

Faculty of Informatics

Faculty of Informatics - Papers

University of Wollongong

Year 2005

DICE: Internet delivery of immersive
voice communication for crowded virtual
spaces

P. Boustead* F. Safaei[†]

M. Dowlatshahi[‡]

*University of Wollongong, boustead@uow.edu.au

[†]University of Wollongong, farzad@uow.edu.au

[‡]University of Wollongong, mehran@uow.edu.au

This article was originally published as: Boustead, P, Safaei, F & Dowlatshahi, M, DICE: Internet delivery of immersive voice communication for crowded virtual spaces, IEEE Virtual Reality 2005 Proceedings (VR2005), 12-16 March 2005, 35-41. Copyright IEEE 2005.

This paper is posted at Research Online.

<http://ro.uow.edu.au/infopapers/42>

DICE: Internet Delivery of Immersive Voice Communication for Crowded Virtual Spaces

Paul Boustead*
Smart Internet CRC
Telecommunications and IT
Research Institute,
University of Wollongong

Farzad Safaei†
Smart Internet CRC
Telecommunications and IT
Research Institute,
University of Wollongong

Mehran Dowlatshahi‡
Smart Internet CRC
Telecommunications and IT
Research Institute,
University of Wollongong

ABSTRACT

This paper develops a scalable system design for the creation, and delivery over the Internet, of a realistic voice communication service for crowded virtual spaces. Examples of crowded spaces include virtual market places or battlefields in online games. A realistic crowded audio scene including spatial rendering of the voices of surrounding avatars is impractical to deliver over the Internet in a peer-to-peer manner due to access bandwidth limitations and cost. A brute force server model, on the other hand, will face significant computational costs and scalability issues. This paper presents a novel server-based architecture for this service that performs simple operations in the servers (including weighted mixing of audio streams) to cope with access bandwidth restrictions of clients, and uses spatial audio rendering capabilities of the clients to reduce the computational load on the servers. This paper then examines the performance of two components of this architecture: angular clustering and grid summarization. The impact of two factors, namely a high density of avatars and realistic access bandwidth limitations, on the quality and accuracy of the audio scene is then evaluated using simulation results.

CR Categories: C.2.4 [Computer-Communication Networks]: Distributed Systems—Distributed applications K.8.0 [Personal Computing]: General—Games;

Keywords: immersive voice communication, virtual environments, networked games, virtual reality

1 BACKGROUND

We define a crowded virtual space as an area in a Networked Virtual Environment (NVE) such as a Multiplayer Online Game (MOG) where a large number of avatars congregate. Examples include a virtual conference or workshop, as well as battlefields and market places in MOGs. The focus of this paper is on the provision of an immersive voice communication environment that scales to dense virtual environments and can be cost effectively delivered over the Internet. We call this a Dense Immersive Communication environment (DICE). Currently, the most popular large virtual environments are MOGs. Some of these games support many thousands of players simultaneously (hence are often referred to as *massively* multiplayer games). The dominant method of communication between users in these games is a text-based chat system. Third-party Internet voice applications such as Ventrillo and Team-Speak are sometimes used to add *walkie-talkie* style communications over a shared channel with only one person comfortably talking at a time.

*e-mail:paul@titr.uow.edu.au

†e-mail:farzad@uow.edu.au

‡e-mail:mehran@uow.edu.au

IEEE Virtual Reality 2005
March 12-16, Bonn, Germany
0-7803-8929-8/05/\$20 ©2005 IEEE

Walkie-talkie style voice communications are also being incorporated into many new MOGs especially on the Microsoft Xbox platform. We would like DICE to provide the ability for avatars to interact with a virtual crowd effortlessly. Natural multi-person voice communication is characterized by the presence of multiple simultaneous conversations (except in highly structured and formal situations where a strict protocol of "one person talking at a time" has to be observed). The ability to be peripherally aware of the nearby conversations and join these dynamically is critical to our sense of satisfaction of being in the presence of a crowd. Current audio conferencing services (and similarly a single shared voice channel in MOGs) are not well equipped to cope with multiple parallel conversations because they deliver a simple mix of other voices to each participant. In this mix, everyone is at the same level of loudness and their voice is not spatially related to their virtual location. It is difficult for the listener, therefore, to make sense of this mix if there are many simultaneous speakers. Nevertheless, humans have no problem in unravelling this complexity in the presence of a real crowd, because the audio is in perfect harmony with the visual scene with respect to the spatial distribution of speakers. With DICE, the voices of others in one's vicinity are heard in harmony with their visual representation (location, distance and spatial placement with respect to the listener). Each participant can hear a realistic and personalised spatial mix of voices in their 'hearing range' and this mix is dynamically changed as people move within the virtual environment (and consequently in and out of each others hearing range). DICE will provide an important capability in situations where multi-party voice communication is either the primary purpose of gathering in the virtual space or at least an important means to achieve the actual goal

