Game Changer? How VoIP Is Impacting the Way We Play
Author

John Halloran

Department of Computing and the Digital Environment, Coventry University, UK
Abstract

Recently, computer games producers have integrated Voice over Internet Protocol (VoIP) into
distributed multiplayer games, allowing gamers playing at a distance to talk to each other ‘ear-to-ear’ in an audio-conference-like setting. How does being
able to talk to one another in this manner affect the gaming experience? A longitudinal study of a group of adults playing a multiplayer team game is
presented. Our analysis looks at how the players used VoIP talk to interact with each other in the virtual game world. We found that VoIP represents
talk in ways that differ both to face-to-face talk and to text-mediated communications, and this leads to new forms of multiplayer gameplay: VoIP
audio representations interact with, and mediate, the graphical materials of the game world in ways that can generate problems to be overcome for players,
but also provide new opportunities. In particular, our findings show how players used VoIP to coach each other in the early stages of playing together,
and then later on to successfully coordinate more complex game playing. For both, distinctive forms of collaboration made
possible by VoIP were found. On the basis of our findings, we consider how VoIP can be further integrated with graphical representations to enhance
the user experience in distributed multiplayer games.
Keywords: Voice over IP; computer games; multiplayer games; user study; computer-mediated communication; virtual
environments; face-to-face communication.
Introduction

Many genres of computer games now exist. The number of players can vary: there are single-player
games like Tetris and Solitaire; multiplayer games, including car races and poker; and ‘massively multiplayer role play games’ (MMORPGs), for
example Second Life, which allow thousands of people to play at the same time in virtual ‘real world’ environments. Their graphic design, linked to
the ways they have been designed to be played, is equally diverse. Some, including MUDs and MOOs are entirely text-based, while a game like Pac-Man
is controlled via the use of simple 2D graphical interfaces. Other games, including Tomb Raider or The Sims, are acted out in rich 3D virtual
environments. In such environments, the player is typically represented by an ‘avatar’, a 3D figure in the game. The player’s experience can vary
depending on whether their perspective is first- or third-person: in ‘first-person shooters’ (FPSs), for example Unreal Tournament or Half-Life,
the player is part of the action, looking out through the eyes of their avatar, while shooting a gun or driving a car; while third person games, such as
Resident Evil or Gears of War, depict the player’s avatar from a variety of angles. In addition, a number of different methods have been
developed to enable players to communicate with one another in the virtual spaces. A well-established communications method is
text. Text communications range across unix-based commands, through chatroom conversations, to graphical ‘bubbletalk’, where text messages from
player to player appear in speech bubbles. More recently, the ability for players to talk to each other simultaneously while playing a game over audio
connections using headphones and a microphone has been made possible through the use of Voice over Internet Protocol (VoIP).
Given that there are now so many options available to designers to create gaming experiences, questions arise concerning the advantages and
disadvantages of using different forms of interaction, representation and communication. In this paper, we are concerned with one such form: the use
of VoIP in contemporary multiplayer games which involve collaboration in teams. Our fundamental research question is how talking through VoIP
influences and shapes gameplay. In particular, how far and in what ways does audio communication in the form of VoIP support gameplay? Our focus is on
the following aspects: how players use it to learn the game and coach each other, signifying and explaining to one another what they are doing and where they are
going; what team strategies need to be adopted in order to win the game; and how to coordinate joint actions. In addition, we are interested in how talking
‘ear-to-ear’ through VoIP might differ from talk in face-to-face interaction, as well as text communications in games; particularly whether there
are inherent properties of VoIP that might present challenges to gamers, but also support novel kinds of interactions not found in
face-to-face or text-based game settings.

A longitudinal study was carried out over a four-month period. We looked at
how VoIP was used by a pool of adult players when learning and playing various games on Microsoft’s Xbox Live platform, which integrates
VoIP with multiplayer games. Microsoft’s latest console, Xbox 360, uses the Live platform to implement VoIP, and Sony and Nintendo are also
producing VoIP solutions for their own platforms. Thus, VoIP is very much the latest development in communications, set to
become standard for console gaming.

We present detailed findings from one of the team-based games that was
played most frequently, Return to Castle Wolfenstein (published by Activision). To address our research questions, we carried out two related
analyses. The first was a quantitative analysis of the types of utterances spoken by the participants over the period, to establish how much and what kinds of
talk occurred, variation across individuals, and development over time. A complementary qualitative analysis was undertaken to look at how talk supports
gameplay. This includes characteristics of VoIP as an audio representation of voices; how this form of audio representation supports the interaction of
gamers; how it works in relation to the different graphical and textual representations that are used in the game; and how it resources interaction in
ways that remove some of the problems of text-based interaction but raise others, as well as allowing for new forms.
It was found that despite a range of issues and challenges, VoIP talk was effectively used to support players in understanding, producing, operating, and
integrating different kinds of graphical representations in order to construct interaction and make sense of the game. In particular, VoIP
communications were found to be effective in supporting players in coaching each other, and for coordination purposes where people need to work
together in teams. We examine the development of talk and gameplay, as well as their relationship for this game genre, and discuss design implications.
to further support players using VoIP for this type of game.
Background

Talk has always been important in playing games, serving a variety of functions: discussing and
reinterpreting rules in children’s playground games (Hughes, 1983); announcing a hand in Poker, and bluffing by false announcement (Hayano,
1982); discussing handicaps in golf (Goffman, 1959); and sociable chat in bowling (Putnam, 2000). Other common practices include calling and
exhorting in football; disputing calls in tennis; narrating action in children’s games; teaching other people how to play; congratulating and
Celebrating wins; upbraiding and criticising failures; commenting and commentating. Talk also extends beyond the playing of games to where a game
is reflected on afterwards, or at the same time – ‘metagame’ talk (Garfield, 2000), or can move beyond the game into other kinds of sociable talk. It seems
appropriate, therefore, to provide a means by which distributed game players can talk live to one another while playing, to help them learn and play, and to
enhance the social experience.

However, distributed multiplayer games differ to co-located physical games.
Hence the kinds of interaction that they support are quite different. Players are not co-located but geographically separated. This means that
the space in which the game is played is commonly a 3D virtual environment, and the characters within it are not the players themselves but
avatars representing them. Until recently, people could not verbally talk to each other in these settings, but interacted via text communication. With the
arrival of VoIP, players can now talk to each other using headsets with microphones. One obvious benefit of using VoIP in distributed multiplayer
games is that it means that players do not have to switch between controlling their avatar and typing in text when wanting to communicate
(http://www.christine.net/2006/03/the_impact_of_v.html). Talking leaves hands free to get on with moving the avatar or controlling weapons, for example,
which is important for games that are fast-paced and where timing and movement are key to winning. But can distributed multiplayer
games which allow players to talk in this fashion offer the variety, spontaneity and usefulness of talk found in co-located games? What are its benefits and limitations?
Before turning to current literature on this, we will lay some groundwork concerning how face-to-face communication works, and relate to issues with
text-based communications in games. Both are important for understanding issues and requirements for voice-
based communications in games.
Text Communications in Games

A key feature of face-to-face communication is ‘tight coordination’ (Sacks et al,
1974): the phenomenon whereby conversation is marked by almost instantaneous switches between speakers without gaps or interruptions. Small
gaps and overlaps can occur, but these are of the order of microseconds. A major reason is that in face-to-face interaction, the unfolding of an utterance is
both visible and audible. Because the individual words in a sentence can be heard and interpreted before the sentence has finished, listeners can
anticipate what is likely to be said, and construct responses in advance, avoiding gaps and overlaps. This has two key effects. First, it enables speakers to
focus on the same topic; and, second, it means that speakers can construct turns, so that they follow each other in timely ways without interrupting, or
leaving pauses (which only occur if needed).

