2009 Conference in Games and Virtual Worlds for Serious Applications

It’s Talk, But Not as We Know It: Using VoIP to Communicate in War Games

John Halloran
Serious Games Institute / Department of Computing and the Digital Environment
Coventry University
Coventry, UK
John.Halloran@coventry.ac.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. What effect does this have on gameplay? A longitudinal study of a multiplayer team game is presented. Our analysis looks at how the players (all adults) used VoIP to resource their interaction with each other in the virtual game world. We found that VoIP represents talk in ways that differ both to face-to-face communication and to text-mediated communications. VoIP audio representations interact with, and mediate, the graphical materials of the game world in distinctive and unusual ways which can generate problems to be overcome for players. But they also provide clear benefits for learning and coordination, which are found neither in face-to-face or text communication. We conclude by considering the implications of these problems and benefits for design.

Keywords: VoIP; multiplayer games; user study; computer-mediated communication; human-computer interaction; virtual environments.

I. INTRODUCTION

Until recently, multiplayer gamers communicated by means of text messages, but Voice over Internet Protocol (VoIP) now makes it possible for gamers to talk to each other using headphones and a microphone. The use of VoIP may well become standard for communications in games: Microsoft’s Xbox Live was the first proprietary gaming package to integrate VoIP, and its latest, Xbox 360, uses the Live platform. In addition, Sony and Nintendo plan to integrate VoIP into their own products. Since VoIP for games is clearly expanding, the purpose of this paper is to contribute to understanding how it behaves as a communications medium for games, and what such an understanding implies in terms of games design.

A longitudinal qualitative user study was carried out over a four-month period. We looked at how a pool of adult gamers used VoIP on Xbox Live. Here, we present detailed findings from the game that was played most frequently, Return to Castle Wolfenstein. Our main research question was how talking through VoIP influences and shapes gameplay. Our focus is twofold. First, we analyse how VoIP is used by players to learn games, and to coordinate action. Second (and relatedly), we examine how VoIP communication in games differs from face-to-face talk, and from text communications. In particular, we show that while the properties of VoIP present challenges to gamers, it has benefits for gameplay, supporting novel forms of interaction not found either in face-to-face or text communications. On the basis of our findings we discuss how far VoIP-supported games need to be redesigned to obviate the problems, and to further support its benefits.

II. BACKGROUND

Talk has always been important in playing games, serving a variety of functions. These include discussing and reinterpreting rules in children’s playground games [8]; announcing a hand in Poker, and bluffing by false announcement [6]; discussing handicaps in golf [4]; and sociable chat in bowling [11]. 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.

However, in distributed multiplayer games, players are not co-present, but geographically separated. The game is played in a 3D virtual environment, where the players are not present in person, but are represented by avatars. And, until recently, instead of talking, players needed to use text in order to communicate. Unsurprisingly, this differs to face-to-face communications in a number of ways.

A. Text Communicattons in Games

Research into text communications in games reveals three important issues: multithreading, labelling and spatialisation.

1) Multithreading

In face-to-face communication, 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 [12]. This has two 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

978-0-7695-3588-3/09 $25.00 © 2009 IEEE
DOI 10.1109/VS-GAMES.2009.36

ways without interrupting, or leaving pauses (which only occur if they are necessary).

In contrast, text messages in many computer games do not appear word-by-word, 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. Curtis [2] points out that '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'. This may be problematic where speakers all need to concentrate on one topic.

The appearance of multiple simultaneous conversations – known as 'multithreading' - may cause confusion, and can represent a problem where there is a need to focus on the same topic (see e.g., [10]). However, in recent games that use text-based communications, including those that use 'bubbletalk' (see below), utterances unfold word-by-word, and this supports focus and turn-taking [1], helping remove the problems.

2) *Labelling and Spatialisation*

In face-to-face communication where discussion is happening amongst several people, who is speaking is obvious, by virtue of the distinctive timbre of the voice, the unique appearance of the speaker, and the co-variation of speech with mouth movements. In contrast, text looks the same regardless of speaker, and lip movement does not apply. In purely text-based games with no graphical world, the identity of a speaker is given by the labelling of messages with the player's 'gamertag' (in-game player nickname). In games which do feature graphical worlds, the player is represented by an avatar which is labelled with the gamertag (which often floats above it). Text messages are labelled with the same gamertag. Therefore, labelling supports players using text communications to work out who is speaking.

