of the gameworld; it is within this frame that I experience myself as a game character such as Lara Croft. Conway and Trevillian emphasise again how gameplay is characterised by rapid shifting between frames: "the human player, as *Dasein*, maintains the capability to switch levels moment to moment, literally between milliseconds in some instances, such as playing in a crowded street on a mobile phone: immersed in the Character World I tap my thumb against the screen to make a dialog choice, I switch to Social World and check the street name, then to Operative World and check the score, and so on." [9].

Significantly for the application of frame analysis in this paper, Conway [8] draws attention to Goffman's emphasis on the various ways frame structures can break down, leading to a loss of shared understanding, and embarrassment and awkwardness at misunderstood activity; and related to this, the ongoing effort of players to mark, repair and maintain frame boundaries. Conway studies the response cries (such as "ouch!" and "no!") of people playing videogames, and interprets these cries as social performances intended to manage the appearance of the player to spectators. Along with the example from Fine above, this observation shows how players' speech acts, including the rich combination of verbal and non-verbal cues, present an important signifier of frame-shifts and is a promising site for investigation. The aim of this paper is to apply and elaborate this frame analysis for the situation of voice-controlled games, exploring how frame-shifts, both intentional and inadvertent, constitute a vital aspect of the gameplay experience and are therefore an important area of consideration for the game designer.

Verbal and Non-verbal Voice Interaction

Not all voice interfaces use speech as an input. While there have been numerous configurations of speech-based voice interaction in games [2], the most popular and commercially successful types of voice interaction games have been those that eschew speech recognition, namely karaoke games that respond to vocal pitch and shout-to-move platformer games that only respond to voice volume [1]. Researchers have established the viability of game controls based on pitch [17, 19, 33], volume [19], breath [34, 36] and tonguing [19]. Following Igarashi and Hughes [19], we refer to this whole category as *non-verbal voice interaction*, and we refer to voice interaction that uses speech recognition as *verbal voice interaction*, or as *voice command* when the speech is in the form of a command. *Voice control* refers to the use of verbal or non-verbal voice inputs to control the game system, as distinct from *manual control* such as mouse and keyboard. The difference between verbal and non-verbal voice interaction for games has received little empirical study, with one small-scale test of a prototype game [33] being the only study we are aware of.

3 METHODOLOGY

We selected three games to be played in the study (see Figure 1). We chose to use commercially-published games rather than custom-built prototypes to ensure that our study reflects an accurate experience of real-world, high-fidelity game design, complete with a fully realised aesthetic gameworld—an element that has been lacking in previous HCI research on voice interaction games [2], and which we would have been unable to achieve in a research prototype. We considered dozens of candidate games, and excluded those that were too complex or too simple to play for a 15 minute period, those that used voice interaction as a secondary feature rather than a primary modality, and those that were not playable by manual input as well as voice input. We selected the following three games as the ones that offered the best contrasting experience of different voice interaction game design patterns [2], as they cover both verbal and non-verbal voice input, both PC and mobile platforms, and three distinct game genres. Below we briefly summarise salient differences between the games:

The Howler is a physics puzzle game, controlled wholly by the volume of the player's voice. Any sustained sound causes the player's avatar (a hot air balloon) to rise, while the absence of sound causes it to fall. Manual controls use a mouse click instead of sound. At different altitudes, the wind pushes the balloon either left or right. The player must use vertical and horizontal momentum to navigate a series of obstacles.

Air Traffic Control Voice (ATCV) is an air traffic control simulator, in which the player directs aeroplanes to arrive at and depart from airport runways without crashing. It uses realistic air traffic control vocabulary for commands, such as: "Learjet two one golf, turn left heading zero six zero". Manual controls use touchscreen gestures such as tapping a plane to select it and swiping a semi-circle to change its heading.

Tom Clancy's EndWar (EndWar) is a real-time strategy game in a modern military setting. It uses 'who-what-where' commands [2] such as "unit one, move to Delta" and "all tanks, attack hostile four". A menu of commands appears when the player presses the voice input key. Manual controls follow the standard scheme for PC strategy games: left mouse click to select units, right mouse click to send them to a point.

	Participant A	**Participant B**
The Howler (single player)	Voice control	Manual control
	Manual control	Voice control
Post-game interview		
EndWar (single player)	Voice control	Manual control
	Manual control	Voice control
EndWar (multiplayer)	Voice control	Manual control
	Manual control	Voice control
Post-game interview		
ATCV (single player)	Voice control	Manual control
	Manual control	Voice control
Post-game interview		
Comparative interview		

Table 1: Study outline. Game order rotated between sessions.

Recruitment

We recruited 24 participants (14 identified as female, 10 as male) through posters and online bulletin board messages at the University of Melbourne. The majority were students or staff of the university. All participants received a gift voucher for participating. Ages ranged from 18 to 47, with an average age of 27. All participants had used a voice assistant such as Siri or Cortana, and seven had used a smart speaker such as Alexa. Seventeen used some form of voice-controlled technology at least a few times a year. Eight indicated they had played a voice-controlled videogame, and only one played voice-controlled games more than once a year. Sixteen played videogames on at least a weekly basis, and 18 played videogames in the company of other people at least a few times a year.

The study was run over 12 sessions of 90 minutes each, with a pair of participants in each session. Nine of the pairs knew each other prior to the study, and the other three were pairs of strangers. We used pairs as our interest lay in studying gameplay in a social setting. Previous studies have found that paired user testing is an effective alternative to single user testing for social usage of technology, as it can facilitate more discussion at a similar level of quality with less prompting by the researcher and provide a more relaxed and enjoyable experience for the participants [32, 37]—which is particularly significant for the study of gameplay.

In a pilot session, one participant was not able to get the speech commands to function adequately due to their accent. Following this, we updated the recruitment forms to only allow participants who were able to declare "I can make speech recognition systems understand me most of the time". All the remaining participants were able to use speech recognition successfully with at least one of the games in the study. However, three participants were unable to use voice commands in ATCV, and three different participants were unable to use voice commands in EndWar, due to the speech recognition systems failing to recognise their accents reliably (this group included both native English speakers and non-native English speakers). In these cases, participants played the game using manual control only, and gave comments based on their impression of the game and their observation of their co-participant using voice control. We discuss the decision not to exclude these participants in the Limitations section. All of the participants were able to use the non-verbal voice controls of The Howler.

Study Procedure

The study room was furnished to look similar to a domestic living room, with sofas, bookshelves, potted plants, a television and a desk with two computers. The lead researcher welcomed each pair of participants, explained the study and introduced each game in sequence. The participants played all three games in single-player mode, and played EndWar in a competitive multiplayer match against each other (see Table 1). The order of the games was varied between sessions in a Latin square design to reduce order effects. During each gameplay session, one participant at a time played with voice control while the other played with manual control, swapping halfway through. (Simultaneous use of voice control was tested in a pilot session, but caused too much interference due to utterances intended for one microphone being picked up by the other.)

After each game, the lead researcher conducted a short (around three minutes) semi-structed interview with the participants. Questions focused on their initial reactions, what they had been thinking about or concentrating on while playing the game, and whether they felt a sense of adopting an in-game identity ("Who did you feel like you were while playing that game?"). At the end of the third and final interview, the researcher asked participants to compare and rank the games they had played according to how much they *enjoyed* playing, how *comfortable* they felt, how much they felt *in control*, and how much they felt *focused*. These questions were chosen to elicit discussion of the degree and nature of their engagement with the game and with the social context. Participants were asked each question in regard to the voice interaction mode of play, and encouraged to explain if their responses would differ for the manual control mode. Participants were also asked in what setting they would be most likely to play the game again. All interviews were conducted with both participants simultaneously, which generated productive back-and-forth discussions between the participants in each session. To conclude, participants were asked about their experience with voice interaction technology and videogames.

Sessions were recorded by video camera, webcam and screen-capture on the desktop computers. Video recordings were transcribed using a professional transcription service. The lead researcher took notes throughout each session, which were used to inform the questions asked in the interviews.

Analysis

We conducted a thematic analysis on the interview recordings following Braun and Clarke's six-part model of thematic analysis [5]. During the first stage—familiarisation with the data—the lead researcher catalogued the main social frames that were apparent in the interview comments, identifying four frames. Another member of the research team noted that these frames reflected Conway and Trevillian's SOC model [9], with which the lead researcher had not been familiar. An evaluation of the SOC model was conducted, which we describe in the Results section. Based on this, the SOC model was taken as a primary framework and sensitising concept [3] for the thematic analysis.

We defined a small set of *a priori* codes for the interview data, focused on comparisons between the different games and modalities. The majority of codes were created inductively during the coding process, to describe unanticipated themes in participants' explanations of their experiences of voice interaction. The interviews were divided between two researchers, who independently reviewed each transcript from their set of interviews alongside its video recording, correcting any transcription errors and simultaneously developing the codes to label it. Once all the interviews had been coded once, the research team reconvened to compare their codes and confirm that the nascent themes were consistent across both sets of interviews. Following this, the two coders swapped data sets and conducted a second round of coding on the interviews that they had not yet analysed, adding to or revising the codes from the first round and further developing the themes. Once this was complete, the themes were reviewed, refined, defined and mapped in accordance with Braun and Clarke [5].

The research team also conducted an initial review of both the video recordings and the lead researcher's notes from the sessions. The behaviours and comments that we observed were largely accounted for in the interviews, and we concluded that participants' own explanations of their actions provided more insight than our interpretations of their actions for the purpose of understanding their subjective experience. Therefore, a structured analysis of the gameplay recordings was deemed unnecessary.

4 RESULTS

The first section below contains a summary of how participants compared the three games in terms of enjoyment, control, comfort and focus. This is intended to provide high-level context for the qualitative results that follow, and may be skipped over. The primary results are contained in the subsequent sections: 'Evaluating and Revising the SOC Model', which describes how our analysis validates and builds upon a frame analytic model of gameplay [9], and 'Themes of Player Experience', which describes how participants' accounts of voice interaction gameplay revealed competing and conflicting social frames.

Comparisons Between Games

This section summarises how participants ranked the games on several criteria, and what they focused on whilst playing each one. These questions were asked as part of a semi-structured interview that focused on encouraging discussion and explanation rather than ensuring comprehensiveness. Accordingly, the quantitative results do not provide statistically reliable measurements that can be used to compare the games, but rather an indication of the general sentiments of participants—as context for the qualitative findings that follow. Participants who did not provide a clear ranking of all three games for a question were not pressed to do so; for example, they were not asked to differentiate games that they considered to be equal. In addition, participants who were not able to use speech recognition[1] for EndWar or ATCV are excluded from the rankings for those games. We provide the quantitative results in the format (X/Y) to indicate the number of participants who gave an answer (X) out of the total who answered the question (Y).

When asked "Which of the three games did you enjoy the most and enjoy the least with voice control?", the largest numbers of participants stated that they enjoyed EndWar the most (12/19), ATCV as the middle (10/19), and The Howler

[1] See Recruitment and Limitations sections for further discussion.

the least (11/23). The Howler was polarising, however, as a substantial minority of participants (8/23) enjoyed it the most out of the games they played.

When asked "During which games did you feel most in control and least in control while you were using voice control?", the majority of participants ranked EndWar highest (13/19), ATCV as the middle (10/19), and The Howler lowest (15/21).

When asked "Which games did you feel most comfortable playing and least comfortable playing with voice control?", the largest number ranked EndWar highest (11/19) and ACTV as the middle (9/19). Again The Howler was polarising, with (10/22) ranking it as the least comfortable, (7/22) ranking it as the most comfortable, and only (5/22) ranking it as the middle.

For the question "During which game did you feel most focused on the game itself, as opposed to what was happening around you, while you were using voice control?", results were somewhat more split. For EndWar, slightly more participants felt most focused (7/15) than least focused (5/15). For ATCV, similar proportions felt most focused (4/14) and least focused (3/14), with the rest ranking it in the middle. As for The Howler, the majority of participants felt least focused (10/17), but a few felt most focused (5/17).

The enjoyment of EndWar was attributed to its greater sense of control. Participants found its voice commands easy to understand and remember due to their familiar terminology and the on-screen menu of command phrases. ATCV lacked a visible command menu and used unfamiliar phrases such as "descend maintain two thousand", which increased the difficulty of its voice commands. The Howler's lack of constraints on voice input was cited by both those who enjoyed it most and those who enjoyed it least. Those who preferred The Howler praised its non-verbal voice interaction as offering a low barrier to operation and enabling an unusual style of control. Those who disliked The Howler explained that the non-verbal voice interaction was uncomfortable in a way that they did not enjoy, and which distracted from their focus on the game.

When asked "What were you thinking about or concentrating on during the game you just played?", three participants (3/22) said that they were more focused on what they sounded like while playing The Howler than the actual gameplay. Several others reported being somewhat distracted by self-consciousness about their voice, although their main focus remained on the game itself. All participants reported being more focused on the gameplay than how they sounded for both EndWar and ATCV. Among those who reported concentrating on the gameplay, there was a split between those who focused on strategy (such as where to send units or how to approach a certain level) and those who focused on function (such as what intonation or volume to use to improve their control). For EndWar, the majority said that they concentrated on strategy (16/20) over function (4/20). For ATCV, there was an even split between strategy (8/17) and function (9/17). For The Howler, half of the participants concentrated on strategy (11/22), with the remainder focused on function (8/22) or how they sounded (3/22).

Perhaps corresponding with this, most participants (11/18) said that The Howler was best suited to being played in a social context. Only a minority said the same about EndWar (5/15) and ATCV (1/13), which were generally agreed to be best played "locked up in a room by myself" (P24). While The Howler provoked the greatest sense of concern about the perceptions of other people, it also had the potential to create an enjoyable shared social feeling, as described in the thematic analysis.

Evaluating and Revising the SOC Model

An exploratory analysis was conducted on the interviews, prior to the full thematic analysis, by one of the researchers. The researcher sorted the interview comments into categories based on their frame of reference, and came up with the following four frames: social, functional, strategic and imaginary. When the full research team discussed this result, it was noted that these four frames were similar to the three levels of the SOC model [9], a framework that the researcher who developed the categories was not familiar with. We re-examined the interview data against this framework, and determined that our "social" and "imaginary" frames were an exact fit for the Social World and Character World in the SOC model, and that our "functional" and "strategic" frames reflected different keys within the Operative World. We take this as an independent validation of the SOC model.

Accordingly, we have adopted a revised version of the SOC model as our primary framework for interpreting the results. Our revision is to split out the Operative World frame into two distinct frames, the Functional World and Strategic World (see Figure 2), which are described below. We refer to the revised model as the "SFSC model" to reflect these changes.

The *Functional World* frame was active when participants thought about themselves as *users* of the game system. It could be characterised as the answer to the question "How are you controlling the game?" When attending to this frame, participants focused on the interaction between their physical behaviour and the game state, such as when they concentrated on slowing their speech down to a pace that the game's speech recognition could follow.

The *Strategic World* frame was active when participants approached the game as *strategists*, concentrating on how to manipulate the dynamics of the gameplay to achieve their intentions and goals. It could be characterised as the answer to the question "What are you trying to do in the game?" When attending to this frame, participants thought about their actions within the context of the systemic relationships between game objects, with less attention paid to the manner

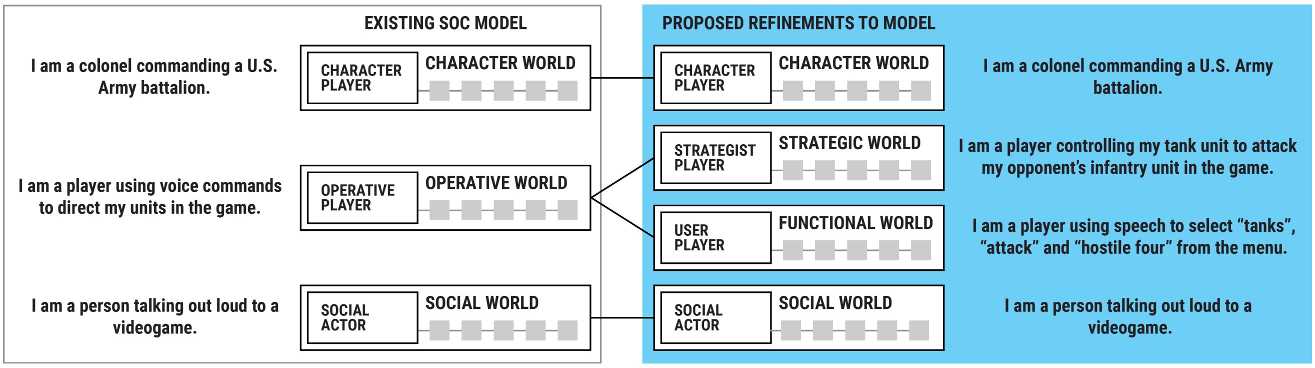

Figure 2: Proposed refinements to the SOC model of the game event [9].

in which they physically affected the game state. An example would be during a competitive multiplayer match, when a player contemplated where to send their units to achieve a pincer movement against their opponent's units.

The Functional World was much more salient to participants when they were using voice controls than when they were using manual controls. Many participants stated or implied that they did not need to be "actively thinking" (P15) about manual controls, especially those who had previous experience with similar game controls. This suggests a reason why the SOC model does not distinguish between functional and strategic levels within its Operative World frame: because experienced players can quickly master the controls of most ordinary games to the extent that they do not need to attend consciously to them, Functional World considerations tend to recede out of their awareness. However, when the control scheme for a game is unfamiliar or especially challenging—as voice controls often are for most players, and manual controls often are for novice players—the Functional World can be a dominant frame that players must attend to. Players oscillate their attention between functional and strategic frames of thinking, just as they oscillate their attention between pretence and reality [13, 18].

The next section explores how oscillations of attention play out across all four of the frames of the SFSC model (Social World, Functional World, Strategic World and Character World) during voice interaction gameplay. The thematic analysis demonstrates the necessity of considering all four frames to understand players' experience, particularly when they are engaging in voice interaction.

Themes of Player Experience

We developed seven themes to describe the effects that participants consistently described in their accounts of voice interaction gameplay: (1) Thematically appropriate voice commands increase engagement with the Character World, (2) Vocalisations that sound odd in the Social World disrupt engagement with the Character World, (3) All vocalisations are interpreted as meaningful communication, (4) Players remember voice command phrases by meaning, not always by wording, (5) Speech intended for the Social World can infringe upon the Functional World, (6) Voice commands encourage concentration on the Strategic World, and (7) Voice commands feel disconnected from their effects in the Strategic World.

The first five themes all describe forms of interplay between the Social, Functional, Strategic and Character Worlds, which could be productive or disruptive. The last two themes describe ways in which voice commands felt different to manual control, and how these differences influenced players' ability to engage with the strategic world of the game. We unpack some of the implications for game design in the Discussion section.

Theme 1: Thematically appropriate voice commands increase engagement with the Character World

Across nearly all sessions, participants said they felt more engaged with the Character World when they used verbal voice commands compared to when they used manual controls. They described the game as feeling more "immersive" or "real" with voice commands—more like a world and less like "just a game". P14 said that as someone with a tendency to anthropomorphise things, "I would just get lost in this."

This feeling was generally attributed to how closely voice commands fit the idea of what was happening in the Character World. In doing so, it evoked a strong sense of stepping into the role of the player-character: "It gave me the feeling that I'm commanding troops, it gave me the feeling that I am an actual commander." (P20) Most participants reported a sense of identification with a player-character persona when using voice commands in EndWar and ATCV, although not in The Howler. Participants pointed out the distinctive language that these two games used for their voice commands, which demarcated their voice commands as speech 'in character' due to its difference from their ordinary speech. Participants sought to amplify this distinction by speaking in the manner that they imagined their persona would speak:

> I was trying to imitate those people who work in towers. Their intonation, how they speak. (P20)

The games' responses to voice commands further reinforced the impression that its characters were perceiving the player as the player-character. For example, P5 commented that their units' verbal acknowledgements (such as "Okay, roger captain") felt more "natural" in response to voice commands than when they came in response to mouse clicks.

The immersive quality of voice commands was universally seen as a positive. It was by far the most commonly praised aspect of voice interaction, and considered the most interesting reason to use voice control. P17 said that voice commands turned the strategy game EndWar into a kind of role-playing game, which "just makes it more sort of authentic. Like you're playing the part." This immersiveness was contrasted with manual control, which "makes it feel more like a game" (P10). P24 said that with manual controls, "I don't feel like I'm the person doing it, "cause I'm just tapping around."

Part of the enjoyment of voice command was that it gave players a sense of personal authority. As P16 described it: "With the voice control, I felt like a commander or something. Like having that power." This did not necessarily mean that the player was better able to control the game; in most cases, the manual controls were described as more responsive, precise and reliable. But by providing a stronger sense of inhabiting the player-character persona, voice also provided a stronger sense of having the player-character's authority.

Theme 2: Vocalisations that sound odd in the Social World disrupt engagement with the Character World

All of the participants said they felt uncomfortable using voice interaction in at least one of the games. Participants consistently described this discomfort as coming from their own thoughts about what people who were not involved in the game would think about the sounds they were making. Nearly all of the comments about this discomfort referred to the idea of a hypothetical outside observer, from whose perspective the participant risked a loss of face and social status.

> This isn't a game where I would play it in front of people, in general. The other games, sure it's weird talking to a computer, but at least you can understand what the person's saying. It's not garbled. (P9)

Self-consciousness was far more pronounced with The Howler than the other games, due to its lack of voice commands. Participants experimented with verbal and non-verbal vocalisations to control The Howler, and generally found that non-verbal vocalisations (such as sustained vowel sounds) provided better control. This created a dilemma, as it meant the vocalisations that made the most sense in the functional world did not make sense and did not convey meaning in the social world, and *vice versa*. This drew players' attention away from the Strategic and Character Worlds, and towards the Functional and Social Worlds that were in conflict.

> I was more aware of my surroundings because I had this feeling of: I'm doing something that doesn't make sense if you weren't playing the game with me. I had, like, I was in the game but also like: "What do I sound like?" (P3)

The two games with voice commands did not create discomfort and self-awareness to the same extent as The Howler. Participants attributed this to an understanding that an outside observer could quickly understand what they were doing if they overheard voice commands, as opposed to non-verbal vocalisations: "If it were meaningful words coming out they would think, 'Oh, she's playing a game.' " (P24) However, this was dependent on the voice commands creating a sensible impression of a game scenario. At times when this broke down—such as when a participant was obliged to repeat themselves several times in a row because the game did not understand their speech—participants' preoccupation with the hypothetical outside observer returned, and they once again lost their sense of engagement with the Character World.

Part of the discomfort with non-verbal voice interaction was a fear of being seen as mentally incompetent. Participants described feeling "silly" or "stupid" due to the meaninglessness of their vocalisations, and some said that their own non-verbal vocalisations reminded them uncomfortably of young children or people with neurological disorders.

However, not all participants disliked the self-consciousness brought on by non-verbal voice interaction. Three participants rated The Howler as the game that they felt least comfortable playing but also the game they enjoyed the most. P14 commented: "I think that's kind of its charm, right? It kind of makes you feel a little bit uncomfortable." In a social situation, the fact that non-verbal voice interaction drew players' attentions towards the Social World could be a benefit.

Theme 3: Players struggle not to interpret vocalisations as meaningful communication

A common difficulty with non-verbal voice interaction was the difficulty of perceiving voice as non-speech. It was not only that third-party observers were expected to interpret non-verbal voice as nonsensical speech, as described in Theme 2—participants' perception of their *own* speech was also a source of trouble. Hearing their own voice made participants self-aware and self-conscious about what they appeared to be "saying".

Many participants exhibited more difficult deciding what to say to the game when using non-verbal voice interaction than when using speech commands. For these participants, the lack of constraints on their speech input did not free them from thinking about what to say, but instead caused them to think even harder about what utterances to use. A typical approach was to try out several different vocalisations, usually beginning with a repeated phrase (such as "go up, go up") or a stream-of-consciousness instruction ("okay now go up up over this bit yes good ok back down a little now"); if this did not provide enough control, participants moved on to a repeated phoneme ("upupupupup") or a sustained vowel sound ("uuuuuuuuh"); eventually settling on a vocalisation that provided a balance of control over the game while minimising incongruity in the social world.

As these examples indicate, participants often chose vocalisations that were derived from words that meaningfully represented their intentions. This shows that they were retaining an understanding of their vocalisations as *voice commands* rather than as merely *acoustic sounds* (or in terms of the SFSC model, as actions in the Strategic World rather than the Functional World). This confusion was shown clearly when participants needed to switch between making the balloon go up and making it go down. Instead of making noise when the balloon should rise and being silent when it should fall, participants would say "up up up!" when it should rise and "down down down!" when it should fall—which of course only made the balloon rise further.

> It was going too high, and I'm going, "Stop stop stop stop!" That's just making it fly higher. I would have to get my brain out of the mode for that. (P14)

What this illustrates is that participants had adopted a faulty mental model of how the non-verbal voice interaction worked, imagining (albeit unconsciously) that the balloon was responding to their speech content rather than their voice volume. This model was persuasive at first, because it resolved the tension between the Social World understanding of speech as meaningful communication and the Functional World effects of the game responding to their voice. But it achieved this by inaccurately perceiving the game as a voice command system, which was demonstrated when the function and verbal meaning of the player's voice input were no longer aligned.

The same dissonance was apparent at times between utterances in the Functional World and the Character World. Participants described voice interaction in The Howler as being "ridiculous" because it did not relate to anything that was happening in that game's Character World, whereas the Character Worlds of ATCV and EndWar made it "much easier to grasp what was going on" (P18) in relation to voice. This was not an inherent quality of non-verbal voice input, however, as some non-verbal vocalisations were suggested that could make sense:

> I feel like it would have felt better if it was better tied into the sort of narrative of the game. If you were trying to distract soldiers, or something, and you'd yell to get their attention. (P10)

The key distinction was that voice interaction should have a meaningful purpose that was discernable in the context of the game's Character World.

Theme 4: Players remember voice command phrases by meaning, not always by wording

Participants learned the voice commands of each game quickly, and in general had little difficulty recalling what voice commands were available. However, they often mixed up the phrasing of a command or used a wrong word, as in saying "call tower" instead of "contact tower". This was particularly common in ATCV, as it used air traffic control jargon that was unfamiliar to the participants and it lacked a visible menu of voice commands.

> If it were a real situation, I would probably get away with missing a few commands. Like the pilot would obviously still get what is "70 degrees to your right." (P23)

Even in EndWar, with its relatively familiar terminology and the on-screen menu of available commands, participants often accidentally substituted a synonym for a command phrase, such as "go to" instead of "move to". In some cases the game accommodated this by design, and responded as though it heard the correct word, but in other cases the command failed to register. The participant was not always aware that their phrasing was the source of the problem when this happened.

These participants were quite capable of remembering the commands by their meaning, and of coming up with words that represented that meaning, which would be sufficient for communication in the social world or the Character World. However, the intermediary of the Functional World was unable to recognise these commands, despite them having meaning-content that was the same as voice commands that it could recognise.

Theme 5: Speech intended for the Social World can infringe upon the Functional World

A frequent difficulty arose in voice interaction when the game picked up on and responded to utterances that were not intended as voice inputs. This happened most often with The Howler, because it was less constrained in the vocalisations it reacted to. When the balloon flew too close to the upper boundary of the screen, participants often reacted with laughter or a yelp of surprise—which prompted the balloon to accelerate upwards, causing them to fail the level. This also happened when a participant made an offhand comment to someone else in the room, or shouted a response cry [8] as described in Theme 3. In each of these cases, an utterance that was meant for the Social World infringed upon the Functional World due to the game's lack of ability to discriminate between the two, and the effects of this accident propagated into the Strategic World of the gameplay.

This is analogous to the "Midas touch problem" [20] in gaze input systems. The Midas touch problem refers to the inability of an eye-tracking system to distinguish between eye movements that are intentional or relevant to the user's goal and those that are random or unrelated to the user's goal. In this respect, non-verbal voice interaction is most similar to gaze input, in that it shares the problem that any vocalisation the player makes is treated as an input. Voice commands avoid this for the most part, since only certain words are treated as inputs, but may still suffer from accidental activations when the player says something that matches or is similar to one of these commands. The two voice command games in our study both minimise this risk by including a push-to-talk key, so that the game only listens for voice commands after the key is pressed. As a result, neither game is controllable solely by voice: they are multimodal at least to the extent that voice input has to be activated by a manual button-press.

In two instances, something similar to this "Midas touch problem for voice" was triggered intentionally. This happened when participants leaned over and shouted something at their co-participant's microphone, taking advantage of its inability to distinguish between the voice of each player. This action disrupted the normally linear relationship between a player, their interface and their avatar, as one player took temporary control of the other player's avatar through the 'wrong' interface, and deliberately rekeyed all three agents to a new frame of prankish stolen control, if only for a few moments.