Computer games have often relied on text-based communications among
players. This differs from face-to-face communication in many ways: one is that utterances typed at a keyboard do not appear as they are produced, but only
after they have been completed and entered: this is especially the case for early games which are entirely text-based. As a result, turn-taking may not
be observed: players often type messages simultaneously, as if they were all speaking at once. A result is the emergence of multiple topics (see e.g.,
Brown and Bell, 2004b; Taylor, 2002; Curtis, 1992; Manninen, 2003; Muramatsu and Ackerman, 1998; Wright et al., 2002). In discussion of one of the
earliest text-based games, ‘LambdaMOO’, Curtis (1992) points out how the appearance of utterances only after their completion introduces delays into text-
based conversation that can seem unnatural. Players do not wait for utterances to appear, but continue with their own, so that multiple conversations are held at
the same time: ‘it is very rare for there to be only one thread of discussion; during the pause while one player is typing a response, the other player commonly
thinks of something else to say and does so, introducing at least another level to the conversation, if not a completely new topic' (op. cit., p. 13). This
phenomenon, known as ‘multithreading’, is quite different to conversations held in the physical world, where listeners are able to anticipate what speakers’
utterances are going to be because they can hear them as they unfold word-by-word.
However, while multithreading may reflect problems with focussing and turn-taking, there are also benefits. O’Day et al (1998), for example, point
out that multithreading in educational MOOs supports. They explain how students were able to discuss four dimensions of a learning problem
simultaneously. In addition, text-based conversations are persistent in ways that face-to-face conversations are not. There is a ‘history trail’ of conversation such
that previous utterances can be re-read (e.g., Becker and Mark, 1998), and this may compensate for any confusion arising from multithreading. Hence,
multithreading may only be problematic where speakers all need to concentrate on one topic. Where many conversations are needed, it is not a major
issue. Even in settings like LambdaMOO where multithreading is not necessarily desirable, Curtis (op. cit.) makes the observation that players
get used to it, and ‘handle the multiple levels smoothly’. He also points out that in face-to-face conversation, topic changes can be regarded as
interruptions that may not be accepted because of the discontinuity they produce. In a virtual environment where continuity is not necessarily expected – and
may not be required - topic changes and deviation from face-to-face turn-taking behaviour are much more acceptable.
Some recent games have been designed to more closely emulate face-to-face communication. They support speakers communicating via text in
focussing on the same topic. In these games, typed messages appear on screen word by word. In addition, rather than being entirely text-based, they provide
a combination of virtual environments and text windows. Brown and Bell (2004a) discuss an MMORPG called ‘There’. Here, utterances appear in
speech bubbles which appear over players’ heads. The writers point out that ‘bubbletalk’ results in utterances that are produced and interleaved
more akin to those found in face-to-face interaction: in other words, turn-taking is observed. This is associated with less multi-threading. Furthermore, word-by-
word text messages can support more focussed interactions with objects, thus supporting collaboration.
Like multithreading, bubbletalk has some disadvantages, but it also has characteristics which provide opportunities for communication which are
not afforded by face-to-face settings. Because the locations of the speech bubbles of an avatar co-vary with the movement of that avatar, they provide
a strong visual cue as to who is speaking. In crowded environments, this makes it possible to communicate across distances without trouble,
and to hold focussed conversations with distant players while many intervening conversations may also be going on. This is not possible in face-to-
face settings. Thus, while bubbletalk helps remove the problematic multithreading that can occur where focussed discussion is needed, it also
supports multithreading as required - where different sets of people need to hold separate conversations at the same time. However, in very busy environments,
there may be issues. Bubbles can occlude parts of the scene, so that the smaller, less generally visible objects may be harder to collaboratively
interact with. Bubbles may also occlude other bubbles, making discussion harder to follow. However, for other game-like virtual environments designed in
the same way as ‘There’, for example HabboHotel, these issues are less important because the activities are less goal-based. Here, bubbles are effective at
supporting the less focussed, more non-committal conversation that occurs.
Talk and its Context

The affordances of bubbletalk arise from a specific relationship between communication
and its context. This context is the virtual environment of the game world. The relationship allows (1) tight coordination, (2) communication at-a-
distance, and (3) multithreading where needed. These three affordances arise by virtue of the word-by-word appearance, spatial
positioning (including co-variance with avatar movement), and visibility of text communications when they are embedded in speech bubbles.
The relationship between communication and context is also crucial for multiplayer games which use voice communications rather than text. Before
turning to this, we need to understand more about the relationship between communication and its context in face-to-face interaction.
The context for talk, be it face-to-face, text or audio, is integral to how things are referred to and understood. To talk about something in the real world depends on
a context that is meaningful, and this often implies physical co-presence, mutual visibility and shared reference (e.g., Clark, 1996; Garfinkel,
1967; Goffman, 1959). According to Garfinkel (1967), it is important that the parties to social interaction can see and describe the same things,
and understand those things in similar ways. People have to be able to observe others’ actions, and the objects to which they may refer, while also
recognizing what these actions and objects mean. This also implies that action has to be produced in such a way that it is readily comprehensible by others.
Clark (1996) has shown how talk and its context are not only related, but intimately connected. In particular, talk and its physical context are
mutually constitutive in terms of meaning: settings make talk meaningful and talk makes settings meaningful. Clark claims that language is a form of
‘joint action’: ‘one carried out by an ensemble of people acting in coordination with one another’ (Clark, 1996: 3). He discusses examples of
conversation where utterances have little meaning without shared reference to an environment that is being worked on and changed as
talk unfolds. Conversations, including Clark’s example of that between a shop assistant and a customer, are supported and made meaningful by the physical
arrangement of the environment, the orientation of the people involved towards each other, and their shared views onto items of
importance. The co-presence of these with talk allows interaction to occur in ways that allow quick resolution of ambiguity, for example about what is
being bought; and anticipation, for example of the need to provide change.

This has important implications for conditions
that need to be in place before people can use talk to work together. Schmidt (2002) has argued that ‘seamlessness’ of talk is commonly found in
collaborative work, implying a meaningful context. Greatbatch et al (1993) show how social interaction in everyday collaborative settings,
including consultations between doctors and patients, is a complex choreography of physical context, movement and gesture, where mutual
visibility and shared views support focus and turn-taking. This work, as well as that of Clark and Garfinkel, shows that the physical context of talk is highly
important. However, it also implies an intentional context: that speakers have goals which are the reason for talking in the first place, and the physical context is
key in helping them to form and accomplish those goals.

The ability to communicate effectively, then, depends on particular relations
between talk and its context. When particular cues are absent, problems concerning multithreading and turn-taking can arise. This has important
implications for the design of virtual environments including multiplayer games. In order to design games that support meaningful communication,
an issue is how to support the interplay between talk and its physical context in ways which make both meaningful. It is not only important to provide cues
to allow anticipation of utterances, but also to consider whether people can see the same things and understand them in consistent ways.
VoIP Communications in Games

We have seen that text communications in games differ to face-to-face
communications, in both problematic and advantageous ways. Problems in games that are entirely text-based include lack of focus and multiple
topics associated with unwanted multithreading when utterances only appear when completed. Advantages include the ability to discuss different
aspects of a topic simultaneously. In games that use bubbletalk to integrate text with a virtual environment, advantages include visibility of other
players as well as objects; shared reference when speakers can see and refer to the same thing; and communication with others at a distance, even where
there are intervening conversations (this latter is impossible in face-to-face interaction).
In addition, text communications - regardless of type - make it easy to recognise who is speaking. When there are many people in a game, this
means that a player is able to address the right person, can put a face to a speaker when addressed by them, and can link what they are saying to what they are
referring to. This is because utterances are labelled with the speaker’s name. Face-to-face interaction also makes it easy to perceive who is speaking: voices are
different, they co-vary with the position of the speaker, and speech is synchronised with facial and other movements. In addition, face-to-face interaction
supports visibility and recognition of speakers and objects, because they are visually present, as well as differentiated.
Knowing who is speaking in team-based games is important in order to be able to collaborate effectively (Halloran et al, 2004). If a player is asked
or told to do something by another, it is difficult to respond appropriately if the identity of the speaker is not known. Knowing who is speaking is also
important for purposes of reference. If, in a team game, a player announces, for example, ‘the enemy are round the back’, knowing which speaker said this
establishes their position and can disambiguate the location they are referring to. However, research into VoIP-enabled games shows that it can be hard to know
who is speaking (Halloran et al, 2004; Gibbs et al, 2004; Wadley et al, 2003). This problem occurs for three reasons: (1) the characteristics of voices
and utterances when represented through VoIP; (2) issues with the labelling of utterances with the speaker’s name; and (3)
similarities in avatar appearance and behaviour.