Face-to-face communications are spatialised: voices move where speakers move, in space. This is another cue to establishing who is speaking. In many games which use text communications, text strings all appear in the same place so are not spatialised. So labelling is essential. However, some modern games - for example 'There', a massively multiplayer role-play game (MMORPG) - use 'bubbletalk' [1]. Here, messages appear in speech bubbles which appear over players' heads, so that utterances are spatialised. Because the location 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 both to perceive what everyone is saying, and to communicate across distances, holding discussions with distant players while many intervening conversations may also be taking place. 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.

B. *VoIP Communications in Games*

Knowing who is talking is important. Our previously published research has shown that it is crucial in team-based games, in order to be able to collaborate effectively [5]. 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. However, working out who is speaking – 'speaker disambiguation' – presents problems.

VoIP represents spoken utterances in distinctive ways that differ from face-to-face communications. All voices are represented at the same amplitude, regardless of the distance of avatars from each other. In addition, VoIP represents voices monaurally, not in stereo. Hence, VoIP-represented voices have all positional information removed and do not co-vary with the position of avatars. An obvious cue to resolving this is the individual timbre of voices. However, 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. This has the effect of making same-sex voices sound similar.

These issues mean that there is as much a need to label VoIP utterances with gamertags as there is in games which use non-spatialised text communications (i.e., non-bubbletalk games). However, it is difficult to see how a spoken utterance could be labelled in such a way (automated voice-labels could obscure the utterance itself; asking players to prefix all their utterances with their gamertag might soon be forgotten), which means that labelling has to rely on graphical resources. As we will see below, graphical implementations of such labelling in Xbox Live are limited and the problem of speaker disambiguation has not been fully addressed. This is one of the major issues facing gamers using VoIP.

III. The Study

A. *Aim*

The aim of our study was to explore how VoIP-mediated communication was used by gamers, and how it shapes the way distributed multiplayer games are played. A particular emphasis was on how gameplay is influenced by the properties of VoIP as an audio representation. We were concerned to understand how these audio representations interact with the graphical representations found in games, and how the resulting ensemble is used by players to produce successful gameplay.

B. *Design*

A group of 10 adults aged between 20 and 48 took part. Seven of these were male and three female. We classified players regarding expertise. 1 male and 2 females were classed as novices; 1 female as intermediate; and the other 6 males as advanced. The members of the group were largely unknown to each other before the study.

Each participant was equipped with a broadband connection, an Xbox Live console, an Xbox controller, and an Xbox Communicator (headphone with microphone). Several games were made available that they could choose to play, including Gotham Racing, Midtown Madness, Ghost Recon: Island Thunder, and Return to Castle Wolfenstein.

The participants gamed together once a week for 10 weeks at a fixed time, for 60 minutes. The most popular game was Return to Castle Wolfenstein, which was played for five of the ten weeks. This is a fast moving game with a World War II theme and featuring two teams, ‘Axis’ and ‘Allied’. Members of a team can hear and talk to members of that team only; not the other. The teams compete to meet a variety of objectives which 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.

C. Method

We used an adapted form of virtual ethnography [7] in which we observed our players in their rooms. Hence, they were all aware of our presence as observers but not as participants - virtual ethnography is usually the other way round. One advantage was that we could record games from the viewpoint of the players, rather than our own; another was that we could record more than one perspective.

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. We captured both screen and audio. Recording two different audiovisual perspectives onto simultaneous game events helped us to think about similarities and differences across different groupings of participants in games.

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. For a variety of reasons, we were not always successful. 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. The 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 and content of talk. Our qualitative analysis was designed to find out more about how talking with VoIP shapes and resources gameplay.

IV. FINDINGS (1): QUANTITATIVE ANALYSIS

Our quantitative analysis established that there were broadly similar amounts of talk across the five sessions. Figure 1 presents the number of words spoken per minute (WPM) versus the number of utterances per minute (UPM) across the sessions. 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.

	Session 1	Session 2	Session 3	Session 4	Session 5	All
WPM	76.73	63.74	71.25	72.27	84.53	73.7
UPM	10.81	9.46	11.27	11.18	12.31	11

Figure 1. Amount of talk by session

Figure 1 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 of a similar nature regardless of who happens to be playing with whom.

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 could be 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. Figure 2 shows this co-variance. The low WPM/UPM is 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.