Theme 6: Voice commands encourage concentration on the Strategic World

When comparing voice commands to manual controls, participants commented that the games felt more complex when playing with voice, and that they had to pay more conscious attention to the game state than they did when using manual controls. This was partly a matter of needing to think through the steps involved in voice commands—either because they were less familiar or because they were more complex—which resulted in more attention being paid to each action:

> You have to think what action you want to do. Then you have to think, how is that represented in the menu? And then you have to say it. So that's more complicated. (P4)

This did not appear to impede the participants' ability to focus on the strategic frame. Instead, focusing on the voice command options seemed to facilitate a greater sense of awareness of the game state and the strategic options that were available to them. Multiple participants made comments such as:

> You are more in the game when you use the voice, in the sense that you actually have to pay attention to the unit numbers, and what they're doing. Whereas when you play with the mouse, you sent them at some random target. So I had a better overview when I was using the voice. (P19)

It is unclear whether this sense of concentration reflected a real heightened awareness of the game state. It could be that participants had to invest more conscious thought to use voice commands and this mental work merely provided the feeling of being more engaged, without any actual increase in strategic awareness. For example, P17 described "thinking a lot harder" with voice commands but finding manual controls more "fluid" and intuitive. Nevertheless, the subjective experience for most participants was that voice commands facilitated their awareness of the Strategic World.

Theme 7: Voice commands feel disconnected from their effects in the Strategic World

Voice commands were associated with a sense of distance from the action in comparison to manual controls. Participants described feeling that they had "more direct control of what's happening" (P4) when using manual controls, whereas with voice commands there was a sense that "I wasn't really controlling the units" (P1). Although the units had the same level of autonomy in both conditions, their autonomy felt higher with voice commands:

> When you're playing with your hand, you have to control everything. But when you're telling the airplane, it seems it finds its way towards the runway. (P20)

The sense of disconnection from the action was partly attributable to the spatial and temporal distance between the voice input and the game's response. With manual controls, participant always saw the active unit because their mouse cursor or finger was on it when it was selected; but with voice commands the active unit could be off-screen or at an unknown location when it was selected. And unlike mouse clicks and taps, which were responded to instantaneously, voice commands could take around a second to generate a response. When there was a lot of activity on-screen, it could be difficult to tell which actions were precipitated by the voice commands and which were the result of the game AI. Participants reported needing to look at their units after giving a command to see whether any of them were responding.

> I didn't even check if the command was accepted. I just kept yelling out commands. And so I wouldn't know if it actually accepted some of it. (P19)

This uncertainty was greatest in EndWar, as its movable 3D camera perspective meant that the response to many commands happened off-screen.

5 DISCUSSION

When interviewed about their experience of playing a selection of voice interaction games, participants in our study spoke about multiple overlapping frames that shifted in and out of focus, reflecting the "oscillating nature of engrossment" with game frames described by [13]. We identified four distinct types of frames based on the interviews: social, functional, strategic and imaginary. These four frames reflect Conway and Trevillian's Social-Operative-Character model of the game event [9], with their Operative World frame divided into the two distinctive sub-frames of function and strategy. We have therefore proposed a revision of their model that incorporates four frames rather than three: Social World, Functional World, Strategic World and Character World, or SFSC model (Figure 2).

Our thematic analysis of the interviews generated seven ways in which these frames were evoked or disrupted for participants during voice interaction gameplay. We found that both the major concerns and the major sources of enjoyment that players described could be accounted for in terms of interactions between frames. In general, gameplay was most enjoyable when frames were well aligned (Themes 1, 4 and 6), and most disrupted when frames were perceived to be in tension (Themes 2, 3, 5 and 7).

This points to the largest non-technical problem for voice interaction games, which is the difficulty of establishing a comfortable alignment between a player's Social World and the game's Character World. Our study echoes Rico and Brewster's finding that "the imagined interpretations of others" [31] influences the perceived acceptability of voice interaction, and particularly non-verbal voice interaction. This is a challenge for design because it has less to do with how players engage with the game itself than it does with how players' actions in engaging with the game appear outside of the game, in the Social World—the frame in which the game designer has no control and little influence. It is also something of a catch-22: failing to align voice commands with the character identity creates a dissonance between the player and their character [7], but aligning voice commands with the character identity makes the player's speech more anomalous in the Social World. In ordinary gameplay, players ameliorate social tensions using their voice as a communication back-channel [12], such as by indicating role distance [15] between their primary identity as a social actor and the behaviour they are engaging in to play the game [8]. Voice interaction impedes this back-channel by requiring players to reserve their voice for the game.

However, participants alluded to a solution to this problem. The voice inputs that were considered the least socially inappropriate were those that an outside observer could immediately recognise as game inputs; participants preferred these over voice inputs that did not convey an apparent purpose. Designers can accommodate this by ensuring that voice commands unambiguously communicate their purpose as game inputs. One way to approach this would be to give the player-character persona in the game a distinctive speaking style, and ensure that voice commands align with this style. This would increase the tension between the Social World and the Character World frames, but in a way that reinforced the boundary between the two frames, to allay the player's apprehensions that their actions might be misinterpreted as something other than game-playing.

Barriers to engagement with the Functional, Strategic and Character Worlds

Fine [13] and Conway and Trevillian [9] describe a player's engagement with a game as a process of *upkeying* [16]. When I play a game, I am always first and foremost myself, a human being in the primary frame of the Social World; to enter into the other frames I must make some effort, and this effort must be supported by the other objects and people within the frame. Upkeying in the SFSC model is sequential and hierarchical: I must engage with the game functionally as a *user* before I can fully enter the role of a *strategist*, and I must engage as a *strategist* before I can fully enter the role of my *character*. (Note that I need not be highly calculating or deliberate to engage as a strategist, as long as I am attending in some intentional way to the systemic relationships between objects within the game.) To upkey successfully takes both my own participation and the co-operation of the game: if the controller does not respond to my button presses, my attempt to upkey myself to the Functional World as a user is foiled, and I am downkeyed out of the Functional World.

In these terms, a voice *interface* is a technology that seeks to upkey a Social World action (speech) into a Functional World action (voice input). A voice interaction *system* is one that seeks to upkey this Functional World action (voice input) into a Strategic World action (voice control). And a voice interaction *game* is one that seeks to upkey this Strategic World action (voice control) into a Character World action (an air traffic controller's instructions). This upkeying faces resistance both from the technology, which may fail to process my voice input correctly, and from other people in the Social World, who may distract me from my engrossment with the Character World that allows me to experience my speech as upkeyed.

Upkeying to...	Resisted when...
Character World	Voice inputs feel incongruous with the Character World. Voice inputs feel incongruous with the Social World.
Strategic World	Voice inputs feel disconnected from their effects. Player has a faulty mental model of the voice interaction system.
Functional World	Player has difficulty remembering command phrases. System picks up on unintended voice inputs.

Table 2: Tensions that interfere with a player's ability to upkey their engagement.

Here the SFSC model reveals its practical use as a model for thinking about game design, and particularly voice interaction game design. We have seen that tensions between and within particular frames prevented our participants from becoming fully engaged with different aspects of the game—in other words, prevented them from upkeying. By stepping back and looking at the SFSC model as a whole, we can map out barriers to upkeying at each level, as we show in table 2. That is not to imply that these tensions are always negative; a game that wishes to draw focus to the Social World rather than the Character World may use certain tensions to downkey gameplay intentionally. However, mapping frame tensions in this way provides an implicit priority chart for games that wish to evoke the Strategic World and Character World, as it highlights that lower-level tensions must be addressed before higher-level tensions can be effectively resolved.

Reduced sense of agency

Participants reported feeling as though units in the game had more autonomy when they used voice commands compared to when they used manual controls (Theme 7), even though the two control modalities activated the same actions. This reflected a general sense of reduced personal agency with voice commands. Based on participants' comments, we hypothesised that this sense of diminished agency was caused in part by the spatial and temporal distance of voice commands—by which we mean their ability to refer to units and locations anywhere in the gameworld—and their somewhat variable response delay due to the need for speech processing. This enlarged the gulf of evaluation [28] for voice commands compared to manual controls, and left some players uncertain of the outcome of their commands.

The sense of diminished agency may be a fundamental characteristic of voice inputs. One measure of perceived agency is "intentional binding", a phenomenon in which users perceive the time delay between their own action and its effect to be shorter than it really is. In an experimental study, Limerick et al. [21] found that voice commands created the inverse effect to intentional binding: users systematically overestimated the time delay between their voice input and the computer's response. They concluded that "Voice interfaces will feel less responsive and as a result users may experience a reduced sense of ownership or responsibility for the outcomes of their actions" [21]. Although our study contained no measurement of intentional binding, our participants' comments were consistent with Limerick et al.'s conclusions.

It is interesting to note that this diminished sense of agency was not reflected in a diminished sense of authority. Rather, the reverse was true. As we have reported, participants felt less in control *as a user* but more in control *as a character* when using voice. This speaks to the distinction between the Functional World and Social World frames in the player experience, and the need to consider them separately.

Difficulty remembering command phrasing

The observation that participants could remember command functions more easily than command phrases (Theme 4) points to an important qualitative difference in how speech operates in the Functional World compared to the Social, Strategic and Character Worlds. In the latter frames, speech operates on the human model of language production, in which meaning precedes phrasing, and there are many ways to represent the same meaning in different words. As one participant pointed out, a human listener would know that "turn right zero seven zero degrees" means essentially the same thing as "turn seventy degrees to your right", even if these have different connotations, because the phrases have an underlying meaning that is consistent. Participants recalled the underlying meaning of each command and, as in ordinary Social World speech, constructed a sentence to represent it.

In the voice interface, and therefore in the Functional World, the phrasing of a command precedes its meaning—or rather its function. The machine does not have a sense of the underlying meaning of the words, only an algorithm to recognise specific combinations of sounds. This is in contrast to the participants, who struggled not to perceive any utterance as communicating meaning (as described in Theme 3). To become proficient, the player has to let go of their intuitive understanding of voice commands as being *words that map to meanings* and learn to think of them instead as *phrases that* *map to functions*. In this sense, the player has to configure themselves to the machine [38], turning the familiar activity of speech into an unfamiliar new behaviour.

However, the Functional World is only one of the frames to which the player is attending, and the voice command itself has multiple identities across these different frames: at the same time as it is an operative phrase independent of meaning, it is also a character's speech, a statement of strategic intent, and an utterance with meaning to anyone who overhears it.

Limitations

Using contemporary speech recognition systems to study voice interaction introduces some necessary trade-offs. One that we confronted was whether to exclude from the study participants who could not make themselves reliably understood by one of the voice command games. As we are not seeking to establish comparable results across controlled study conditions, we decided to include these participants. Excluding them would introduce a particular bias to the results, minimising the importance of functional considerations and presenting an unrealistically positive picture of the state of speech recognition in recent videogames.

Although the study room was decorated as a domestic space, with sofas, bookshelves, potted plants and a television, it was still a room in a university building rather than a true home environment. The presence of a researcher and another player is also likely to have influenced the participants' experience. On top of this, the participants were playing all three games for the first time (with the exception of one participant who had played EndWar several years ago). We are confident that our findings represent concerns and issues that would appear in other settings, but we have no doubt that a more naturalistic observation of voice interaction gameplay, in a home environment and over longer-term usage, would reveal findings that we have not discovered here.

6 CONCLUSION

In this paper we have argued that frame analysis is a productive lens for understanding the player experience of voice interaction, a game modality that has proved challenging for designers and researchers but has lacked theoretical attention. We have provided a frame analysis account of observations and interviews with 24 participants who played three different voice interaction games, which included both verbal and non-verbal forms of voice control as well as manual controls. Through a thematic analysis informed by Conway and Trevillian's [9] frame-analytic model of the game event, we developed seven themes that characterised the experience of voice gameplay. We also proposed a revision of the SOC model to account for the difference between functional and strategic frames that was apparent in voice interaction gameplay.

Our frame analysis found that voice commands were associated with an increased sense of taking on a Character World persona, and particularly the persona of an authority figure. However, voice interactions that were incongruent with the Social World, such as non-verbal voice inputs, disrupted the player's engagement with the Character World. Participants intuitively processed and remembered utterances as meaningful communication, which sometimes added confusion to their understanding of voice inputs in terms of their operative function. Speech that was intended for the Social World caused unintended outcomes in the game system, particularly during non-verbal voice interaction. Finally, we observed that participants felt more engaged with the Strategic World when using voice commands, but at the same time felt that they had less direct control compared to when they used manual controls. Based on these findings, we have explored several design considerations for voice interaction games, and provided a summary of the tensions that impose barriers to upkeying voice interaction at each level.

ACKNOWLEDGMENTS

We acknowledge the Australian Commonwealth Government and the Microsoft Research Centre for Social NUI for their support on this project.

REFERENCES

[1] Fraser Allison, Marcus Carter, and Martin Gibbs. 2017. Word Play: A History of Voice Interaction in Digital Games. *Games and Culture* (2017). https://doi.org/10.1177/1555412017746305

[2] Fraser Allison, Marcus Carter, Martin Gibbs, and Wally Smith. 2018. Design Patterns for Voice Interaction in Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '18)*. ACM, New York, NY, USA.

[3] Glenn A. Bowen. 2006. Grounded Theory and Sensitizing Concepts. *International Journal of Qualitative Methods* 5, 3 (2006), 12–23. https://doi.org/10.1177/160940690600500304

[4] Sheryl Brahnam and Antonella De Angeli. 2012. Gender affordances of conversational agents. *Interacting with Computers* 24, 3 (2012), 139–153. https://doi.org/10.1016/j.intcom.2012.05.001

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[6] Gordon Calleja. 2007. Revising Immersion: A Conceptual Model for the Analysis of Digital Game Involvement. In *DiGRA '07 - Proceedings of the 2007 DiGRA International Conference: Situated Play*, Vol. 4. http://www.digra.org/wp-content/uploads/digital-library/07312.10496.pdf

[7] Marcus Carter, Fraser Allison, John Downs, and Martin Gibbs. 2015. Player Identity Dissonance and Voice Interaction in Games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '15)*. ACM, New York, NY, USA, 265–269. https://doi.org/10.1145/2793107.2793144

[8] Steven Conway. 2013. Argh!: An exploration of the response cries of digital game players. *Journal of Gaming & Virtual Worlds* 5, 2 (2013), 131–146. https://doi.org/10.1386/jgvw.5.2.131_1

[9] Steven Conway and Andrew Trevillian. 2015. "Blackout!'" Unpacking the Black Box of the Game Event. *Transactions of the Digital Games Research Association* 2, 1.

[10] Sebastian Deterding. 2018. Alibis for Adult Play: A Goffmanian Account of Escaping Embarrassment in Adult Play. *Games and Culture* 13, 3 (2018), 260–279. https://doi.org/10.1177/1555412017721086
[11] Brian R. Duffy and Karolina Zawieska. 2012. Suspension of disbelief in social robotics. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. 484–489. https://doi.org/10.1109/ROMAN.2012.6343798
[12] Starkey Duncan. 1972. Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology* 23, 2 (1972), 283–292. https://doi.org/10.1037/h0033031
[13] Gary Alan Fine. 1983. *Shared Fantasy: Role Playing Games as Social Worlds* (1 ed.). University of Chicago Press.
[14] Gordon Fletcher and Ben Light. 2011. Interpreting digital gaming practices: SingStar as a technology of work. http://usir.salford.ac.uk/17245/
[15] Erving Goffman. 1961. *Encounters: Two Studies in the Sociology of Interaction.* Bobbs-Merrill. OCLC: 710786.
[16] Erving Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience.* Harvard University Press, Cambridge, MA, USA.
[17] Perttu Hämäläinen, Teemu Mäki-Patola, Ville Pulkki, and Matti Airas. 2004. Musical computer games played by singing. In *Proc. 7th Int. Conf. on Digital Audio Effects (DAFx'04), Naples.*
[18] Mitchell Harrop, Martin Gibbs, and Marcus Carter. 2013. The Pretence Awareness Contexts and Oscillating Nature of Coaching Frames. In *Proceedings of the 2013 DiGRA International Conference: DeFragging Game Studies*, Vol. 7. http://www.digra.org/wp-content/uploads/digital-library/paper_45.pdf
[19] Takeo Igarashi and John F. Hughes. 2001. Voice As Sound: Using Non-verbal Voice Input for Interactive Control. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology (UIST '01).* ACM, New York, NY, USA, 155–156. https://doi.org/10.1145/502348.502372
[20] Robert J. K. Jacob. 1990. What You Look at is What You Get: Eye Movement-based Interaction Techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90).* ACM, New York, NY, USA, 11–18. https://doi.org/10.1145/97243.97246
[21] Hannah Limerick, James W. Moore, and David Coyle. 2015. Empirical Evidence for a Diminished Sense of Agency in Speech Interfaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15).* ACM, New York, NY, USA, 3967–3970. https://doi.org/10.1145/2702123.2702379
[22] Jonas Linderoth. 2012. The Effort of Being in a Fictional World: Upkeyings and Laminated Frames in MMORPGs. *Symbolic Interaction* 35, 4 (2012), 474–492. https://doi.org/10.1002/symb.39
[23] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16).* ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288
[24] Clifford Nass. 2004. Etiquette Equality: Exhibitions and Expectations of Computer Politeness. *Commun. ACM* 47, 4 (2004), 35–37. https://doi.org/10.1145/975817.975841
[25] Clifford Nass and Kwan Min Lee. 2000. Does Computer-generated Speech Manifest Personality? An Experimental Test of Similarity-attraction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00).* ACM, 329–336. https://doi.org/10.1145/332040.332452
[26] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. https://doi.org/10.1111/0022-4537.00153
[27] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices. *Journal of Applied Social Psychology* 27, 10 (1997), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x
[28] Donald A. Norman. 1986. Cognitive engineering. In *User Centred System Design*, Donald A. Norman and Stephen W. Draper (Eds.). L. Erlbaum Associates, Hillsdale, NJ, USA, 31–61.
[29] Daniel Pargman and Peter Jakobsson. 2008. Do you believe in magic? Computer games in everyday life. *European Journal of Cultural Studies* 11, 2 (2008), 225–244. https://doi.org/10.1177/1367549407088335
[30] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214
[31] Julie Rico and Stephen Brewster. 2010. Gesture and Voice Prototyping for Early Evaluations of Social Acceptability in Multimodal Interfaces. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI '10).* ACM, New York, NY, USA, Article 16, 9 pages. https://doi.org/10.1145/1891903.1891925
[32] Tara Shrimpton-Smith, Bieke Zaman, and David Geerts. 2008. Coupling the Users: The Benefits of Paired User Testing for iDTV. *International Journal of Human-Computer Interaction* 24, 2 (2008), 197–213. https://doi.org/10.1080/10447310701821558
[33] Adam J. Sporka, Sri H. Kurniawan, Murni Mahmud, and Pavel Slavík. 2006. Non-speech Input and Speech Recognition for Real-time Control of Computer Games. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility (Assets '06).* ACM, New York, NY, USA, 213–220. https://doi.org/10.1145/1168987.1169023
[34] Misha Sra, Xuhai Xu, and Pattie Maes. 2018. BreathVR: Leveraging Breathing As a Directly Controlled Interface for Virtual Reality Games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* ACM, New York, NY, USA, Article 340, 12 pages. https://doi.org/10.1145/3173574.3173914
[35] Jaakko Stenros. 2010. Playing the System: Using Frame Analysis to Understand Online Play. In *Proceedings of the International Academic Conference on the Future of Game Design and Technology (Futureplay '10).* ACM, 9–16. https://doi.org/10.1145/1920778.1920781
[36] Paul Tennent, Duncan Rowland, Joe Marshall, Stefan Rennick Egglestone, Alexander Harrison, Zachary Jaime, Brendan Walker, and Steve Benford. 2011. Breathalising Games: Understanding the Potential of Breath Control in Game Interfaces. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology (ACE '11).* ACM, New York, NY, USA, Article 58, 8 pages. https://doi.org/10.1145/2071423.2071496
[37] Daniel Wildman. 1995. Getting the Most from Paired-user Testing. *Interactions* 2, 3 (1995), 21–27. https://doi.org/10.1145/208666.208675
[38] Steve Woolgar. 1990. Configuring the user: the case of usability trials. *The Sociological Review* 38 (1990), 58–99. Issue S1. https://doi.org/10.1111/j.1467-954X.1990.tb03349.x

8 Discussion

This chapter presents a holistic discussion of the four studies in the thesis. It begins with a review of how each study addressed one or more of the research questions posed in Chapter 2. Section 8.2 presents the first of two major contributions for the thesis as a whole: a general model of gameplay, based on frame analysis (Goffman, 1974), that accounts for the social tensions of voice gameplay and provides a guiding framework for voice game design. Section 8.3 presents the second major contribution of the thesis: a typology of five distinct genres of voice gameplay that reflect different goals, different design patterns and different concepts of voice interaction, and consequently evoke different player experiences. The core ideas of these two contributions were initially introduced in Chapters 6 and 7, but the constraints of the *Thesis with Publication* format meant that those chapters were unable to discuss them in the context of the rest of the thesis. This chapter incorporates the findings of all four studies to develop these core ideas into more complete theoretical frameworks. Section 8.4 considers the implications that the thesis findings may have for the design of smart speakers, voice assistants and virtual and augmented reality experiences. This is followed by a discussion of the limitations of the thesis a reflection upon the value that historical analysis provided in placing the research findings into a proper context. Finally, the chapter notes some possibilities for future research informed by this thesis.

8.1 Research questions and findings

This thesis began with the identification of a rising class of activity: voice-operated interaction with characters and environments (VOICE). Prior research has shown that voice interaction holds an intuitive appeal to users as it strongly evokes the impression of conversing with an intelligent entity (Carter et al., 2015; Jia et al., 2013; Luger & Sellen, 2016; Nass & Brave, 2005; Purington et al., 2017). This means that there is a natural affinity between voice interaction and game design, as fields in which the illusion of virtual "life" inside the computer system is a central factor. However, there has been no body of academic literature to understand how, when and why voice interaction has been used in videogames, or to provide a theoretical understanding of its consequences for gameplay. With that in mind, this thesis addressed the overall research question:

> **Main RQ** How does voice-operated interaction with characters and environments (VOICE) influence the player experience of digital games?

To answer this question I conducted four studies in sequence, each of which addressed a specific research question that looked at one element of the topic. Each of these studies has been included in the thesis in the form of a published article. This section summarises how each study has answered one or more of the specific research questions, and concludes with a discussion of how all four studies together have addressed the main research question stated above.

Study 1: Natural Language Gameplay

The initial focus of the research was on conversational interaction in gameplay, rather than voice gameplay. Accordingly, Chapter 4 explores text-based interaction between a player and a game character with artificial intelligence (AI)—actually a Wizard of Oz prototype—in *Minecraft*. The aim of this study was to investigate how players engage with a game character using natural language, and the assumptions and expectations that influence their behaviour. This aim is reflected in the first research question of the thesis:

> **RQ1** How do players make sense of and interact with a game character that responds to natural language text?

This study found that conversational interaction was a source of enjoyment and interest for players even though it introduced a number of challenges, particularly in regards to concerns about ambiguity in instructions or questions leading to unintended consequences. Participants varied across two dimensions in their responses to the character: engagement, or the extent to which they showed interest in interacting with the character, and anthropomorphism, or the extent to which they treated the character as an intelligent entity capable of making judgements. The study found that players drew their expectations about the capabilities of a talking AI character from analogies to human social behaviour (both in the physical world and in online games), other videogames and depictions of AI in fiction.

As an overall theme of the results, participants found natural language dialogue to be a somewhat difficult and unwieldy way to engage with a game character, but still viewed it favourably due to its capacity to imbue the character with a greater sense of being alive and aware than it had when it did not use the text dialogue. Some participants expressed an interest in spoken interaction as a way to increase that sensation, which contributed to the shift in focus to voice interaction in the remainder of the thesis.

In Chapter 4, I had attempted to study natural language interaction in an open-ended way, by presenting players with a loosely designed “sandbox” game scenario that did not put many constraints on their behaviour and then observing what they did and said in response. However, the results made it clear that even this did not provide a clean slate for research, as the participants talked about how they were influenced by a range of factors, including their past experience with *Minecraft* and other games, the constraints of the interface, the scenario and task that they were presented with, and so on. For the subsequent research on voice gameplay, I determined that it would be more informative to begin with a thorough understanding of how voice interaction has been used in videogames, and to ground later studies in this reality rather than starting with a supposed "clean slate" for what the game design might look like. I reasoned that an approach based in real-world game design would be better able to pick up on factors that may be missed by a "clean slate" approach, and that may have been overlooked in prior research. The influence of the imaginary gameworld is one such factor that would turn out to be a critical determinant of the player experience.

Study 2: Voice Game History

Following Chapter 4, I decided that my approach to studying voice gameplay should be grounded in an understanding of real-world game design. However, as far as I could determine, no substantive sources existed that provided a picture of voice game design, beyond discussions of individual games. This was not only the case in the academic literature, but also in the broader writings and publications of game designers, critics and game hobbyists. Voice games had, in essence, not been identified and discussed as a unified category.

Chapter 5 addressed this research gap with a chronologically complete history of voice games. It covered the origins of the format and how it has developed over the course of videogame history. This account of voice game history was created by collating and analysing information about hundreds of published videogames with voice interaction features. The study contributes to a broader body of literature on videogame histories, taking what Suominen (2016) describes as a "genealogical" approach to the topic by developing an overview of the lineages and subcategories of voice interaction games over time. A goal of the study was to understand what might have prompted or constrained the appearance of voice games in the past, by observing patterns in where, when and in what format they had appeared. This is reflected in the second research question of the thesis:

> **RQ2** What is the historical lineage of voice game design and what are the major influences that have shaped it?

Chapter 5 identified seven distinct phases in the history of voice interaction games. Each phase consists of a time period in which voice games emerged in a distinctive way. These phases are chronologically overlapping, but distinguished by factors such as the geographic region in which voice games were produced, the game genres that were most represented, and the technological platforms that were used to afford voice gameplay. Based on this pattern of phases, I argue that the production of voice games has been highly reactive to the constraints of the hardware and software platforms on which videogames have primarily been played during each time period.

Such constraints were diminishing rapidly at the time the study was conducted, due to the increasing ubiquity of microphone-enabled gaming devices. The published article noted the emergence of an incipient boom in small-scale voice games produced by independent game developers, and predicted that this form of voice game design would continue to proliferate. This prediction was borne out in the subsequent study, detailed in Chapter 6, which found that a record number of voice games were published in the year that followed the prediction.

In comparing phases of voice interaction game design, Chapter 5 identified an apparently characteristic difference in the voice game design of Western and Japanese game developers. Voice games developed in Japan have been much more likely to afford two-way communication between the player and a character, in a manner that presents the two as relative equals in status; building a personal relationship with a game character has often been a goal of these games, whether overtly or implicitly. This trend is exemplified by "digital companion" or "virtual pet" games, such as *Seaman*, a game in which the player takes care of a talking fish in a virtual aquarium and answers its questions about their own personal life and beliefs. In contrast, games developed in North America and Europe have been much more likely to afford voice commands, often in military strategy scenarios. These games present the player as an authority figure issuing one-way instructions to characters who follow orders without question. And yet the single most popular form of voice games identified in the study—the karaoke videogame—reflected neither of these camps. In karaoke videogames, the focus is on performing for an audience of human onlookers rather than for virtual characters. The distinction between these styles of voice game design provided an early indication of several of the key findings of the thesis: that there are multiple fundamentally different genres of voice gameplay; that the impression of a living Character World can be a primary source of the appeal of voice games; and that the configuration of the relationship between player, character and onlookers is especially significant to the experience of voice gameplay.

Study 3: Voice Game Design Patterns

Chapter 5 contributed an overview of the lineages and underlying factors that have shaped voice interaction games as an overall format. However, there remained a gap in the literature with regards to understanding voice game design. While researchers had looked at specific novel applications of voice games, there had been no extensive investigations into the design of published voice games. Chapter 6, addressed this research gap through a comprehensive analysis of voice games and the creation of a pattern language for voice game design, to answer the third research question of the thesis:

> **RQ3** What design patterns are currently implemented in voice games?

The study method began with an extensive process of online research, which produced a database of 471 digital games, with a written description of the characteristics and voice affordances of each. This represented a comprehensive and systematic approach at compiling a complete census of all voice games that had been published in any form—including a number of voice games that were published in the form of a detailed description in an academic paper.[17] Through an iterative process of inductive coding, I developed 25 design patterns that describe the main variations in voice game design. These patterns range from direct descriptions of affordances, such as SHOUT TO SHOOT, to more abstract patterns describing the structure and framing of voice interaction, such as SPEAK AS A CHARACTER.