One reason why speakers can be hard to identity is that voices in VoIP in
multiplayer games can sound similar. It can be hard to recognize and discriminate between other players’ voices and where the utterance is coming
from unless the players know each other beforehand (Halloran et al., 2003, 2004). It can also be difficult to pick out a consistent speaker where
several people are speaking. Wadley et al. (2003) found that not knowing who is talking can make players reluctant to talk using VoIP. These
issues are related to the implementation of audio communications via VoIP in games, which leads to particular forms of
representation of voices and utterances.

The VoIP channel in games is allocated an IP layer that is much thinner than the
graphics layer. This can lead to degradation, including breakup, even with fast servers where there is no graphics lag, with VoIP conversations
ending up sounding like ‘military radios’ or ‘intercoms’ (www.von.com). This contrasts with face-to-face communications where
voices are not processed or degraded in any way. In addition, during and between utterances, there is a ‘black background’ which removes ambient
cues from voices, including the kind of room they are in and other audio information including key taps or music: these may be useful in differentiating
a number of co-present speakers in audio conferences. Further, all voices are represented at the same amplitude, so that the relative loudness of
voices is established only by the volume at which a player speaks (and not, for example, by distance). In addition, headsets have a single (rather than double)
earpiece which means all voices are delivered to the same ear: VoIP represents voices monaurally. Hence, VoIP-represented voices have positional information
removed and, in contrast to face-to-face talk, and bubbletalk, do not co-vary with the position of speakers (i.e., for games, avatars).
Where there are a number of voices, particularly of the same sex, identifying who is speaking can become confusing not just because VoIP can make them sound
similar, but also because of issues with labelling. In common with pre-VoIP multiplayer games, VoIP-enabled games feature a label known as a ‘gamertag’
(a player’s in-game nickname). This appears with his/her avatar, usually floating above it. However, in contrast to text communications in games,
player’s utterances, since they are auditory and not graphical, are not labelled with the gamertag. Thus, there is no immediate visual means of linking
an utterance with an avatar. This becomes an important issue where VoIP already makes it difficult to identify speakers by reducing the distinctiveness of voices, as
well as removing proximity and positional cues.

However, some VoIP-enabled games feature graphical tools to show that
there is speech activity from certain players. An example is an animated ‘loudspeaker’ icon which is associated with a gamertag. Games that offer this
include Midtown Madness and Gotham Racing (both published by Microsoft Game Studio). A screenshot from the latter appears as Figure 1:
Figure 1 Avatars, Gamertags and Speech Icons in ‘Gotham Racing’ (Microsoft Game Studio)
Please see Figure 1 in full PDF version
This screenshot shows two avatars (cars) with persistent gamertags, ‘DarkBluePhoenix’ and ‘DucDarkAngel’. At the bottom centre is
an animated speaker icon, and the ‘DucDarkAngel’ gamertag. This indicates that DucDarkAngel is speaking.
To work out who is speaking in these games involves recognising an avatar’s gamertag, and then looking for a loudspeaker icon elsewhere on the
screen, which is labelled with the same gamertag. In some games (e.g., Gotham Racing) loudspeaker icons and associated gamertags only appear when players
are speaking, making it possible for them to be several, or none. For others, including Midtown Madness, a list of all the players’ gamertags appears.
at all times, necessitating visual search through the list to see if that player’s loudspeaker icon is active. Identifying who is speaking, then, involves relating an
audio representation (the utterance) to a chain of graphical representations (the avatar, the gamertag, and the loudspeaker icon). Thus, speaker identification
in VoIP-supported games may require more cognitive effort than in text-based games.
In addition, not all VoIP-supported games behave in the same way as Gotham Racing and Midtown Madness. For example, in Return to Castle
Wolfenstein, gamertags are not persistently attached to avatars, but only appear when a weapons sighting is pointed at the avatar. In addition, when a player is
speaking, a loudspeaker icon appears on a compass at the bottom centre of the screen to show the direction of the speaker, but it is not labelled with
the gamertag (Figure 2.1). To link an utterance with an avatar, then, requires resolving direction of the speaker icon with an onscreen avatar. Where
there is more than one, this may become confusing; also, the avatar may not be visible in the current scene (they may be behind a wall, for example). While
a speaker icon above a speaking avatar does appear given a proximity of less than (approximately) five metres (Figure 2.2),
this is lost at greater distances.
Figure 2 Avatars, Gamertags and Speech Icons in ‘Return to Castle Wolfenstein’ (Activision)
Please see Figure 2 in full PDF version
Gamertags are not persistent but appear within a certain range of weapon sighting (the circle at the centre of the image in both screenshots). At the
bottom centre of the screen is a compass (‘N’ indicating north) and a speaker icon. NB the avatar is a blurred figure left of centre;
Close proximity of a speaking avatar to player triggers a speaker icon over the avatar.
Talk in VoIP-enabled games behaves in a different ways to talk in face-to-face settings, and this can make it difficult to work out who is speaking. The context of
talk - in games, a virtual world - also behaves differently. Issues with visibility and shared reference can arise from non-mutual perspectives,
generic avatar movements, and visual similarity of avatars.

Non-mutual perspectives arise when players’ views
onto the same world are different, even though they may be in the same place in that world. This can be because they are looking in different directions, but can
also be due to view management, which is done by means of ‘flying cameras’. These allow players to view the world from different angles,
including behind and above themselves (Ducheneaut and Moore, 2004). When this happens, players may not be able to see each other, or to recognize
objects that other players are looking at and discussing. This may create problems for shared reference. Against this, however, Brown and Bell
(2004) point out that where virtual objects are large they can create interest that draws avatars close and enables views to be shared.
Avatar movements are attenuated compared to the richness of real people’s movements. Speech-related movements of the mouth are absent, as well as facial
expressions, creating further challenges in identifying who is speaking. All avatars use the same fixed repertoire of movements, and they are
not as detailed, nuanced or individualised as in face-to-face interaction. Thus, avatar gestures may lose some of their impact in terms of, for example,
drawing attention or pointing things out. Gestures may also be difficult to time, for example a bow gesture to show required respect can
appear late because of a control-to-action lag (Ducheneaut and Moore, 2004). Hence, interactors in virtual environments may have to emphasize their
gestures more than in the real world to achieve the shared reference. Avatars can also appear to the players to be visually similar. While in some
games (for example, The Sims), avatars are customisable for appearance and can be made highly distinctive (Brown and Bell, 2004a;
Taylor, 2002), in others, such as war games, the members of the same teams all wear the same uniforms. This can make it difficult to work out who is who,
exacerbating the voice differentiation problems of VoIP.
The Study

Aims and Objectives

The aim of our study was to explore how VoIP-mediated
communication was used and shapes the way distributed multiplayer games are played. A particular emphasis was on how this is influenced by
the properties of VoIP as an audio representation, its interaction with the graphical representations found in games, and how the resulting ensemble of
audio and graphical representations is operated and integrated by players to produce successful and enjoyable play.
Design

A group of 10 adults aged between 20 and 48 took part in the study. Seven of these were male and three
female. They included a couple who lived together and gamed in the same room, and two housemates who gamed in separate rooms. Otherwise, the
members of the group were unknown to each other.