The scalable and cost effective provision of DICE is a significant challenge. At each instant and for each avatar, the audible range of its voice has to be determined (based on the loudness of voice and audio propagation characteristics of the virtual environment such as presence of barriers and walls). This voice will have to be included in the audio scene of all avatars within this audible range. For example, using a peer-to-peer model, this audio packet has to be multicast to all these avatars' clients. The clients, in turn, receive multiple audio streams from everyone in their 'hearing range' and need to spatially render this audio scene by placing the source of these audio streams in conformity with the relative location of avatars in the virtual environment. As avatars move, the clients must dynamically reconfigure their multicast trees. This model has significant cost and scalability issues. In crowded spaces, the client's Internet access bandwidth will limit how many audio streams they can receive. Also for each client (and the underlying network) to maintain highly dynamic multicast trees is computationally expensive.

Several research projects have studied different architectures for voice communications in virtual environments including [2], [5] and [4]. These studies, however, have not focused on high density virtual environments. The delivery of a dense immersive audio environment using peer-to-peer, central server and several distributed server architectures has been quantitatively studied in [3] [7][6][8]. These works conclude that a distributed server architecture is nec-

essary in order to meet the delay requirements of the service when the users are spread over a wide geographical area. However, there are several important issues that were not addressed, including: (i) how to partition computations between the clients and a set of distributed servers; (ii) how to cope with access bandwidth limitations of clients; and (iii) how to improve the computational scalability of the server with due attention to the accuracy of the delivered audio scenes for crowded spaces. In this paper, we will provide a scalable design based on the following observation. From the perspective of each avatar, individual audio sources are of varying importance. Here we define the *interactive zone* as the immediate vicinity of the avatar where active communicative interaction may take place, while the *background zone* is the region outside the interactive zone stretched to the limits of hearing range. Within the interactive zone quality requirements are more stringent when compared to the background zone. Hence, we can tolerate more inaccuracy in spatial rendering for the background zone.

This paper proposes a novel technique that divides audio scene creation tasks between servers and clients in conjunction with a technique that we will call *angular clustering* to reduce the access bandwidth requirements. Only simple computations, namely weighted audio mixing operations, are performed on the audio streams in the servers. Scalability to dense environments is enabled by a technique called grid summarisation. In essence this architecture trades off accuracy for computational complexity and scalability. In order to examine accuracy, this paper presents a scene accuracy metric based upon a concept called *acceptable angular error* and discusses the impact of our architecture on scene accuracy. We have implemented our DICE architecture for a trial of this service which is currently underway. An initial discussion of these trials will be presented, further details on the trial and associated user studies will be presented in future publications.

The rest of this paper is organised as follows: Section 2 presents our design requirements and assumptions. Section 3 describes our proposed architecture and summarization techniques. In section 4 simulation results are presented to examine the performance of various aspects of our architecture. The testbed is described in Section 5. Section 6 discusses the results and briefly explains future directions of this work.

2 DESIGN REQUIREMENTS, PERCEIVED BOTTLENECKS AND ASSUMPTIONS

The DICE service was designed to facilitate natural communications in large-scale virtual environments with users distributed around the Internet. The following design requirements, perceived bottlenecks and assumptions were taken into account when designing DICE.

C1 - Scale to large virtual environments with many avatars.

This will allow the addition of a natural voice communication service to large scale virtual environments such as online games, trade shows and international conferences.

C2 - Scale to dense (crowded) regions in the virtual world.

The cost of a brute force solution would increase significantly with an increase in the average number of talkers within hearing range of avatars. The aim of the DICE service is for resource requirements to scale less than linearly with an increase in the number of avatars within hearing range.

C3 - Scale to a large geographical spread of participants.

Many of the large multiplayer games such as Everquest attract customers from wide geographical regions. It is important to take this large geographical spread into account in the design of the service so as not to excessively degrade performance in terms of delay and resource usage. More

detailed examinations of this issue can be found in [3][7], which discuss the need for distributed servers and different server placement techniques.

C4 - Cope with limited access bandwidth for each participant.

In the design of the service we assume limitations on access bandwidth of clients. In the trial of DICE all the participants have an ADSL access bandwidth of 512k downstream and 256k upstream. However, a large proportion of this access bandwidth is used by the virtual environment, which is a fast paced first person shooter.

C5 - Distant sources require less spatial accuracy. In essence we assume that people can tolerate some degree of inaccuracy in placement of speaking sources when the sources are not in the immediate vicinity of the listener. This major assumption of this work is currently being tested with a group of trial participants (approx. 50 people). Initial reactions from users do not mention the placement inaccuracies.