The primary contribution of Chapter 6 is the pattern language itself. It serves as a directory of game design solutions and a vocabulary for describing and conceptualising theoretical uses of voice interaction in games. However, analysis of the patterns revealed three further insights. The first insight is that there has been a distinct separation between games that are operated by speech (verbal voice input) and games that are operated by volume, pitch or other vocal characteristics (non-verbal voice input). The second insight is that certain design patterns have tended to appear in combination; the clustering of design patterns within certain games reveals five distinctive genres of voice game, described in Section 8.3. The third insight is that researchers have focused their attention on a narrow subset of voice game design patterns, which are among the least used patterns in commercial voice games. This indicates a separation between what academics have been studying and what players have been playing.

Study 4: Voice Gameplay

Chapter 6 provided an understanding of voice game design, and how voice games have differed in terms of their design. In particular, it noted a distinction between verbal and non-verbal voice gameplay. However, it left open the questions of how players experience these forms of voice gameplay, and how differences in design affect the player experience. A few prior studies had looked at voice gameplay experiences, including one that I conducted with co-authors in the early stages of my candidature (Carter et al., 2015). However, a review of the literature had found no studies that systematically compared verbal and non-verbal voice gameplay, or compared voice gameplay to non-voice gameplay in a non-prototype videogame. Consequently, Chapter 7 addressed these two research questions:

RQ4 What is the distinctive nature of voice gameplay?

RQ5 How do players experience verbal and non-verbal voice gameplay?

Adult participants were recruited in pairs to play a set of three videogames—alternately using voice input and manual input (touchscreen or keyboard and mouse)—and were then interviewed about the experience. I used thematic analysis (Braun & Clarke, 2006) to develop an account of the players' experiences,

[17] The complete database of voice games is included in Appendix B.

with frame analysis (Goffman, 1974) as an interpretive framework. Several of the themes reflect differences between verbal and non-verbal voice gameplay.

A main contribution of Chapter 7 was the SFSC Model[18] of gameplay, an iteration and empirical validation of a theoretical model developed by Conway and Trevillian (2015). The SFSC Model describes four frames in which actions have meaning as part of the gameplay event: Social World, Functional World, Strategic World and Character World. The results of the study confirm that voice interaction can facilitate a rich experience of engagement with a lively imagined gameworld, but also point to the possibility of tensions between frames that can disrupt that engagement. Altogether, the findings provide a diagnosis for several of the most common complaints about voice gameplay, and a basis for game designers to avoid them.

Main RQ: Understanding voice game design and gameplay

This thesis aims to address the following overarching research question:

> **Main RQ** How does voice-operated interaction with characters and environments (VOICE) influence the player experience of digital games?

Each of the results chapters explores an aspect of this question and reveals an additional layer of complexity. Chapter 4 demonstrates that natural language presents difficulties to players, both in their ability to express thoughts unambiguously and in their ability to differentiate the reality of artificial intelligence from the illusion. Chapter 5 shows that voice game design has been highly contingent on the technical constraints of popular videogame platforms as well as cultural conceptions about the types of relationships players have with virtual characters. Chapter 6 documents the breadth of voice game design and finds that there five fundamental types of voice gameplay, each with a different experiential goal. Finally, Chapter 7 finds that players do not take a singular meaning from voice gameplay, but instead continually evaluate the meaning of their experience through the lens of four different socially constructed frames of meaning simultaneously. Collectively, these studies show that the success of any one voice game is dependent on a complex interplay of factors. A narrow study of what makes voice game design successful in one context will not necessarily translate to a different place, time, game genre or social setting.

Of course, the need for a multi-faceted understanding is not unique to voice game design. Several scholars have argued for the necessity of studying videogames at multiple levels, due to the inherent complexities of the form (e.g. Aarseth, 2003; Björk, 2008). Konzack (2002) proposes a model of analysis that

[18] The acronym SFSC refers to the four frames of the SFSC Model: Social World, Functional World, Strategic World and Character World. This follows the naming convention of the original SOC Model (Conway & Trevillian, 2015), which refers to three frames: Social World, Operative World and Character World.

considers a videogame across seven distinct layers, from its underlying hardware and programming, to its functionality, gameplay mechanics and semiotic meanings, through to its intertextual referentiality and socio-cultural milieu. In Table 8.1, I summarise the high-level findings of this thesis in relation to Konzack's model, showing how it has addressed six out of these seven layers.

Table 8.1 How the thesis addresses Konzack's (2002) seven layers of analysis

	Natural Language Gameplay (Study 1, Chapter 4)	**Voice Game History** (Chapter 5, Study 2)	**Voice Game Design Patterns** (Chapter 6, Study 3)	**Voice Gameplay Experience** (Chapter 7, Study 4)
Hardware		Platforms have shaped and constrained the form and popularity of voice game design		
Program code				
Functionality		Speech recognition has improved steadily, but ambitions for speech-based interaction have always outstripped system capabilities	Games use both verbal and non-verbal voice input with varying degrees of complexity	Voice and manual input engender different responses from players. Likewise for verbal and non-verbal voice input
Gameplay	Various difficulties stem from the basic problem of specificity in natural language interaction	Popular styles of voice gameplay have differed across time and place	Games use voice input for dialogue, selection, navigation, control and performance actions	Voice gameplay exhibits tensions between four social frames, which can be productive or disruptive
Meaning	Players infer different kinds of intelligence from a game character that responds to language	Popular genres and settings for voice gameplay have differed across time and place	Games alter the meaning of voice actions through their fictional framing.	Responses to voice interaction are strongly influenced by the ways in which game fiction shapes its meaning
Referentiality	Players interpret voice gameplay with reference to their experiences of other games, media and human social behaviour			Players are acutely conscious of the need to act in a way that observers can recognise as a familiar frame
Socio-culture	Voice interaction is paradoxically idealised and distrusted compared to text-based interaction with technology	Voice gameplay has been presented in systematically different ways in Japan and Western countries		

8.2 A theoretical model of voice gameplay

The following sections address the main research question through a discussion of the two major theoretical contributions of this thesis. The first contribution is a general model of gameplay, entitled the SFSC Model, which provides a theoretical basis for understanding the dynamics of voice gameplay (Section 8.2). The second contribution is a typology of five genres of voice gameplay, each of which reflects a fundamentally different form of voice activity (Section 8.3). These contributions are discussed in light of all four studies in the thesis, and their implications for design and future research are presented.

In Chapter 7, I developed a general model of gameplay that iterated upon the prior theoretical model proposed by Conway and Trevillian (2015). As the model was presented in a published paper with a strict page limit, I could not incorporate the full context of the thesis research into the discussion of the model in that chapter. In this section, I will develop the model further in light of that additional context.

Section 8.2.1 begins with a review of the SFSC Model as it has been described so far, with its four frames sequenced by the order in which each one transforms the meaning of the player's action from the previous frame (Goffman, 1974). It then introduce a second dimension to the model: a continuum that runs from social explanations of activity to instrumental explanations of activity. This second dimension allows us to see that the instrumental meanings of voice gameplay have received the lion's share of prior research, whereas the social meanings have received relatively little attention. Section 8.2.2 discusses how designers and researchers can use the SFSC Model to predict player responses to voice gameplay, based on an analysis of a game's design prior to any user testing.

8.2.1 The SFSC Model

Chapter 7 described the development of a model of gameplay based on frame analysis (Goffman, 1974) and a prior model of gameplay as an event consisting of multiple frames of meaning (Conway & Trevillian, 2015). It established that players work to maintain four frames of meaningful activity during voice gameplay. Following the naming conventions set by Conway and Trevillian (2015), these frames were named the Social World, Functional World, Strategic World and Character World, and the overall model as the SFSC Model. Within each frame, the player takes on a distinct role: respectively the Social Actor, User Player, Strategist Player and Character Player. This section elaborates on the SFSC Model by reconsidering the dimensionality of its four frames, and discussing the implications of the model for research and design with reference to voice game design patterns.

One of the core findings of Chapter 7 is that the player engages with the four frames of the SFSC Model in an implicit sequence. The Social World is the highest priority frame, in the sense that the player is always attending to it to some degree, and a threat or disruption to the player's role in this frame (such as the intrusion of a judgemental bystander) is liable to disrupt their engagement with the other frames, by "pulling them out" of their roles in the Functional, Strategic and Character Worlds. The Functional World is the second priority, as a disruption in this frame is likely to disrupt the player's engagement with the Strategic and Character Worlds, but less likely to disrupt their engagement with the Social World. And so on, with the Strategic World coming next and the Character World last as a frame that does not show a tendency to disrupt any of the frames before it.

Table 8.2 Frames of the SFSC Model and corresponding player identity roles

Frame	Player identity role	Upkey	Downkey
Character World	Character player		
Strategic World	Strategist player		
Functional World	User player		
Social World	Social actor		

This priority-ordering of frames matches the phenomena that Fine (1983), Linderoth (2012) and Conway and Trevillian (2015) describe as "upkeying" and "downkeying". Goffman (1974) originally used these terms to describe accidents and slippages in social frames. Game studies scholars have adopted "upkeying" to describe how players work at transforming the meaning of activity in lower-level (primary) frames into higher-level (secondary) frames, aided by the game design, and "downkeying" to describe how players drop these transformed meanings in response to adverse circumstances. The Social World is the "primary framework" (Goffman, 1974, p. 21): the ordinary experience that we commonly regard as everyday reality. The Functional, Strategic and Character Worlds are secondary frameworks: transformations and re-transformations of the primary framework to have a new meaning, signalled by intelligible indicators and conventions that Goffman refers to as "keys" (Goffman, 1974, p. 40), which in a videogame context include aspects of the audiovisual feedback of the videogame system itself. Conway and Trevillian (2015) arrange their SOC Model as a vertical stack, with the Social World at the bottom and the Character World at the top, to represent both the directionality of "upkeying" and "downkeying" between frames and the fact that upkeying gameplay requires progressively greater efforts from players and game objects acting in concert. They explain a player's engagement with the frames of gameplay as an escalating process: "the more one upkeys, the more one commits to the role within the Game Event; the more one downkeys, the more one assumes disinterest in the focused encounter … for example withdrawing as

Character Player but remaining as Social Actor." (Conway & Trevillian, 2015, pp. 77–78) This escalation is reflected in the results of the empirical study of voice gameplay (Chapter 7), which show that players remain engaged with the primary frame of the Social World, but advance and withdraw their engagement with the secondary frames. Accordingly, the SFSC Model retains the vertical structure of the SOC Model, in which the frames are arranged in ascending order from the Social World to the Character World, as depicted in Table 8.2.

The order of the SFSC Model reflects an escalating engagement with the *diegesis* of the game. The term diegesis has been adapted from literary and film theory to describe "the game's total world of narrative action" (Galloway, 2006, p. 7), or what might be termed its gameworld imaginary. In the Character World frame, players orient themselves most completely to the diegetic content of the game, and act as though the gameworld is a living world in which characters have minds and emotions. In the Strategic World frame, players concern themselves with the gameworld in a more abstracted sense, treating its characters as pieces on a gameboard rather than as living entities. In the Functional World frame, players engage with the game interface as a non-diegetic entity, as an object outside of the "magic circle" (Salen & Zimmerman, 2004, p. 95) of the gameworld imaginary, which receives inputs directly from the player's physical body—another non-diegetic element of the game event. Finally, in the Social World frame, players are oriented towards the wider social world of their everyday experience. This sequence shows that the frames in the SFSC Model are placed in order of the extent to which they engage with the game's diegesis. Upkeying in the SFSC Model is a process of progressively orienting the action towards the diegetic gameworld, and away from an untransformed social context.

However, there is another logical way to order the frames of the SFSC Model: according to the type of *activity* each one is most concerned with. The activity dimension describes the orientation of each frame towards either social action or instrumental action. At the social end are actions that are intended to enact relationships between people (real or imagined), whereas at the instrumental end are actions that are intended to manipulate objects. When the four frames of the SFSC Model are sorted on this dimension, the Social World comes first as it is the frame most concerned with social relations between real people. The Character World comes next, as it also evokes social relations, albeit social relations whose meaning is bounded by the diegetic gameworld; between the player-character and other game characters (some of whom may be the avatars of other players). The Strategic World falls on the instrumental side of the activity dimension, as this is the frame in which the player considers game characters as playing pieces to be acted upon for the ludic purpose of achieving game objectives. Finally, the Functional World is almost purely instrumental, as for the purposes of this frame the player is concerned with neither the people around them nor the virtual lives of the gameworld inhabitants, but only with operating the interface. Comparing the four frames in terms of their orientation to social or instrumental activity reveals that this activity dimension forms an orthogonal axis to the diegesis dimension, revealing a different set of similarities and differences between frames than has previously been described in Chapter 7.

Table 8.3 Frames of the SFSC Model classified by activity type and diegesis

	Non-diegetic	Diegetic
Social activity	Social World	Character World
Instrumental activity	Functional World	Strategic World

Simplifying the diegesis and activity dimensions into two binary oppositions allows them to be represented on a two-by-two matrix (Table 8.3). This representation calls attention to two important dynamics in voice gameplay. Firstly, it illustrates the tension between approaching the game as an essentially instrumental system and approaching the game as a site of social meaning, which is the source of what my co-authors and I have termed “identity dissonance” (Carter et al., 2015) in voice gameplay: a phenomenon in which players reject the use of voice commands when those commands are expressed in terms of the Strategic World, and not aligned to the player’s identity role in the Character World. Identity dissonance indicates an excessive focus on the design of voice gameplay as an instrumental activity, and insufficient consideration of voice gameplay as a meaningful social activity.

Secondly, Table 8.3 calls attention to the commonality between the Social and Character Worlds as frames of social activity. The unidimensional SFSC Model (Chapter 7) puts these two frames at opposite ends of its sequence; however, when considered in terms of activity, the Social and Character Worlds are more similar to each other than they are to the Functional and Strategic Worlds. In these socially-oriented frames, speech is an act of communication whose power is its ability to change the mind of a thinking entity that hears and understands it, and possibly sway them into a new course of action. In the instrumentally-oriented frames, by contrast, speech is an act of control whose power is to directly act upon the game system, through direct triggers rather than through true comprehension of the words’ lexical meaning. The tension between these two different conceptions of speech was expressed in the findings of the qualitative interview study (Chapter 7), such as in the participants’ intuitive sense that game characters should understand voice inputs that had the right meaning even if they did not use the prescribed phrasing (as when a player said “call tower” instead of the correct voice command “contact tower”).

Table 8.3 also provides a clearer framework for understanding which aspects of voice gameplay have received the most study to date. As detailed in the literature review (Chapter 2), most HCI research on voice interaction concerns issues of instrumental activity: the usability of specific voice interfaces for particular ends

(Clark et al., 2019). Research on games reflects this, as the majority of prior literature has focused on the development and evaluation of novel voice input schemes or serious games for purposes such as language learning and speech therapy. In recent years, more research has been conducted on the social dimensions of voice interaction in general, concerning subjects such as the social acceptability of voice interaction in public settings and the influence of virtual character design on responses to voice assistants; but studies that extend this social perspective to voice games are scarce, as Chapter 2 revealed.

8.2.2 Design applications of the SFSC Model

The SFSC Model provides a theoretical basis for predicting what players will expect from voice interaction features in games. Whereas the voice game design patterns detailed in Chapter 6 suggest options for the high-level configuration of voice affordances, the SFSC Model provides an evaluative structure for considering how specific implementations of voice gameplay will be received, by examining the meaning of voice actions within each frame, and looking for tensions between those frames. Table 8.4 outlines the elements to consider for any given voice affordance.

Table 8.4 Elements of voice gameplay according to the SFSC Model

Frame	Who speaks	Who listens	What understanding is expected
Character World	Character player	Non-player character	Meaning of spoken words in context of the Character World
Strategic World	Strategist player	Game	Meaning of spoken words in context of the interface
Functional World	User player	Device	Matching sounds or words to system functions
Social World	Social actor	Onlooker	Meaning of spoken words in context of the Social World

This table can be used to evaluate voice game design at any level of abstraction, from an overall game to a single voice command (see Table 8.5). In any case, it is important to consider the design element in terms of all four levels of the model, independently and in combination, as the research in this thesis shows that it is the contradictions between frames that causes the greatest tension, rather than an inability of players to adapt to incongruity within any individual frame.

Table 8.5 Evaluating a verbal voice action with the SFSC Model

Action: Giving a voice command in *There Came an Echo*: "Syl, move to Alpha One"			
Design patterns: WHO-WHAT-WHERE COMMANDS, SELECT BY NAME, NAME AN ACTION, WAYPOINTS			
Frame	**Who speaks**	**Who listens**	**What it means**
Character World	"Sam"	"Syl"	Asking Syl to move to location Alpha One
Strategic World	Player as the player	Game as a strategic challenge	Move unit designated Syl to location designated A1
Functional World	Player as a user	Game as an interface	A series of preconfigured keywords
Social World	Player as self	Any onlookers	Player is talking to a game character

The model expresses the difficulty of voice game design, as each additional layer imposes a new constraint. It may be relatively easy to devise voice inputs that satisfy two frames at once, such as a vocalisation that sounds appropriate to onlookers (Social World) and is easy to express in a way that the game system can consistently recognise (Functional World). But it is not so easy to devise voice inputs that satisfy all four frames at once—sound appropriate for the social context, function in a way that minimises cognitive load, express the player's strategic intentions and match the player's in-game persona—while also supporting interesting gameplay.

However, that is not to say that the four frames must be in harmony for a game design to be successful. The history study in Chapter 5 and the design patterns study in Chapter 6 show that there are a number of approaches to voice game design that use tensions between frames productively to create enjoyable player experiences. In terms of the voice game genres introduced in Chapter 6, the Voice Command and Voice Conversation forms of gameplay are generally based upon aligning all four frames to give the player a seamless experience of involvement in a well-developed Character World. This creates the kind of enjoyable player experience that Hunicke et al. (2004) call the fantasy aesthetic, and Ermi and Mäyrä (2005) call imaginative immersion. In contrast, Voice Performance games are generally designed to engage the player with the Social World meaning of their actions, and often create deliberate tensions between frames to achieve this. For example, the VOLUME CONTROL design pattern—used in games such as *Yasuhati* and *The Howler*—requires the player to shout erratically to succeed at the game's challenges; a behavioural stricture for the Functional, Strategic and Character World frames that is completely incongruous with normal behavioural strictures of the Social World frame (see Table 8.6). In the presence of onlookers, this creates a shared humorous spectacle that accords with what Hunicke et al. (2004) call the "fellowship aesthetic" and Lazzaro (2009) calls "people fun". The next section describes these voice game genres in greater detail, including the ways in which each genre tends to employ the four levels of gameplay described by the SFSC Model.

Table 8.6 Evaluating a non-verbal voice action with the SFSC Model

Action: Yelling "aaah!" to control a hot air balloon in *The Howler*			
Design patterns: VOLUME CONTROL			
Frame	**Who speaks**	**Who listens**	**What it means**
Character World	No clear speaker	A hot air balloon	No coherent meaning
Strategic World	Player as the player	Game as platform puzzle	Louder equals higher
Functional World	Player as a user	Game as interface	Louder equals higher
Social World	Player as self	Onlooker	Unclear meaning

8.3 Five genres of voice gameplay

The previous section approached VOICE in games as an overall phenomenon, and described a single theoretical model to understand it in all of its forms. This section discusses the differences between those various forms. It begin with a review of the typology introduced in Chapter 6, which comprises five genres of voice gameplay. Whereas the initial development of that typology drew entirely upon the results of the design patterns study, this section synthesises a more detailed account of each genre from the findings of all four studies in the thesis, as well as prior literature.

The process of identifying voice game genres involved three main steps, as reported in Chapter 6. The first step was an inductive analysis of 471 published voice games, through which I developed a vocabulary of 25 unique patterns for voice game design. The second step was a re-examination of the voice game database to find sets of patterns that have consistently appeared together in the same game, following Björk et al.'s (2003, p. 190) proposal that game genres may be defined by the set of design patterns that they consistently employ. The third step was a comparison of the voice games that had been studied in academic research against the voice games that were commercially published, as well as a comparison of games with verbal voice input against games with non-verbal voice input. From these steps, I identified five distinct clusters of voice games with their corresponding design patterns. The first two clusters (Voice Command and Voice Conversation) consisted of games that were designed around giving commands to or engaging in conversation with the game characters. A third cluster (Voice Practice) consisted of games that were designed to guide or motivate player to practice specific vocal actions; this corresponded closely with both "serious games" (see Abt, 1970) and games that were developed for academic research. The fourth cluster (Voice Performance) consisted of games that were based around social spectacle and the idea of the player putting on some kind of performance for an audience; the majority of non-verbal voice games fit into this cluster. Finally, there were a small number of left-over games and two design patterns that fit none of the above patterns: SPELLCASTING and BREATH PHYSICS, which are distinctive as patterns in which the player affects the virtual environment directly with their voice, rather than communicating with characters within it. Consequently, the fifth and smallest cluster of voice gameplay (Voice Incantation) was defined according to this distinction.

Each one of these clusters represents a voice game genre, as indicated in parentheses in the previous paragraph. The genre names reflect the overall type of vocal activity upon which its gameplay is based, following Apperley's proposal that scholarly definitions of game genres should reflect the "nonrepresentational, specifically interactive, characteristics of video games" (2006, p. 7)., I have named each genre of voice game for the type of vocal activity that it involves: Voice Practice, Voice Performance, Voice Command, Voice Conversation and Voice Incantation. Gameplay in Voice Practice and Voice Performance focuses primarily on the Social World frame. Gameplay in Voice Command, Voice Conversation and Voice Incantation focuses primarily on the Character World frame; the genres are distinguished by the way in which the player's voice engages the characters and objects in the Character World. The design patterns that constitute each genre are shown in Table 8.7.

Defining genres is an inherently imprecise exercise, but it is an instructive one in this case, because it shows how different voice games are simulating different forms of activity, and have different basic conceptions of how the player's voice should affect both the gameworld and the everyday social-world context. A designer or researcher must understand these underlying aims before they can properly evaluate an existing voice game or determine how to improve upon it. The differences make it plain that there is no single ideal for what voice games should be. It is possible for a game to belong to multiple genres—for instance, the strategy game *There Came an Echo* employs both Voice Command and Voice Conversation forms of gameplay—but each one places a different demands upon the player's actions, with regards to the four frames of gameplay described in the SFSC Model, and these demands can be directly contradictory.

So far, the thesis has described the five genres of voice game only in terms of their design patterns, based upon the research in Chapter 6. The sections that follow explain each genre in greater detail, drawing upon the historical understanding developed in Chapter 5, the player experience insights developed in Chapter 7 and the SFSC Model of voice gameplay developed in Chapters 7 and 8. Each genre is given a title, a brief summary, one or more illustrative examples, a breakdown of the design patterns that underpin it, an evaluation of its dynamics within the context of the SFSC Model, a discussion of its gameplay implications and challenges, and historical notes where relevant.

Table 8.7 Genres of voice gameplay and associated design patterns

Voice gameplay genre	**Associated design patterns**
Voice Practice	SPEAK AS THE PLAYER PRONUNCIATION PITCH CONTROL
Voice Performance	SPEAK AS THE PLAYER or SPEAK AS A CHARACTER KARAOKE VOLUME CONTROL SHOUT TO SHOOT
Voice Command	SPEAK AS A CHARACTER SITUATED COMMANDER FLOATING COMMANDER WHO-WHAT-WHERE COMMANDS SELECT BY NAME SELECT BY TYPE NAME AN ACTION NAME A BEHAVIOUR WAYPOINTS RELATIVE DIRECTIONS ABSOLUTE DIRECTIONS
Voice Conversation	SPEAK AS A CHARACTER CHOICE OF OPTIONS QUESTION AND ANSWER SCRIPTED CONVERSATION UNSCRIPTED CONVERSATION
Voice Incantation	SPEAK AS A CHARACTER SPELLCASTING BREATH PHYSICS OVERHEARD NOISE

8.3.1 Voice Practice

A substantial number of voice games are serious games (see Abt, 1970), designed to support vocal exercises for a serious purpose such as language learning or speech rehabilitation. In these games, the gameplay provides a structure and a feedback loop to guide the player through an activity, such as pronouncing a word correctly (Lan et al., 2014; Navarro-Newball et al., 2014; Tan et al., 2013) or sustaining a vowel sound at a sufficient volume (Lopes et al., 2016). This form of gameplay is named Voice Practice, to indicate that its purpose is the learning and repetition of the vocal act itself, rather than any kind of enjoyment or meaningful experience the player may gain from playing. An enjoyable player experience may be an important part of the design, but usually for the purpose of sustaining the player's engagement with the activity, rather than for the enjoyment as an end in itself. In this respect, Voice Practice reflects what Carter et al. (2014) describe as the operative paradigm of games research. Voice games that exemplify Voice Practice include *Flappy Voice* (Lan et al., 2014), a game controlled by vocal loudness, designed to facilitate treatments for childhood apraxia of speech; *sPeAK-MAN* (Tan et al., 2013), an adaptation of *Pac-Man* in which the player can defeat enemies by speaking words chosen by a speech therapist; and the *Tactical Language and Culture Training System* (W. L. Johnson & Valente, 2009), a serious game developed for the United States Marine Corps to practice speaking and interacting with people in foreign countries.

Design patterns commonly used for Voice Practice are PRONUNCIATION and PITCH CONTROL. In comparison to other types of voice gameplay, the player is more likely to SPEAK AS THE PLAYER rather than SPEAK AS A CHARACTER, reflecting the fact that the focus is on the player's own self-development.

In terms of the SFSC Model, the gameplay is typically focused on activities in the Functional World frame, as the correct operation of the game's functions is the ultimate goal of the game. While many of these games include virtual characters and environments to evoke a Character World frame and gameplay challenges to evoke a Strategic World frame, these elements tend to be thinly developed when compared to other types of voice games.

Since many Voice Practice games are the product of academic research, there is a uniquely large amount of literature on this type of game. A number of prior studies have found specific Voice Practice games to be functional, fit for purpose and generally enjoyable (see Section 2.4). However, as these studies have focused on evaluating the usability and effectiveness of their specific voice game, the literature contains little information about the player experience of Voice Practice gameplay beyond these high-level indications.

8.3.2 Voice Performance

The second genre of voice gameplay again makes the player's vocalisations the focus of the experience, but this time for entertainment rather than for any serious operative purpose. Voice Performance creates a spectacle out of voice interaction, by encouraging players to produce vocal performances that are ridiculous, ostentatious or amusing in some other way. This kind of gameplay draws much of its appeal from spectatorship and the reactions of an audience (Fletcher & Light, 2011), so it is nearly always found in games that are explicitly designed for group play. Voice games that exemplify Voice Performance include the karaoke game *SingStar*, the shout-controlled arcade game *PowerFall* (Vieira et al., 2014) and the voice-acting simulation game *Yoake no Mariko*.

Design patterns commonly used are KARAOKE, VOLUME CONTROL and SHOUT TO SHOOT. This type of gameplay is closely associated with non-verbal voice input: a high proportion of games that use non-verbal voice input are designed for this kind of gameplay, and nearly all games designed for Voice Performance use non-verbal voice input exclusively.

In terms of the SFSC Model, the activity in Voice Performance is heavily focused on the Social World, due to its grounding in spectatorship and social play. The Character, Strategic and Functional Worlds tend to be simple and accessible, to avoid pulling the player's attention too deeply into the details of what is happening on the screen. Take for example the game *Guitar Hero World Tour*: although it gives the player a virtual persona and narrative as a member of a rock band climbing the ladder of fame, it provides little beyond this surface to add depth to the Character World; and while its singing-based gameplay can be challenging to perform, it is conceptually simple (match the pitch indicated on screen) and so requires little attention to the Strategic or Functional World beyond the immediate act that the player is performing. As a result, the experience of gameplay is rooted "in the room" rather than "in the gameworld".

One version of Voice Performance was included in the player experience study (Chapter 7), in the form of *The Howler*, a game with completely non-verbal voice interaction. To be successful in this game, players found that they had to make unusual sounds that did not resemble normal speech. The clearest finding of the study was that this non-verbal voice gameplay made the Social World frame much more salient to players compared to verbal voice gameplay. For some players, this was a source of enjoyment, as they found humour or a transgressive thrill in the act of making themselves "uncomfortable" in front of other people. Other players did not enjoy *The Howler* in the context of the study; however, many acknowledged that it could be more enjoyable in a different context, such as with friends at a private party. For the other games in the study, based on Voice Command, players were more relaxed about the social context in which they were played. This indicates the particular importance of the Social World frame to the success or failure of Voice Performance gameplay.

