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

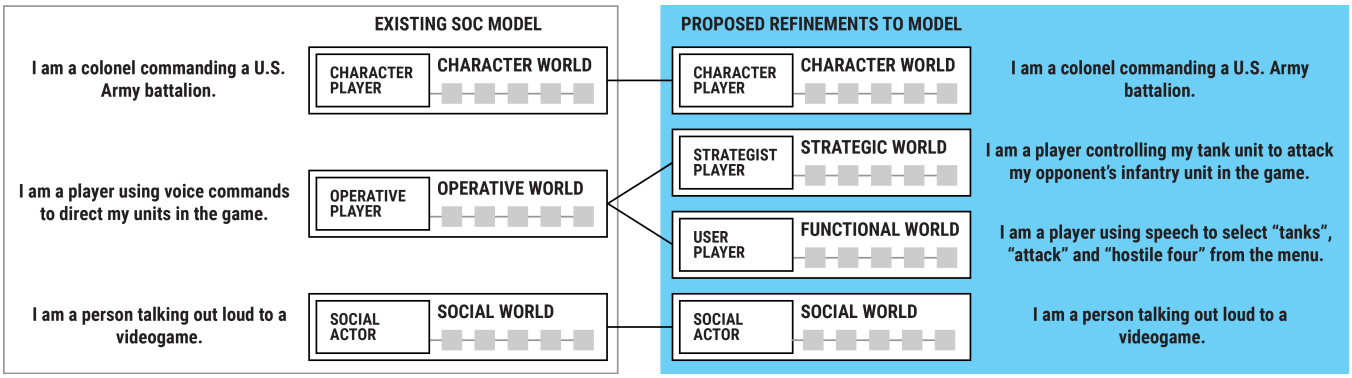

Figure 2: Proposed refinements to the SOC model of the game event [9].

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 the 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