3 SYSTEM ARCHITECTURE

This paper proposes an architecture that divides DICE tasks between the clients' computing devices and a (set of) server(s) called Scene Creation Servers (SCS). Note that *scene* refers to the *communication audio scene* and not the visual scene throughout this paper unless explicitly mentioned. By a suitable splitting of functions, we can take advantage of existing off-the-shelf functionality in the clients (for example, available virtual/physical speaker systems and rendering hardware on the clients) and meet the design criteria specified above. Essentially the split immersive communications system performs scene simplification in the SCSs and transmits the lower bandwidth simplified scene to clients (addressing C4). The clients then render the simplified scene using the metadata provided by the SCS and any rendering technique.

The architecture of the system is shown in Figure 1. The major functional blocks are:

B1 Scene creation server(s) (SCS)

B2 Audio clients

B3 Control server(s)

B4 VE Clients

B5 VE Server

The audio client (B2) sends a single mono audio stream (captured from a microphone) to one or more SCSs. The audio client receives up to a fixed number (K) of mono audio streams with associated positional meta-data from one SCS server. This positional meta-data is used to spatially place (using client's available spatial audio rendering techniques) each of the received mono-audio streams to construct an audio scene that corresponds, as accurately as possible, to the visual component of the VE. The trivial case is where K is a large number (greater than or equal to the number of avatars in one's hearing range). In this case, we can send every audio stream directly to the client to be spatially placed and the job of the SCS is merely forwarding and multicasting of the streams. However K will be small in practice due to limitations of the access bandwidth and the fact that this bandwidth is also used by the VE state information exchange. In our implementations we have restricted K to 4 to ensure good responsiveness of the application over widely available access bandwidths such as ADSL. In this case, the scene creation servers are used to enrich the small number of available streams. We refer to each of these streams as angular cluster summaries or cluster summaries. Each stream is

Client supports any speaker system.
(Examples include stereo headphones,
5.1/6.1/7.1 speakers,
3D speakers systems, and
3D HRTF headphones)

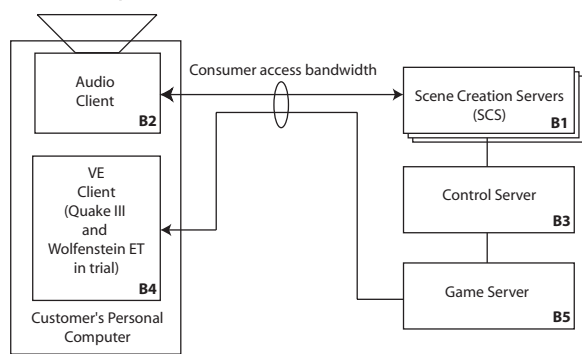


Figure 1: Basic functional elements of DICE

also tagged with the information about the spatial placement of its content, which we refer to as the centre of mass of the cluster (angular cluster). Each cluster summary represents the weighted mix of the audio of all the avatars in a segment of the hearing range around the listener (examples are shown in Figure 2). The weight is applied to the mixing operation to ensure the variations in volume of the individual sources due to distance are maintained. The exact mixing operations and the makeup of the K streams to be sent to the audio clients are determined by the control server. For a large (addressing C1) and geographically distributed virtual world we assume that several separate SCSs will be used. Previous work [3] has shown that it is advantageous to distribute these servers geographically based upon the physical locations of the user base to improve delay performance for interactive voice communications.

The VE client and VE server (B4 and B5) are outside the scope of this paper. Our prototype uses an existing VE client and VE server. We do however, need information from the VE server to correctly render the audio communications scene for each avatar. We propose to obtain this information by monitoring this information at the VE server (possibly by snooping state information update packets or through a defined Application Programming Interface). The required information includes: coordinates of all avatars, status information pertaining to each avatar (dead or alive, mute on/off, special abilities with respect to hearing focus and range allowed by the rules of the application), and pertinent information about the environment that might affect the audio propagation or characteristics (presence of barriers, acoustic characteristics, etc.).

The control server (B3) uses the information from the VE server and information about each of the users available resources (including the access bandwidth) to configure the audio servers (B1). The control server performs the following tasks:

1. Determine which SCS to use to construct the scene for a particular avatar. This decision will be made based upon the respective location of users in the virtual and physical worlds.
2. Configure the scene creation servers to create the audio streams for each avatar. This consists of (a) determining the makeup of the K audio streams that will be sent to the audio client, and (b) calculating the centre of mass of each stream.

We call this process *angular summarisation*.