8.3.3 Voice Command

The third type of gameplay, Voice Command, focuses on the experience of giving verbal instructions and seeing them carried out by characters (or occasionally objects) in the gameworld. In most cases, the player is given an in-game persona as some kind of authority figure who instructs obedient non-player characters, such as an army general, spaceship captain or football team manager. Games that exemplify Voice Command include the sci-fi tactics game *There Came an Echo*, the basketball game *NBA2K16* and the spaceship shooter game *Voice Commander*. In the history of voice game design (Chapter 5), Voice Command gameplay has been more common in the West compared to Japan.

This type of gameplay commonly uses a combination of overlapping design patterns focused on voice commands. The player is positioned as either a SITUATED COMMANDER or FLOATING COMMANDER, and their voice inputs include the ability to NAME AN ACTION or NAME A BEHAVIOUR. In more complex examples, the voice inputs may be structured as WHO-WHAT-WHERE COMMANDS, allow the player to call upon a specific character or unit through SELECT BY NAME or SELECT BY TYPE, and direct them around the gameworld with WAYPOINTS, RELATIVE DIRECTIONS or ABSOLUTE DIRECTIONS.

The extensively combinatorial structure of these design patterns reflects the fact that Voice Command is focused on the mechanical aspects of gameplay, or in terms of the SFSC Model, the Strategic World frame. Voice inputs are meaningful first and foremost as strategic instructions, which must be used in the right combinations at the right time to direct the gameplay successfully. Characters within this type of game are usually defined primarily by the strategic function that they provide—for example, football teammates whose attributes are their ability to sprint, pass and tackle. These abilities are often expressed in a persona that makes the abilities easier to remember, such as a famous real-world footballer. In this way, the meaning of objects in the Character World is made to overlap with their meaning in the Strategic World, so that each frame supports engagement with the other.

The player experience study (Chapter 7) suggests that the Functional World is primarily a source of difficulty in Voice Command. Games with a focus on this type of gameplay typically define their voice commands in terminology that allows strategic intent to be expressed in a way that suits the Character World, as in the air traffic control jargon in *ATC Voice*. The ideal player experience this points towards is one in which the player could adopt the mindset of their character, state their instructions and see the game respond appropriately with little need to think about the Functional World. Considerations such as remembering which exact keyword to use in which exact combination are a distraction from the core intended gameplay experience, rather than part of it. This is expressed in one of the themes of the player experience study: “Players remember voice command phrases by meaning, not always by wording”. That is, the intention (Strategic World) of the command is more readily brought to mind than the performance (Functional World) of the command.

The player experience study also reveals a difference between the subjective experience of Voice Command and controlling units with manual controls. Even when voice commands and manual commands are functionally the same, players feel less in control of the action with voice commands and have more of a sense that the characters under their command are exercising independent judgement, as though those units have more autonomous AI (see also Limerick et al., 2015). At the same time, players feel a greater sense of personal authority and status, and some players perceive themselves to be more focused on the Strategic World of game actions. This seems to reflect an intuition about the kind of control that is possible with one's voice versus one's hands: the former exercised through influence over other thinking, autonomous humans, and the latter exercised through direct manipulation of objects that may not have a mind of their own. A character obeying a verbal instruction implies that it has understood, and whereas a character obeying a manual instruction does not.

8.3.4 Voice Conversation

The fourth type of gameplay simulates the experience of holding a conversation with a living character. Instead of using voice as a control method to make the character carry out an action, Voice Conversation uses voice as a communication method to allow the player and the character to establish mutual understanding and rapport. The player's Strategic World goal may be to convince the character to do something, but this act of convincing treats the character as a thinking entity with a mind of its own to influence, as opposed to a piece on a game-board that will automatically follow the player's commands. This type of gameplay emphasises the experience of being listened to rather than simply being obeyed. Voice games that exemplify Voice Conversation include the virtual pet games *Seaman* and *Hey You, Pikachu!* Many games alternate between Voice Conversation and Voice Command, such as the stealth survival game *Plan Be* and the role-playing action game *Mass Effect 3*. In the history of voice game design (Chapter 5), Voice Conversation gameplay has been more common in Japan compared to the West.

This type of gameplay is often structured as a QUESTION AND ANSWER exchange. The player is usually given a CHOICE OF OPTIONS as part of a SCRIPTED CONVERSATION, although some games allow for UNSCRIPTED CONVERSATION. As the focus is on the Character World, the player will normally SPEAK AS A CHARACTER rather than as themselves, which differentiates this kind of gameplay from the experience of voice interaction with non-game conversational agents. The dialogue often takes place in a turn-based sequence in which any other real-time gameplay aspects are suspended; for example, *Mass Effect 3* alternates dialogue scenes with action scenes, to ensure that the player can focus on the dialogue while it is occurring. This serves to simplify the Strategic World during conversation to reduce its salience and keep the focus on the Character World.

The player experience of Voice Conversation was not studied directly in the frame analysis study (Chapter 7). However, that study found that voice commands generate an increased sense that non-player characters are alive and intelligent, and the text-based dialogue study (Chapter 4) found that players perceive a similarly high degree of life and intelligence in characters that converse through text, even if their responses are very basic. In addition, the sources analysed for the research on voice game history (Chapter 5) and design patterns (Chapter 6) included player accounts of emotionally moving Voice Conversation gameplay (e.g. Keogh, 2012). Collectively, this suggests that Voice Conversation has the potential to facilitate a high level of engagement and emotional involvement with the Character World, possibly beyond the level of text-based dialogue.

8.3.5 Voice Incantation

The fifth type of gameplay is one of the least common in voice games to date, but one that may be a promising area for further game design exploration based on the findings of this thesis. That is Voice Incantation: the use of the player's voice as a force that affects the virtual environment of the gameworld. In this type of gameplay, the player is not communicating information to a character, but creating an effect directly through the power of their voice. It has been used most often in games with a fantasy setting, such as *Skyrim: Legendary Edition*, *In Verbis Virtus*, *Chant Savant* and *The Broken Seal*, and occasionally in games with a science fiction setting, such as *Mass Effect 3*.

SPELLCASTING is the archetypal design pattern for Voice Incantation, as it most clearly reflects the metaphor in which the voice itself has power. BREATH PHYSICS is also a form of this type of gameplay, in that the player's voice (or breath) exerts an effect on the game environment through the physical movement of air. OVERHEARD NOISE is an edge case; in that pattern, the player's voice creates a noise within the gameworld that characters can notice and react to. That is arguably an example of Voice Incantation rather than Voice Command or Voice Conversation, because the effect of the player's voice is to generate noise in the environment rather than to communicate with a character *per se*. However, individual cases of OVERHEARD NOISE may fit more readily into one of the other genres.

There is no published research on the player experience of Voice Incantation games, as far I am aware. However, it appears to be a promising area for voice game design because it can avoid or resolve the frame tensions that this thesis has identified as disruptive to the player experience. These tensions arise due to conflicts between the meaning of an act of voice interaction in different frames; for example, a voice command means one thing in the Character World (an act of meaningful communication between conscious people), but something entirely different in the Functional World (an act of performing a designated sound to produce a direct systemic response). Voice Incantation gameplay and the SPELLCASTING design pattern resolve this tension: in both the Character World and the Functional World, voice input is understood as an act of performing a specific word of power to activate a direct response in the gameworld. The potential meanings of these words of power are arbitrarily limited to the schema of the game's spellcasting vocabulary, and therefore do not suggest Strategic World variables beyond those prescribed by the Character World (whereas ordinary language suggests additional possibilities, as noted in Theme 4 of Chapter 7). And, as magical incantations can be arbitrarily defined, it allows more leeway for a game designer to ameliorate the difficult conflict between the Character and Social Worlds, in which voice commands that sound too unusual for the player's identity in the Social World are disruptive to engagement with the Character World, but voice commands that are not suited to the player's role in the gameworld are less effective at increasing engagement with the Character World. Accordingly, Voice Incantation is a type of voice gameplay that merits further study. Encouragingly, its most notable appearances in published games have all come within the past few years, showing that designers have recently taken an interest in exploring the potential of this genre of voice game.

8.4 Implications beyond videogames

So far, this chapter has described a model of voice gameplay and a set of voice game genres that relate to conventional screen-based videogames. Only a small amount of the data on which this thesis is based was drawn from other sources, such as smart speaker games and virtual reality games (see Section 8.6 for a further discussion on this point). However, the conclusions of this research do have wider implications. Studying videogames is useful not only for its own sake, but also because videogames are a novelty-driven subdomain of technology in which interaction modalities are popularised sometimes years before they become common in non-game interfaces.[19] This section discusses the implications of the thesis for three neighbouring areas of design: smart speaker games, voice assistants and virtual and augmented reality applications.

8.4.1 Implications for smart speaker game design

As of 2019, smart speaker platforms such as Alexa and Google Home are among the most high-profile voice user interfaces (VUIs). These systems typically consist of a physical speaker equipped with far-field microphones, not necessarily connected to a screen, operating a VUI that processes voice inputs through cloud-

[19] See for example the Nintendo DS, a handheld game console that preceded the iPhone by three years as a popular mass-market consumer device with a touchscreen interface.

based speech recognition, and responds in the voice and persona of a virtual assistant. Research on domestic use shows that they are typically placed in communal family areas of a house such as the kitchen or living room, and most commonly used for functions such as playing music and initiating web searches (Adobe, 2019; Bentley et al., 2018). An ethnomethodological study of Amazon Echo users (Porcheron et al., 2018) found the smart speaker was often used in a dynamic multi-person social context, in which members of the household would intersperse comments to each other and to the smart speaker, often collaborating or competing to issue commands to the system.

The SFSC Model suggests a number of challenges for game design in this context. The lack of a visual display is an impediment to establishing depth and complexity in both the Character World frame, due to the inability to visually depict an inhabited gameworld that responds to the player's actions, and the Strategic World frame, as the complexity of the situation that can be presented is effectively limited to what a voice can describe within the timeframe of human working memory and attention. The social context of use further constrains the Character and Strategic World frames, both because the player's attention is split between the game and the other people present, and because a heightened salience of the Social World will tend to decrease players' willingness to commit themselves as heavily to a Character World role (as described in Chapter 7 and previously in this chapter). The social context of use also represents the potential upside for smart speaker game design: removing the screen and enabling input from anybody present, without the need for a controller, minimises the elements that would distract players and onlookers from each other in a traditional videogame. Therefore, game design that supports interactions between players in the Social World frame with minimal intrusion of Strategic and Character World elements is the most straightforwardly promising avenue for smart speaker gameplay.

The majority of game design patterns identified in Chapter 6 are not easily replicated in an audio-only game. The patterns SPEAK AS A CHARACTER, SITUATED COMMANDER and FLOATING COMMANDER require the player to have a sense of an identity within the Character World, but this would seem to be more difficult to establish in a game that cannot depict the player's avatar or their viewpoint on the gameworld. Three patterns describe voice inputs that refer to visible text or objects: SCRIPTED CONVERSATION, WAYPOINTS, RELATIVE DIRECTIONS and KARAOKE. While not impossible to implement in an audio-only game, these patterns do not seem well suited to the format. Several other design patterns imply voice inputs that produce an immediate, direct effect on the gameworld: SHOUT TO SHOOT, VOLUME CONTROL, PITCH CONTROL, BREATH PHYSICS and SPELLCASTING. Again, while it would be possible to describe the outcomes of these player actions audibly, this does not seem well suited to the format. The design pattern OVERHEARD NOISE is predicated on the player staying silent, or selectively using their voice to attract attention during stealth gameplay that otherwise requires them to remain unnoticed; restricting the player's voice in this way would be impractical for a platform in which voice is the main input modality.

The two patterns that appear best suited to the audio-only format are QUESTION AND ANSWER and UNSCRIPTED CONVERSATION. QUESTION AND ANSWER provides a clear turn-taking structure to guide the dialogue in the absence of visual feedback, and UNSCRIPTED CONVERSATION removes the need for visual feedback that tells the player what they can say. However, where unscripted inputs are not practical, CHOICE OF OPTIONS is one clear way to provide guidance to the player on what they can say, although it is limited by the fact that the options would need to be either standardised or listed by the voice interface at each new decision point. WHO-WHAT-WHERE COMMANDS is a practical design patter for voice inputs, but as mentioned above, the complexity of the outcomes is constrained to what the voice interface can describe in the timeframe before the player begins to forget what has been said.

This evaluation is broadly supported by the analysis of 22 Amazon Alexa games described in Chapter 6. In accordance with the above, QUESTION AND ANSWER was one of the most commonly used patterns, and WHO-WHAT-WHERE COMMANDS was used in a small number of cases, while patterns of more direct action (such as VOLUME CONTROL), patterns that imply a response to visual objects (such as KARAOKE) and patterns that are grounded in the player-character position (such as SITUATED COMMANDER) do not appear despite being common in screen-based videogames. CHOICE OF OPTIONS was the most commonly used pattern of all, providing direct instructions to players about how to interact with the game, at the cost of slowing gameplay down by repeatedly requiring players to listen to a verbal list of options. I hypothesise that players may prefer the inverse pattern, UNSCRIPTED CONVERSATION, but that this does not appear in more games due to the difficulty of supporting unpredictable player voice inputs.

While a comprehensive review of smart speaker games has not been conducted, it is clear that one of the most common genres of game on these platforms is the trivia game. This fits the requirements described above, of a form of gameplay that supports engagement players in the Social World frame through minimally intrusive gameplay elements that do not seek to establish a complex Strategic World or Character World. It is notable, however, that this form of gameplay does not strongly satisfy any of the five types of voice gameplay identified in Section 8.3. The thesis as a whole suggests that the current generation of smart speakers is highly limited in its ability to deliver a lasting and enjoyable experience of voice gameplay, although the combination of a smart speaker platform with some form of visual display would greatly alleviate these limitations.

8.4.2 Implications for voice assistant design

The beginning of this thesis noted that videogames represent an instructive domain for non-game voice user interfaces (VUI) because of the central role that virtual characters play in most VUI. As prior research attests—and Chapters 4 and 7 support—conversational interaction and especially voice interaction are associated with anthropomorphism: the tendency for users to respond to devices that "speak" as though they are human beings (Cowan et al., 2015; Nass & Brave, 2005). This effect is a double-edged sword for the user experience, as the sensation of speaking to an intelligent character can increase initial user engagement with VUI, but often leads to disappointment over the longer term as the user's inflated expectations are not met (Cowan et al., 2017; Luger & Sellen, 2016).

The SFSC Model suggests an alternative way of looking at this issue: common VUI such as Siri and Alexa create the impression of a virtual character, but not the impression of a virtual world in which they are situated. Consider Cortana, the voice assistant whose persona is based on a character from the *Halo* videogame series. The videogame version of Cortana exists in the Character World of the *Halo* games, and has an identity that is meaningfully grounded in that context. Cortana must demonstrate a façade of intelligence and awareness of her Character World to appear to be coherent with it, but only in certain highly constrained ways. There is a limited set of objects and events that a player would expect Cortana to react to: scripted dialogue from other characters, moments when the player accomplishes an objective, perhaps the actions of in-game enemies. The illusion of a gameworld as a "real" place does not depend on the player being able to interact with every character or object within it in all the ways that they could if it was physically real; on the contrary, the strict limitations on how the player can interact with characters and objects helps to maintain the coherence of the gameworld as a symbolic text (Allison, 2016). However, when the character of Cortana is taken out of the Character World of the game and converted into a voice assistant in a smartphone, her context is entirely changed. Now she is an actor in the Social World of ordinary human interaction—partly because she is not represented on the device screen as having a virtual environment around her, and partly because her role is to communicate about facts, objects and events that exist in the everyday world. The baseline of expectations for actors in the everyday world is infinitely higher than the expectations of actors in a gameworld. Even non-human animals show a degree of situational awareness and mental acuity that is utterly beyond the reach of a system such as Cortana. Thus, even though the voice assistant version of Cortana uses far more sophisticated AI than the videogame version of Cortana, the latter maintains its impression of intelligence far more consistently than the former.

It would be instructive to test whether the user experience of voice assistants and other non-game conversational agent characters could be improved by positioning them within a clearly delineated Character World. Presenting Siri as an actor inside a virtual world who speaks to your player-character, rather than an actor in the everyday world that speaks to you directly, may better signify its nature as a character whose abilities are only symbolically representative of actual intelligence. Initial research to explore this hypothesis could include a comparison between user responses to non-game voice assistants, such as Siri, and voice assistants in videogames, such as ANNA (Automated Natural Navigation Assistant) in the open-world racing game *Forza Horizon 3*.

Feminist scholars have pointed to a second and more insidious problem with the anthropomorphisation of conversational agents. The conversational agents in the most popular VUI are presented as deferential female "assistants", either exclusively or by default. This has caused some to question whether common VUI design practices are taking advantage of or contributing to sexist stereotypes about the role of women in society as submissive, conciliatory, undemanding and "other" (Brahnam & Weaver, 2015; Guzman, 2016; Woods, 2018). That issue is not directly apparent in this thesis—since videogames have disproportionately represented male characters (Williams et al., 2009), the majority of game characters to whom a player may speak are male.[20] However, the thesis does show that one of the main sources of enjoyment in voice interaction is the thrill of exercising authority over subservient characters. This is directly apparent in the player experience research (Chapter 7) and indirectly apparent in the way that Western voice game design, in particular, has often depicted fantasies of command over others in militaristic scenarios (Chapters 5 and 6). I do not consider it problematic for players to enjoy stepping into the role of an authority figure in the context of a videogame. However, the fact that this is a source of enjoyment does lend credence to the suggestion that users of draw some of their enjoyment of non-game VUI from the experience of wielding power over a subservient character. Whether that is likely to influence the attitudes or behaviour of people towards other humans over the long term is unknown, and beyond the scope of this thesis. However, this thesis provides one indicator to suggest the question of power dynamics between users and conversational agents merits further scrutiny.

[20] That is not to say that there is no trace of sexism in voice game design. Take for example *N.U.D.E.@ Natural Ultimate Digital Experiment*, a virtual companion game in which the player uses voice input to teach a robot to perform basic tasks and recognise the names of objects. While the gameplay itself is arguably neutral, the combination of the title and the fact that the robot is visually depicted as a young woman clearly implies that the marketing for the game attempted to generate interest through sexualisation of the main character. However, overtly sexist examples such as this represent only a small fraction of the larger set of voice games, in which female characters were notable more for their scarcity than for their alignment with sexist tropes.

8.4.3 Implications for virtual and augmented reality

Voice input has been recognised as a potentially useful interaction modality for virtual and augmented reality (VAR) interfaces. This is due in part to the difficulty of using physical controls while wearing a VAR headset. Microsoft's HoloLens augmented reality (AR) system, Facebook's Oculus Rift virtual reality (VR) system and Samsung's Gear VR system all accept voice input. Each of these systems' voice affordances are presented as a menu of actions that the player can activate by saying the associated keyword or phrase, in some cases only when that phrase appears on the screen.[21] In the language of voice game design patterns (Chapter 6), these systems are employing a combination of CHOICE OF OPTIONS, SELECT BY NAME and NAME AN ACTION as their core model of voice interaction.

All of the voice game design patterns identified in Chapter 6 are potentially applicable to VAR systems. To take one example, WAYPOINTS is an effective solution to the problem of how to indicate space in a virtual environment; this design problem is just as relevant to VAR as it is to videogames, and VAR imposes greater restrictions on solutions such as amplifying the movement of the viewpoint (Wilson et al., 2018) or using a mouse as a physical pointer. Other voice game design patterns, such as OVERHEARD NOISE and BREATH PHYSICS, are effective ways to give the player a sense of physical presence in the virtual world; the sense of presence is the core concern of VAR systems, even more so than it is for videogames. Several voice game design patterns have begun to appear in VAR scenarios. Sra et al. (2018) have designed a series of VR games controlled by BREATH PHYSICS; *The Broken Seal* employs SPELLCASTING in VR to create an immersive experience of Voice Incantation; and *Star Trek: Bridge Crew* allows the player to command a Starfleet ship in VR using WHO-WHAT-WHERE COMMANDS.

Indeed, all five types of voice gameplay identified in Section 8.3 are also manifestly applicable to VAR game design. This is clear in the case of Voice Practice, Voice Command, Voice Conversation and Voice Incantation, as each of these types of gameplay is primarily concerned with the interaction between the player and the game system. Voice Performance may be a less obvious candidate due to its focus on social interaction between people as social actors in the Social World frame, which may appear to be impeded by a VAR headset. However, some of the most popular existing VR games put the lie to this, by making voice communication between players "inside" and "outside" the virtual reality the basis of game mechanics. See for instance *Keep Talking and Nobody Explodes*, a game in which a player wearing a VR headset must defuse a bomb in the virtual world by following instructions given to them by other players who can see a bomb defusal manual but not the bomb itself. While this is not a case of VOICE (voice-operated interaction with characters and environments) gameplay, because the players' voices do not operate or interact with the game elements

[21] Microsoft refers to this as the "see it, say it" model (Zeller et al., 2019).

directly, it shows the potential for voice-driven gameplay in VAR even with a focus on Social World performance.

With that having been said, the SFSC Model suggests a reason why voice interaction may be better suited to AR than VR. Naïvely, one might assume the opposite: that since voice interaction tends to improve the player's sense of involvement in the Character World unless disrupted by embarrassment in the Social World, and VR increases the player's sense of presence in a Character World while reducing their awareness of the Social World, voice interaction would be a perfect fit for VR. However, this overlooks the fact that the Social World is the primary framework (Goffman, 1974, p. 21) of the game event, and the other frames of the SFSC Model are transformations (Goffman, 1974, p. 41) of that primary framework. That means that the player does not lose their sense of participation in or disengage their awareness of the Social World when they involve themselves in the Character World. VR is likely to increase the user's concerns about the Social World, because they cannot monitor their physical environment for the presence or reactions of onlookers. In contrast, AR allows the player to engage in Character World activities while maintaining a simultaneous awareness of their Social World context, which should enable them to actively work at reducing the tensions between these frames.

8.5 Limitations

The limitations of the individual studies have been discussed in Chapter 3 (Methodology) and in the Results chapters. However, there are three limitations of the research project as a whole that are important to consider here: the lack of a consistent disciplinary basis across the four studies; the reliance on English-language sources and English-speaking players; and the use of pre-existing published games rather than custom-made research prototypes. This section addresses the extent to which these factors constrain the conclusions that can be drawn from this research.

Each of the four studies in the thesis has employed a research method rooted in a different disciplinary background: media archaeology from media studies; pattern languages from architecture, via design research and game studies; frame analysis from sociology; and lab-based user studies and Wizard of Oz prototyping from HCI. The rationale for this methodologically pluralist approach has been outlined in Chapter 3, and the benefits that it brought to the research have been discussed earlier in this chapter. However, this approach also foregoes some of the benefits of consistency. Using one method repeatedly would allow the later studies to iterate more directly upon the earlier ones. It would allow me as the researcher to improve my own technical proficiency with that singular approach, such that the procedural quality of the final study could naturally be expected to exceed that of the first. Accordingly, one might suppose that this research has foregone a level of depth and sophistication that could have been achieved in any one of its specific lines of inquiry. I have judged the trade-off to be worthwhile, as the pluralist approach has enabled me to develop a more comprehensive account of voice games, in which key observations are triangulated from multiple perspectives. However, I hope to see future research projects take a more sustained approach to the topic, and to delve more deeply into the specific complexities of voice game design that this thesis has uncovered.

In contrast to its methodological pluralism, the thesis is limited by a lack of diversity in the language of its subjects and sources. The player studies in Chapters 4 and 7 were conducted exclusively in English, although not all participants were native English speakers. The literature that it draws upon is nearly all in English and comes primarily from English-speaking countries. For Chapters 5 and 6, I accessed some sources in other languages through an online machine translation system, and I strove to catalogue every non-English-language voice game that I could identify as part of my analysis. Nevertheless, the research remains heavily rooted in the Anglophone perspective on voice interaction. This is a particularly serious consideration for the study of voice interaction, which may have a material impact on the findings. There could be inherent characteristics of different languages that change the experience of voice gameplay. It is possible that the differences between Japanese and Western game development described in Chapter 5 have been influenced by differences in the native languages (given that English was the primary language for nearly all of the Western games). It is also possible that the tendency to perceive intelligence in talking virtual characters is influenced by the language in which they speak—or the shared culture of Anglophone countries could influence these factors, rather than the language itself. With these factors in mind, the results of this research should be generalised to other languages with caution.

In addition to being English speakers, the participants in the player studies were largely drawn from what Heinrich et al. (2010) have dubbed the WEIRD societies: Western, educated, industrialised, rich and democratic. While this is true of a great deal of academic literature, research has shown that people from WEIRD societies are often not representative of humanity in general (Henrich et al., 2010). As a result, the findings of this thesis may not be entirely representative of non-English-speaking people from outside the WEIRD societies.

A final consideration is that all of the studies were conducted on or with pre-existing videogames, rather than games developed specifically for the research. Even the natural language study, which presented players with a constructed gameplay scenario, used the game *Minecraft* as the basis for this construct. This was a deliberate choice, as Chapter 3 explains: since prior research had indicated that the diegetic gameworld was an important factor in the experience of voice gameplay, I concluded that it was important for this thesis to look at games that had highly developed diegetic worlds, which would have been impractical to achieve in custom-built research prototypes. This decision resulted in rich insights into the player's reaction to the gameworld, reflected throughout the Results and Discussion sections. However, it gave me less control over the research stimuli than I would have had if I designed and created prototype voice games expressly for the research. I have been cautious about making overly definitive comparisons between the voice games in this thesis, as any such comparison must acknowledge that the games vary in more ways than ones I have described. For example, the three games used in the voice-based gameplay study (Chapter 7) differed in terms of their visual style, music, story and mission structure. A player study that held these and other variables constant between conditions while altering only one or two variables of interest would allow for more direct and definitive statements to be made about the effect of such specific variables.

8.6 Reflection on the value of historical analysis in HCI research

Many doctoral theses in HCI include some consideration and acknowledgement of the history of the topic. Often this is done primarily through the literature review. Relatively few, however, engage in a full historical analysis of the subject as a central part of the research process, as this thesis has done in Chapter 5. This section, I reflect upon my reasons for engaging in such an extensive historical study, and the way in which it has provided a solid contextualising foundation for the remainder of the research.

The initial motivation for studying voice gameplay was to understand a difficult area of design that had received relatively little academic scrutiny (as outlined in Chapter 2). I had already begun the research with an exploratory lab study on natural language interaction (Chapter 4), and considered extending this line of inquiry with further lab-based testing. However, something that had become apparent to me in the process of conducting that first study was that the problem I was pursuing was not yet clearly defined. There were plenty of indications that voice-based and language-based interaction were challenging in a videogame setting, but attempting to replicate the problem in a study in a way that was representative of real-world design required a series of decisions—about context of use, about technical platforms, and especially about the design of gameplay itself—for which no clear guide existed. It was also not yet clear which problems were most significant. My colleagues and I had previously found that "identity dissonance" (Carter et al., 2015) was an issue in voice gameplay, but that study had focused on only four videogames from a narrow time period on a single gaming platform, making it an uncertain guide to the larger design space.

The design history study (Chapter 5) was undertaken to ensure that the research would begin from a thorough understanding of voice games as they exist in the world. It revealed the specific ways in which voice games have been successful—

more successful than they were frequently portrayed in popular narratives—and the ways in which they have failed to live up to aspirations. By conducting this study, I immersed myself in voice game design, scrutinising every voice game that I could identify. The knowledge I gained in this process informed the remainder of the thesis in several productive ways.

First, the historical study provided insights into the importance of platform constraints. From this, I was able to see that the most important technical constraints on voice game design had already been effectively addressed, with the proliferation of microphone-equipped platforms on the one hand and the increasing availability of effective speech recognition systems on the other. The basic technological components of voice game development were already readily available to game developers to deploy. Consequently, I oriented the remainder of the research away from technical considerations and towards more experiential aspects of voice game design.

Secondly, it equipped me as a researcher with an encyclopaedic knowledge of voice game design that I took forward into the later studies. The historical study gave me a much better appreciation for the breadth of voice game design and the significance of its varieties than I had beforehand. For example, it showed that although speech-based forms of voice interaction have captured much of our imagination and attention, non-verbal forms of voice interaction have had a greater cultural impact in practice. Recognising the existence of significant subtypes of voice interaction provided the motivation for the design pattern analysis that followed (Chapter 6). In fact, the historical analysis effectively served as the first iterative loop of data collection and analysis for the design pattern study. It also informed the design of the player experience study (Chapter 7), as it led me to prioritise a research approach that could compare different forms of voice and non-voice interaction, rather than studying voice gameplay in a less controlled and more naturalistic context.