Each participant was equipped with broadband internet access at home, an
Xbox Live console, an Xbox controller, and an Xbox Communicator - a headphone with microphone allowing voice communications. Several
games were made available that they could choose to play, from a range of genres. These included Midtown Madness and Gotham Racing (race
games); and Ghost Recon and Return to Castle Wolfenstein (war games/FPSs). We ran all sessions from a fast server to eliminate lag problems.
The participants gamed together once a week for 10 weeks at a fixed time, for 60 minutes. The game that was played the most by the participants was Return to
Castle Wolfenstein, which our analysis focuses on. It was played for five of the ten weeks. It is a fast moving team game that involves two teams at war.
One team is ‘Axis’, and the other ‘Allied’. Players choose which team to join and can switch teams during a session. Members of a team can hear and talk
to members of that team only; not the other. The teams have an objective to meet, and the winning team is the one that achieves their objective first.
According to game and level, objectives can vary in nature and difficulty, from capturing a certain number of flags, through destroying a submarine, to stealing
gold and delivering it to a waiting jeep.
Analytic Method

For this study we used an adapted form of virtual ethnography (Hine, 2000), a method often used to
study computer-mediated communication in virtual communities and environments. A researcher is a participant in a virtual world, recording the
interaction in some way, while also making use of logs, instruction manuals and so on. An important feature is that the observation is usually
unknown to those observed. In contrast, we chose not to take part in playing the computer games but to sit in the participant’s rooms while
they played. Hence, they were all aware of our presence as observers but not as participants (virtual ethnography is usually the other way round). An
advantage was that we could record games from the viewpoint of the players, rather than our own - and, in addition, make recordings from more
than one point of view. The method also allowed us to ask our participants questions and interview them afterwards.
We observed and video-recorded 2 of the 10 participants per gaming session, rotating around the group with the aim of recording each participant.
at least once. A camera was set up to capture the screen (either ambiently or direct screen capture, depending on the circumstances), and a small, lightweight tie
microphone was attached to their Xbox Communicator headphones to record the audio. Recording two different viewpoints onto the same
event, and recording two ‘views’ onto the audio conference, also helped us to think about similarities and differences across different groupings of
participants in games, which proved useful in terms of understanding how far games get played similarly, or not, according to who is playing.
We tried, as far as possible, to ensure that the two observed players played on different teams, and again as far as possible, to record both ‘sides’ of every game,
i.e. the two different teams including the two (mutually exclusive) audio conferences. We were not always successful due to both observed players
occasionally being on the same team, recordings starting at different times (so that we recorded one side only because the recording of the other was
still being set up; in fact in one week, there was no capture for one side), or because the observed player had not necessarily joined the game when the
other players had started playing. For ‘Return to Castle Wolfenstein’, we recorded 54 games in total, and were able to record both sides for 33 of these.
This means, in total, there are 87 transcripts (both sides of 33 games, i.e. 66; and one side of 21 games).
Our findings are largely based on transcripts of the video and audio recordings of gameplay, but data we also draw on includes transcripts of interviews,
and talk before and after each session (between the observed person and the observer). The gameplay transcripts were analysed using a coding scheme to
identify kinds of talk. Two complementary analyses were carried out: quantitative and qualitative. In the quantitative analysis, we
examined the amount of talk, i.e., how much talk there was in different sessions; what players interacted with what other players; and the content of
the talk. Our qualitative analysis was designed to find out how talking with VoIP shapes and resources gameplay.
Findings

Quantitative Analysis: Patterns of Talk
We analysed who played with whom to get an idea of whether pairs played together all the time or they played on different teams. Table 1 shows who played
with whom, using fictitious gamer names. It can be seen that many of the participants played with the same player for over 50% of the games (54 in
total), while some pairings never occurred. (NB some players’ participation was lower than others’ across the sessions. Weepy, for example, participated in
three of the five sessions; Thomas in one session only).
Table 1: Pairwise Interactions in ‘Return to Castle Wolfenstein’
Games Played
Please see Table 1 in full PDF version
We then examined how much talk took place during the games and how it changed over time. Table 2 presents the number of words spoken per minute.
(WPM) versus the number of utterances per minute (UPM) across the 5 sessions (which took place in 5 separate weeks). The reason for looking at both
WPM and UPM was to find out whether there were variations in utterance ‘densities’, i.e. how long or short utterances were, as well as their frequency.
This might indicate, for example, the need for certain utterances to be lengthier (due to demands of the game, say); that certain individuals produce
longer utterances than others (perhaps indicating individual differences or differences in game role); or that at certain times, talk needed to proceed faster.
Table 2 Average Words per Minute (WPM) and Average Utterances per Minute (UPM) by Session
Please see Table 2 in full PDF version
Table 2 shows that the WPM and UPM averages are similar across the sessions. Each of the individual values was close to the average of all the
values on each count. In addition, the ratio of utterances versus words per minute by session was similar, approximately 1:7 (which means that on
average each utterance consisted of 7 words). These results suggest that there were similar amounts of talk per session regardless of who played
with whom, and what level of game was played. The differences between sessions are not to do with different patterns of talk (e.g. longer or shorter
utterances) but with the same pattern speeding up (Session 5), or slowing down (Session 2). Hence, what gets talked about and how much talk there is, is
not dependent on individuals but is likely to be of a similar nature regardless of who is playing.
However, it is important to recognise that the individual session values are averages of all the games within a session. To check that the findings on
utterance densities and individual differences were correct, within individual sessions we did average WPM and average UPM by game. There was quite
large variation. What is striking is that there is still a clear covariance between WPM and UPM: the lower one of these values, the lower the other, again
suggesting the same pattern of a given average number of words per utterance – and this again despite variation in who was on the team from game
to game. The low WPM/UPM ratio occurs where people know the game and it is easy: there is less need for talk to coordinate and organise the
game, so that utterances are more widely spaced.

The transcripts showed that there was very little temporally overlapping
talk, suggesting it followed the rules of face-to-face conversation much more than text-based communication. In particular, it was able to
support anticipation of utterances.

We then coded the utterances at two further levels of analysis, to
examine what kinds of talking were taking place. The first classified the utterances in terms of three codes: ‘game’, ‘meta-game’, and ‘outgame’, in order to
get a better understanding of what was being talked about when playing the game. The ‘game’ code refers to an utterance that directly relates to the
current state of play for the current instance of the game, for example, “Look out behind you”. The ‘meta-game’ code was used to refer to comments about
a game over several different instances (which could include different levels), e.g., “This is faster moving than the other level”, and that reflected
general attitudes or knowledge derived from repeated experience, e.g., “I don’t like the way you have to pull the trigger when you want to speak in this game”
The ‘outgame’ code was used to refer to an utterance about things other than the game, e.g., “How’s the weather where you are?”. The second level
of analysis labelled the same utterances in terms of the particular action being referred to, such as ‘give instruction’, ‘give information’, ‘require
information’ or ‘comment’. This was intended to give an indication of how the topic of conversation changed over time, and, in particular, to determine
how much was spent on learning or coaching, and how much on coordination. Each utterance was coded at the two levels. For example, the utterance “I’ve
got a grenade and I’ve got a gun” was coded (i) ‘game’ and (ii) ‘give information’.

Game-based utterances accounted for 90% of the
total, meta-game utterances for 9% and outgame utterances for only 1%. The analysis indicates that the vast majority of the talk was about playing the
game; participants were caught up in the moment of the current game, such that they had little time to talk about anything else. The little amount of outgame
talk was ‘small talk’, such as a player saying what he was having for dinner, while the small amount of metagame talk was mainly about how to use the console.
The second level coding scheme revealed the kinds of activities that were taking place. These were classified in terms of information, instruction,
action and ‘other’. Information codes can be commentary on what is happening, etc. Instruction codes concern telling someone, or finding out,
how to do something. General instructions (GIVE_INST) are action-oriented but not about how to take a specific action, whereas GIVE_INST_ACT
are always about how to take specific actions. Action codes also include information, and are also instructive. They assume players already know how
to do something. Table 3 (over) gives the second level coding scheme. All the codes are expansions of the ‘game’ code (‘G’) at the first level, and are illustrated
with examples from the transcripts.
Table 4 Proportion of Action Types Based on Second Level Analysis

Please see Table 4 in full PDF version
Table 4 (above) summarises the average percentage of instruction, information, action and other types of utterances per session. It can be seen
that the majority of utterances are information types, ranging from 50-60%. Action utterances range from 14% to 27% per session and increase over
Sessions 1 to 4 before dropping back to 17% in Session 5. What is most striking from the findings is the rapid decrease in use of instruction utterances,
which fall from 18% in Session 1 to 0.3% in Session 4.