The remainder of this paper will introduce an angular clustering algorithm and will measure the performance of the entire system in terms of audio scene accuracy. We also present a zone based summarization technique to improve the computational scalability of the weighted mix operations performed by the SCS.

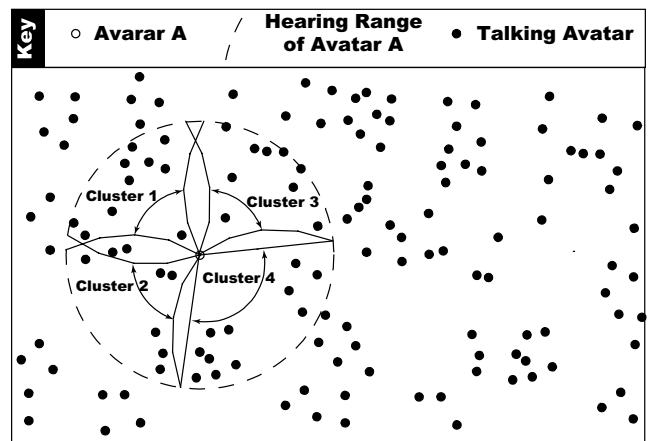


Figure 2: Angular Clustering

3.1 Angular Clustering Algorithm

As stated before, the audio servers will send K mono streams to each client. Each stream is a weighted mix of the audio of all the avatars in a segment of the scene surrounding the listener. Each cluster mix is then spatially placed as a single audio source emanating from a specific location called the centre of mass of the cluster. Because the audio mix is coming from a single point as opposed to the location of individual avatars, any clustering will introduce both angular and distance errors. Using assumption C5 we assume that the perceptual impact of errors will be higher for nearby avatars as opposed to distant avatars.

Here, we introduce an angular clustering method called the Distance Weighted Clustering algorithm that creates two types of clusters *interactive zone clusters* and *background zone clusters*. Interactive zone clusters are created for at least the closest M avatars (where $M \leq K$). After the creation of the M interactive zone clusters K - M background zone clusters are created to cover the large angular gaps in the communications scene that are not covered by the interactive zone clusters.

The interactive zone clusters are created by selecting the next nearest avatar to the listener which has not been clustered before as the centre of mass of a new cluster. In the example in Figure 3 a scene is being created for avatar A_o . Avatar A_k is the closest avatar to A_o and therefore a cluster is created that is centered around the angle θ_k . Another avatar A_j at distance D_{oj} from A_o is considered to be in the cluster centered around θ_k if the angle θ_{jk} is less than an acceptable value ϵ_{oj} . This acceptable value is defined as a function of the distance between the listener (in this case A_o) and the speaker (A_j). If this distance is large then we assume that a larger angular error can be tolerated by the listener. We choose a linear increase in angular error from E_{min} to E_{max} with increasing distance.

$\epsilon_{oj} = E_{min} + E_{max}D_{oj}/D_{max}$ Where D_{max} is the hearing range of avatar A_o . The acceptable angular error of avatars within hearing range of A_o is represented by the area denoted *region of acceptable error* in Figure 3.

The algorithm to create an audio scene for an avatar can be summarized as follows:

1. Choose the closest avatar which has not been included in an existing cluster. If that avatar fits into an existing cluster then place it in that cluster otherwise create a new cluster centred around that avatar.
2. Repeat 1 until M clusters have been created

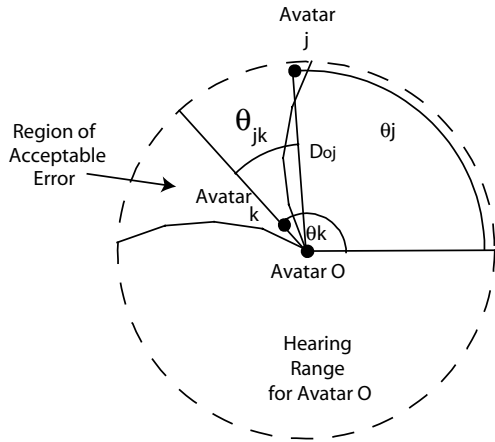


Figure 3: Interactive zone cluster

3. Find the largest angular gap between clusters and create a new cluster in the centre of that gap.
4. Repeat 3 until a total of K clusters have been created.
5. Place all remaining avatars within the best of the K clusters (least angular error)

It should be noted that we do not include avatar orientation in this algorithm since we assume that each avatar is an audio point source, and that the resultant scene for each avatar can be individually rotated to suit the listeners orientation. This rotation of the final scene may be performed in the client or the server. The advantage of performing this operation on the client is that it can be done quickly in response to a rotating avatar.