The third insight was that, despite the temporal trends in voice game design, there were certain consistent ideal forms of voice gameplay that designers had been striving towards from the earliest years to the present. The core underlying ideas about voice gameplay seemed to differ not so much between time periods or between individual designers as between geographic regions. This implied that our aspirations for voice gameplay might have less to do with the specific abilities of the technology (which has changed substantially over time) than it does with the cultural context in which the technology has arisen. Narratives about talking automata, which date back to ancient mythology, were supercharged in the twentieth century through science fiction stories such as *2001: A Space Odyssey* and *Star Trek*. For the research, this meant that it would not be enough to study whether people were successful or satisfied with the experience of playing a voice game; it was necessary to understand the interpretive framework that players used to imagine what was going on in the interaction. This informed the use of frame analysis in Chapter 7, with its central question of “what is it that is going on here?”

Finally, the historical perspective put recent developments in voice interaction in a long-term context. This has turned out to be important, because this thesis was researched and written over a period of substantial change for voice technologies. When I began to research the topic in 2015, the Microsoft Kinect was the most prominent platform in voice gameplay, having recently prompted the inclusion of voice elements in an unprecedented range of big-budget, high-profile console videogames. These voice elements had received mixed reactions (Carter et al., 2015), reflecting the inherent difficulty of voice game design, but the Kinect remained a widespread gaming platform with a player base in the tens of millions, having been dubbed the “fastest selling consumer device” (‘Microsoft Kinect “fastest-selling device on record”’, 2011) in a Guinness World Record a few years earlier. By contrast, smart speakers were a comparatively niche product with uncertain prospects: Amazon had only recently launched its Alexa platform with the Amazon Echo, and Google Home was more than a year away from launch.

As I write this in 2019, the Kinect has been officially retired from the Xbox product line for three years. No other dedicated videogame console has yet replaced it with similar voice interaction features. In this period, voice game design has fallen into a lull on dedicated videogame consoles, following the platform-driven ebb and flow that Chapter 5 identified as a recurring trend. Meanwhile, however, smart speakers have grown to become a large international market, with the Amazon Alexa alone selling more than 100 million units to date (Adobe, 2019). This has led to a proliferation of voice games on smart speakers, and a renewed interest in voice interaction from HCI researchers (Clark et al., 2019). At the same time, there has been a sharp increase in the number of small-budget voice games created by independent game developers for smart speakers, mobile phones and personal computers (documented in Chapters 5 and 6). Accordingly, the smart speaker appears to have currency as the voice interface of the moment, while phones and personal computers maintain their status as the workhorses of voice interaction in everyday life. Naturally, academic and popular interest has turned towards these platforms, and they dominate recent HCI literature on voice interaction technology while older platforms such as the Kinect recede into obscurity.

This raises a question: would this thesis have been more useful to future designers if it had focused on the newer platforms, such as smart speakers, rather than older platforms and older videogames? The historical study shows why this is not the case. It reveals that voice technology platforms have been inconstant, with no single platform remaining prominent for more than a few years. Each platform shaped the form that voice game design took for the duration of its popularity, but the longer view shows that voice game design has always returned to a stable set of genres. These voice game genres are ultimately oriented not towards the technology they happen to use, but towards the more stable lodestar of the player's experience. The purpose of voice gameplay—and perhaps voice interaction in other contexts as well—is to provide meaningful experiences that satisfy the desires and perspectives of the human beings who use it. Technological platforms are attempts to fulfil these experiences; to focus on the technology of the moment is to get this the wrong way around. When we seek to understand voice game design from the perspective of human experience, there is more to learn from technologies whose heyday has passed than from technologies that have not yet settled into a clear place in our homes and cultures.

8.7 Future research directions

This thesis lays a theoretical foundation for future research on voice-operated interaction with characters and environments (VOICE). This section will briefly outline some promising avenues for follow-up studies in the context of virtual and augmented reality (VAR) experiences and smart speaker applications.

8.7.1 Future research in virtual and augmented reality

Virtual and augmented reality (VAR) is one of the most promising use contexts for VOICE, due to the perceived immersiveness of voice interaction and the practical difficulties of using physical control devices when the user's view is obscured. Although this thesis has focused on VOICE in the context of traditional videogames, with only a few VAR games reviewed in Study 2 and Study 3, most of the design patterns and player experience findings appear to be transferrable to VAR experiences. Many of the voice game design patterns solve for practical functions that could apply to any spatial virtual environment—such as how to issue a movement instruction for an agent in a dynamic three-dimensional space. This applies not only to ludic VAR experiences, but to other forms of VAR as well. Accordingly, this thesis opens up a range of possibilities for extending its research to VAR. Each of the major findings would be interesting to explore in a VAR context, from considering how the SFSC Model of voice gameplay is altered when players are in virtual reality, to evaluating usability issues for each of the major voice gameplay genres in VAR.

8.7.2 Future research in smart speakers

This thesis has briefly explored VOICE in the context of smart speakers, in Chapter 6. The smart speaker games in that dataset used a subset of the design patterns found in screen-based voice games, appearing to be more constrained in the variety of design patterns that can be deployed on these platforms due to the lack of visual and manual interface elements. Extending this analysis with a full-scale study of smart speaker game design patterns would provide a clearer indication of the avenues that game developers are exploring in this novel platform, and highlight promising opportunities to adapt design patterns from screen-based games to the audio-only format. In particular, the act of navigating a gameworld appears to present a greater challenge for smart speaker games due to the lack of a visual display—possibly reflected in the finding that smart speaker games commonly afford movement through the design pattern of ABSOLUTE DIRECTIONS, which is rarely used in screen-based videogames.

Chapters 7 and 8 presented several frameworks for understanding the design and player experience of different forms of voice gameplay, but none of these has so far been evaluated for smart speakers. A study of smart speaker gameplay using frame analysis would provide an opportunity to examine whether the player experience has similar dynamics to those reported in Chapter 7. I would hypothesise that players systematically experience a greater focus on the Social World frame during smart speaker gameplay than during screen-based gameplay, as players are typically more inclined to orient their attention towards each other when there is no screen depicting a gameworld for them to orient towards instead. This hypothesis is consistent with prior research that has found smart speaker usage to be highly integrated into the fabric of social life in the home (Porcheron et al., 2018). Research along these lines will help to determine whether smart speaker games represent a substantial opportunity for game design with immersive Character World elements, as seen in voice games such as *Hey You, Pikachu!* and *There Came an Echo*, or are likely to remain limited to more Social World-based forms of gameplay, such as quizzes and chatbots.

8.8 Conclusion

The novelist William Gibson once remarked that "we can't see our culture very well, because we see with it." (Gibson & Northam, 2015) Once we have incorporated an assumption, practice or belief into our thought processes, we find it hard to imagine how we saw the world without it. Consider, then, how tremendously difficult it must be to separate ourselves from our intuitive understanding of speech. We cannot remember a time before we had begun to learn to speak. Speech plays a role in nearly everything we have ever done. It is a precondition to the very notion of culture. It is the basis of the language that we use to constitute our conscious thoughts. To paraphrase Gibson: we cannot hear our speech very well, because we hear with it.

Voice user interfaces make this most utterly familiar of activities unfamiliar. Because we each have such a deep-rooted and nuanced understanding of what voice is, even small voice cues evoke rich impressions of virtual life, which makes voice interaction a powerful experience. But that same deep-rooted and nuanced understanding makes our expectations for voice interaction immensely difficult to satisfy—and makes it hard for us to conceive of voice as something other than the kind of speech that we have been learning since birth. As speech recognition systems improve, the challenge for voice interaction design becomes less a matter of technical engineering and more a matter of human psychology.

Central to this challenge is the question of how far a voice-operated system should try to leverage existing human notions about voice and speech, and how far it should try to alter them. This thesis has explored both sides of that decision, through an investigation of voice interaction in games. It has documented games that present voice interaction with all of its associated trappings: a conscious interlocutor, a relationship, a spatial context, meaningful communication and mutual understanding—none of which are truly present, but each of which can be represented. Conversely, it has documented games that break the familiar conventions, making voice interaction playful by providing an alien experience of what is ordinarily deeply familiar. The thesis unites these forms under the banner of *voice-operated interaction with characters and environments* (VOICE), a term broad enough to encompass all forms of vocal activity while simultaneously emphasising the importance of the imaginative elements (characters and environments) to voice gameplay.

The first major contribution of this thesis is a theoretical model of player experience that explains the major experiential dynamics of voice gameplay—both positive and negative. This model, based in frame analysis, describes the four frames of player experience as the product of two dimensions: diegetic/non-diegetic and social/instrumental activity. The model provides a theoretical basis to distinguish between instrumental elements of voice gameplay, which have received extensive study, and social elements of voice gameplay, which have received comparatively little attention beyond the works included in this thesis.

The second major contribution is an account of the five core genres of voice gameplay. Each genre represents a distinct type of voice activity. In this chapter, I have detailed each one in terms of its underlying action metaphor, its constituent design patterns and its known design considerations with reference to the SFSC Model.

The findings of this thesis are not the last word on voice gameplay or voice game design. Rather, they are a starting point. They point to the inherent complexity of voice interaction, and the myriad of ways in which subtle changes in presentation and framing can drastically alter the way that it is experienced. As voice interfaces become more common, and complex voice-operated interactions with characters and environments proliferate on more devices, the need for contextualised research on VOICE experiences will only increase. This thesis plants a flag in a territory that has been a hinterland of game design and game studies, but where unique experiences lie waiting to be developed.

9 References

Aarseth, E. (2001). Computer Game Studies, Year One. *Game Studies*, *1*(1).

Aarseth, E. (2003). Playing Research: Methodological approaches to game analysis. *Proceedings of the Digital Arts and Culture Conference*, 28–29.

Abt, C. C. (1970). *Serious Games*. Viking Press.

Adobe. (2019). *State of Voice Assistants 2019* (Adobe Digital Insights 2019). https://www.slideshare.net/adobe/state-of-voice-assistants-2019

Al Hashimi, S. (2007). Preferences and Patterns of Paralinguistic Voice Input to Interactive Media. *Proceedings of the 12th International Conference on Human-Computer Interaction: Intelligent Multimodal Interaction Environments*, 3–12. http://dl.acm.org/citation.cfm?id=1769590.1769592

Alexander, C., Ishikawa, S., Silverstein, M., Jacobson, M., Fiksdahl-King, I., & Angel, S. (1977). *A Pattern Language: Towns, Buildings, Construction*. Oxford University Press.

Allison, F. (2016). Are You Out of Your Mind? Focalization in Digital Games. *Transactions of the Digital Games Research Association*, *2*(3), 31–60.

Allison, F., Carter, M., & Gibbs, M. (2017). Word Play: A History of Voice Interaction in Digital Games. *Games and Culture*, 1555412017746305. https://doi.org/10.1177/1555412017746305

Allison, F., Carter, M., Gibbs, M., & Smith, W. (2018). Design Patterns for Voice Interaction in Games. *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, 5–17. https://doi.org/10.1145/3242671.3242712

Allison, F., Luger, E., & Hofmann, K. (2018). How Players Speak to an Intelligent Game Character Using Natural Language Messages. *Transactions of the Digital Games Research Association*, *4*(2), 1–50. https://doi.org/10.26503/todigra.v4i2.88

Allison, F., Newn, J., Smith, W., Carter, M., & Gibbs, M. (2019). Frame Analysis of Voice Interaction Gameplay. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 393:1–393:14. https://doi.org/10.1145/3290605.3300623

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, *35*(4).

Anderson, J. N., Davidson, N., Morton, H., & Jack, M. A. (2008). Language learning with interactive virtual agent scenarios and speech recognition: Lessons learned. *Computer Animation and Virtual Worlds*, *19*(5), 605–619. https://doi.org/10.1002/cav.265

Apperley, T. (2006). Genre and game studies: Toward a critical approach to video game genres. *Simulation & Gaming*, *37*(1), 6–23. https://doi.org/10.1177/1046878105282278

Aylett, M. P., Kristensson, P. O., Whittaker, S., & Vazquez-Alvarez, Y. (2014). None of a CHInd: Relationship Counselling for HCI and Speech

Technology. *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, 749–760. https://doi.org/10.1145/2559206.2578868

Banks, J. (2015). Object, Me, Symbiote, Other: A social typology of player-avatar relationships. *First Monday*, *20*(2). https://doi.org/10.5210/fm.v20i2.5433

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, *41*(1), 164–171. JSTOR.

Bayliss, P. (2007). Beings in the Game-world: Characters, Avatars, and Players. *Proceedings of the 4th Australasian Conference on Interactive Entertainment*, 4:1–4:6. http://dl.acm.org/citation.cfm?id=1367956.1367960

Benjamin, L. (2006). Historical Evidence: Facts, Proof and Probability. In D. G. Godfrey (Ed.), *Methods of Historical Analysis in Electronic Media* (pp. 25–46). Routledge.

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. (2018). Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, *2*(3), 91:1–91:24. https://doi.org/10.1145/3264901

Bessière, K., Seay, A. F., & Kiesler, S. (2007). The ideal elf: Identity exploration in World of Warcraft. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, *10*(4), 530–535. https://doi.org/10.1089/cpb.2007.9994

Bingham, W. V. D., & Moore, B. V. (1959). *How to Interview* (4th ed.). Harper.

Björk, S. (2008). Games, Gamers, and Gaming: Understanding Game Research. *Proceedings of the 12th International Conference on Entertainment and Media in the Ubiquitous Era*, 64–68. https://doi.org/10.1145/1457199.1457213

Björk, S., & Holopainen, J. (2004). *Patterns in Game Design*. Charles River Media, Inc.

Björk, S., Lundgren, S., & Holopainen, J. (2003). Game Design Patterns. *Proceedings of the 2003 DiGRA International Conference: Level Up*, *2*.

Blackwell, A. F. (2015). HCI As an Inter-Discipline. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 503–516. https://doi.org/10.1145/2702613.2732505

Bødker, S. (2015). Third-wave HCI, 10 Years Later—Participation and Sharing. *Interactions*, *22*(5), 24–31. https://doi.org/10.1145/2804405

Bødker, S. (2006). When Second Wave HCI Meets Third Wave Challenges. *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles*, 1–8. https://doi.org/10.1145/1182475.1182476

Bogost, I., & Montfort, N. (2009). Platform studies: Frequently questioned answers. *Digital Arts and Culture*.

Bopp, J. A., Müller, L. J., Aeschbach, L. F., Opwis, K., & Mekler, E. D. (2019). Exploring Emotional Attachment to Game Characters. *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 313–324. https://doi.org/10.1145/3311350.3347169

Borchers, J. O. (2001). A Pattern Approach to Interaction Design. *AI & SOCIETY*, *15*(4), 359–376. https://doi.org/10.1007/BF01206115

Brahnam, S., & De Angeli, A. (2012). Gender affordances of conversational agents. *Interacting with Computers*, *24*(3), 139–153. https://doi.org/10.1016/j.intcom.2012.05.001

Brahnam, S., & Weaver, M. (2015). Re/Framing Virtual Conversational Partners: A Feminist Critique and Tentative Move Towards a New Design Paradigm. In A. Marcus (Ed.), *Design, User Experience, and Usability: Users and Interactions* (pp. 172–183). Springer International Publishing. https://doi.org/10.1007/978-3-319-20898-5_17

Brand, J. E., Jervis, J., Huggins, P. M., & Wilson, T. W. (2019). *Digital Australia 2020*. IGEA.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Bromley, S. (2018). Interviewing players. In A. Drachen, P. Mirza-Babaei, & L. Nacke (Eds.), *Games User Research* (pp. 163–173). Oxford University Press.

Budiu, R., & Laubheimer, P. (2018a). Intelligent Assistants: Creepy, Childish, or a Tool? Users' Attitudes Toward Alexa, Google Assistant, and Siri. *Nielsen Norman Group*. https://www.nngroup.com/articles/voice-assistant-attitudes/

Budiu, R., & Laubheimer, P. (2018b). Intelligent Assistants Have Poor Usability: A User Study of Alexa, Google Assistant, and Siri. *Nielsen Norman Group*. https://www.nngroup.com/articles/intelligent-assistant-usability/

Caillois, R. (2001). *Man, Play and Games*. University of Illinois Press.

Caroux, L., Isbister, K., Le Bigot, L., & Vibert, N. (2015). Player–video game interaction: A systematic review of current concepts. *Computers in Human Behavior*, *48*, 366–381. https://doi.org/10.1016/j.chb.2015.01.066

Carter, M. (2015). Emitexts and Paratexts: Propaganda in EVE Online. *Games and Culture*, *10*(4), 311–342. https://doi.org/10.1177/1555412014558089

Carter, M., Allison, F., Downs, J., & Gibbs, M. (2015). Player Identity Dissonance and Voice Interaction in Games. *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, 265–269. https://doi.org/10.1145/2793107.2793144

Carter, M., Downs, J., Nansen, B., Harrop, M., & Gibbs, M. (2014). Paradigms of Games Research in HCI: A Review of 10 Years of Research at CHI. *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play*, 27–36. https://doi.org/10.1145/2658537.2658708

Carter, M., Gibbs, M., & Arnold, M. (2012). Avatars, Characters, Players and Users: Multiple Identities at/in Play. *Proceedings of the 24th Australian*

Computer-Human Interaction Conference, 68–71. https://doi.org/10.1145/2414536.2414547

Cassell, J. (2000, April). Embodied Conversational Interface Agents. *Communications of the ACM*, *43*(4), 70–78. https://doi.org/10.1145/332051.332075

Cassell, J. (2001). Embodied Conversational Agents: Representation and Intelligence in User Interfaces. *AI Magazine*, *22*(4), 67–67. https://doi.org/10.1609/aimag.v22i4.1593

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., & Yan, H. (2001). More than just a pretty face: Conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, *14*(1–2), 55–64. https://doi.org/10.1016/S0950-7051(00)00102-7

Charmaz, K. (2014). *Constructing Grounded Theory* (2nd ed.). SAGE Publications.

Chen, Y., Naveed, A., & Porzel, R. (2010). Behavior and Preference in Minimal Personality: A Study on Embodied Conversational Agents. *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, 49:1–49:4. https://doi.org/10.1145/1891903.1891963

Chou, L. (2013). *The Voice in the Garden: An experiment in combining narrative and voice input for interaction design* [Masters thesis]. University of Southern California.

Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., & Cowan, B. (2019). The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers*. https://doi.org/10.1093/iwc/iwz016

Cler, G. J., Mittelman, T., Braden, M. N., Woodnorth, G. H., & Stepp, C. E. (2017). Video Game Rehabilitation of Velopharyngeal Dysfunction: A Case Series. *Journal of Speech, Language, and Hearing Research*, *60*(6S), 1800–1809. https://doi.org/10.1044/2017_JSLHR-S-16-0231

Coavoux, S., Boutet, M., & Zabban, V. (2017). What We Know About Games: A Scientometric Approach to Game Studies in the 2000s. *Games and Culture*, *12*(6), 563–584. https://doi.org/10.1177/1555412016676661

Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). *Voice User Interface Design*. Addison-Wesley Professional.

Collins, K. (2013). *Playing with Sound: A Theory of Interacting with Sound and Music in Video Games*. MIT Press.

Collins, K. (2011). Making Gamers Cry: Mirror Neurons and Embodied Interaction with Game Sound. *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound*, 39–46. https://doi.org/10.1145/2095667.2095673

Conconi, A., Ganchev, T., Kocsis, O., Papadopoulos, G., Fernández-Aranda, F., & Jiménez-Murcia, S. (2008). PlayMancer: A Serious Gaming 3D Environment. *2008 International Conference on Automated Solutions for Cross Media Content and Multi-Channel Distribution*, 111–117. https://doi.org/10.1109/AXMEDIS.2008.29

Consalvo, M. (2007). *Cheating: Gaining Advantage in Videogames*. MIT Press.

Consalvo, M. (2017). When paratexts become texts: De-centering the game-as-text. *Critical Studies in Media Communication*, *34*(2), 177–183. https://doi.org/10.1080/15295036.2017.1304648

Conway, S., & Trevillian, A. (2015). "Blackout!" Unpacking the Black Box of the Game Event. *Transactions of the Digital Games Research Association*, *2*(1). https://doi.org/10.26503/todigra.v2i1.42

Copier, M. (2003). The other game researcher: Participating in and watching the costruction of boundaries in game studies. *Proceedings of the 2003 DiGRA International Conference: Level Up*. DiGRA '03. http://www.digra.org/wp-content/uploads/digital-library/05163.46510.pdf

Cote, A., & Raz, J. G. (2015). In-depth interviews for games research. In P. Lankoski & S. Björk (Eds.), *Game Research Methods: An Overview* (pp. 93–116). ETC Press.

Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., & Beale, R. (2015). Voice Anthropomorphism, Interlocutor Modelling and Alignment Effects on Syntactic Choices in Human-computer Dialogue. *International Journal of Human-Computer Studies*, *83*(C), 27–42. https://doi.org/10.1016/j.ijhcs.2015.05.008

Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Earley, D., & Bandeira, N. (2017). 'What Can I Help You with?': Infrequent Users' Experiences of Intelligent Personal Assistants. *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 43:1–43:12. https://doi.org/10.1145/3098279.3098539

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed.). SAGE Publications.

Creswell, J. W., & Miller, D. L. (2000). Determining Validity in Qualitative Inquiry. *Theory Into Practice*, *39*(3), 124–130. https://doi.org/10.1207/s15430421tip3903_2

Cross, N. (1999). Design Research: A Disciplined Conversation. *Design Issues*, *15*(2), 5–10. JSTOR. https://doi.org/10.2307/1511837

Cross, T. (2017, December 26). Welcome to the golden age of board games—Brought to you by the internet. *Australian Financial Review*. https://www.afr.com/life-and-luxury/arts-and-culture/welcome-to-the-golden-age-of-board-games--brought-to-you-by-the-internet-20171219-h07a36

Crotty, M. (1998). *The Foundations of Social Research: Meaning and Perspective in the Research Process* (1st ed.). SAGE Publications.

Culley, K. E., & Madhavan, P. (2013). A note of caution regarding anthropomorphism in HCI agents. *Computers in Human Behavior*, *29*(3), 577–579. https://doi.org/10.1016/j.chb.2012.11.023

Dahl, D. A. (2013). Natural Language Processing: Past, Present and Future. In A. Neustein & J. A. Markowitz (Eds.), *Mobile Speech and Advanced Natural Language Solutions* (pp. 49–73). Springer New York.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies—Why and how. *Knowledge-Based Systems*, *6*(4), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

Davis, K. H., Biddulph, R., & Balashek, S. (1952). Automatic Recognition of Spoken Digits. *The Journal of the Acoustical Society of America*, *24*(6), 637–642. https://doi.org/10.1121/1.1906946

De Angeli, A., & Brahnam, S. (2008). I Hate You! Disinhibition with Virtual Partners. *Interacting with Computers*, *20*(3), 302–310. https://doi.org/10.1016/j.intcom.2008.02.004

Derboven, J., Huyghe, J., & De Grooff, D. (2014). Designing Voice Interaction for People with Physical and Speech Impairments. *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, 217–226. https://doi.org/10.1145/2639189.2639252

Deterding, S. (2009). The Game Frame: Systemizing a Goffmanian Approach to Video Game Theory. *Proceedings of the 2009 DiGRA International Conference: Breaking New Ground: Innovation in Games, Play, Practice and Theory*. DiGRA '09.

Deterding, S. (2017). The Pyrrhic Victory of Game Studies: Assessing the Past, Present, and Future of Interdisciplinary Game Research. *Games and Culture*, *12*(6), 521–543. https://doi.org/10.1177/1555412016665067

Dingler, T., Lindsay, J., & Walker, B. N. (2008, June). Learnability of Sound Cues for Environmental Features: Auditory Icons, Earcons, Spearcons, and Speech. *2008 Proceedings of the International Conference on Auditory Display*. International Conference on Auditory Display, 2008. https://smartech.gatech.edu/handle/1853/49940

Dix, A., Ramduny-Ellis, D., Rayson, P., Onditi, V., Sommerville, I., & Mackenzie, A. (2003). Finding Decisions Through Artefacts. *Proceedings of Human Computer Interaction, International 2003*, 78–82.

Dormans, J. (2013). Making Design Patterns Work. *Proceedings of the Second Workshop on Design Patterns in Games, DPG*. Fourth Workshop on Design Patterns in Games (DPG 2015), Pacific Grove, CA.

Dourish, P. (2004). *Where the Action Is: The Foundations of Embodied Interaction*. MIT Press.

Dow, S., Mehta, M., Harmon, E., MacIntyre, B., & Mateas, M. (2007). Presence and Engagement in an Interactive Drama. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1475–1484. https://doi.org/10.1145/1240624.1240847

Duarte, E. F., & Baranauskas, M. C. C. (2016). Revisiting the Three HCI Waves: A Preliminary Discussion on Philosophy of Science and Research Paradigms. *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems*, 38:1–38:4. https://doi.org/10.1145/3033701.3033740

Duggan, M. (2015). *Gaming and Gamers*. Pew Research Centre. https://www.pewresearch.org/internet/2015/12/15/gaming-and-gamers/

Duval, J., Rubin, Z., Segura, E. M., Friedman, N., Zlatanov, M., Yang, L., & Kurniawan, S. (2018). SpokeIt: Building a Mobile Speech Therapy Experience. *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 50:1–50:12. https://doi.org/10.1145/3229434.3229484

Ermi, L., & Mäyrä, F. (2005). Fundamental components of the gameplay experience: Analysing immersion. *Proceedings of the 2005 DiGRA International Conference: Changing Views: Worlds in Play*. DiGRA '05.

Fernández-Aranda, F., Jiménez-Murcia, S., Santamaría, J. J., Gunnard, K., Soto, A., Kalapanidas, E., Bults, R. G. A., Davarakis, C., Ganchev, T., Granero, R., Konstantas, D., Kostoulas, T. P., Lam, T., Lucas, M., Masuet-Aumatell, C., Moussa, M. H., Nielsen, J., & Penelo, E. (2012). Video games as a complementary therapy tool in mental disorders: PlayMancer, a European multicentre study. *Journal of Mental Health (Abingdon, England)*, *21*(4), 364–374. https://doi.org/10.3109/09638237.2012.664302

Fernström, M. (2003). *Sound objects and human-computer interaction design* (D. Rocchesso & F. Fontana, Eds.; pp. 33–46).

Fine, G. A. (1983). *Shared Fantasy: Role Playing Games as Social Worlds* (1st ed.). University of Chicago Press.

Fisher, R. J. (1993). Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research*, *20*(2), 303–315. https://doi.org/10.1086/209351

Fletcher, G., & Light, B. (2011). *Interpreting digital gaming practices: SingStar as a technology of work*. European Conference on Information Systems, Helsinki, Finland. http://usir.salford.ac.uk/17245/

Fontana, A., & Frey, J. H. (1994). Interviewing: The Art of Science. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 361–376). SAGE Publications.

Fontana, A., & Frey, J. H. (2005). The Interview: From Neutral Stance to Political Involvement. In *The Sage handbook of qualitative research, 3rd ed* (pp. 695–727). SAGE Publications.

Franinović, K., & Serafin, S. (2013). *Sonic Interaction Design*. MIT Press.

Fraser, J., Papaioannou, I., & Lemon, O. (2018). *Spoken Conversational AI in Video Games: Emotional Dialogue Management Increases User Engagement*. 179–184. https://doi.org/10.1145/3267851.3267896

Friberg, J., & Gärdenfors, D. (2004). Audio Games: New Perspectives on Game Audio. *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*, 148–154. https://doi.org/10.1145/1067343.1067361

Fuertes, J. N., Gottdiener, W. H., Martin, H., Gilbert, T. C., & Giles, H. (2012). A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology*, *42*(1), 120–133. https://doi.org/10.1002/ejsp.862

Galloway, A. R. (2006). *Gaming: Essays on Algorithmic Culture* (NED-New edition, Vol. 18). University of Minnesota Press; JSTOR. https://www.jstor.org/stable/10.5749/j.ctttss5p

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design Patterns: Elements of Reusable Object-oriented Software*. Addison-Wesley Longman Publishing Co., Inc.

Genette, G. (1997). *Paratexts: Thresholds of interpretation*. Cambridge University Press.

Genette, G., & Maclean, M. (1991). Introduction to the Paratext. *New Literary History*, *22*(2), 261–272. JSTOR. https://doi.org/10.2307/469037

Gergle, D., & Tan, D. S. (2014). Experimental Research in HCI. In *Ways of Knowing in HCI* (pp. 191–227). Springer New York.