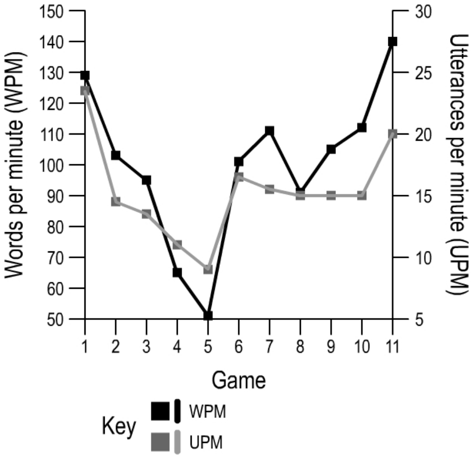

Figure 2. Amount of talk by game

These findings also show that there was a good deal of talk, suggesting that it was important for players to be able to do so.

To find out more about what the talk concerned, we developed a two-level coding scheme. The first level classified the players' utterances in terms of three codes: G (game), MG (metagame), and OG (outgame). The G 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 MG code was used to refer to comments that reflect general attitudes or knowledge derived from repeated experience, for example, "This is faster-moving than the other level". The OG code was used for utterances about things other than the game, e.g., "How's the weather where you are?". The second level of the coding scheme labelled utterances according to the particular action being referred to. Codes at this level are too numerous to list here, but fell into four clusters. The 'instruction' cluster included codes like GIVE_INST (give instruction, e.g. "We need to collect five flags") and REQ_INST_ACT (request instruction on how to act, e.g. "How do you do an airstrike"). The 'information' cluster, similarly, featured codes like GIVE_INF (e.g. "The documents are on the table") and REQ_INF (require information, e.g. "Where are you?"). The 'action' cluster included REQ_ACT (require action, e.g. "Take that flag"), CONF_ACT (confirm that action will be taken, e.g. "OK I am taking the flag now"), and REQ_STOP_ACT (require action to stop, e.g. "Don't shoot me, I am on your side"). The 'other' cluster included codes like ADDR (address another player) or COMM (make a comment). Thus, each utterance was coded at the two levels. For example, "I've got a grenade and I've got a gun" was coded G: GIVE_INF.

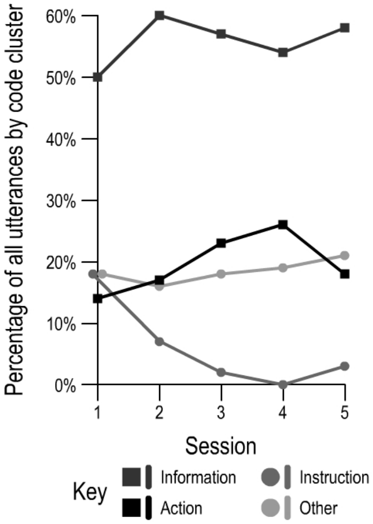

Figure 3. Utterance types over time

In terms of the Level 1 codes, the G code accounted for 90% of all utterances, MG for 9%, and OG for only 1%. This indicates that the vast majority of the talk was about playing the current game; players had little time to talk about anything else.

Figure 3 shows the average percentage of 'instruction', 'information', 'action' and 'other' codes per session. It can be seen that the majority of utterances concern information, ranging from 50-60%. Utterances relating to action 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.

The 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 - that is, coaching and learning - 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.

V. Findings (2): Qualitative Analysis

A qualitative analysis was carried out to find out more about the two major types of activity identified: coaching and learning, and coordination. In this section, we look at a number of key vignettes which illustrate how coaching and learning, and coordination, were achieved through the use of VoIP, focussing on the challenges and benefits VoIP generates.

A. *Coaching and Learning*

1) *Toasting your team*

A basic challenge for novice players is to work out who else is on their team. If this is not clear, the player may act inappropriately. Thus, in Session 1, Weepy (all names changed) repeatedly turned her flamethrower on her team-mates, 'toasting' them.

In the vignette discussed here, Weepy is faced with the simultaneous presence of a number of avatars and voices in an unfamiliar game. The relationships between these - which voice belongs to which avatar - are too difficult for Weepy to resolve. 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. But this 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 not just because it assumes prior knowledge of what team she belongs to, or because uniforms look similar regardless of team or role. She is also confused due to the representational design of Return to Castle Wolfenstein.