3.2 Grid Summarization

To create each cluster summary, the SCS has to perform a weighted audio mixing operation. If the number of avatars in a cluster is large, this computation may become expensive to complete in real-time in a server supporting a large number of users. The main goal of grid summarization is to reduce the number of participating audio streams in the mixing operation especially in the densely populated scenes with many simultaneously talking avatars (addressing C1 and C2).

We divide the virtual space using a square grid into 'cells', as shown in Figure 4. Each avatar resides in a cell which we refer to as its home. The hearing range of each avatar will cover more than one cell. In Grid Summarization, the server is required to produce a linear mix (not weighted) of all the voice signals in each of its allocated cells. We refer to the mixed audio stream of each cell as the cell summary. Each summarization server calculates the centre of mass for each of its cells based on the distribution of talking avatars within them. Center of mass of a cell is obtained by averaging the coordinates of the talking avatars residing in that cell.

By using a cell summary in place of its constituent avatars when the totality of the cell falls in the background zone, the computation load on the SCS is reduced, however, we also introduce more errors in spatial placement. The extent to which the SCS computational load is reduced depends upon the size of the cells in comparison to avatar hearing range. This relationship is examined later in this paper.

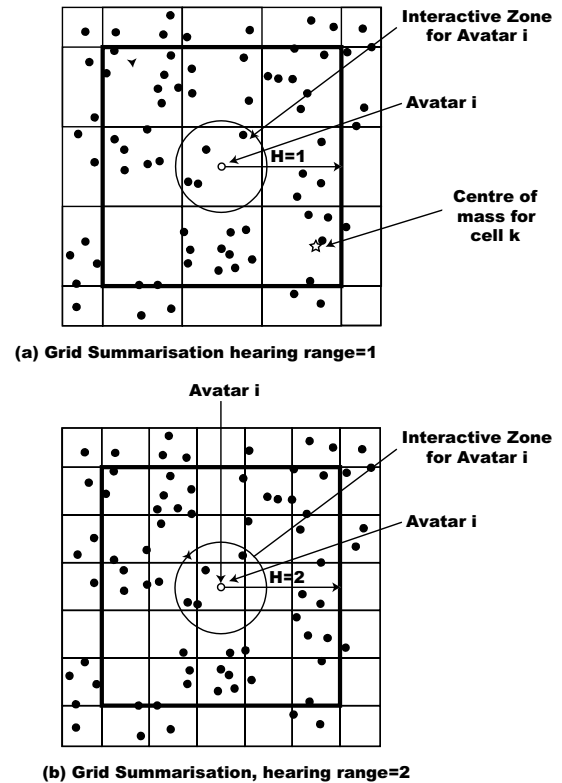


Figure 4: Grid Summarisation

4 SIMULATION RESULTS

The DICE architecture was simulated to measure the performance in terms of angular error, distance error, computational resource requirements in the server and scalability to increasing avatar density. The virtual world was modelled by an area of X by X meters. This virtual world was populated with N avatars in a uniform random distribution of x, y coordinates. The virtual world is then divided into grid summarization zones. No obstructing objects are assumed in the virtual environment and all sound sources are assumed to be omni directional. Each avatar has a square hearing range and can hear all avatars in H surrounding zones as shown in Figure 4 (a) for $H = 1$ and Figure 4 (b) for $H = 2$. When $H = 1, 2$ and 3 this effectively divides the hearing range of the avatars into 9, 25, and 49 cells respectively. A square hearing range was chosen to reduce processing costs in determining the inclusion of talking avatars into hearing ranges of other avatars. Each avatar has on average T talking avatars within its hearing range. The results were taken for a snapshot of avatar locations. Computation cost is calculated in terms of the number of stream mixes required to create each audio scene. The computational cost of the control algorithm is not modelled in this simulation, this cost is dependant on the speed of movement of avatars.

The distance error that is introduced by grid based summarisation is measured as the difference between the actual distance of a talking avatar h in cell K from the listening avatar i and the distance of the centre of mass of cell K from the intended avatar i . Angular error is a result of both grid and angle summarisation and is measured as the angular difference between where the sound source j is rendered from the perspective of avatar i and the actual location of that avatar if summarisation was not performed. We use the notion of an acceptable angular error (ϵ) as defined in the distance

weighted clustering description in Section 3.1. We define deviation from acceptable angular error (V_{ojk}), for rendering of a speaker A_j in Figure 3 from the perspective of a listener A_o rendered in a cluster centered at angle θ_k error, to be 0 degrees if the angular error (θ_{jk}) is less than acceptable (ϵ_{oj}). Otherwise V_{ojk} is calculated as the difference between the acceptable ϵ_{oj} and the actual angular error (θ_{jk}).

$$V_{ojk} = \begin{cases} 0 & \text{if } \theta_{jk} < \epsilon_{oj} \\ \theta_{jk} - \epsilon_{oj} & \text{otherwise} \end{cases} \quad (1)$$

The common parameters used in the simulation are: minimum acceptable error $E_{min} = 15$ degrees, the maximum acceptable error $E_{max} = 45$ degrees. The number of streams to be sent to the audio client (K) is set to 4 and the number of interactive zone clusters (M) is set to 3. Interactive distance of each avatar is assumed to be equal to 10% of the hearing range. The virtual world comprises of 100 by 100 grid cells, i.e. 10000 grid cells.