Gernsback, H. (1916, June). An electric pup that heeds your call. *The Electrical Experimenter*, *4*(2), 104.

Gibson, W., & Northam, B. (2015, July 29). By Hand & Brain. *Medium*. https://medium.com/hand-brain/william-gibson-57e75f87e525

Gieryn, T. F. (1983). Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review*, *48*(6), 781–795. JSTOR. https://doi.org/10.2307/2095325

Gilbert, R., Thadani, V., Handy, C., Andrews, H., Sguigna, T., Sasso, A., & Payne, S. (2014). The psychological functions of avatars and alt(s): A qualitative study. *Computers in Human Behavior*, *32*, 1–8. https://doi.org/10.1016/j.chb.2013.11.007

Gillies, M. (2018). Creating Virtual Characters. *Proceedings of the 5th International Conference on Movement and Computing*, 22:1–22:8. https://doi.org/10.1145/3212721.3212835

Glas, R. (2015). Vicarious Play: Engaging the Viewer in Let's Play Videos. *Empedocles: European Journal for the Philosophy of Communication*, *5*(1), 81–86. https://doi.org/10.1386/ejpc.5.1-2.81_1

Glaser, B., & Strauss, A. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Publishing Company.

Goffman, E. (1961). *Encounters: Two Studies in the Sociology of Interaction*. Bobbs-Merrill.

Goffman, E. (1974). *Frame Analysis: An Essay on the Organization of Experience*. Harvard University Press.

Goldstein, M., Alsiö, G., & Werdenhoff, J. (2002). The Media Equation Does Not Always Apply: People are not Polite Towards Small Computers. *Personal and Ubiquitous Computing*, *6*(2), 87–96. https://doi.org/10.1007/s007790200008

Goodrich, M. A., & Schultz, A. C. (2007). Human–Robot Interaction: A Survey. *Foundations and Trends in Human–Computer Interaction*, *1*(3), 203–275. https://doi.org/10.1561/1100000005

Gorniak, P., & Roy, D. (2005). *Speaking with your Sidekick: Understanding Situated Speech in Computer Role Playing Games*. 57–62.

Gould, J. D., Conti, J., & Hovanyecz, T. (1982). Composing Letters with a Simulated Listening Typewriter. *Proceedings of the 1982 Conference on Human Factors in Computing Systems*, 367–370. https://doi.org/10.1145/800049.801813

Grammenos, D., Savidis, A., Georgalis, Y., & Stephanidis, C. (2006). Access Invaders: Developing a Universally Accessible Action Game. *Computers Helping People with Special Needs*, 388–395. https://doi.org/10.1007/11788713_58

Grammenos, D., Savidis, A., & Stephanidis, C. (2009). Designing universally accessible games. *Computers in Entertainment (CIE)*, *7*(1), 8.

Grammenos, D., Savidis, A., & Stephanidis, C. (2005). *UA-Chess: A universally accessible board game*. *7*.

Grammenos, D., Savidis, A., & Stephanidis, C. (2007). Unified Design of Universally Accessible Games. *Universal Access in Human-Computer Interaction. Applications and Services*, 607–616. https://doi.org/10.1007/978-3-540-73283-9_67

Green, P., & Wei-Haas, L. (1985). *The Wizard of Oz: A tool for rapid development of user interfaces* (Technical Report UMTRI-85-27).

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Speech acts* (Vol. 3, pp. 41–58). Academic Press.

Grudin, J. (2012). A Moving Target—The Evolution of Human-Computer Interaction. In J. A. Jacko (Ed.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications* (3rd ed., pp. xxvii–lxi). CRC Press.

Guba, E. E., & Lincoln, Y. S. (1994). Competing Paradigms in Qualitative Research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105–117). SAGE Publications.

Guzman, A. L. (2016). Making AI Safe for Humans: A Conversation with Siri. In R. W. Gehl & M. Bakardjieva (Eds.), *Socialbots and Their Friends: Digital Media and the Automation of Sociality*. Taylor & Francis.

Hämäläinen, P., Mäki-Patola, T., Pulkki, V., & Airas, M. (2004). Musical Computer Games Played by Singing. *Proc. 7th Int. Conf. on Digital Audio Effects (DAFx'04)*. 7th International Conference on Digital Audio Effects (DAFx'04), Naples, Italy.

Harada, S., Wobbrock, J. O., & Landay, J. A. (2011). Voice Games: Investigation into the Use of Non-speech Voice Input for Making Computer Games More Accessible. *Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction - Volume Part I*, 11–29.

Harrison, S., Sengers, P., & Tatar, D. (2011). Making epistemological trouble: Third-paradigm HCI as successor science. *Interacting with Computers*, *23*(5), 385–392. https://doi.org/10.1016/j.intcom.2011.03.005

Harrison, S., Tatar, D., & Sengers, P. (2007). The Three Paradigms of HCI. *Proceedings of CHI 2007*.

Hart, C. (2017). Getting Into the Game: An Examination of Player Personality Projection in Videogame Avatars. *Game Studies*, *17*(2). http://gamestudies.org/1702/articles/hart

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–83; discussion 83-135. https://doi.org/10.1017/S0140525X0999152X

Hill, R. W., Gratch, J., Marsella, S. C., Swartout, W., & Traum, D. (2003, June). Virtual Humans in the Mission Rehearsal Exercise System. *Kunstliche Intelligenzi (KI) (Special Issue on Embodied Conversational Agents)*. http://ict.usc.edu/pubs/Virtual%20Humans%20in%20the%20Mission%20Rehearsal%20Exercise%20System.pdf

Hoffmann, L., Krämer, N. C., Lam-Chi, A., & Kopp, S. (2009). Media Equation Revisited: Do Users Show Polite Reactions Towards an Embodied

Agent? *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, 159–165. https://doi.org/10.1007/978-3-642-04380-2_19

Hornbæk, K. (2013). Some Whys and Hows of Experiments in Human–Computer Interaction. *Foundations and Trends in Human-Computer Interaction*, *5*(4), 299–373. https://doi.org/10.1561/1100000043

Hornbæk, K., & Oulasvirta, A. (2017). What Is Interaction? *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5040–5052. https://doi.org/10.1145/3025453.3025765

Huang, X., Baker, J., & Reddy, R. (2014). A Historical Perspective of Speech Recognition. *Communications of the ACM*, *57*(1), 94–103. https://doi.org/10.1145/2500887

Huhtamo, E., & Parikka, J. (2011). Introduction: An Archaeology of Media Archaeology. In E. Huhtamo & J. Parikka (Eds.), *Media Archaeology: Approaches, Applications, and Implications* (1st ed., pp. 1–21). University of California Press.

Huizinga, J. (1949). *Homo Ludens: A study of the play-element in culture*. Routledge & Kegan Paul.

Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A Formal Approach to Game Design and Game Research. *Proceedings of the AAAI Workshop on Challenges in Game AI*.

Iacovides, I., Aczel, J., Scanlon, E., & Woods, W. (2013). Making sense of game-play: How can we examine learning and involvement? *Transactions of the Digital Games Research Association*, *1*(1). https://doi.org/10.26503/todigra.v1i1.6

Igarashi, T., & Hughes, J. F. (2001). Voice As Sound: Using Non-verbal Voice Input for Interactive Control. *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*, 155–156. https://doi.org/10.1145/502348.502372

Isbister, K., & Doyle, P. (2002). Design and Evaluation of Embodied Conversational Agents: A Proposed Taxonomy. *AAMAS '02: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems*. AAMAS '02, New York.

Isbister, K., & Nass, C. (2000). Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, *53*(2), 251–267. https://doi.org/10.1006/ijhc.2000.0368

Janicki, A., & Wawer, D. (2013). Voice-driven computer game in noisy environments. *International Journal of Computer Science and Applications*, *10*(1), 31–45. Scopus.

Jentsch, M., Biermann, M., & Schweiger, E. (2019). Talking to Stupid?!? Improving Voice User Interfaces. *Mensch Und Computer 2019 - Usability Professionals*. https://doi.org/10.18420/muc2019-up-0253

Jewell, M. (2011). Integrating Game Mechanics and Pedagogy: The Design and Production of Prospero, an Exploratory Speech Production Game. In D. Gouscos & M. Meimaris (Eds.), *Proceedings of ECGBL 2011*. Acad Conferences Ltd.

Jia, H., Wu, M., Jung, E., Shapiro, A., & Sundar, S. S. (2013). When the Tissue Box Says 'Bless You': Using Speech to Build Socially Interactive

Objects. *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 1635–1640. https://doi.org/10.1145/2468356.2468649

Johnson, D., & Gardner, J. (2009). Exploring Mindlessness as an Explanation for the Media Equation: A Study of Stereotyping in Computer Tutorials. *Personal Ubiquitous Comput.*, *13*(2), 151–163. https://doi.org/10.1007/s00779-007-0193-9

Johnson, W. L. (2007). Serious use of a serious game for language learning. *Frontiers in Artificial Intelligence and Applications*, *158*, 67.

Johnson, W. L., & Beal, C. R. (2005). *Iterative Evaluation of a Large-Scale, Intelligent Game for Language Learning*. 290–297.

Johnson, W. L., & Valente, A. (2009). Tactical Language and Culture Training Systems: Using AI to Teach Foreign Languages and Cultures. *AI Magazine*, *30*(2), 72. https://doi.org/10.1609/aimag.v30i2.2240

Johnson, W. L., & Wu, S. (2008). Assessing Aptitude for Learning with a Serious Game for Foreign Language and Culture. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Intelligent Tutoring Systems* (pp. 520–529). Springer. https://doi.org/10.1007/978-3-540-69132-7_55

Jolin, D. (2016, September 25). The rise and rise of tabletop gaming. *The Observer*. https://www.theguardian.com/technology/2016/sep/25/board-games-back-tabletop-gaming-boom-pandemic-flash-point

Jonze, S. (2013). *Her* [Cinema]. Warner Bros. Pictures.

Jørgensen, K. (2007). *'What are Those Grunts and Growls Over There?' Computer Game Audio and Player Action* [PhD thesis]. Copenhagen University.

Jørgensen, K. (2009). 'I'm overburdened!' An Empirical Study of the Player, the Avatar, and the Gameworld. *Proceedings of the 2009 DiGRA International Conference: Breaking New Ground: Innovation in Games, Play, Practice and Theory*. DiGRA '09.

Jørgensen, K. (2013). *Gameworld Interfaces*. MIT Press.

Kahn, P. H., Freier, N. G., Kanda, T., Ishiguro, H., Ruckert, J. H., Severson, R. L., & Kane, S. K. (2008). Design Patterns for Sociality in Human-Robot Interaction. *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, 97–104. https://doi.org/10.1145/1349822.1349836

Kelley, J. F. (1984). An Iterative Design Methodology for User-friendly Natural Language Office Information Applications. *ACM Transactions on Information Systems*, *2*(1), 26–41. https://doi.org/10.1145/357417.357420

Kelley, J. F. (1983). An Empirical Methodology for Writing User-friendly Natural Language Computer Applications. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 193–196. https://doi.org/10.1145/800045.801609

Keogh, B. (2012, December 8). A Sum of Parts: Use Your Words. *Gameranx*. https://gameranx.com/features/id/11325/article/a-sum-of-parts-use-your-words/

Keogh, B. (2016, April 6). You can't ignore the cultural power of video games any longer [Text]. *ABC News*. https://www.abc.net.au/news/2016-04-06/keogh-why-you-should-care-about-video-games/7303744

Kirke, A. (2018). When the Soundtrack Is the Game: From Audio-Games to Gaming the Music. In Duncan Williams & N. Lee (Eds.), *Emotion in Video Game Soundtracking* (pp. 65–83). Springer International Publishing. https://doi.org/10.1007/978-3-319-72272-6_7

Kiselkov, S., Kováčová, B., Kozmon, M., Litvová, L., Poláčik, M., Tekeľ, J., & Kovárová, A. (2008). Bloodlezz: Simulation of the Life of a Mosquito. *Poster Presented at the Slovak University of Technology in Bratislava Informatics and Information Technologies Student Research Conference (IITSRC)*.

Klimmt, C., Hefner, D., & Vorderer, P. (2009). The Video Game Experience as 'True' Identification: A Theory of Enjoyable Alterations of Players' Self-Perception. *Communication Theory*, *19*(4), 351–373. https://doi.org/10.1111/j.1468-2885.2009.01347.x

Knijnenburg, B. P., & Willemsen, M. C. (2016). Inferring Capabilities of Intelligent Agents from Their External Traits. *ACM Transactions on Interactive Intelligent Systems*, *6*(4), 28:1–28:25. https://doi.org/10.1145/2963106

Knox, W. B., Glass, B. D., Love, B. C., Maddox, W. T., & Stone, P. (2012). How Humans Teach Agents. *International Journal of Social Robotics*, *4*(4), 409–421. https://doi.org/10.1007/s12369-012-0163-x

Konzack, L. (2002). Computer Game Criticism: A Method for Computer Game Analysis. *Computer Games and Digital Cultures Conference Proceedings*, 89–100.

Kostoulas, T., Mporas, I., Kocsis, O., Ganchev, T., Katsaounos, N., Santamaria, J. J., Jimenez-Murcia, S., Fernandez-Aranda, F., & Fakotakis, N. (2012). Affective speech interface in serious games for supporting therapy of mental disorders. *Expert Systems with Applications*, *39*(12), 11072–11079. https://doi.org/10.1016/j.eswa.2012.03.067

Krause, M., Smeddinck, J., & Meyer, R. (2013). A Digital Game to Support Voice Treatment for Parkinson's Disease. *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 445–450. https://doi.org/10.1145/2468356.2468435

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology* (2nd ed.). SAGE Publications.

Kumar, A., Reddy, P., Tewari, A., Agrawal, R., & Kam, M. (2012). Improving Literacy in Developing Countries Using Speech Recognition-supported Games on Mobile Devices. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1149–1158. https://doi.org/10.1145/2207676.2208564

Lallée, S., Yoshida, E., Mallet, A., Nori, F., Natale, L., Metta, G., Warneken, F., & Dominey, P. F. (2010). Human-robot cooperation based on interaction learning. In *From motor learning to interaction learning in robots* (pp. 491–536). Springer.

Lan, T., Aryal, S., Ahmed, B., Ballard, K., & Gutierrez-Osuna, R. (2014). Flappy Voice: An Interactive Game for Childhood Apraxia of Speech Therapy. *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play*, 429–430. https://doi.org/10.1145/2658537.2661305

Lang, H., Klepsch, M., Nothdurft, F., Seufert, T., & Minker, W. (2013). Are Computers Still Social Actors? *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 859–864. https://doi.org/10.1145/2468356.2468510

Lankoski, P., & Björk, S. (2007). Gameplay Design Patterns for Believable Non-Player Characters. *Proceedings of the 2007 DiGRA International Conference: Situated Play*, *4*.

Lankoski, P., Johansson, A., Karlsson, B., Björk, S., & Dell'Acqua, P. (2011). AI Design for Believable Characters via Gameplay Design Patterns. In *Business, Technological and Social Dimensions of Computer Games: Multidisciplinary Developments* (p. 15). IGI Global. https://doi.org/10.4018/978-1-60960-567-4.ch002

Lazzaro, N. (2009). Why We Play: Affect and the Fun of Games. In A. Sears & J. A. Jacko (Eds.), *Human-Computer Interaction: Designing for Diverse Users and Domains* (p. 284). CRC Press.

Lea, W. A., & Shoup, J. E. (1979). *Review of the ARPA SUR Project and Survey of Current Technology in Speech Understanding.* (No. N00014-77-C–0570; p. 162). Office of Naval Research.

Lee, E. J., Nass, C., & Brave, S. (2000). Can Computer-generated Speech Have Gender?: An Experimental Test of Gender Stereotype. *CHI '00 Extended Abstracts on Human Factors in Computing Systems*, 289–290. https://doi.org/10.1145/633292.633461

Lee, E.-J. (2008). Gender Stereotyping of Computers: Resource Depletion or Reduced Attention? *Journal of Communication*, *58*(2), 301–320. https://doi.org/10.1111/j.1460-2466.2008.00386.x

Lee, K.-F. (1988). Word Recognition in Large Vocabularies: On large-vocabulary speaker-independent continuous speech recognition. *Speech Communication*, *7*(4), 375–379. https://doi.org/10.1016/0167-6393(88)90053-2

Lewis, C. (1982). *Using the 'Thinking-aloud' Method in Cognitive Interface Design* (No. RC9265). IBM TJ Watson Research Center.

Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). An Overview of Noise-robust Automatic Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(4), 745–777. https://doi.org/10.1109/TASLP.2014.2304637

Li, X. (2013). *Coping with negative emotions through breath control* [Masters thesis]. University of Southern California.

Limerick, H., Moore, J. W., & Coyle, D. (2015). Empirical Evidence for a Diminished Sense of Agency in Speech Interfaces. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3967–3970. https://doi.org/10.1145/2702123.2702379

Linderoth, J. (2012). The Effort of Being in a Fictional World: Upkeyings and Laminated Frames in MMORPGs. *Symbolic Interaction*, *35*(4), 474–492. https://doi.org/10.1002/symb.39

Linek, S. B., Gerjets, P., & Scheiter, K. (2010). The speaker/gender effect: Does the speaker's gender matter when presenting auditory text in multimedia messages? *Instructional Science*, *38*(5), 503–521. https://doi.org/10.1007/s11251-009-9115-8

Liu, Y., Goncalves, J., Ferreira, D., Xiao, B., Hosio, S., & Kostakos, V. (2014). CHI 1994-2013: Mapping Two Decades of Intellectual Progress Through Co-word Analysis. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3553–3562. https://doi.org/10.1145/2556288.2556969

Loaiza, D., Oviedo, C., Castillo, A., Portilla, A., Álvarez, G., Linares, D., & Navarro, A. (2013). *A video game prototype for speech rehabilitation*. 2013 5th International Conference on Games and Virtual Worlds for Serious Applications, VS-GAMES 2013. Scopus. https://doi.org/10.1109/VS-GAMES.2013.6624218

Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., Sorsche, P., Hirsch, D., Li, Q., & Martinez, A. (2018). Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science*, 0961000618759414. https://doi.org/10.1177/0961000618759414

Lopes, M., Magalhães, J., & Cavaco, S. (2016). A Voice-controlled Serious Game for the Sustained Vowel Exercise. *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology*, 32:1–32:6. https://doi.org/10.1145/3001773.3001807

Luger, E., & Sellen, A. (2016). 'Like Having a Really Bad PA': The Gulf Between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297. https://doi.org/10.1145/2858036.2858288

Luo, Z., Westerman, D., & Banks, J. (2019). Extending the self: Player-avatar relations and presence among U.S. and Chinese gamers. *Journal For Virtual Worlds Research*, *12*(3). https://doi.org/10.4101/jvwr.v12i3.7395

Lv, Z., Esteve, C., Chirivella, J., & Gagliardo, P. (2017). Serious game based personalized healthcare system for dysphonia rehabilitation. *Pervasive and Mobile Computing*, *41*, 504–519. https://doi.org/10.1016/j.pmcj.2017.04.006

Lv, Z., Esteve, C., Chirivella, J., & Gagliardo, P. (2015). Clinical Feedback and Technology Selection of Game Based Dysphonic Rehabilitation Tool. *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*, 253–256. http://dl.acm.org/citation.cfm?id=2826165.2826207

Martin, P. (2018). The Intellectual Structure of Game Research. *Game Studies*, *18*(1). http://gamestudies.org/1801/articles/paul_martin

Marwick, A. (1989). *The Nature of History* (3rd ed.). Macmillan Education.

Marwick, A. (2001). *The New Nature of History: Knowledge, Evidence, Language* (1st ed.). Palgrave.

Maulsby, D., Greenberg, S., & Mander, R. (1993). Prototyping an Intelligent Agent through Wizard of Oz. *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, 277–284. https://doi.org/10.1145/169059.169215

Mäyrä, F. (2008). *An Introduction to Game Studies*. SAGE Publications.

Mäyrä, F. (2014, April 29). Game Studies: A polyphonic discipline? *Frans Goes Blog*. https://fransmayra.fi/2014/04/29/game-studies-a-polyphonic-discipline/

Mayring, P. (2004). Qualitative Content Analysis. In U. Flick, E. von Kardoff, & I. Steinke (Eds.), *A Companion to Qualitative Research* (pp. 266–269). SAGE Publications.

McAleer, P., Todorov, A., & Belin, P. (2014). How Do You Say 'Hello'? Personality Impressions from Brief Novel Voices. *PLOS ONE*, *9*(3), e90779. https://doi.org/10.1371/journal.pone.0090779

McCarthy, J., & Wright, P. (2007). *Technology As Experience*. The MIT Press.

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), 72:1–72:23. https://doi.org/10.1145/3359174

McGee, K. (2007). Patterns and Computer Game Design Innovation. *Proceedings of the 4th Australasian Conference on Interactive Entertainment*, 16:1–16:8. http://dl.acm.org/citation.cfm?id=1367956.1367972

McGraw, I., Gruenstein, A., & Sutherland, A. (2009). A self-labeling speech corpus: Collecting spoken words with an online educational game. *INTERSPEECH-2009*. 10th Annual Conference of the International Speech Communication Association.

McGraw, I., & Seneff, S. (2008). Speech-enabled card games for language learners. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, *2*, 778–783. Scopus.

McGraw, I., Yoshimoto, O., & Seneff, S. (2009). Speech-enabled Card Games for Incidental Vocabulary Acquisition in a Foreign Language. *Speech Communication*, *51*(10), 1006–1023. https://doi.org/10.1016/j.specom.2009.04.011

McMillan, S. J. (2000). The Microscope and the Moving Target: The Challenge of Applying Content Analysis to the World Wide Web. *Journalism & Mass Communication Quarterly*, *77*(1), 80–98. https://doi.org/10.1177/107769900007700107

Melcer, E., Nguyen, T.-H. D., Chen, Z., Canossa, A., El-Nasr, M. S., & Isbister, K. (2015). Games Research Today: Analyzing the Academic Landscape 2000-2014. *Proceedings of the 10th International Conference on the Foundations of Digital Games (FDG 2015)*, *17*, 20.

Microsoft Kinect 'fastest-selling device on record'. (2011, March 10). *BBC News*. http://www.bbc.com/news/business-12697975

Miller, J., & Glassner, B. (2010). The 'Inside' and the 'Outside': Finding Realities in Interviews. In D. Silverman (Ed.), *Qualitative Research* (3rd ed., pp. 125–139). SAGE Publications.

Moskvitch, K. (2017, February 16). The machines that learned to listen. *BBC Future*. https://www.bbc.com/future/article/20170214-the-machines-that-learned-to-listen

Munteanu, C., Jones, M., Oviatt, S., Brewster, S., Penn, G., Whittaker, S., Rajput, N., & Nanavati, A. (2013). We Need to Talk: HCI and the Delicate Topic of Spoken Language Interaction. *CHI '13 Extended*

Abstracts on Human Factors in Computing Systems, 2459–2464. https://doi.org/10.1145/2468356.2468803

Murad, C., & Munteanu, C. (2019). 'I don't know what you're talking about, HALexa': The case for voice user interface guidelines. *Proceedings of the 1st International Conference on Conversational User Interfaces*, 9:1–9:3. https://doi.org/10.1145/3342775.3342795

Murad, C., Munteanu, C., Cowan, B. R., & Clark, L. (2019). Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing*, *18*(2), 33–45. https://doi.org/10.1109/MPRV.2019.2906991

Mustaquim, M. M. (2013). Automatic speech recognition: An approach for designing inclusive games. *Multimedia Tools and Applications*, *66*(1), 131–146. Scopus. https://doi.org/10.1007/s11042-011-0918-7

Nagy, M. (2006). Penguin Quart—Slovak Digit Speech Recognition Game Based on HMM. *Artificial Intelligence Applications and Innovations*, 179–186. https://doi.org/10.1007/0-387-34224-9_21

Nanjo, H., Mikami, H., Kunimatsu, S., Kawano, H., & Nishiura, T. (2009). A Fundamental Study of Novel Speech Interface for Computer Games. *ISCE: 2009 IEEE 13th International Symposium on Consumer Electronics, Vols 1 and 2*, 291–293.

Nass, C., & Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. MIT Press.

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81.

Nass, C., Moon, Y., & Carney, P. (1999). Are People Polite to Computers? Responses to Computer-Based Interviewing Systems. *Journal of Applied Social Psychology*, *29*(5), 1093–1109. https://doi.org/10.1111/j.1559-1816.1999.tb00142.x

Nass, C., Moon, Y., & Green, N. (1997). Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices. *Journal of Applied Social Psychology*, *27*(10), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers Are Social Actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. https://doi.org/10.1145/191666.191703

Navarro-Newball, A. A., Loaiza, D., Oviedo, C., Castillo, A. D., Portilla, A., Linares, D., & Álvarez, G. I. (2014). Talking to Teo: Video game supported speech therapy. *Entertainment Computing*, *5*(4), 401–412. Scopus. https://doi.org/10.1016/j.entcom.2014.10.005

Neto, J. N., Silva, R., Neto, J. P., Pereira, J. M., & Fernandes, J. (2011). Solis' Curse—A Cultural Heritage Game Using Voice Interaction with a Virtual Agent. *2011 Third International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, 164–167. https://doi.org/10.1109/VS-GAMES.2011.31

Neuman, W. L. (2014). *Social Research Methods: Qualitative and Quantitative Approaches* (7th ed.). Pearson Education.

Norman, D. A. (1983). Some Observations on Mental Models. In D. Gentner & A. L. Stevens (Eds.), *Mental Models* (pp. 7–14). Lawrence Erlbaum Associates.

Norte, S., & Lobo, F. G. (2008). *Sudoku Access: A Sudoku game for people with motor disabilities*. 161–167. Scopus. https://doi.org/10.1145/1414471.1414502

O'Connor, J. J. M., & Barclay, P. (2017). The influence of voice pitch on perceptions of trustworthiness across social contexts. *Evolution and Human Behavior*, *38*(4), 506–512. https://doi.org/10.1016/j.evolhumbehav.2017.03.001

O'Donovan, J., Ward, J., Hodgins, S., & Sundstedt, V. (2009). *Rabbit Run: Gaze and voice based game interaction*. Eurographics Ireland Workshop, December.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776–783. https://doi.org/10.1037/h0043424

Pan, Y., & Stolterman, E. (2013). Pattern Language and HCI: Expectations and Experiences. *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 1989–1998. https://doi.org/10.1145/2468356.2468716

Parikka, J. (2012). *What is Media Archaeology?* Polity Press.

Parikka, J., & Suominen, J. (2006). Victorian Snakes? Towards A Cultural History of Mobile Games and the Experience of Movement. *Game Studies*, *6*(1). http://gamestudies.org/0601/articles/parikka_suominen

Parisi, D. (2008). Game Interfaces as Bodily Techniques. *Handbook of Research on Effective Electronic Gaming in Education*.

Parisi, D. (2013). Shocking Grasps: An Archaeology of Electrotactile Game Mechanics. *Game Studies*, *13*(2). http://gamestudies.org/1302/articles/parisi

Parker, J. R., & Heerema, J. (2008). Audio Interaction in Computer Mediated Games. *International Journal of Computer Games Technology*, *2008*, 1:1–1:8. https://doi.org/10.1155/2008/178923

Paul, C. A. (2010). Process, Paratexts, and Texts: Rhetorical Analysis and Virtual Worlds. *Journal For Virtual Worlds Research*, *3*(1). https://doi.org/10.4101/jvwr.v3i1.804

Phan, T. (2017). The Materiality of the Digital and the Gendered Voice of Siri. *Transformations*, *29*.

Picca, D., Jaccard, D., & Eberlé, G. (2015). Natural Language Processing in Serious Games: A state of the art. *International Journal of Serious Games*, *2*(3). https://doi.org/10.17083/ijsg.v2i3.87

Poláček, O. (2014). *Designing Text Entry Methods for Non-Verbal Vocal Input* [PhD thesis, Czech Technical University in Prague]. https://dspace.cvut.cz/handle/10467/28194

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice Interfaces in Everyday Life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 640:1–640:12. https://doi.org/10.1145/3173574.3174214

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). 'Alexa is My New BFF': Social Roles, User Satisfaction, and Personification of the Amazon Echo. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2853–2859. https://doi.org/10.1145/3027063.3053246

Qingtang, L., Yang, W., Linjing, W., Jingxiu, H., & Peng, W. (2015). Design and Implementation of a Serious Game Based on Kinect. *2015 International Conference of Educational Innovation through Technology (EITT)*, 13–18. https://doi.org/10.1109/EITT.2015.10

Qvarfordt, P., Jönsson, A., & Dahlbäck, N. (2003). The Role of Spoken Feedback in Experiencing Multimodal Interfaces As Human-like. *Proceedings of the 5th International Conference on Multimodal Interfaces*, 250–257. https://doi.org/10.1145/958432.958478

Rabiner, L., & Juang, B.-H. (2008). Historical Perspective of the Field of ASR/NLU. In J. Benesty, M. M. Sondhi, & Y. Huang (Eds.), *Springer Handbook of Speech Processing* (pp. 521–538). Springer. https://doi.org/10.1007/978-3-540-49127-9_26

Reddy, D. R., Erman, L. D., & Neely, R. B. (1973). A model and a system for machine recognition of speech. *IEEE Transactions on Audio and Electroacoustics*, *21*(3), 229–238. https://doi.org/10.1109/TAU.1973.1162456

Reeves, B., & Nass, C. (1996). *The Media Equation: How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge University Press.