The percentage of instruction utterances in
Session 5 is also negligible. What this indicates is, broadly, that while information utterances remained the same throughout the sessions,
the proportion of action utterances increased while the proportion of instruction-based utterances decreased. This suggests that after Session
2, the players did not need to request or to give instructions. Rather, the main concern was information and action: in
other words, the players had learned how to play.
Table 3 Second Level Coding Scheme

Please see Table 3 in full PDF version
The two main findings from the quantitative analyses – first, that the amount of talk was broadly the same across the sessions, and second, that the
proportions of action and instruction utterances varied over these sessions - indicate that the participants changed how they played and how they
talked about it over time, spending more of the early sessions asking for and giving instructions to each other than in the later sessions, where more time
was spent on talk to support the players in coordinating their actions in order to win the game.
Qualitative Analysis: The Interplay of VoIP Talk and Graphical Representations in the Game
The quantitative analyses showed how much talk took place over the study, what it was about, and how it changed. A qualitative analysis was then
performed on the kinds of talk that the participants engaged in to understand, produce, operate, and integrate the different kinds of graphical
representations that enabled them to progress with the games. In particular, we look in more detail at the two major types of activity identified:
coaching, and coordination; and how the latter emerges out of, and can depend on, the former. It examines how talk changes as the participants become more
experienced and play games at increasing levels of difficulty. The focus is on how the participants integrate what appears on screen with what is being
said, and how talk constructs the gameplay but is also constrained by the demands of the game, its design, and the behaviour of VoIP as an
audio representation. To illustrate the nature of these, vignettes are presented which show what the work players have to do to overcome issues to
do with speaker disambiguation, non-mutual perspectives, object reference, and shared meaning.
Coaching

The transcripts for the beginning session revealed much evidence of coaching taking place, where one
player instructed another of what to do to progress with a game. Here we discuss three examples of coaching, two of which
were successful, while the other was unsuccessful.
How to Deliver Ammunition (1)

In the following example, the objective is to capture a set of five flags, in different
positions in a virtual city landscape. This requires players to take particular roles. One of these is soldier. Soldiers have a range of important
weapons not available to the other roles, but they have limited ammunition. Another role is lieutenant, and a key part of this role is supplying ammunition to
the soldiers and other team members (including him/herself) at various intervals. A further role is being the medic who is responsible for ‘giving
health’ (rather like battery power that players can run low on), and bringing wounded team members ‘back to life’. There was much evidence of the more
experienced players telling the less experienced ones about the different roles each must adopt to play the game, what they needed to do in that role, and how to
change roles. In the following excerpts, Mars explains to his lieutenant, Di, how to deliver ammunition:
1. Mars Right you’ve got a green pack in your weapons list

2. Di O-oh... Someone’s shot me
3. Mars  Press A and B to go through your weapons pack

4. Di  Press A and B is it
5. Mars A and B yeah and you’ll cycle through your weapons, and somewhere you should have a, er, I’ve been set on fire,
you should have, you should have some ammunition, and just drop em on the floor and we can pick em up
we’ve got extra ammo

This excerpt shows that the game is fast moving, with both players being attacked
during this interaction. One effect of this is that Mars gives a summary of how to deliver ammunition, rather than detail concerning identification of
ammunition and how it is dropped. There is another interruption caused by Mars’ awareness of the current status of the game
(the other team are about to win):

6. Di Drop it on the floor
7. Di    Oh Yes I’ve got the ammo, how do I drop it?

8. Mars   Um just shoot it as if it was a weapon
and it’ll fall on the floor

9. Mars I’m gonna um run off and take some
flags cos they’ve nearly won

10. Mars Can someone get that flag that’s in our
area and we can win

At utterance 9 (above) Mars, whose avatar has previously been co-located
with Di’s, moves away from her to try to capture a flag. 12 seconds later, Di has not delivered ammunition. This time, Mars instructs her, telling her specifically what
to do, and offering feedback in response to her actions:

11. *Mars Di you need to drop some ammo for everybody so change*
your weapon until it changes into a green backpack

12. Di Where is everyone
13. Mars  Behind you

14. Di  OK I drop it just by pressing ‘A’

15. Di  OK shall I drop it
16. Mars  No no not that that’s a grenade, no no, keep going next weapon, Yeah that’s it drop that
17. Di  By pressing

18. Mars  Just shoot

19. Mars  That’s it. And again. That’s it.
OK. As soon as you see anybody on our team you need to drop those for us.
20. *Di OK*

These exchanges span much of the game they are excerpted from. They concern an interaction
between Mars and Di where Di learns effectively to play the role of lieutenant and deliver the ammunition. This is accomplished through an
interaction between talk, avatar proximity and mutual visibility: both Mars and Di can see each other, and both are able to refer to a mutually visible ‘weapons
pack’ (carried by Di), as well as its contents.

It is important to note that before the study, Mars and Di already knew each other:
thus, their voices were mutually familiar such that there was no issue with the mutual recognition of voices in the VoIP audio conference (consistent with
findings reported at www.christine.net/2006/03/the_impact_of_v.html). They are also different-sex voices. During the exchanges, neither player
had any problem in relating utterances to avatars and other graphical game events to construct a meaningful interaction.
In the following excerpt, a player, Lancelot, is alone in a building, when he
suddenly hears a voice asking him for ammunition. He cannot see any other avatars but responds appropriately:
1. Buzz Lancelot, Lancelot

2. Lancelot Yeah?
3. Buzz    Can you give us some ammo mate?

4. Lancelot  Some ammo?
5. Buzz        Yeah

6. Lancelot  I would if I could find it
Lancelot cannot see either the ammunition or the person he needs to give it to. This is actually something he is carrying, but as a novice, he does not
at this stage know this. So he has to find the ammunition, which involves realizing he himself is carrying it, locate Buzz, which involves
recognising his avatar as belonging to the player who has made the request, and then pass the ammunition to him. The interaction continues:
7. Buzz  Do you know how to give out ammo?

8. Buzz  Obviously not
During these utterances, an avatar appears on screen. There is no other avatar simultaneously on screen and no other voice. This is followed by:
9. Buzz If you press your change weapon button

10. Buzz Lancelot
11. Lancelot Yeah

12. Buzz Press your change weapon button and you get
13. Lancelot Oh there’s a pod, there’s a pod isn’t there

14. Buzz Yeah there’s a pod
15. Lancelot Ah that’s what it is

16. Buzz Press that

17. Lancelot Got you
18. Buzz  One more

19. Buzz  One more

20. Buzz  And one for luck
Lancelot  Ah Yes, great

Utterance 15 confirms that Lancelot was unaware that he was carrying ammunition. Once the pod
has been identified as carrying the ammunition, Buzz instructs Lancelot to ‘press that’ and this results in a parcel of ammunition being thrown at Buzz’s
avatar, who picks it up. This action is repeated three times, and throughout, Buzz’s avatar gestures to Lancelot to continue as he says ‘and again’. Thus, there
is constant feedback establishing that the avatar visible and voice audible belong to the same player. Screenshots from this
interaction appear as Figure 3
Figure 3 Recognising and interacting with a Speaker in ‘Return to Castle Wolfenstein’ (Activision)
Please see Figure 3 in full PDF version
.1 Empty scene with Lancelot’s weapon visible (Coincides with utterance 1 from transcript);
.2 Appearance of avatar (utterance 7);

.3 Discovery of ammunition, housed in
'pod' visible at bottom right (utterance 13);

.4 Delivery of ammunition - in pack lying on floor
This episode, like the example of Mars and Di, shows the effective
construction of an interaction: effective in that it supports a key event that needs to happen in the game (delivering ammunition). Unlike Mars
and Di, the central issue for Lancelot was the integration of avatar and utterances to recognise the individual he was interacting with. The fact
that there is only one other player visually and auditorily present throughout means Lancelot knows which avatar to interact with while hearing
Buzz’s various utterances. The fact that Buzz’s avatar provides feedback through picking up the ammunition also establishes that this voice belongs to this avatar.
Having linked voice with avatar, Lancelot continues to interact with Buzz as the game progresses, by following him and asking...
him questions about geography and strategy.