4.1 Scalability

The scalability of processing costs of the service to dense environments is examined in Figure 5. The density of avatars in the virtual world was changed to vary the average number of avatars in hearing range of other avatars between 10 and 100. The upper limit of 100 was chosen to clearly show the scalability to very dense environments, even though it may be considered high for current applications. The graph shows the number of mix operations required in the SCS with no grid summarisation and when the hearing range is divided into 9, 25 and 49 divisions ($H = 1, 2$ and 3). The number of mix operations with no summarisation increases linearly with increasing number of avatars in hearing range. This value is easily calculated as $A - K$ mix operations. When grid summarisation is used the number of mix operations increases less than linearly. When the hearing range is divided into a 9 cells ($H = 1$) the grid summarisation approach requires 50% less mix operations (than required with no grid summarisation) with 20 avatars within hearing range and 66% less mix operations with 50 avatars within hearing range. Higher values of H reduce the performance of summarisation. When $H = 3$ summarisation reduces number of mix operations by 50% (compared with no grid summarisation) when there are on average 82 avatars within hearing range. These results indicate that grid summarisation is only effective (assuming the average number of avatars within the hearing range is less than 50 in a practical scenario) if the grids are very large in proportion to the hearing range. If grid squares are large then the architecture will enable scaling to large densities of avatars without significant cost increase. However large grid summarisation squares will lead to additional distance and angular errors that are described in the next sub-section.

4.2 Angular and Distance Errors

Figure 6 shows the percentage of avatars that can be placed in angular summaries within the acceptable angular error (as defined in Section 3.1) over a range of values for T . The performance of the distance weighted clustering algorithm is examined with three values of H ($H = 1, 2$ and 3). For a virtual environment with $T = 10$ the percentage of avatar placements within acceptable angular error is 92% for $H = 1$ and 95.1% for $H = 3$. At very high numbers of avatars within hearing range ($T = 100$) the percentage of avatar placements within acceptable angular error for $H = 1$ and $H = 3$ is 77.8% and 84.2%.

Figure 7 shows the cumulative distribution of deviation from acceptable angular error V . The graph shows V for talkers in the interactive zone (closest 10% of avatars) and background zone (furthest 90% of avatars) for $H = 1$ and $H = 3$. The choice of H has

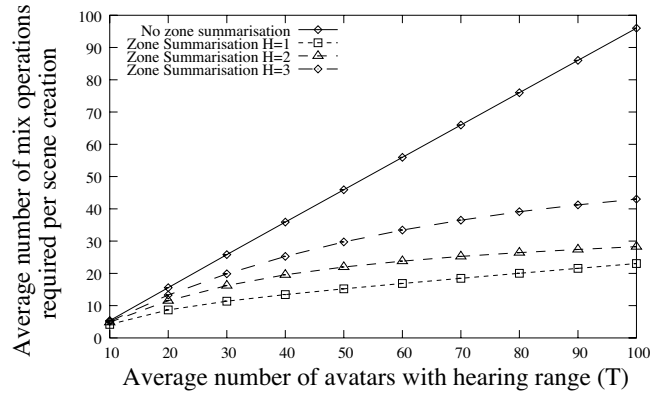


Figure 5: Average number of mix operations required to create an audio scene

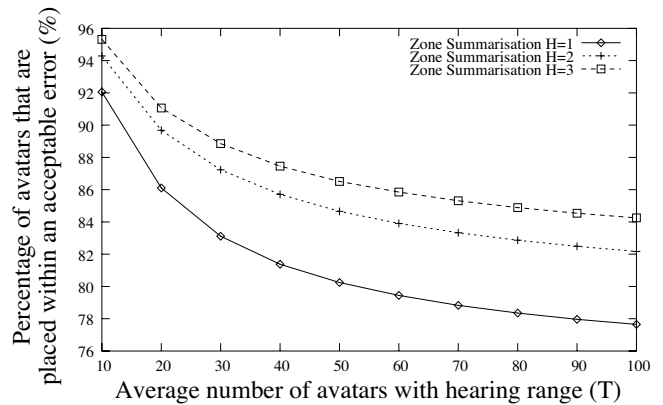


Figure 6: Average number of avatars placed in audio scenes within acceptable error