Rehm, M. (2008). "She is just stupid"—Analyzing user–agent interactions in emotional game situations. *Interacting with Computers*, *20*(3), 311–325. https://doi.org/10.1016/j.intcom.2008.02.005

Rico, J., & Brewster, S. (2010). Gesture and Voice Prototyping for Early Evaluations of Social Acceptability in Multimodal Interfaces. *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, 16:1–16:9. https://doi.org/10.1145/1891903.1891925

Riek, L. (2012). Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, *1*(1), 119–136. https://doi.org/10.5898/JHRI.1.1.Riek

Rivera, G., Hullett, K., & Whitehead, J. (2012). Enemy NPC Design Patterns in Shooter Games. *Proceedings of the First Workshop on Design Patterns in Games*, 6:1–6:8. https://doi.org/10.1145/2427116.2427122

Röber, N., & Masuch, M. (2005). Playing Audio-only Games: A compendium of interacting with virtual, auditory Worlds. *Proceedings of the 2005 DiGRA International Conference: Changing Views: Worlds in Play*. DiGRA '05.

Roy, M. (2016). *La réalité virtuelle pour l'apprentissage des langues: L'influence de l'immersion, de l'interaction et de la présence sur la communication en langue étrangère : une étude exploratoire auprès d'adolescents apprenant le français ou l'allemand* [PhD thesis, University of Strasbourg]. http://www.theses.fr/2016STRAC036

Rubin, Z., & Kurniawan, S. (2013). Speech Adventure: Using Speech Recognition for Cleft Speech Therapy. *Proceedings of the 6th International Conference on Pervasive Technologies Related to Assistive Environments*, 35:1–35:4. https://doi.org/10.1145/2504335.2504373

Ryan, E. B., & Sebastian, R. J. (1980). The effects of speech style and social class background on social judgements of speakers. *British Journal of Social & Clinical Psychology*, *19*(3), 229–233. https://doi.org/10.1111/j.2044-8260.1980.tb00348.x

Ryan, J. O., Kaltman, E., Mateas, M., & Wardrip-Fruin, N. (2015, June). What We Talk About When We Talk About Games: Bottom-Up Game Studies Using Natural Language Processing. *Proceedings of the 10th International Conference on the Foundations of Digital Games*.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, *50*(4), 696–735. https://doi.org/10.2307/412243

Salen, K., & Zimmerman, E. (2004). *Rules of Play: Game Design Fundamentals*. MIT Press.

Sangin, M. (2018). Observing the player experience: The art and craft of observing and documenting Games User Research. In *Games User Research* (pp. 175–188). Oxford University Press.

Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). 'Hey Alexa, What's Up?': A Mixed-Methods Studies of In-Home Conversational Agent Usage. *Proceedings of the 2018 Designing Interactive Systems Conference*, 857–868. https://doi.org/10.1145/3196709.3196772

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Sharp, H., Preece, J., & Rogers, Y. (2019). *Interaction Design: Beyond Human-Computer Interaction* (5th ed.). John Wiley & Sons, Inc.

Silva, A., Mamede, N., Ferreira, A., Baptista, J., & Fernandes, J. (2011). Towards a Serious Game for Portuguese Learning. *Serious Games Development and Applications*, 83–94. https://doi.org/10.1007/978-3-642-23834-5_8

Silvervarg, A., Raukola, K., Haake, M., & Gulz, A. (2012). The Effect of Visual Gender on Abuse in Conversation with ECAs. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Intelligent Virtual Agents* (pp. 153–160). Springer. https://doi.org/10.1007/978-3-642-33197-8_16

Smith, G. (2006). *Erving Goffman*. Routledge.

Sonne, T., & Jensen, M. M. (2016). ChillFish: A Respiration Game for Children with ADHD. *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*, 271–278. https://doi.org/10.1145/2839462.2839480

Sporka, A. J., Kurniawan, S. H., Mahmud, M., & Slavík, P. (2006). *Non-speech Input and Speech Recognition for Real-time Control of Computer Games*. *2006*, 213–220. Scopus. https://doi.org/10.1145/1168987.1169023

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., & Waters, K. (1996). When the Interface is a Face. *Human-Computer Interaction*, *11*(2), 97–124. https://doi.org/10.1207/s15327051hci1102_1

Sra, M., Xu, X., & Maes, P. (2018). BreathVR: Leveraging Breathing As a Directly Controlled Interface for Virtual Reality Games. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 340:1–340:12. https://doi.org/10.1145/3173574.3173914

Starner, T., Leibe, B., Singletary, B., & Pair, J. (2000). MIND-WARPING: Towards Creating a Compelling Collaborative Augmented Reality Game. *Proceedings of the 5th International Conference on Intelligent User Interfaces*, 256–259. https://doi.org/10.1145/325737.325864

Sudnow, D. (1979). *Pilgrim in the Microworld*. Warner Books.

Suits, B. (1978). *The Grasshopper: Games, Life and Utopia*. University of Toronto Press.

Suominen, J. (2016). How to Present the History of Digital Games: Enthusiast, Emancipatory, Genealogical, and Pathological Approaches. *Games and Culture*. https://doi.org/10.1177/1555412016653341

Sutton-Smith, B. (1997). *The Ambiguity of Play*. Harvard University Press.

Tan, C. T., Johnston, A., Ballard, K., Ferguson, S., & Perera-Schulz, D. (2013). sPeAK-MAN: Towards Popular Gameplay for Speech Therapy. *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death*, 28:1–28:4. https://doi.org/10.1145/2513002.2513022

Tennent, P., Rowland, D., Marshall, J., Egglestone, S. R., Harrison, A., Jaime, Z., Walker, B., & Benford, S. (2011). Breathalising Games: Understanding the Potential of Breath Control in Game Interfaces. *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*, 58:1–58:8. https://doi.org/10.1145/2071423.2071496

Teófilo, L. F., Nogueira, P. A., & Silva, P. B. (2013). GEMINI: A Generic Multi-Modal Natural Interface Framework for Videogames. In Á. Rocha, A. M. Correia, T. Wilson, & K. A. Stroetmann (Eds.), *Advances in Information Systems and Technologies* (pp. 873–884). Springer.

Thornberg, R. (2012). Informed Grounded Theory. *Scandinavian Journal of Educational Research*, *56*(3), 243–259. https://doi.org/10.1080/00313831.2011.581686

Torre, I., Goslin, J., White, L., & Zanatto, D. (2018). Trust in Artificial Voices: A ‘Congruency Effect’ of First Impressions and Behavioural Experience. *Proceedings of the Technology, Mind, and Society*, 40:1–40:6. https://doi.org/10.1145/3183654.3183691

Toys that obey your voice. (1916, December). *Popular Science Monthly*, *89*, 718–719.

Tse, E., Greenberg, S., Shen, C., & Forlines, C. (2007). Multimodal multiplayer tabletop gaming. *Computers in Entertainment*, *5*(2). Scopus. https://doi.org/10.1145/1279540.1279552

Umanski, D., Avni, Y., & Rinott, M. (2017). Sonora: Inclusive Voice Play For Children With Various Abilities. *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*, 369–374. https://doi.org/10.1145/3130859.3131324

Vieira, M. F. G., Fu, H., Hu, C., Kim, N., & Aggarwal, S. (2014). PowerFall: A Voice-controlled Collaborative Game. *Proceedings of the First ACM*

SIGCHI Annual Symposium on Computer-Human Interaction in Play, 395–398. https://doi.org/10.1145/2658537.2662993

Wadley, G. (2011). *Voice in virtual worlds* [PhD thesis]. The University of Melbourne.

Wadley, G., Carter, M., & Gibbs, M. (2015). Voice in Virtual Worlds: The Design, Use, and Influence of Voice Chat in Online Play. *Human–Computer Interaction*, *30*(3–4), 336–365. https://doi.org/10.1080/07370024.2014.987346

Wadley, G., & Gibbs, M. R. (2010). Speaking in Character: Voice Communication in Virtual Worlds. In *Online Worlds: Convergence of the Real and the Virtual* (pp. 187–200). Springer, London. https://doi.org/10.1007/978-1-84882-825-4_15

Warren, T. (2014, April 2). The story of Cortana, Microsoft's Siri killer. *The Verge*. http://www.theverge.com/2014/4/2/5570866/cortana-windows-phone-8-1-digital-assistant

Whitenton, K., & Budiu, R. (2018). The Paradox of Intelligent Assistants: Poor Usability, High Adoption. *Nielsen Norman Group*. https://www.nngroup.com/articles/intelligent-assistants-poor-usability-high-adoption/

Williams, Dmitri, Martins, N., Consalvo, M., & Ivory, J. D. (2009). The virtual census: Representations of gender, race and age in video games. *New Media & Society*, *11*(5), 815–834. https://doi.org/10.1177/1461444809105354

Williamson, J. R. (2011). Send Me Bubbles: Multimodal Performance and Social Acceptability. *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, 899–904. https://doi.org/10.1145/1979742.1979513

Wilson, G., McGill, M., Jamieson, M., Williamson, J. R., & Brewster, S. A. (2018). Object Manipulation in Virtual Reality Under Increasing Levels of Translational Gain. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 99:1–99:13. https://doi.org/10.1145/3173574.3173673

Woods, H. S. (2018). Asking more of Siri and Alexa: Feminine persona in service of surveillance capitalism. *Critical Studies in Media Communication*, *35*(4), 334–349. https://doi.org/10.1080/15295036.2018.1488082

Yannakakis, G. N. (2012). Game AI revisited. *Proceedings of the 9th Conference on Computing Frontiers*, 285–292. https://doi.org/10.1145/2212908.2212954

Yannakakis, G. N., & Togelius, J. (2015). A panorama of artificial and computational intelligence in games. *IEEE Transactions on Computational Intelligence and AI in Games*, *7*(4), 317–335.

Yee, N., & Bailenson, J. (2007). The Proteus Effect: The effect of transformed self-representation on behavior. *Human Communication Research*, *33*, 271–290.

Yoon, G., & Vargas, P. T. (2016). I See Myself through My Avatar: Evidence from East Asia and North America for Acculturated Effects on Virtual Representation Preference. *Asiascape: Digital Asia*, *3*(1–2), 79–96. https://doi.org/10.1163/22142312-12340049

Yuan, X., & Fan, J. (2011). Design and implementation of voice controlled Tetris game based on Microsoft SDK. *2011 International Conference on Multimedia Technology*, 275–278. https://doi.org/10.1109/ICMT.2011.6001825

Zagal, J. P., Tomuro, N., & Shepitsen, A. (2012). Natural Language Processing in Game Studies Research: An Overview. *Simulation & Gaming*, *43*(3), 356–373. https://doi.org/10.1177/1046878111422560

Zajc, M., Starčič, A. I., Lebeničnik, M., & Gačnik, M. (2018). Tablet game-supported speech therapy embedded in children's popular practices. *Behaviour & Information Technology*, *0*(0), 1–10. https://doi.org/10.1080/0144929X.2018.1474253

Zanbaka, C., Goolkasian, P., & Hodges, L. (2006). Can a Virtual Cat Persuade You?: The Role of Gender and Realism in Speaker Persuasiveness. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1153–1162. https://doi.org/10.1145/1124772.1124945

Zeller, M., McCulloch, J., Bray, B., Hak0n, Ong, S., Schonning, N., & rwinj. (2019, February 24). *Voice input*. Microsoft Docs. https://docs.microsoft.com/en-us/windows/mixed-reality/voice-input

Zeng, X., Lu, P., Zhang, M., & Wang, Y. (2005). Multi-modal game player interface system based on video and speech. *Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao / Journal of Computer-Aided Design and Computer Graphics*, *17*(10), 2353–2358. Scopus.

10 Appendix A

This appendix includes the forms required for the inclusion of published material according to The University of Melbourne guidelines for a *Thesis with Publication*.[22]

THE UNIVERSITY OF MELBOURNE

Declaration for a thesis with publication

PhD and MPhil students may include a primary research publication in their thesis in lieu of a chapter if:

- The student contributed greater than 50% of the content in the publication and is the "primary author", ie. the student was responsible primarily for the planning, execution and preparation of the work for publication
- The student has approval to include the publication in their thesis from their Advisory Committee
- It is a primary publication that reports on original research conducted by the student during their enrolment
- The initial draft of the work was written by the student and any subsequent editing in response to co-authors and editors reviews was performed by the student
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis

Students must submit this form, along with *Co-author authorisation forms* completed by each co-author, when the thesis is submitted to the Thesis Examination System: *https://tes.app.unimelb.edu.au/*. If you are including multiple publications in your thesis you will need to list each publication on this form. Further information on this policy is available at: *gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication*

A. STUDENT'S DECLARATION

I declare that:

- the information below is accurate
- the publication(s) below meets the requirements to be included in the thesis
- The advisory committee has met and agreed to the inclusion of the publication(s) in the student's thesis
- All co-authors of the publication(s) have reviewed the information below and have agreed to its veracity.

Co-Author Authorisation forms for each co-author are attached.

Student's name	Student's signature	Date *(dd/mm/yy)*
Fraser Allison		14/01/20

PRINCIPAL SUPERVISOR'S DECLARATION

Supervisor's name	Supervisor's signature	Date *(dd/mm/yy)*
Assoc. Prof. Martin Gibbs		14/01/2020

B. PUBLICATION DETAILS *(to be completed by the student)*

Click on the "+" button in the bottom right corner to enter multiple publications

Full title	How players speak to an intelligent game character using natural language messages		
Authors	Fraser Allison, Ewa Luger, Katja Hofmann		
Student's contribution (%)	80%	Volume/page numbers	Vol. 4, No. 2, pp. 1-50
Journal or book name	Transactions of the Digital Games Research Association (ToDIGRA)		
Status	☐Accepted and In-press ☒ Published ☐ In progress		Date accepted/ published 07/11/2018

[22] The University of Melbourne. (2019). *Graduate Research Training Policy (MPF1321)*. Retrieved from https://policy.unimelb.edu.au/MPF1321

B. PUBLICATION DETAILS *(to be completed by the student)* Click on the "+" button in the bottom right corner to enter multiple publications			
Full title	Word Play: A history of voice interaction in digital games		
Authors	Fraser Allison, Marcus Carter, Martin Gibbs		
Student's contribution (%)	80%	Volume/page numbers	Published online: doi:10.1177/1555412017746305 (Print publication in press)
Journal or book name	Games and Culture		
Status	☐Accepted and In-press ☒ Published ☐ In progress		Date accepted/ published 14/12/2017

B. PUBLICATION DETAILS *(to be completed by the student)* Click on the "+" button in the bottom right corner to enter multiple publications			
Full title	Design patterns for voice interaction in games		
Authors	Fraser Allison, Marcus Carter, Martin Gibbs, Wally Smith		
Student's contribution (%)	80%	Volume/page numbers	pp. 5–17
Journal or book name	Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play		
Status	☐Accepted and In-press ☒ Published ☐ In progress		Date accepted/ published 28/10/2018

B. PUBLICATION DETAILS *(to be completed by the student)* Click on the "+" button in the bottom right corner to enter multiple publications			
Full title	Frame analysis of voice interaction gameplay		
Authors	Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, Martin Gibbs		
Student's contribution (%)	80%	Volume/page numbers	pp. 393:1–393:14
Journal or book name	Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems		
Status	☐Accepted and In-press ☒ Published ☐ In progress		Date accepted/ published 04/05/2019

Co-author authorisation form

THE UNIVERSITY OF MELBOURNE

All co-authors must complete this form. By signing below co-authors agree to the listed publication being included in the student's thesis and that the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

In cases where all members of a large consortium are listed as authors of a publication, only those that actively collaborated with the student on material contained within the thesis should complete this form. This form is to be used in conjunction with the *Declaration for a thesis with publication form*.

Students must submit this form, along with the *Declaration for thesis with publication form*, when the thesis is submitted to the Thesis Examination System: *https://tes.app.unimelb.edu.au/*

Further information on this policy and the requirements is available at:
gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication

A. PUBLICATION DETAILS *(to be completed by the student)*		
Full title	How Players Speak to an Intelligent Game Character Using Natural Language Messages	
Authors	Fraser Allison, Ewa Luger, & Katja Hofmann	
Student's contribution (%)	80%	
Journal or book name	Transactions of the Digital Games Research Association (ToDIGRA)	
Volume/page numbers	Vol. 4, No. 2, pp. 1-50	
Status	☐ Accepted and In-press ☒ Published	Date accepted/published 7 Nov 2018

B. CO-AUTHOR'S DECLARATION *(to be completed by the collaborator)*

I authorise the inclusion of this publication in the student's thesis and certify that:

- the declaration made by the student on the *Declaration for a thesis with publication form* correctly reflects the extent of the student's contribution to this work;
- the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

Co-author's name	Co-author's signature	Date *(dd/mm/yy)*
Ewa Luger		07.01.19

THE UNIVERSITY OF MELBOURNE

Co-author authorisation form

All co-authors must complete this form. By signing below co-authors agree to the listed publication being included in the student's thesis and that the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

In cases where all members of a large consortium are listed as authors of a publication, only those that actively collaborated with the student on material contained within the thesis should complete this form. This form is to be used in conjunction with the *Declaration for a thesis with publication form*.

Students must submit this form, along with the *Declaration for thesis with publication form*, when the thesis is submitted to the Thesis Examination System: *https://tes.app.unimelb.edu.au/*

Further information on this policy and the requirements is available at:
gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication

A. PUBLICATION DETAILS *(to be completed by the student)*

Full title	How Players Speak to an Intelligent Game Character Using Natural Language Messages
Authors	Fraser Allison, Ewa Luger, & Katja Hofmann
Student's contribution (%)	80%
Journal or book name	Transactions of the Digital Games Research Association (ToDIGRA)
Volume/page numbers	Vol. 4, No. 2, pp. 1-50
Status	☐ Accepted and In-press ☒ Published — Date accepted/published 7 Nov 2018

B. CO-AUTHOR'S DECLARATION *(to be completed by the collaborator)*

I authorise the inclusion of this publication in the student's thesis and certify that:

- the declaration made by the student on the *Declaration for a thesis with publication form* correctly reflects the extent of the student's contribution to this work;
- the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

Co-author's name	Co-author's signature	Date *(dd/mm/yy)*
Katja Hofmann		15/01/20

THE UNIVERSITY OF MELBOURNE

Co-author authorisation form

All co-authors must complete this form. By signing below co-authors agree to the listed publication being included in the student's thesis and that the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

In cases where all members of a large consortium are listed as authors of a publication, only those that actively collaborated with the student on material contained within the thesis should complete this form. This form is to be used in conjunction with the *Declaration for a thesis with publication form*.

Students must submit this form, along with the *Declaration for thesis with publication form*, when the thesis is submitted to the Thesis Examination System: *https://tes.app.unimelb.edu.au/*

Further information on this policy and the requirements is available at: *gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication*

A. PUBLICATION DETAILS *(to be completed by the student)*

Full title	Word play: A history of voice interaction in digital games
Authors	Fraser Allison, Marcus Carter, Martin Gibbs
Student's contribution (%)	80%
Journal or book name	Games and Culture
Volume/page numbers	Published online: doi:10.1177/1555412017746305. (Print edition in press.)
Status	☐ Accepted and In-press ☒ Published ☐ In progress — Date accepted/published: 14 Dec 2017

B. CO-AUTHOR'S DECLARATION *(to be completed by the collaborator)*

I authorise the inclusion of this publication in the student's thesis and certify that:

- the declaration made by the student on the *Declaration for a thesis with publication form* correctly reflects the extent of the student's contribution to this work;
- the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

Co-author's name	Co-author's signature	Date *(dd/mm/yy)*
Marcus Carter		13/11/19

The University of Melbourne
CRICOS Provider Number: 00116K

Last Updated 19 October 2017

THE UNIVERSITY OF MELBOURNE

Co-author authorisation form

All co-authors must complete this form. By signing below co-authors agree to the listed publication being included in the student's thesis and that the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

In cases where all members of a large consortium are listed as authors of a publication, only those that actively collaborated with the student on material contained within the thesis should complete this form. This form is to be used in conjunction with the *Declaration for a thesis with publication form*.

Students must submit this form, along with the *Declaration for thesis with publication form*, when the thesis is submitted to the Thesis Examination System: *https://tes.app.unimelb.edu.au/*

Further information on this policy and the requirements is available at:
gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication

A. PUBLICATION DETAILS *(to be completed by the student)*		
Full title	Word play: A history of voice interaction in digital games	
Authors	Fraser Allison, Marcus Carter, Martin Gibbs	
Student's contribution (%)	80%	
Journal or book name	Games and Culture	
Volume/page numbers	Published online: doi:10.1177/1555412017746305. (Print edition in press.)	
Status	☐ Accepted and In-press ■ Published ☐ In progress	Date accepted/published 14 Dec 2017

B. CO-AUTHOR'S DECLARATION *(to be completed by the collaborator)*

I authorise the inclusion of this publication in the student's thesis and certify that:

- the declaration made by the student on the *Declaration for a thesis with publication form* correctly reflects the extent of the student's contribution to this work;
- the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

Co-author's name	Co-author's signature	Date *(dd/mm/yy)*
Martin Gibbs		14/01/20

The University of Melbourne
CRICOS Provider Number: 00116K

Last Updated 19 October 2017

THE UNIVERSITY OF MELBOURNE

Co-author authorisation form

All co-authors must complete this form. By signing below co-authors agree to the listed publication being included in the student's thesis and that the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

In cases where all members of a large consortium are listed as authors of a publication, only those that actively collaborated with the student on material contained within the thesis should complete this form. This form is to be used in conjunction with the *Declaration for a thesis with publication form*.

Students must submit this form, along with the *Declaration for thesis with publication form*, when the thesis is submitted to the Thesis Examination System: *https://tes.app.unimelb.edu.au/*

Further information on this policy and the requirements is available at:
gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication

A. PUBLICATION DETAILS *(to be completed by the student)*		
Full title	Design patterns for voice interaction in games	
Authors	Fraser Allison, Marcus Carter, Martin Gibbs, Wally Smith	
Student's contribution (%)	80%	
Journal or book name	Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play	
Volume/page numbers	pp. 5–17.	
Status	☐ Accepted and In-press ■ Published ☐ In progress	Date accepted/published 28 Oct 2018

B. CO-AUTHOR'S DECLARATION *(to be completed by the collaborator)*

I authorise the inclusion of this publication in the student's thesis and certify that:

- the declaration made by the student on the *Declaration for a thesis with publication form* correctly reflects the extent of the student's contribution to this work;
- the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

Co-author's name	Co-author's signature	Date *(dd/mm/yy)*
Marcus Carter		13/11/19

The University of Melbourne
CRICOS Provider Number: 00116K

Last Updated 19 October 2017

Co-author authorisation form

THE UNIVERSITY OF MELBOURNE

All co-authors must complete this form. By signing below co-authors agree to the listed publication being included in the student's thesis and that the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

In cases where all members of a large consortium are listed as authors of a publication, only those that actively collaborated with the student on material contained within the thesis should complete this form. This form is to be used in conjunction with the *Declaration for a thesis with publication form*.

Students must submit this form, along with the *Declaration for thesis with publication form*, when the thesis is submitted to the Thesis Examination System: *https://tes.app.unimelb.edu.au/*

Further information on this policy and the requirements is available at:
gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication

A. PUBLICATION DETAILS *(to be completed by the student)*

Full title	Design patterns for voice interaction in games
Authors	Fraser Allison, Marcus Carter, Martin Gibbs, Wally Smith
Student's contribution (%)	80%
Journal or book name	Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play
Volume/page numbers	pp. 5–17.
Status	☐ Accepted and In-press ■ Published ☐ In progress — Date accepted/published: 28 Oct 2018

B. CO-AUTHOR'S DECLARATION *(to be completed by the collaborator)*

I authorise the inclusion of this publication in the student's thesis and certify that:

- the declaration made by the student on the *Declaration for a thesis with publication form* correctly reflects the extent of the student's contribution to this work;
- the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

Co-author's name	Co-author's signature	Date *(dd/mm/yy)*
MArtin Gibbs		14/01/20

The University of Melbourne
CRICOS Provider Number: 00116K

Last Updated 19 October 2017

THE UNIVERSITY OF MELBOURNE

Co-author authorisation form

All co-authors must complete this form. By signing below co-authors agree to the listed publication being included in the student's thesis and that the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

In cases where all members of a large consortium are listed as authors of a publication, only those that actively collaborated with the student on material contained within the thesis should complete this form. This form is to be used in conjunction with the *Declaration for a thesis with publication form*.

Students must submit this form, along with the *Declaration for thesis with publication form*, when the thesis is submitted to the Thesis Examination System: *https://tes.app.unimelb.edu.au/*

Further information on this policy and the requirements is available at:
gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication

A. PUBLICATION DETAILS *(to be completed by the student)*

Full title	Design patterns for voice interaction in games	
Authors	Fraser Allison, Marcus Carter, Martin Gibbs, Wally Smith	
Student's contribution (%)	80%	
Journal or book name	Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play	
Volume/page numbers	pp. 5–17.	
Status	☐ Accepted and In-press ☒ Published ☐ In progress	Date accepted/published 28 Oct 2018

B. CO-AUTHOR'S DECLARATION *(to be completed by the collaborator)*

I authorise the inclusion of this publication in the student's thesis and certify that:

- the declaration made by the student on the *Declaration for a thesis with publication form* correctly reflects the extent of the student's contribution to this work;
- the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

Co-author's name	Co-author's signature	Date *(dd/mm/yy)*
Wally Smith		07/01/2020

The University of Melbourne
CRICOS Provider Number: 00116K

Last Updated 19 October 2017

THE UNIVERSITY OF MELBOURNE

Co-author authorisation form

All co-authors must complete this form. By signing below co-authors agree to the listed publication being included in the student's thesis and that the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

In cases where all members of a large consortium are listed as authors of a publication, only those that actively collaborated with the student on material contained within the thesis should complete this form. This form is to be used in conjunction with the *Declaration for a thesis with publication form*.

Students must submit this form, along with the *Declaration for thesis with publication form*, when the thesis is submitted to the Thesis Examination System: *https://tes.app.unimelb.edu.au/*

Further information on this policy and the requirements is available at:
gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication

A. PUBLICATION DETAILS *(to be completed by the student)*

Full title	Frame analysis of voice interaction gameplay
Authors	Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, Martin Gibbs
Student's contribution (%)	80%
Journal or book name	Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems
Volume/page numbers	pp. 393:1–393:14.
Status	☐ Accepted and In-press ■ Published ☐ In progress — Date accepted/published: 4 May 2019

B. CO-AUTHOR'S DECLARATION *(to be completed by the collaborator)*

I authorise the inclusion of this publication in the student's thesis and certify that:

- the declaration made by the student on the *Declaration for a thesis with publication form* correctly reflects the extent of the student's contribution to this work;
- the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

Co-author's name	Co-author's signature	Date *(dd/mm/yy)*
JOSHUA NEWN		20/01/20

The University of Melbourne
CRICOS Provider Number: 00116K

Last Updated 19 October 2017

THE UNIVERSITY OF MELBOURNE

Co-author authorisation form

All co-authors must complete this form. By signing below co-authors agree to the listed publication being included in the student's thesis and that the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

In cases where all members of a large consortium are listed as authors of a publication, only those that actively collaborated with the student on material contained within the thesis should complete this form. This form is to be used in conjunction with the *Declaration for a thesis with publication form*.