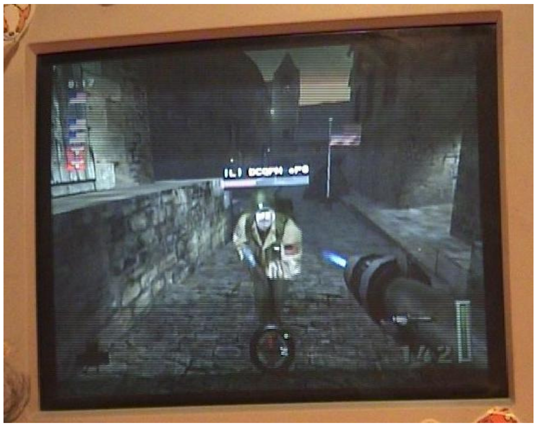

Figure 4. Problems identifying a team member

Figure 4 shows Weepy's viewpoint (hers is the weapon - a flamethrower- in the foreground). In the middle of the screen is an avatar with a tan and green uniform. At the same time, a voice, which belongs to this avatar, tells her what colour her team's uniforms are; and (after this screenshot) that he has been 'toasted'. All these are resources to help Weepy establish that this avatar is a member of her team, and that she is taking inappropriate action in relation to it. However, she is unable to link the voice to the avatar. Although the avatar carries Xlr8's gamertag, his voice - which is one among many - is not labelled. There is one resource to make this connection: the compass at the bottom centre of the screen. When an avatar speaks, an icon on this compass flashes. But this requires (a) noticing the compass (but only one player of the 10 said they ever made any use of it); and (b) understanding how the location of an avatar maps to a direction on the compass. Because of these representational issues, Weepy is unable to link the voice complaining about 'toasting' to the avatar in front of her, and so behaves inappropriately. Thus, in situations where there are many avatars and speakers, speaker disambiguation can be a problem. However, under certain conditions, learning can happen more smoothly, as the following vignette shows.

2) *How to deliver ammunition*

In the following excerpt, a novice 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. Following utterances 1-6, an avatar appears on screen. There is no other avatar simultaneously on screen and no other voice can be heard. This is followed by:

```
7  Buzz       If you press your change weapon
              button
8  Buzz       Lancelot
9  Lancelot   Yeah
10 Buzz       Press your change weapon button and
              you get
11 Lancelot   Oh there's a pod, there's a pod
              isn't there
12 Buzz       Yeah there's a pod
13 Lancelot   Ah that's what it is
14 Buzz       Press that
15 Lancelot   Got you
16 Buzz       One more
17 Buzz       One more
18 Buzz       And one for luck
19 Lancelot   Ah Yes, great
```

Utterance 13 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 5 (the view is Lancelot's).

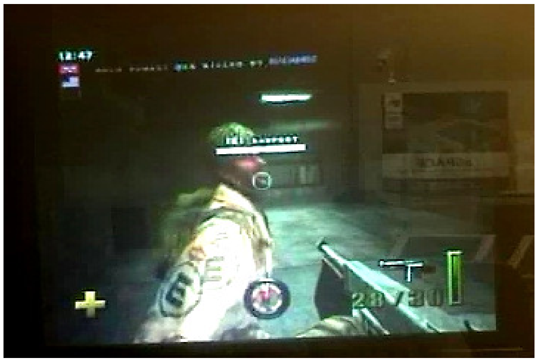

Figure 5. Successful interaction

Why is Lancelot, a novice, able to interact successfully and appropriately with a team mate where Weepy was not? This episode shows the effective construction of an interaction: effective in that it supports a key event that needs to happen in the game (delivering ammunition). Like Weepy, the central issue for Lancelot is the integration of avatar and utterances to recognise the individual he is 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. Thus, self-contained pairwise interactions with VoIP obviate the need for further support for linking utterances to avatars.

B. *Coordination*

Speaker disambiguation is a major issue for learning in this VoIP-supported game. But, as our quantitative analysis shows, talk moved from learning to coordination. At this stage of the game, novices had learned to play. A notable part of the process of moving from learner to more expert player is the reduction of dependency on other speakers for purposes of action. Instead of referring to each other, which involves the need for mutual identification, players begin to refer to and organise objects in the environment in order to achieve this.

1) *Organising an ammunition 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 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, in contrast to the organisation of ammunition delivery in the preceding vignette, 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, he dumps it at a particular location and players can go to this location and collect ammo as and when they need it. In order to be able to do this, players need a range of game knowledge. In addition to knowing what 'ammo' and 'health packs' are, there needs to be knowledge of a shared reference point: the relevant flag which is known to be in the vicinity - 'round by the flag'; 'at the flag'. In addition to this knowledge, Weepy, in contrast to the example of her confusion given above, has developed a strategy to overcome problems involved in not knowing who is speaking - broadcasting and waiting for responses: 'anybody'; 'anyone'.