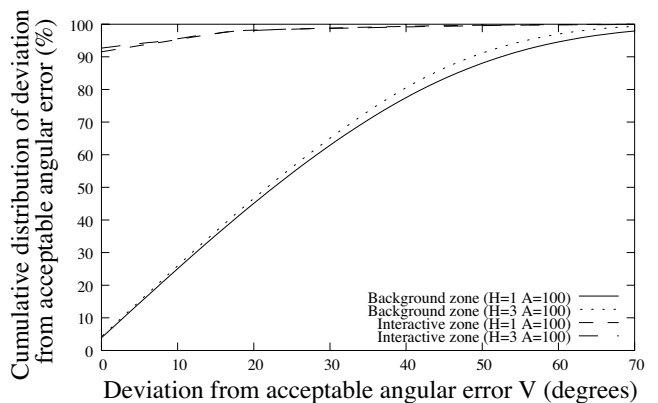


Figure 7: Cumulative distribution of deviation from acceptable angular error V

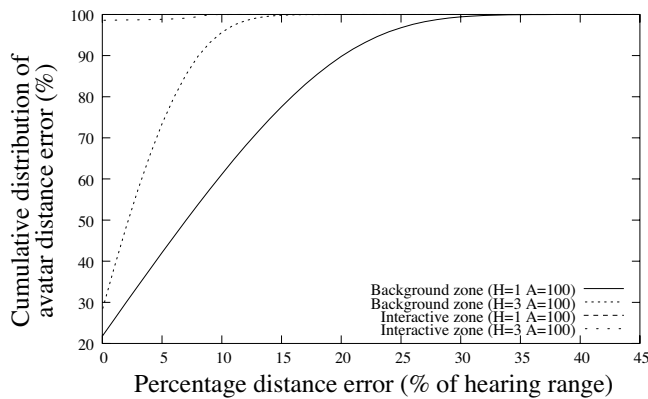


Figure 8: Cumulative Distribution of Distance Error

little effect on V since the angular clustering algorithm is the major influencing factor on angular error. Increasing the accuracy of grid summarisation only reduces distance error. Deviation from acceptable angular error is below 20 degrees for over 98% of avatars within the interactive zone. Deviation from acceptable angular error for the background zone is significantly higher with 50% of avatars being placed with more than 20 degrees deviation. Some 5% of avatar placements have greater than 60 degrees deviation from acceptable for the background zone. This large background zone error is a result of selecting $N = 4, M = 3$ meaning that the cluster selection is heavily weighted towards the interactive zone. Due to the small number of clusters it is likely that at least a few of the distant avatars will not be contained within a cluster and will therefore have an angular error of greater than E_{max} .

Figure 8 shows the cumulative distribution of distance error due to grid summarisation. The value of H has a significant impact upon distance error for the background zone and negligible impact upon the interactive zone. When $H = 1$ approximately 90% of the avatars in the background zone are rendered within 20% of their actual distance. When $H = 3$ approximately 90% of the avatars in the background zone are placed within 7% of their actual distance.

5 TESTBED

A DICE testbed has been implemented at three locations in Australia with servers in Wollongong, Sydney and Launceston. This testbed includes implementations of the audio client, scene creation server, and control server. The clustering and summarisation algorithms proposed in this paper are fully implemented in the testbed. Existing virtual environments are used to generate the visual side of the service. The virtual environments currently being used are Quake III and Enemy Territory which are popular examples of network games. These virtual environments are attached to the DICE service by writing a small server based modification using publicly available "mod kits"¹. The server modification simply extracts the position of each avatar and sends this along with other status information including if the player's avatar is dead or alive. The chosen virtual environments are not ideal as they only support up to 32 players in each virtual world and plans are underway to get access to virtual environments that support significantly more than 32 players. However, even with limited numbers of participants in one virtual world, the current DICE testbed is useful for verifying usability of the service. The servers being used are based on single 1.2Ghz Intel Celeron processors. It is interesting to note that the first SCS implementation can easily create audio scenes for 32

¹Mod kits allow third parties to change the game - mod kits are popular since they extend the commercial lives of games

players on this old server. Initial measurements on the server load indicate that the server may be able to support up to 100 players - however physical tests have only been conducted so far with 32 participants.