This differs from the example with Mars and Di in a key respect: part of the
interaction involves finding out which avatar is making the request to Lancelot, where the interacting pair has already been mutually identified in the example of
Mars and Di. The important factor appears to be the absence of other voices and avatars. The fact that there is only one other player visually and auditorily
present throughout means Lancelot knows which avatar to interact with while hearing Buzz’s various utterances. Buzz’s avatar provides feedback
through gesturing at Lancelot to throw, and picking up the ammunition, also establishes that this voice belongs to this avatar.
It also demonstrates the use of gamertags to get the attention of other players: It opens with Buzz calling ‘Lancelot, Lancelot’. As an experienced player, Buzz is
able to do this by virtue of several things. He recognizes that Lancelot is a lieutenant on his team. Team members are recognised by a general
style of uniform (for example colour); but also by the fact that their particular uniforms are slightly different (with a longer coat than other
roles). At the same time, he is able to work out what the lieutenant is called by passing a weapons sighting over him to reveal his gamertag. This allows him
to call the right person to get the necessary interaction started. It also allows the linkage of a voice to the avatar and gamertag, so that the three
representations attaching to a player - voice, avatar, gamertag - are associated.

These excerpts demonstrate how VoIP can
be used successfully to progress through a game, with pairs being able to communicate readily with one another and know where the other is and
what they are doing. However, it is easy for novices in settings where there are several other players to get confused concerning who they are
interacting with, not only because VoIP makes voices sound similar, but also because avatars can look similar not only within teams, but across different
teams. The third example (below) is of unsuccessful coaching, where Weepy, a novice, has difficulty associating avatars’ dress with the right team, and
this results in her inappropriately attacking her team-mates.
Who’s on My Team? (3)

A basic challenge for novice players is to work out who else is on their team, and who is not. This is essential
for appropriate behaviour, including not attacking - or attacking. Uniforms are a crucial cue to who is friendly and who is an enemy. Even if it is not
known what gamertag or voice relates to what avatar, recognising its uniform correctly helps ensure correct behaviour. However, acting effectively
depends on more than this: Frequently; the novice may be instructed in some way: to deliver ammunition (as we have seen), or to follow. These instructions involve
interacting with particular avatars and are verbal, so they depend on relating a voice to an avatar - the same challenge of speaker
recognition as for the first two examples.

In the following example (screenshots appear as Figure 4, over), the
simultaneous presence of a number of avatars and voices in an unfamiliar game is too confusing for Weepy to resolve when direct requests are made to
her by team-mates. The excerpt starts with Weepy trying to establish who is on her team:
1. Weepy So what colour are my team wearing?
2. Buzz    Yeah anyone that’s green, or tan

3. Weepy   Anyone that’s green
This establishes that Weepy’s team-mates are wearing green/tan uniforms. This is followed by Xlr8 giving other information:
4. Xlr8

Your Axis has got the long black jackets on. And someone’s just toasted me, on my own team.
Xlr8 here explains what the enemy team uniform looks like. It assumes Weepy knows that she is on the Allied team. At the same time, Xlr8 complains that
someone - Weepy, in fact - has attacked him with a flamethrower. Weepy responds that she feels it was probably her, and
announces that she is confused:

5. *Weepy*  Oh, was that me?

6. *Weepy*  OK I’m confused
The advice Weepy receives about how to tell teams apart is confusing because it assumes prior knowledge of what team she belongs to. But it is confusing not
just because Xlr8 assumes this knowledge, or for reasons of ambiguity of utterances and similarities of dress across different teams, but also because of
the representational design of Return to Castle Wolfenstein.
Figure 4 Attempting to identify team members in ‘Return to Castle Wolfenstein’ (Activision)
Please see Figure 4 in full PDF version
.1 Weepy starts to fire at an avatar (utterance 1 from the transcript);

.2 Weepy hears a voice advising colour of her
team's uniform (utterances 2 and 3);

.3 Weepy fires at an avatar, and hears a voice
complaining of being ‘toasted’ (utterance 4);

.4 Weepy remains unable to distinguish which avatars
belong to her team and which d not (utterance 6)

Frames .1, .3 and .4 above all clearly show Xlr8’s gamertag. They also show
someone wearing a tan and green uniform. At the same time, Xlr8’s voice tells Weepy that he is being burned as well as what her team’s uniforms look like.
All these are resources to help Weepy establish that this avatar is a member of her team. However, unlike the examples above with Di and Mars, and Buzz and
Lancelot, she is unable to link the voice to the avatar. The compass at the bottom of the screen is the only available resource for achieving this. It has a
speaker icon which shows when someone speaks, and the location of this person is mapped to the compass. However (as we saw above), the icon is not
labelled with a gamertag. When asked, only one of the 10 players, Mars, said he made any use of this.
All three of these vignettes show limited use of weapons scoping for gamertag, and little use of the compass to try to work out who is talking. Rather,
it appears that certain conditions need to be in place to clarify the relationship between graphical and audio representations more
immediately. In these circumstances, the implementation of talk via VoIP in Xbox Live seems to work most effectively for pairwise interactions that
are self-contained between two players, working together. Both speaker identification and object focus are effectively supported, and there is
feedback as well as tight coordination in terms of talk. This is evidence of shared meaning, a feature of the coupling of talk to physical contexts found in
face-to-face talk also found with VoIP. This can be for players who know each other (as in the example of Di and Mars), as well as players who do not. The
conditions may be less favourable where there are too many people in the vicinity, making speaker ambiguity a real problem which prevents a player
being able to make sense of the game; interact with graphical objects (here avatars and weapons) appropriately; or elicit responses to help them
learn what they need to know.

Where the conditions for speaker identification, object reference, and
shared meaning were in place, the kind of ambiguity experienced by Weepy in the above example did not arise to such an extent. Another example from our
data shows that where Weepy was in a self-contained pairwise interaction (i.e. one with no other avatars or voices) similar to those between
Mars and Di, and Lancelot and Buzz, the interaction was also effective, for similar reasons. This resembles the immediately preceding example, but
concerns Weepy learning from Buzz how to attend to a particular door through which enemy attacks are mounted, and how to deal with these attacks.
One further important aspect of effective VoIP-supported pairwise interactions for coaching purposes is what happens in terms of player
perspectives. Non-mutual perspectives can occur in games due to view management protocols, or simply because players are looking in different
directions. Return to Castle Wolfenstein also has view management protocols to allow players to see themselves from behind, as well as to view the scene
behind them (as in a rear view mirror). Episodes in Return to Castle Wolfenstein where there needs to be object focus in order to interact – as
represented in the first two vignettes above – oblige players to develop views onto those objects and for both players to know the other is attending to them.
The effect of this in our study was that in cases requiring interaction around objects, the default first-person view (as shown in all our screenshots from...
the game) was almost exclusively the only one used, tending to encourage mutual perspectives.
Coordination

To be able to play games at higher levels of difficulty requires good coordination between the players.
The objectives of a game at a higher level include attacking/defending a submarine; stealing documents from an underground room and
taking them to a tower to transmit them; stealing gold from a crypt, and taking it to a getaway truck. Thus they involved players needing to attack or defend
a single objective and then subsequently two.

As the players progressed through the games they spent more time
coordinating team action and their movement through the virtual world in relation to specific locations and objectives. They issued and responded
to utterances related to action much more, reflecting the need to coordinate team activity to achieve objectives. Three examples are presented
here, concentrating in particular on features of the later gameplay: effective coordination even with non-mutual perspectives; reduced need for speaker
identification; and an increase in the use of shared knowledge as a coordination device.
Moving as a Team

Achieving an objective requires considerable coordination that is primarily oriented towards
specific locations. Teams can move as large groups, split off as pairs, and return back together again. For example, during a more advanced game in week 3, a
team consisting of Mars, Di, Lancelot and Reevez, congregated at their starting point, a trench. They needed to find their way to an underground
documents room, steal some documents, then climb to the top of the same building to a transmitter room to transmit the contents of the document.
Accessing the documents room required them to enter the building through its roof. The following interaction started out with the four players moving
along the trench together, mutually visible, and holding a four-way conversation:
1. Lancelot Ah right so I’ve come I’m following Reevez in the trench now
2. Mars

So does everyone wanna

3. Reevez

Does someone wanna lead that knows the way
4. Mars Yeah follow me then. Watch out for this sniper

5. Lancelot Di. Di
6. Di Are you gonna follow me, along the trench

7. Di Yeah
By the end of this conversation, the team has split into two pairs: Mars and Reevez - who have moved forward faster - and Di and Lancelot. In the
following excerpt, the two pairs are no longer mutually visible. Mars and Reevez start to climb a ladder to the top of the building, while Lancelot
and Di remain in the trench. However, despite the loss of mutual visibility, the two pairs can still hear and talk to each other. This is because the VoIP
represents voices as co-present regardless of avatar proximity:

8. Mars Here this way this way, Reevez
10. Mars  And then up this ladder at the end. It's the fastest way to go

11. Reevez  Cool
The excerpt above shows Mars and Reevez holding a conversation relating to their immediate concerns, with no speech from the other pair. However, the
two pairs remain able to interact verbally, as the following excerpt shows:

12. Di I think I’m lost
13. Mars  Are you two lost already

14. Di    I’m not sure
15. Lancelot Well we’ve got to get up the top

16. Mars Can you see, can’t even see
where you are. OK

From this point, two clearly separate discussions ensue, one between Mars and
Reevez; the other between Di and Lancelot. The conversation between Reevez and Mars is italicized to distinguish the two:
17. Lancelot We’ve got to get up the top, here we go we’re going up the steps now, all
18. Di  Yep. I’m right behind you

the way up to the top
19. Reevez In here?