The development of shared knowledge about objectives among players removes the need for speaker disambiguation. Here, the coordination of delivery of ammunition is organised by attaching this to a location that is known by all. This contrasts with the example of Lancelot and Buzz discussed above, where, to coordinate the same function, the two speakers needed to identify each other. This suggests that the problems presented by VoIP in terms of speaker disambiguation are obviated when players develop strategies that remove this need as they become more expert.

2) *Moving as a team*

Achieving a team objective requires considerable coordination that is primarily oriented towards specific locations. Teams can move as large groups, split off as pairs, and get back together again. For example, during a game in Session 3, a team consisting of Mars, Di, Lancelot and Reevez, congregate at their starting point, a trench. They need 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 requires them to enter the building through its roof. The following interaction starts 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 Lancelot Yeah

By the end of this exchange, 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 VoIP represents voices as equally loud, and equally present, regardless of avatars' proximity to each other.

8 Mars Here this way this way, Reevez back back back, jump up here
9 Reevez OK
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 the way up to the top
18 Di Yep. I'm right behind you
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 at 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. Here, the communication is across the whole team:

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

In this vignette, the two pairs are able to hold two separate coherent discussions about different locations. Both pairs are able to 'tune out' the other conversation, but to monitor it at the same time, and to rejoin as necessary, similar to the cocktail party phenomenon reported on in psychological studies of dual attention (see e.g., [9]).

The four players, in holding two separate discussions, are effectively multithreading - there are two different discussions. However, the threads do not cross-cut each other, and there is no interruption. In other words, the four speakers observe turn-taking rules as if they were involved in a single conversation, rather than two clearly separate threads. The reason why turn-taking is happening across the whole group, and not just within its two subgroups, appears to be that the group needs to monitor its activity as a whole in order to achieve the objective: each subgroup needs to be able to hear what the other is doing. Additionally, cross-cutting speech would make both threads incomprehensible. It also 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 visual perspectives in order to collaborate around key objects and locations (documents, rooms): the support offered by VoIP for mutual awareness of team members in different places doing different things, allows them to work together even where what they can see is completely different. Both multithreading and differing player perspectives have been reported as problems in the games literature. In direct contrast, this vignette shows that these are both beneficial, supporting coordinated activity at a distance, when VoIP is the communications tool.

VI. Discussion and Conclusions

Our study has shown that despite the technical problems associated with current forms of VoIP, much use was made of it, and it has clear value for coaching and learning, and coordination. Players have developed new kinds of behaviour that capitalise on the distinctive features of VoIP in order to realize benefits. Here we discuss these limitations and benefits further, and consider implications for design revisions to VoIP-supported multiplayer games.

The properties of VoIP (lack of spatial and amplitude cues, same-sex voice similarity), together with absence of labels for utterances, 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. To compensate, the provision of additional graphical tools, spatialisation protocols and 'voice masks' (that enable speakers to change the sound of their voices), may help players recognize more easily who is currently speaking.

However, spatialisation of voices (an approach suggested in e.g., [3]) 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 to coordinate their actions when splitting into subgroups: they use VoIP to multithread. In addition, some graphical representations, like the compass tool, seem to lack salience for players and are hardly used. Voice masks were hardly used by the players as they result in exaggerated 'cartoon' voices which were regarded as irritating to listen to. Players who tried using these were quickly requested to switch them off, despite the gains in voice distinctiveness.

This reflects that where certain conditions were in place in the game, speaker disambiguation happened without the need for such support from further graphical tools, voice spatialisation, or voice masks. These include the presence of only two speakers, focussed 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 (of maps, weapons, levels, and so on) tends to make speaker disambiguation less important, and players construct and conduct joint action by means of reference to the environment rather than to each other as specific individuals. This is also supported by organising external reference to objects, including ammunition (through ammo dumps), where the object may initially be 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. This set-up is, indeed, the one used in other Xbox Live games, including Gotham Racing (see Figure 6). More importantly, a solution like spatialisation, in particular, would likely remove a benefit of the current implementation of VoIP: multithreading which supports the coordinated action of a team that has split into subgroups.