Students must submit this form, along with the *Declaration for thesis with publication form*, when the thesis is submitted to the Thesis Examination System: *https://tes.app.unimelb.edu.au/*

Further information on this policy and the requirements is available at:
gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication

A. PUBLICATION DETAILS *(to be completed by the student)*		
Full title	Frame analysis of voice interaction gameplay	
Authors	Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, Martin Gibbs	
Student's contribution (%)	80%	
Journal or book name	Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems	
Volume/page numbers	pp. 393:1–393:14.	
Status	☐ Accepted and In-press ■ Published ☐ In progress	Date accepted/published 4 May 2019

B. CO-AUTHOR'S DECLARATION *(to be completed by the collaborator)*

I authorise the inclusion of this publication in the student's thesis and certify that:

- the declaration made by the student on the *Declaration for a thesis with publication form* correctly reflects the extent of the student's contribution to this work;
- the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

Co-author's name	Co-author's signature	Date *(dd/mm/yy)*
Wally Smith		07/01/2020

The University of Melbourne
CRICOS Provider Number: 00116K

Last Updated 19 October 2017

THE UNIVERSITY OF MELBOURNE

Co-author authorisation form

All co-authors must complete this form. By signing below co-authors agree to the listed publication being included in the student's thesis and that the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

In cases where all members of a large consortium are listed as authors of a publication, only those that actively collaborated with the student on material contained within the thesis should complete this form. This form is to be used in conjunction with the *Declaration for a thesis with publication form*.

Students must submit this form, along with the *Declaration for thesis with publication form*, when the thesis is submitted to the Thesis Examination System: *https://tes.app.unimelb.edu.au/*

Further information on this policy and the requirements is available at:
gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication

A. PUBLICATION DETAILS *(to be completed by the student)*

Full title	Frame analysis of voice interaction gameplay	
Authors	Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, Martin Gibbs	
Student's contribution (%)	80%	
Journal or book name	Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems	
Volume/page numbers	pp. 393:1–393:14.	
Status	☐ Accepted and In-press ☒ Published ☐ In progress	Date accepted/published 4 May 2019

B. CO-AUTHOR'S DECLARATION *(to be completed by the collaborator)*

I authorise the inclusion of this publication in the student's thesis and certify that:

- the declaration made by the student on the *Declaration for a thesis with publication form* correctly reflects the extent of the student's contribution to this work;
- the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

Co-author's name	Co-author's signature	Date *(dd/mm/yy)*
Marcus Carter		13/11/19

The University of Melbourne
CRICOS Provider Number: 00116K

Last Updated 19 October 2017

THE UNIVERSITY OF MELBOURNE

Co-author authorisation form

All co-authors must complete this form. By signing below co-authors agree to the listed publication being included in the student's thesis and that the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

In cases where all members of a large consortium are listed as authors of a publication, only those that actively collaborated with the student on material contained within the thesis should complete this form. This form is to be used in conjunction with the *Declaration for a thesis with publication form*.

Students must submit this form, along with the *Declaration for thesis with publication form*, when the thesis is submitted to the Thesis Examination System: *https://tes.app.unimelb.edu.au/*

Further information on this policy and the requirements is available at:
gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication

A. PUBLICATION DETAILS *(to be completed by the student)*

Full title	Frame analysis of voice interaction gameplay	
Authors	**Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, Martin Gibbs**	
Student's contribution (%)	80%	
Journal or book name	Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems	
Volume/page numbers	pp. 393:1–393:14.	
Status	☐ Accepted and In-press ■ Published ☐ In progress	Date accepted/published 4 May 2019

B. CO-AUTHOR'S DECLARATION *(to be completed by the collaborator)*

I authorise the inclusion of this publication in the student's thesis and certify that:

- the declaration made by the student on the *Declaration for a thesis with publication form* correctly reflects the extent of the student's contribution to this work;
- the student contributed greater than 50% of the content of the publication and is the "primary author" ie. the student was responsible primarily for the planning, execution and preparation of the work for publication.

Co-author's name	Co-author's signature	Date *(dd/mm/yy)*
Martin Gibbs		14/01/20

The University of Melbourne
CRICOS Provider Number: 00116K

Last Updated 19 October 2017

11 Appendix B

Table 11.1 lists the 471 videogames that were surveyed for Chapters 5 and 6. It includes each game's name and year of publication as well as the platforms for which it was developed. Citations are provided after the title for games that were developed as part of an academic research project.

Table 11.1 Voice games surveyed for Chapters 5 and 6

Game	Year	Platform
Voice-Chess (Reddy et al., 1973)	1973	Hearsay-I
Atari Voice Commander	1983	Atari 2600 / Atari 5200
SuperFly	1983	TI-99/4A MBX
Sewermania	1983	TI-99/4A MBX
I'm Hiding	1983	TI-99/4A MBX
Space Bandits	1983	TI-99/4A MBX
Terry Turtle's Adventure	1983	TI-99/4A MBX
Championship Baseball	1983	TI-99/4A MBX
Raid on Bungeling Bay	1985	Famicom
Star Luster	1985	Famicom
Thayer's Quest	1985	RDI Halcyon
Kid Icarus [Hikari Shinwa: Parutena no Kagami]	1986	Famicom
The Legend of Zelda [Zeruda no Densetsu]	1986	Famicom
Doraemon	1986	Famicom
Star Soldier	1986	Famicom
Super Pitfall [Atlantis no Nazo]	1986	Famicom
Takeshi No Chousenjou	1986	Famicom
Echelon	1987	Commodore 64
Kid Niki: Radical Ninja [Kaiketsu Yanchamaru]	1987	Famicom
Apple Town Monogatari: Little Computer People	1987	Famicom
Karaoke Studio	1987	Famicom
Kamen Rider Club	1988	Famicom
Little Ninja Brothers [Super Chinese 2]	1989	Famicom
Bakushou!! Jinsei Gekijou	1989	Famicom
Bakushou!! Jinsei Gekijou 2	1991	Famicom
Command: Aces of the Deep	1995	PC
Hey You, Pikachu!	1998	Nintendo 64
Seaman	1999	DreamCast, PlayStation 2
Densha De Go! 64	1999	Nintendo 64
MIND-WARPING (Starner et al., 2000)	2000	PC
Dream Audition	2000	PlayStation 2

Dream Audition 2	2000	PlayStation 2
Untitled (Igarashi & Hughes, 2001)	2001	PC
Greatest Striker [International League Soccer]	2001	PlayStation 2
Yoake no Mariko	2001	PlayStation 2
Yoake no Mariko Second Act	2001	PlayStation 2
Dream Audition 3	2001	PlayStation 2
Dream Audition Super Hit Disc 1	2001	PlayStation 2
Dream Audition Super Hit Disc 2	2001	PlayStation 2
Game Commander	2002	PC
Microsoft Game Voice	2002	PC
Star Trek: Bridge Commander (with software)	2002	PC
SOCOM	2002	PlayStation 2
Mission Rehearsal Exercise System (Hill et al., 2003)	2003	PC
Lifeline	2003	PlayStation 2
SOCOM 2	2003	PlayStation 2
Deka Voice	2003	PlayStation 2
Karaoke Revolution	2003	PlayStation 2
N.U.D.E.@ Natural Ultimate Digital Experiment	2003	Xbox
Rainbow Six 3	2003	Xbox
SWAT: Global Strike Team	2003	PlayStation 2, Xbox
Manhunt	2003	PlayStation 2, Xbox
Mario Party 6	2004	GameCube
Unreal Tournament 2004	2004	PC
Get On Da Mic	2004	PlayStation 2
Karaoke Revolution 3	2004	PlayStation 2
SingStar	2004	PlayStation 2
SingStar Party	2004	PlayStation 2
Karaoke Revolution 2	2004	PlayStation 2, Xbox
Densetsu no Quiz Ou Ketteisen	2005	GameCube
Mario Party 7	2005	GameCube
Brain Age / Dr Kawashima's Brain Training	2005	Nintendo DS
Nintendogs	2005	Nintendo DS
WarioWare Touched!	2005	Nintendo DS
Phoenix Wright: Ace Attorney	2005	Nintendo DS
Untitled (Zeng et al., 2005)	2005	PC
UA-Chess (Grammenos et al., 2009)	2005	PC
Neverwinter Nights variant (Gorniak & Roy, 2005)	2005	PC
Tactical Language and Culture Training System (W. L. Johnson, 2007)	2005	PC
SOCOM 3	2005	PlayStation 2
SingStar The Dome	2005	PlayStation 2
NASCAR 06: Total Team Control	2005	PlayStation 2, Xbox

Karaoke Revolution Party	2005	GameCube, PlayStation 2, Xbox
Greg Hastings Tournament Paintball Max'd	2005	Xbox
Odama	2006	GameCube
Nintendogs: Dalmatian	2006	Nintendo DS
Phoenix Wright: Ace Attorney - Justice For All	2006	Nintendo DS
Untitled (Tse et al., 2007)	2006	PC
Tetris variant (Sporka et al., 2006)	2006	PC
Access Invaders (Grammenos et al., 2006)	2006	PC
Penguin Quart (Nagy, 2006)	2006	PC
Ford Game	2006	PC
Racing Pitch	2006	PC
SOCOM Combined Assault	2006	PlayStation 2
Karaoke Revolution Country	2006	PlayStation 2
SingStar Anthems	2006	PlayStation 2
SingStar Deutsch Rock-Pop	2006	PlayStation 2
SingStar La Edad de Oro del Pop Espanol	2006	PlayStation 2
SingStar Legends	2006	PlayStation 2
SingStar Rocks!	2006	PlayStation 2
SingStar top.it	2006	PlayStation 2
Rainbow Six Vegas	2006	PlayStation 3, Xbox 360
Apollo Justice: Ace Attorney	2007	Nintendo DS
The Legend of Zelda: Phantom Hourglass	2007	Nintendo DS
The World Ends With You	2007	Nintendo DS
Mario Party DS	2007	Nintendo DS
Phoenix Wright: Ace Attorney - Trials and Tribulations	2007	Nintendo DS
Façade variant (Dow et al., 2007)	2007	PC
Visual Music (Parker & Heerema, 2008)	2007	PC
Microphone Game	2007	PC
Seaman 2	2007	PlayStation 2
Karaoke Revolution: American Idol	2007	PlayStation 2
SingStar Amped	2007	PlayStation 2
SingStar Apres-Ski Party	2007	PlayStation 2
SingStar Back to the 80s	2007	PlayStation 2
SingStar Bollywood	2007	PlayStation 2
SingStar Deutsch Rock-Pop Vol. 2	2007	PlayStation 2
SingStar Die Toten Hosen	2007	PlayStation 2
SingStar Italian Party	2007	PlayStation 2
SingStar Norsk pa Norsk	2007	PlayStation 2
SingStar Pop Hits	2007	PlayStation 2
SingStar Pop Hits 2	2007	PlayStation 2
SingStar R&B	2007	PlayStation 2
SingStar Rock Ballads	2007	PlayStation 2
SingStar Svenska Hits Schlager	2007	PlayStation 2

Sing It: High School Musical	2007	PlayStation 2, Wii
Rock Band	2007	PlayStation 2, PlayStation 3, Xbox 360, Wii
Boogie	2007	Nintendo DS, PlayStation 2, Wii
Rock Revolution	2008	Nintendo DS
Playmancer (Conconi et al., 2008)	2008	PC
Word War (McGraw, Yoshimoto, et al., 2009)	2008	PC
Sudoku Access (Norte & Lobo, 2008)	2008	PC
Terrestrial Invaders (Grammenos et al., 2009)	2008	PC
BloodLezz (Kiselkov et al., 2008)	2008	PC
The Wind Sea Adventure	2008	PC
Disney Sing It	2008	PC, PlayStation 3, PlayStation 2, Wii, Xbox 360
Tom Clancy's EndWar	2008	PC, PlayStation 3, Xbox 360
SingStar 90s	2008	PlayStation 2
SingStar Boy Bands vs Girl Bands	2008	PlayStation 2
SingStar Clásicos	2008	PlayStation 2
SingStar Country	2008	PlayStation 2
SingStar Hottest Hits	2008	PlayStation 2
SingStar Italian Party 2	2008	PlayStation 2
SingStar Operacion Triunfo	2008	PlayStation 2
SingStar Pop Hits 3	2008	PlayStation 2
SingStar Schlager	2008	PlayStation 2
SingStar Store Dromme	2008	PlayStation 2
SingStar SuomiRock	2008	PlayStation 2
SingStar Turkish Party	2008	PlayStation 2
SingStar ABBA	2008	PlayStation 2, PlayStation 3
SingStar Morangos com Acucar	2008	PlayStation 2, PlayStation 3
SingStar Pop Vol. 2	2008	PlayStation 2, PlayStation 3
SingStar Singalong with Disney	2008	PlayStation 2, PlayStation 3
SingStar Summer Party	2008	PlayStation 2, PlayStation 3
SingStar Vol. 2	2008	PlayStation 2, PlayStation 3
SOCOM Confrontation	2008	PlayStation 3
SingStar	2008	PlayStation 3
SingStar Vol. 3	2008	PlayStation 3
Disney Sing It - High School Musical 3: Senior Year	2008	PlayStation 3, PlayStation 2, Wii, Xbox 360
Rock Band 2	2008	PlayStation 3, Xbox 360, Wii
Karaoke Revolution: American Idol Encore 2	2008	PlayStation 3, Xbox 360, Wii
Guitar Hero: World Tour	2008	PlayStation 2, PlayStation 3, Xbox 360, Wii
Karaoke Revolution: American Idol Encore	2008	PlayStation 2, PlayStation 3, Xbox 360, Wii
Boogie Superstar	2008	Wii
Karaoke Joysound	2008	Wii

Lips	2008	Xbox 360
Rainbow Six Vegas 2	2008	PC, PlayStation 3, Xbox 360
Audiotennis	2009	Mobile, PC
Untitled (Nanjo et al., 2009)	2009	PC
Rabbit Run (O'Donovan et al., 2009)	2009	PC
Voice Race (McGraw, Gruenstein, et al., 2009)	2009	PC
VoiceAttack	2009	PC
Voice Command Duck	2009	PC
World Tour by Goalmaniac.com	2009	PC
Micromissile	2009	PC
Prospero – Rosetta World (Jewell, 2011)	2009	PC
HAWX	2009	PC, Xbox 360
SingStar Chartbreaker	2009	PlayStation 2, PlayStation 3
Singstar Latino	2009	PlayStation 2, PlayStation 3
SingStar Made in Germany	2009	PlayStation 2, PlayStation 3
SingStar Mallorca Party	2009	PlayStation 2, PlayStation 3
SingStar Mecano	2009	PlayStation 2, PlayStation 3
SingStar Motown	2009	PlayStation 2, PlayStation 3
SingStar Polskie Hity	2009	PlayStation 2, PlayStation 3
SingStar Pop Edition	2009	PlayStation 2, PlayStation 3
SingStar Queen	2009	PlayStation 2, PlayStation 3
SingStar Studio 100	2009	PlayStation 2, PlayStation 3
SingStar SuomiPop	2009	PlayStation 2, PlayStation 3
SingStar Take That	2009	PlayStation 2, PlayStation 3
SingStar Vasco	2009	PlayStation 2, PlayStation 3
Disney Sing It: Pop Hits	2009	PlayStation 2, PlayStation 3, Wii
Guitar Hero: Van Halen	2009	PlayStation 3, Xbox 360, Wii
Lego Rock Band	2009	PlayStation 3, Xbox 360, Wii
Rock Band: The Beatles	2009	PlayStation 3, Xbox 360, Wii
Karaoke Revolution (2009)	2009	PlayStation 3, Xbox 360, Wii
Guitar Hero 5	2009	PlayStation 2, PlayStation 3, Xbox 360, Wii
Guitar Hero: Metallica	2009	PlayStation 2, PlayStation 3, Xbox 360, Wii
Guitar Hero: Smash Hits	2009	PlayStation 2, PlayStation 3, Xbox 360, Wii
We Sing	2009	Wii
Lips: Canta en Espanol	2009	Xbox 360
Lips: No1 Hits	2009	Xbox 360
Rock Band Reloaded	2010	Mobile
Dragon Gaming Speech Pack	2010	PC
Jet Pass	2010	PC
Gamic	2010	PC
SingStar Chart Hits	2010	PlayStation 2, PlayStation 3

SingStar Die größten Solokünstler	2010	PlayStation 2, PlayStation 3
SingStar Fussballhits	2010	PlayStation 2, PlayStation 3
SingStar The Wiggles	2010	PlayStation 2, PlayStation 3
SingStar Apres-Ski Party 2	2010	PlayStation 3
SingStar Dance	2010	PlayStation 3
SingStar Guitar	2010	PlayStation 3
SingStar SuomiHitit	2010	PlayStation 3
Disney Sing It: Family Hits	2010	PlayStation 3, Wii
Disney Sing It: Party Hits	2010	PlayStation 3, Wii
Guitar Hero: WoR	2010	PlayStation 3, Xbox 360, Wii
Rock Band: Green Day	2010	PlayStation 3, Xbox 360, Wii
Rock Band 3	2010	PlayStation 3, Xbox 360, Wii
Karaoke Revolution: Glee	2010	Wii
We Sing Encore	2010	Wii
We Sing Robbie Williams	2010	Wii
Kinectimals	2010	Xbox 360
Lips: I Love the 80s	2010	Xbox 360
Lips: Party Classics	2010	Xbox 360
Cat Scream [1]	2011	Mobile
Pah!	2011	Mobile
Voice Race	2011	Mobile
Scream'N'Run	2011	Mobile
Nintendogs + Cats	2011	Nintendo 3DS
Solis' Curse (Neto et al., 2011)	2011	PC
Tetris variant (Yuan & Fan, 2011)	2011	PC
Voice Game Controller (Harada et al., 2011)	2011	PC
REAP.PT (Silva et al., 2011)	2011	PC
SingStar SuomiSuosikit	2011	PlayStation 3
Karaoke Revolution: Glee 2	2011	Wii
We Sing Deutsche Hits	2011	Wii
We Sing Down Under	2011	Wii
We Sing Rock!	2011	Wii
We Sing UK Hits	2011	Wii
Karaoke Revolution: Glee 3	2011	Wii, Xbox 360
Dance Central 2	2011	Xbox 360
Halo: Combat Evolved Anniversary	2011	Xbox 360
Forza Motorsport 4	2011	Xbox 360
Kinect Sports: Season Two	2011	Xbox 360
Sing! Karaoke	2012	Mobile
Market Game / Farm Game (Kumar et al., 2012)	2012	PC
What Did You Say???	2012	PC
UltraStar	2012	PC
Binary Domain	2012	PC, PlayStation 3, Xbox 360
SingStar SuomiHuiput	2012	PlayStation 3

Rock Band Blitz	2012	PlayStation 3, Xbox 360
We Sing 80s	2012	Wii
We Sing Pop!	2012	Wii
Sing Party	2012	WiiU
Wii Karaoke U	2012	WiiU
Mass Effect 3	2012	Xbox 360
Voice Fighter	2013	Mobile
ATC Voice	2013	Mobile
Don't Pop It!	2013	Mobile
Edukitty ABC	2013	Mobile
Mario Party: Island Tour	2013	Nintendo 3DS
Phoenix Wright: Ace Attorney - Dual Destinies	2013	Nintendo 3DS
GeMiNI (Teófilo et al., 2013)	2013	PC
sPeAK-MAN (Tan et al., 2013)	2013	PC
Rally Navigator (Janicki & Wawer, 2013)	2013	PC
The Voice in the Garden (Chou, 2013)	2013	PC
Untitled (Krause et al., 2013)	2013	PC
PowerFall (Vieira et al., 2014)	2013	PC
Speech Adventure (Rubin & Kurniawan, 2013)	2013	PC
Blowing Blues (X. Li, 2013)	2013	PC
EVEIL-3D (Roy, 2016)	2013	PC
Skyrim: Legendary Edition	2013	PC, Xbox 360
Ryse: Son of Rome	2013	PC, Xbox One
The Howler	2013	Mobile, PC
SingStar SuomiHelmet	2013	PlayStation 3
Dead Space 3	2013	Xbox 360
Splinter Cell: Blacklist	2013	Xbox 360
Forza Motorsport 5	2013	Xbox One
Rocksmith 2014	2013	Xbox One
Battlefield 4	2013	Xbox One
FIFA 14	2013	Xbox One
Dead Rising 3	2013	Xbox One
Tomb Raider: Definitive Edition	2013	PlayStation 4, Xbox One
Need For Speed Rivals	2013	PlayStation 4, Xbox One
NBA2K14	2013	PlayStation 4, Xbox One
Zoo Tycoon	2013	Xbox 360, Xbox One
Call of Duty: Ghosts	2013	Xbox 360, Xbox One
Crash of Ah	2014	Mobile
Yyaah!	2014	Mobile
Phonics With Phonzy	2014	Mobile
Flappy Voice (Lan et al., 2014)	2014	PC
Talking to Teo (Navarro-Newball et al., 2014)	2014	PC
ControlMyJoystick	2014	PC
Elite: Dangerous	2014	PC

SingStar SuomiBileet	2014	PlayStation 3, PlayStation 4
SingStar Ultimate Party	2014	PlayStation 4
Dance Central Spotlight	2014	Xbox One
Voice Commander	2014	Xbox One
Kinect Sports Rivals	2014	Xbox One
Project Spark	2014	Xbox One
FIFA 15	2014	Xbox One
D4: Dark Dreams Don't Die	2014	Xbox One
Wolfenstein: The New Order	2014	PlayStation 4, Xbox One
NBA2K15	2014	PlayStation 4, Xbox One
Plants vs. Zombies: Garden Warfare	2014	PC, PlayStation 4, Xbox One
Alien: Isolation	2014	PC, PlayStation 4, Xbox One
Forza Horizon 2	2014	Xbox 360, Xbox One
Call of Duty: Advanced Warfare	2014	Xbox 360, Xbox One
Dragon Age: Inquisition	2014	Xbox 360, Xbox One
Bouncy Balls	2014	
Rock Paper Scissors Lizard Spock	2015	Alexa
Earplay	2015	Alexa
The Wayne Investigation	2015	Alexa
The Magic Door	2015	Alexa
Jeopardy!	2015	Alexa
Amazing Word Master Game	2015	Alexa
Articulation Station	2015	Mobile
Mayday! Deep Space	2015	Mobile
Ah! Bird	2015	Mobile
Shouting Blocks	2015	Mobile
Untitled (Qingtang et al., 2015)	2015	PC
VoiceBot	2015	PC
Alien Escape	2015	PC
There Came an Echo	2015	PC
Bot Colony	2015	PC
Neuro Ski	2015	PC
In Verbis Virtus	2015	PC
Mic Frog	2015	PC
Rock Band 4	2015	PlayStation 4, Xbox One
FIFA 16	2015	Xbox One
NBA2K16	2015	PlayStation 4, Xbox One
F1 2015	2015	PlayStation 4, Xbox One
WhereInTheWorld	2016	Alexa
Would You Rather For Family	2016	Alexa
Millionaire Quiz Game	2016	Alexa
Escape the Room	2016	Alexa
Tic Tac Toe	2016	Alexa
Akinator	2016	Alexa

Beat the Dealer (blackjack)	2016	Alexa
Memory Match	2016	Alexa
Hunt the Yeti	2016	Alexa
Dungeon Adventure	2016	Alexa
Vocal Warrior	2016	Mobile
Blow Fly	2016	Mobile
Oáto	2016	Mobile, PC
Sustained Vowel Game (Lopes et al., 2016)	2016	PC
Be Herd	2016	PC
Moon Cypher	2016	PC
Northern Lights	2016	PC
Plan Be	2016	PC
Get Me Out Of Here	2016	PC
Full Voice Throttle	2016	PC
Emoji Scream	2016	PC
Super Lamaze Coach	2016	PC
Stay With Me	2016	PC
Chant Savant	2016	PC
Are You With Us	2016	PC
Dandelion	2016	PC
EnigmBox	2016	PC
Voice of the Citizens	2016	PC
Voice Bounce	2016	PC
Forest Friends VR	2016	VR
Forza Horizon 3	2016	Xbox One
Song Quiz	2017	Alexa
Deal or No Deal Game	2017	Alexa
True or False	2017	Alexa
Volley Trivia Showdown Game	2017	Alexa
Yes Sire	2017	Alexa
Colossal Cave Adventure	2017	Alexa
Crack Me!	2017	Mobile
Gladiation	2017	Mobile
Crew Chief for Project Cars	2017	Mobile
Tamadog	2017	Mobile
Voice Volume Catcher	2017	Mobile
Citymals	2017	Mobile
Outloud Citymals	2017	Mobile
Gold Miner Talk	2017	Mobile
8 Eighth Note - Scream Go	2017	Mobile
Big Mouth	2017	Mobile
Bike Scream	2017	Mobile
Bird Scream	2017	Mobile
Bird Scream Go	2017	Mobile

Cat Scream [2]	2017	Mobile
Chicken Go Scream	2017	Mobile
Chicken Go Scream: Voice Jump	2017	Mobile
Chicken Scream	2017	Mobile
Chicken Scream Go 2017	2017	Mobile
Don't Scream Chicken	2017	Mobile
Don't Stop Eighth Note	2017	Mobile
eight note game adventure 2017	2017	Mobile
Eight Note - Voice Jump	2017	Mobile
Eighth Note	2017	Mobile
Eighth Note - Scream Go 2	2017	Mobile
Eighth Note 3D - Don't Stop!	2017	Mobile
Flapsonic	2017	Mobile
Flippy Bird 2	2017	Mobile
Franky Poof Fly	2017	Mobile
Franky Poof Space	2017	Mobile
Furball	2017	Mobile
Gingerliness	2017	Mobile
Husky Scream Hero	2017	Mobile
Monkey Scream & Sing Go	2017	Mobile
Naughty Bunny Rush	2017	Mobile
New Eighth Super Note	2017	Mobile
Ninja Scream Run	2017	Mobile
Note Bird	2017	Mobile
Oopstacles	2017	Mobile
Pa-pa Jump	2017	Mobile
Parrot Scream	2017	Mobile
Pepe Scream	2017	Mobile
Rambler 2	2017	Mobile
Roaring Tone	2017	Mobile
Robo Run	2017	Mobile
Robot Scream	2017	Mobile
Scream Chicken 2	2017	Mobile
Scream Dog	2017	Mobile
Scream Dog Go: Eighth Note	2017	Mobile
Scream Ducks	2017	Mobile
Scream Flying - Eighth Note Go	2017	Mobile
Scream Game: Flying Bird	2017	Mobile
Scream Go	2017	Mobile
Scream Go Crazy Capsule	2017	Mobile
Scream Go Don't Stop Eighth Note	2017	Mobile
Scream Go Gorilla	2017	Mobile
Scream Go Hero: Eighth Note	2017	Mobile
Scream Go Stickman	2017	Mobile

Scream Go: Amazing Spider Kid	2017	Mobile
Scream Go: Chicken Scream	2017	Mobile
Scream Go: Scream Cowboy	2017	Mobile
Scream Go: Scream Jump Dog	2017	Mobile
Scream Grand Spiral	2017	Mobile
Scream N Shout	2017	Mobile
Scream Puppy	2017	Mobile
Scream Runner	2017	Mobile
Scream Touch Jump	2017	Mobile
Screamy Chicken!	2017	Mobile
Screamy Ski	2017	Mobile
Shout	2017	Mobile
Sleepy Bird	2017	Mobile
Sponge Scream	2017	Mobile
Sub Wars - Voice Control	2017	Mobile
Super Stick Man	2017	Mobile
Surf Man	2017	Mobile
Toubbi Scream Go	2017	Mobile
Untitled	2017	Mobile
Vocal Tamago	2017	Mobile
Yasuhati	2017	Mobile
Yasuhati / With your voice!	2017	Mobile
Yasuhati Cat Scream	2017	Mobile
Yelling Go	2017	Mobile
Yelling GO 2 - Scream Control	2017	Mobile
ZigZag Scream: Blocky Animals	2017	Mobile
Color Scream	2017	Mobile
Crazy Sink Pirates	2017	Mobile
Read With Phonzy	2017	Mobile
Vocaline	2017	Mobile
Mr. Pet VR	2017	Mobile, VR
Mario Party: The Top 100	2017	Nintendo 3DS
Sonora (Umanski et al., 2017)	2017	PC
Untitled (Lv et al., 2017)	2017	PC
Untitled (Cler et al., 2017)	2017	PC
VoiceWarrior	2017	PC
Surrealist Siri	2017	PC
Scream Train	2017	PC
Garden of Silence	2017	PC
Voice of God	2017	PC
Soundquake	2017	PC
Blow or Die!!!	2017	PC
Insane Screaming Mayhem	2017	PC
Jet	2017	PC

Resonance - The Lost Score	2017	PC
Slavery & John Corp	2017	PC
Call of the Zombie	2017	PC
Plurality Spring	2017	PC
Oh No! Seagulls!	2017	PC
Be A Man: Do The Right Thing	2018	Mobile
Scream n Move 18	2018	Mobile
Voice Drone Fly	2018	Mobile
Voice Dunk	2018	Mobile
The Broken Seal	2018	VR
Soundspeed	unk.	PC