20. Mars Yep

21. Lancelot Carry on up up up
22. Di  Up these stairs

23. Mars  Right Reevez, if you go through the main doors I’ll come
through the bottom way, the other way

24. Reevez It’s in here is it
25. Mars  Yeah you just keep going the way you were going. I’m gonna be coming in
from behind them

26. Lancelot Now
27. Lancelot Come up the ladder come on

28. Di OK I'm here

29. Lancelot Right OK
30. Lancelot: Now we have to find the way down.

31. Mars: Reevez I’m just gonna go into...
the back of the document room

32. Reevez  Shit, I’m dying
33. Di Where are we going now

34. Lancelot This is the way down you following me?
35. Reevez  I got in there but I got killed
36. Lancelot Now down the steel steps

37. Di OK
38. Lancelot and that takes you down to where the documents are
During this excerpt, Mars and Reevez stop being concerned about Lancelot and Di’s location, and concentrate on their own actions. However, the fact
that the VoIP audio conference makes all voices equally present means that it is easy for the two groups to cut into each other’s discussions if necessary, as
does Di with utterance 39, below. She does this to establish how far the team has progressed in terms of reaching its first objective:
39. Di

Do we have the documents now?
40. Mars

Yeah I’ve got the documents, I’m racing to the er
This establishes that Mars has retrieved the documents by descending to the documents room. The next step is to climb back to the top of the
building to the transmitter room via some steps. Lancelot becomes confused because he has descended the steps but not seen Mars:
41. Lancelot  We should be able to see you, cos we’re on the steps
42. Mars

No, Oh shit could really do with some cover
43. Di I don’t know where you are
44. Mars  Get up to the top of the building
45. Lancelot  We’re in the document room now

46. Mars  No that’s no good I’ve got
the documents you need to be at the top

47. Di Oh
48. Lancelot  But we didn’t see you, it

49. Mars  There’s two ways to get down to the
room that’s probably the problem

50. Lancelot I see
Hence, the two pairs are able to hold two separate coherent discussions about different locations, despite the issues of similar amplitude and non-
spatialisation of utterances. Both pairs are able to ‘tune out' the other conversation, but to monitor it at the same time, similar to the cocktail party phenomenon.
reported on in psychological studies of dual attention (Cherry, 1953; Kahneman and Treisman, 1984).
The four players, in holding two separate discussions, are effectively multithreading - there are two different discussions. However, the threads do
not interfere with one another. Utterances from the two different conversations are not overlaid, and there is no interruption. In other
words, the four speakers observe turn-taking rules as if they were involved in a single conversation. The reason why turn-taking is happening across the whole
group, and not just within its two subgroups, appears to be because the group as a whole needs to monitor its activity as a whole in order to achieve the
objective. In addition, it seems likely that the lack of confusion between the two separate conversations is due to the tight link between each conversation
and its physical context, each of which is quite different, with one group, for example, referring to going up some stairs (utterance 22 above), while
the other refers to some doors (utterance 23). What this example also shows is that it is not necessary for players to share perspectives in order to
collaborate around key objects and locations (documents, rooms).
b)  

Team Attack on an Objective

Another example of effective coordination is where all players are co-
located but where there is no need for speaker identification in order to take effective action. In the following excerpt a team of 6 players is trying to attack
a submarine defended by the other team. The submarine can only be accessed by blowing up a sealed door, which requires an engineer, one of the
roles players can take (in addition to soldier, lieutenant and medic). Three players – Kat, Buzz and Lancelot – are at a door. The game produces a
text message which advises Lancelot, the observed person, of an important location: ‘You are near the filtration door’. The importance of this is
recognised by Lancelot and Buzz:

1. Kat

You are near the filtration,
something or other?

2. Lancelot Yeah, right, yeah that’s yeah,
An engineer is needed to blow up this door. Kat and Buzz propose that dynamite is needed but
Lancelot explains he only has other types of weapons and tries a bomb but with no effect.
4. Kat  Oh we need dynamite

5. Lancelot  Ah, I might I might have some. Hang on
a minute, what have I got? I’ve got a bomb.

6. Buzz Bombs are no good. No, you
can’t, you can’t
go through the
door
7. Lancelot I’ve got a
grenade and
I’ve got a gun. Two guns. And a knife

8. Kat

I haven’t got any dynamite
9. Lancelot: No. I could try throwing a bomb at it. Stand back.
Buzz and Shimmer then each place dynamite at the foot of the door, telling the others what they are doing and what the others should do:
10. **Buzz** Someone open the door and get the submarine
12. Buzz Double dynamite

13. Reevez Dynamite planted near the filtration door, so that’s
going to, that’s going to blow then

14. Shimmer I’d stand back guys
In this episode, the various players do not need to know who specifically is speaking. Coordination is instead focussed on what needs to be done in terms
of mutually visible objects - i.e., the door that needs to be blown up, and the dynamite that is placed. The actions of Shimmer and Reevez, both engineers who
know they have to place dynamite by the door, reflect that they have implicit knowledge of how the game works such that they do not need to use talk
to coordinate their actions between themselves. However, talk is still used where the other team members with less knowledge attempt to work
out what to do; and for the more experienced players to advise them.

In reverse order, the two examples above show how
VoIP supports coordination both with and without mutual perspectives. What is key to both is that the players have shared knowledge of the objective.
and what needs to be done. This contrasts with the early stages of the game where there was much requesting and giving of instructions. It also
contrasts in that shared knowledge enables players to concentrate on interacting around external objects visible to all rather than needing to interact
with each other, and this removes the need for speaker disambiguation.
c) *Using an Ammo Dump*

This example shows how players deliberately organise external reference so that they do not need to
interact with each other in order to get ammo, freeing them up to concentrate on the game objective. It shows how it is not necessary to resolve who is
speaking or to have mutual perspectives so long as there is shared knowledge of the gamespace or ‘map’, and the objects that populate it. The following
excerpt from the third Return to Castle Wolfenstein session shows how Weepy gets ‘ammo’ by interacting with all the people on her team:
1. Weepy I need ammo anybody got some?