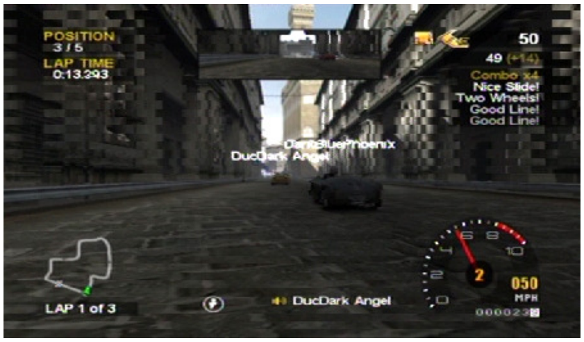

Figure 6. Representational arrangement of Gotham Racing.
The avatar 'DucDark Angel' is the furthest away of the two. This avatar is speaking, indicated by the speaker icon at the bottom of the screen, which carries the same gamertag.

As we saw, one of the key issues in the literature on text-based computer games is multithreading. This sometimes makes it difficult for players to communicate or collaborate effectively with each other, since it is associated with simultaneous discussions of different topics, where collaboration requires all to focus on the same thing. Multithreading in text games happens, in particular, where players have to wait for another's utterances to be completed before they appear onscreen. We found that VoIP brought back the cue of anticipation that is lost in this mode of interaction. However, perhaps counter-intuitively, multithreading was also found in our study. Two independent discussions were found to occur, which supported the coordination of the team whose members were having these two different discussions. However, unlike other forms of multithreading, which can cause problems with turn-taking, we found that VoIP-supported multithreading features turn-taking not just within each thread, but in the discussion as a whole: different discussions do not crosscut. This happens because the team needs to monitor the different discussions of its subgroups in order to achieve its objectives. What is more, VoIP allows this by representing voices at the same amplitude. Thus, VoIP enables a new form of multithreading which supports team coordination at a distance.

This finding shows that VoIP has unanticipated benefits which are important to preserve. It also shows that it may not be necessary, in virtual environments including games, to try to mimic what happens in face-to-face settings, where if a speaker is further way, their voice should be at lower volume and resolution. We propose that spatialising audio or altering amplitude with distance might even be counter-productive, since it could disrupt this new form of multithreading.

The foregoing discussion considers how we should balance problems and benefits in VoIP when considering design revisions. This issue reflects our finding that the content of talk changed over the course of the study from coaching to coordination, 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. Later on, this knowledge forms a basis for more sophisticated behaviour. Some players did not need coaching, including Mars, Buzz, and Reevez, 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 novice players quickly. This ability of VoIP to leverage learning and coaching may support more rapid development of players to expert level, and the design of more challenging games with steeper learning curves.

Acknowledgments

The research reported here was funded by the PACCIT LINK programme under the ESRC/DTI initiative. Thanks to Yvonne Rogers, Geraldine Fitzpatrick, and Paul Marshall, who all made important contributions.

References

[1] Brown, B. and Bell, M., 2004. Social Interaction in 'There'. In Proc CHI'04, 1465-1468.

[2] Curtis, P., 2002. Mudding: social phenomena in text-based virtual realities. Proceedings of the 1992 conference on directions and implications of advanced computing.

[3] Gibbs, M., Wadley, G. and Benda, P., 2006. Proximity-based chat in a first person shooter: using a novel voice communication system for online play. In Proceedings of the 3rd Australasian conference on Interactive entertainment, 96-102.

[4] Goffman, E., 1959. The Presentation of Self in Everyday Life. University of Edinburgh Social Sciences Research Centre.

[5] Halloran, J., Fitzpatrick, G., Rogers, Y. and Marshall, P., 2004. Does it matter if you don't know who's talking? Multiplayer games and voiceover IP. In Proc. CHI 2004, 1215-18.

[6] Hayano, D., 1982. Poker Faces: The Life and Work of Professional Card Players. University of California Press.

[7] Hine, C., 2000. Virtual Ethnography, Sage.

[8] Hughes, L., 1983. Beyond the rules of the game: why are Rooie rules nice? in: F. E. Manning (Ed.), The World of Play. West Point, NY: Leisure Press, pp. 188-199.

[9] Kahneman, D. & Treisman, A., 1984. Changing views of attention and automaticity, in: Parasuraman, R. and Davies, D. R. (Eds.) Varieties of Attention. London: Academic Press.

[10] Muramatsu, J. and Ackerman, M.S., 1998. Computing, social activity, and entertainment: a field study of a game MUD. *Computer Supported Cooperative Work: The Journal of Collaborative Computing*, January, 1998, 7(1), pp. 87-122.

[11] Putnam, R., 2000. Bowling Alone. Simon and Schuster.

[12] Sacks, H., Schegloff, E. A. and Jefferson, G., 1974. A Simplest Systematics for the Organisation of Turn-Taking for Conversation. In Language, 50:696-735.