In order to gauge user reactions to DICE several focus groups have been set up by a commercial partner of the DICE project. The results of the focus groups were positive and will be published shortly. The focus groups were aimed at examining the applicability of the DICE service to games. One interesting outcome of the focus groups was that in general hard core gamers were interested in DICE for reasons in addition to natural social communication. A game with DICE was seen as a much more immersive experience with the possibility of many new tactics including for example, sneaking up on the opposing team to eavesdrop on battle-plans. The focus groups lead to the addition of a DICE walkie-talkie. The DICE walkie-talkie allows players to talk to team-members that are far away, however the voice of the player using the walkie talkie is also heard through the environment. A player may use the DICE walkie-talkie when they are eavesdropping on the enemy - however the player will have to whisper on the walkie-talkie or they may be discovered by the enemy.

The next stage of the project is a trial of the service in conjunction with a major Australian telecommunication service provider. The trial will enable us to gain further insights into how people use the service, and obtain user reactions to audio quality with varying DICE clustering and summarization parameters. In particular we wish to gain an insight, from the perspective of users, into the tradeoff between scene accuracy, computational complexity and bandwidth. At the time of writing a trial is currently underway involving approximately 50 game players. Details of the trial will be published after it is completed. Early results, however, are encouraging, particularly with regard to low values for N (such as $N=4$). Players notice when the avatars very close to them are inaccurately placed, however if the closest few avatars are accurately placed many players do not notice inaccuracies (even large angular errors) in the placement of more distant avatars.

6 DISCUSSION AND CONCLUSIONS

The results presented in this paper indicate that using the DICE architecture in conjunction with the distance weighted clustering algorithm and a small number of audio streams ($K = 4$) we can provide an immersive audio environment over the Internet that is highly accurate in the foreground, with the penalty of reduced accuracy in the background. In order to cost effectively scale to dense crowds, grid summarisation is appropriate if we choose grid sizes that are a large proportion of the hearing range. The simulation results show that if the summary grid squares are $1/9$ the size of hearing range we can save in excess of 50% of server resources when there are 20 or more avatars within hearing range. The savings over the approach without grid summarisation continue to improve as avatar density increases. This introduces a negligible additional angular error (the clustering algorithm is the main source of angular error) and a distance error that is less than 22% of the hearing range for 95% of elements in the audio scene.

The DICE service has so far been implemented as an add-on to an existing network game with users attaching through ADSL connections. The selection of $K = 4$ is vital for performance of the game with this bandwidth limitation. The service is currently the focus of user studies to examine the effect of the trade-off between resources and scene accuracy. We believe that DICE is a promising future Internet application that allows many people to communicate, socialise or work online in a natural manner.

The testbed is enabling us to gain insights into the usability and usefulness of this service. These insights will be the subject of future papers. The current testbed is focussed upon use of DICE in

networked games. However, the testbed is also being used for team meetings with participants around Australia. The next application to be investigated will be virtual classrooms. To facilitate this application a three dimensional map of a university department building has been created. This map has lecture theaters, offices, meeting rooms and tea-rooms.

REFERENCES

- [1] Grenville Armitage. An experimental estimation of latency sensitivity in multiplayer quake 3. In *Proceedings of the 11th International Conference on Networks, Sydney, Australia*, pages 137–141, 2003.
- [2] J. Bolot and F. Parisi. Adding voice to distributed games on the internet. In *Proceedings of the Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies*, pages 480–487, 1998.
- [3] Paul Boustead and Farzad Safaei. Comparison of delivery architectures for immersive audio in crowded networked games. In *Proceedings of the 14th ACM International Workshop on Network and Operating Systems Support for Digital Audio and Video, Kinsale, County Cork, Ireland*, June 2004.
- [4] T. Funkhouser, P. Min, and I. Carlbom. Real-time acoustic modeling for distributed virtual environments. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 365–374, 1999.
- [5] C. Greenhalgh and S. Benford. Massive: A collaborative virtual environment for teleconferencing. *ACM Trans. On Computer-Human Interaction*, 2(3):239–261, September 1995.
- [6] Cong Nguyen, Farzad Safaei, and Paul Boustead. A distributed proxy system for provisioning immersive audio to massively multiplayer online games. In *IEEE International Conference on Local Computer Networks LCN'2004*, 2004.
- [7] Cong Nguyen, Farzad Safaei, and Paul Boustead. A distributed server architecture for providing immersive audio communication to massively multi-player online games. In *IEEE International Conference on Networks ICON'2004*, 2004.
- [8] Cong Nguyen, Farzad Safaei, and Paul Boustead. Performance evaluation of a proxy system for providing immersive audio communication to massively multi-player games. In *IEEE International Workshop on Networking Issues in Multimedia Entertainment NIME'2004*, 2004.
- [9] P. Quax, T. Jehaes, P. Jorissen, and W. Lamotte. A multi-user framework supporting video-based avatars in networked games. In *Proceedings of NetGames 2003*, May 2003.