2. Weepy Can anyone give me ammo?
3. Kat

It’s with the health packs round by the flag I think
4. Buzz  Umm, ammo at the flag

5. Weepy  Cheers
           [approaches flag]
The team lieutenant, Buzz, has in this example created an ‘ammo dump’: rather than delivering ammo to players on request, examples of which appear
above, he dumps it at a particular location and players can go to this location and collect ammo as and when they need it. The four players on the
team (which includes Di) are playing a ‘capture the flag’ game which involves them finding flags in different locations, so they need to split up. This means
that their perspectives differ, but that as long as they know the location of the flag referred to, this can be returned to and ammo picked up.
When there is a number of flags, each team needs to establish which is the relevant one, i.e., which flag is being referred to. Figure 5.1 shows Weepy
broadcasting her request for ammo, where Di is in front of her. In Figure 5.2, Buzz appears next to the flag he is referring to. Kat is also near this flag. Co-
location establishes the necessary reference although the player’s perspectives differ. Meaning is disambiguated both through game
knowledge: knowing what ‘ammo’ and ‘health packs’ are; and being able to refer to a shared reference point (the flag) which is known to be in the vicinity: ‘round by
the flag’; ‘at the flag’. Also, in contrast to the example above of Weepy being unable to identify teammates, this player has developed a strategy to
overcome problems involved in not knowing who is speaking - broadcasting and waiting for responses: ‘anybody’; ‘anyone’.
Figure 5 Using Shared Knowledge to Organize Ammo Exchange in ‘Return to Castle Wolfenstein’ (Activision)
Please see Figure 5 in full PDF version
.1 Broadcasting a request for ammo (utterance 1 in transcript); .2 Establishing the location of the ammo dump
Discussion and Conclusion

Our study has shown that despite the technical problems associated with
current forms of VoIP, players have developed a number of VoIP mechanisms for coordinating their gaming activities and progressing
with the game. The findings revealed the work that needs to be done to couple the talk with the action and graphical representations
in the game, and the problems that can arise.
The Interaction of Audio Communications with Graphical Representations
The importance of the physical context in VoIP games was demonstrated when using audio-based talk to refer to the graphical representations used in the
game (e.g., maps, buildings, avatars, weapons, vehicles, documents). It was essential for the creation of shared meaning, in order to focus talk towards common
goals. In face-to-face games, as with face-to-face communication generally, who is speaking can be clearly and immediately perceived, as can what is
being referred to. In addition, talk can be anticipated because it can be heard as it unfolds. What comes for free in face-to-face communication was
shown to need cognitive effort in multiplayer games. Utterances need to be linked with avatars. Views onto the same objects need to be established. Confusion
can arise in terms of turn-taking and focus in conversation. In text based games, these issues are at least partly compensated for by the fact that
communications are of the same ‘material’ as the rest of the game: graphical. Notwithstanding issues with turn-taking, this can have some intrinsic
advantages for the integration of communications with the other graphical materials of the game, notably avatars: labelling with gamertags,
and (with bubbletalk), spatialisation.
**Speaker Disambiguation**

The technical limitations of VoIP mean that voices sound quite different to those in face-to-face
interaction, positional and ambient cues are absent, they are monaural and at the same amplitude regardless of distance of the speaker (or avatar), and
same-sex voices sound similar. As our qualitative analysis has shown, this can make it hard to relate an utterance with the identity of the speaker. This is
exacerbated by the fact that avatars’ appearance and behaviour can be similar. Linking gamertags and utterances to avatars can be difficult to accomplish.
To compensate, the provision of additional graphical tools, spatialisation protocols and ‘voice avatars’ (that enable speakers to change the
sound of their voices to make them more distinctive), may help players recognize more easily who is currently speaking.
However, spatialisation of voices (an approach suggested by researchers including Gibbs et al., 2006; and Singh and Acharya, 2004) may not necessarily
be a good idea for games like Return to Castle Wolfenstein since players make use of the non-positional, equal amplitude properties of their VoIP-
represented voices when splitting into subgroups. In addition, some graphical representations, like the compass tool, require considerable cognitive
effort to understand and associate with a particular speaker. Voice avatars were hardly used by the players as it results in exaggerated ‘cartoon voices’ which were
regarded as irritating to listen to. The players who tried using voice masks were quickly requested to switch them off, despite the
gains in voice distinctiveness.

The reason for this appears to be that where certain conditions were in place in
the game, speaker disambiguation happened without the need for such support either from further graphical tools or from voice spatialisation, or
voice avatars. These include the presence of only two speakers, focused interaction around an object in a shared perspective, and feedback.
In the later stages of the game, shared knowledge of maps, objectives and roles often reduced the need for speaker disambiguation. For coordinated actions,
there was little evidence of the use of gamertags and other graphical tools to identify speakers, or of the need for spatialisation. The development of game
knowledge (maps, weapons, levels, etc.) tends to make speaker disambiguation less important, and players conduct joint actions by
means of reference to the environment and objects rather than to each other as specific individuals. This is also supported by organising external
reference to objects, including ammo (through ammo dumps), where the object is initially associated with a given player.
This raises the question of how far there is a need for additional audio and graphical representations to help disambiguate voices by associating them with
avatars. It may be that persistent gamertags, which appear at all times alongside avatars, together with the appearance of gamertags with speaker
icons at the bottom of the screen when that avatar speaks, is an optimal set-up.
Multithreading and Turn-Taking

Text-based computer games are characterized by a multithreading mode of
interaction, sometimes making it difficult for players to communicate or collaborate effectively with each other. This happens most where players have to
wait for another’s utterances to be completed before appearing on their screens. We found that VoIP brought back the cue of anticipation that is lost in
this mode of interaction. Furthermore, turn-taking was found to persist even during multithreaded discussions. Two independent discussions
were found to occur, but they did not cross-cut (i.e. the utterances did not happen simultaneously). This appears to be to
support monitoring while also preserving meaning.

Hence, players were able to adapt to the properties of VoIP. A benefit is that
players can coordinate as a team through monitoring while holding separate self-contained discussions. In some situations, this happens because the game
gives rise to clearly distinct channels one of which is relevant to current action while the other is not, so can be disattended. But a disattend channel is still
registered (and may remain relevant to higher order goals), allowing discussion to encompass all the players across both channels if necessary. It
may not be necessary, therefore, to try to mimic what happens in face-to-face settings, where if a speaker is further away, their voice should be at
lower volume and resolution. Furthermore, we propose that spatialising audio or altering amplitude with distance might even be
counter-productive, since it may interfere with or disrupt the mutual monitoring we found was important.
Changing Relationships between Talk and Game

Our findings showed how talk changes over time from coaching to coordination, as
the games progressed, generating different needs in terms of the relationship of VoIP communications to the graphical materials of the game. Novice players
need to coordinate with other team members to learn how to take appropriate action. For example, players like Di and Weepy needed to be told
how to (respectively) perform the role of lieutenant, and defend a position. Later on, this knowledge forms a basis for more sophisticated
behaviour. Some players did not need coaching, including Mars, Buzz, Reevez and Shimmer, but an important property of VoIP is that it allows
experienced players to do this coaching at the start of play. Thus whether or not it is needed by a particular player, VoIP makes it possible to integrate new
and inexperienced players quickly.

Finally, should enhancements to VoIP err towards greater fidelity
with face-to-face talk? Would this make communication more or less natural, given that the context of playing will remain in a 3D virtual
world? Certainly the quality of voices could be improved so that it is easier to disambiguate between them, when there are several people playing the
game. But when there are only 2-4 players in the same vicinity it may not be necessary. Although the talk that results from the introduction of VoIP to
multiplayer games is impoverished compared to face-to-face talk, and even, in some ways, to text, our study has shown that players are able to use this
form of audio conferencing effectively to play a war game that requires teams competing against each other. This is because the players used the graphical
representations in conjunction with their audio-mediated voices in different ways to how they might talk if playing a co-located, physical equivalent
of the game. It also shows how people are good at negotiating interaction in multiple settings whether ‘real’ or ‘virtual’. We need to examine how people
work with, appropriate and interact in different environments, and consider what kind of meaning is generated how, rather than seek to characterise and
build virtual environments that mimic the real world.
Acknowledgments

The research reported here was funded by the PACCIT LINK programme under the ESRC/DTI initiative.
References


One and with Two Ears," Journal of Acoustic Society of America 25, 975--979.

and implications of advanced computing.

Interaction Patterns in a Massively Multiplayer Game," Proc. CSCW’04, 360-369.

"Proximity-Based Chat in a First Person Shooter: Using a Novel Voice Communication System for Online Play," Proceedings of the 3rd Australasian
conference on Interactive entertainment, 96-102.

of Edinburgh Social Sciences Research Centre.

Communication and HCI: An Examination of the Use of Computers in Medical Consultations, "Interacting With Computers, 5, 193-216.

University of California Press.

Hughes, L. (1983). "Beyond
Varieties of Attention.

"Interaction Forms and Communicative Actions in

Muramatsu, J. & Ackerman, M. S. (1998). "Computing,
Social Activity, and Entertainment: A Field Study of a Game MUD, "Computer Supported Cooperative Work: The Journal of Collaborative

Support for Multiplayer Networked Games,"

Wright, T., Boria, E. & Breidenbach, P. (2002). Creative Player Actions in FPS Online Video Games: