Presented at IE'13, Athens, Greece 16-19th July 2013

# Selectable Directional Audio for Multiple Telepresence in Immersive Intelligent Environments

Alfonso Torrejon, Prof. Vic Callaghan, Prof. Hani Hagras
*Intelligent Environments Group, Department of Computer Science and Electronic Engineering*
*University of Essex, Wivenhoe Park, Colchester CO7 9QU, Essex, UK*

{atorree, vic, hani} @ essex.ac.uk

***Abstract* – The general focus of this paper concerns the development of telepresence within intelligent immersive environments. The overall aim is the development of a system that combines multiple audio and video feeds from geographically dispersed people into a single environment view, where sound appears to be linked to the appropriate visual source on a panoramic viewer based on the gaze of the user. More specifically this paper describes a novel directional audio system for telepresence which seeks to reproduce sound sources (conversations) in a panoramic viewer in their correct spatial positions to increase the realism associated with telepresence applications such as online meetings. The intention of this work is that external attendees to an online meeting would be able to move their head to focus on the video and audio stream from a particular person or group so as decrease the audio from all other streams (i.e. speakers) to a background level. The main contribution of this paper is a methodology that captures and reproduces these spatial audio and video relationships. In support of this we have created a multiple camera recording scheme to emulate the behavior of a panoramic camera, or array of cameras, at such meeting which uses the Chroma key photographic effect to integrate all streams into a common panoramic video image thereby creating a common shared virtual space. While this emulation is only implemented as an experiment, it opens the opportunity to create telepresence systems with selectable real time video and audio streaming using multiple camera arrays.**

**Finally we report on the results of an evaluation of our spatial audio scheme that demonstrates that the techniques both work and improve the users' experience, by comparing a traditional omni directional audio scheme versus selectable directional binaural audio scenarios.**

***Directional audio; telepresence; immersive; selectable binaural audio; panoramic audio.***

## I. INTRODUCTION

Existing telepresence systems are designed for inter-system enabled meetings. If more than one system is at either end of the link, then anyone who wishes to express an idea to the whole meeting, or part of it, must to wait for their turn to talk. (i.e. the system only connects one input voice channel at an given time) either intentionally by participant control or compulsory by Voice Activity Detection (VAD) systems. While the approach is suitable for some meetings where there is a common understanding, it often fails for less structured meetings where multiple conversations take place concurrently, thereby slowing discussion because each participant has to wait their turn to talk (i.e. they are forced into a serial rather than parallel process). The proposed system works in a similar way to group-work and networking events where people sitting at different tables discuss their proposals with external listeners or facilitators. While this could be considered as a multiple meeting, instead of a multi table meeting, we took the approach of the well-known "cocktail party effect" [1] where multiple discussions occur at different locations within an immersive environment in which we intend to segregate by speakers or spatial area.

Our helicopter view of the scene display a number of individual meetings held within one event and those attendees to the real main event can move freely, networking and talking between groups within the room.

The method proposed here offers remote attendees the capability of visual and auditory access to the spatial information and the freedom to attend any one meeting at a time, similar to what a real attendee might do in the real space when attending a networking event or collaborative group meeting.

Similar research, GAZE-2 [2], has employed video conferencing systems which try to convey eye contact, providing a hybrid of single and multiple cameras approaches. The work demonstrated that multiparty conversations are much more complicated than their dyadic equivalents, posing problems such as regulation of turn taking.

In a multiparty situation, a natural question arises as to whom will be the next speaker or subject of the conversation, and is the whole group interested in the same subject or could be broken into sub-subjects enabling separate discussions between them? How can those discussions be broken down and fed to a video conferencing system to allow remote listeners to adopt a strategy adopted by local listeners who filter out and defocus unwanted subjects?

Commercial videoconferencing and telepresence systems allow a small number of multiple connections within one unit (usually up to 5), or larger number of locations through more complex multi conferencing units for multiple calls but, generally speaking, all those calls are shared by all at the same time. If a caller wants to establish a discussion with only one of the other participants then all other participants must remain silent or log-off the main meeting in order to reestablish a new call with another interested party.

## II. Experimental Setup

Originally this experiment was designed to be setup with a spherical mirror with a omni directional camera [1], projecting all views onto a doughnut shape image (Fig. 1) that then will be converted onto a 360° panoramic view. Within that panoramic view we would seat the speakers and spatially allocate the audio with a similar converting tool that the one used for pixel conversion (Polar to Cartesian) [1].

**Fig. 1: Spherical lens and resulting 360° image.**

Due to some constraints with the quality of the image and the possibility of interfering the audio experience that is the main subject of this experiment, we replaced the panoramic view with a set of videos stitched together side by side. This stitched image was composed of 6 streams representing the six views completing 360°, similar to the techniques that were used in systems such as FullView ® [5] and Polycom ® (Fig. 2) or, where and when higher resolution was sought, with camera clusters or arrays (Fig. 3) [7][7][7].

**Fig. 2: FullView / Polycom.**

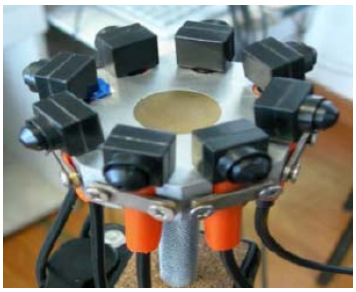

**Fig. 3: Array of multi cameras.**

### A. Recording and assembling streams

In order to emulate those camera clusters (e.g.: FullView/Polycom) we have separately video 6 individuals speaking about different subjects into two sets of scenarios:

a) For the first scenario, the person read 10 sentences about his/her subject (London, Football, Colchester, Internet, University of Essex and Scotland). Henceforth we refer to these as 'Set 1', with a test 1 page and a test 2 page with identical video but different audio method.

b) For the second scenario, 3 of the 6 participants asked questions to the other 3. These were organized as one-to-one conversations, to make it easier to identify who talks to whom. We refer to these as 'Set 2' throughout the paper, and have a test 3 page and test 4 page with identical video but different audio method once again.

The two sets of recordings last for a total of 45 seconds per test page. We have used a plain green fabric background, commonly used for photographic Chroma key effect, to minimize distractions and facilitate easier video editing (see Fig. 4).

During the first set of recordings, a speaker read out the sentences one by one, while occasionally glancing towards positions that other speakers would be expected to be in during the online trials (e.g.: looking to right, left and front with different angles) so as to give viewers the feeling in the activity way taking place in a single unified space.

During the second set of recordings, the person asking questions, and the person replying those questions, are presented facing each other. This increases the sense of realism adding information that helps locate people within the scenario.

A 360° panoramic background picture divided into 6 numbered sections was chosen. A speaker was inserted into each section, thereby enabling the speakers to appear in a common space. These were numbered in order to identify them, for response purposes.


**Fig. 4: Chroma key effect.**

The recordings were converted to FLV files in order to be added to a 360 ° panoramic viewer created with Adobe® AS3. This was done for portability reasons. To move from one stream to another, viewers only need to 'click and drag' on the video.

### B. Panoramic viewer

All streams were ported onto an Adobe Flash® panoramic viewer coded with AS3 and numbered from 1 to 6, so viewers can identify each instance in order to relate to questions on the answer sheets (Fig. 5).

**Fig. 5: Full strip with the 6 videos stitched one after other.**

The window/video size corresponds to the individual stream size (Fig. 6). Viewers were able to 'click and drag' on the video to achieve the effect of space translation, allowing them to watch the stitched videos one after another to complete a 360 degree circle.

**Fig. 6: Panoramic viewer effect.**

Audio was coded as binaural, so when the video travels to the left or right, the audio pan and the volume adjusts in intensity according to the video position thereby adding extra spatial effect and information.

### C. *Web setup*

The experiment was web based and publicly online so that subjects could participate from the comfort of their own computer and place. An initial page introduced the experiment and there were four further test pages.

An initial page was presented with a panoramic player and 360° video. The Panoramic viewing technique was introduced to viewers and the text explained how to 'click and drag' to achieve a panoramic video effect. An example of a question was presented to the user in order to explain the method used for collecting their answers and providing feedback.

For test page 1, in Set 1, all the audio streams contributed equally to the volume, which corresponded to the performance of unidirectional microphone in real environment.

Test page 2, in Set 1, used the same scenario but applied selectable directional audio to decrease all audio volume of frames that were not the focus of the user (the frame the user focused on had a normal sound level). Control of the audio level was linked to panning (i.e. the gaze or focus of the user).

Test page 3, in Set 2, took a different approach by using 3 different scenes. In this 3 speakers were asking questions and the 3 other speakers were replying. To avoid confusion, all 3 questions groups were unrelated to each other. Also, in this test, an audio level for speakers was identical (similar to test 1).

For test number 4, in Set 2, we adopted the same scenario as in test 3 but introduced directional audio effect.

Tests were time-limited, with a timer counting down 45 seconds which was the time allocated to answer the questions. This timer triggered the next s page, so viewers were not permitted to response out of the allotted time. The rationale behind using a timer was to put participants under pressure by limiting their time exposed to the different methods of delivering the audio, avoiding any speech recognition effect (e.g. if they recognized a speakers voice they may, in the worst case, be able to accomplish the goals of the experiment without using our focusing methods, or in the best case, partly use both methods). Our intention was to contrast and compare both methods, existing omni versus our proposed directional method, so questions, answers, speakers and time slot must be exactly the same with the only invariance of the audio delivered to the participant. The design of the experiment tried to avoid participants being exposed for long periods to the same speech pattern, amplitude, frequency spectrum and pitch avoiding any possibility of recognition between sets, so the time slot must be short enough to prevent these experiences being etched into their mind and speech-face associations. Equally we took into consideration strong accents, either local or international.

In order to determine this period we referred to previous research which indicated that this association is fast, occurring in less than 2 min. In particular we used work by von Kriegstein and colleagues who found speakers can improve subsequent speech recognition [9] and how listeners can use the acoustic differences between speakers help to recognize each other by their voice.[10].

### D. *Data collection*

The initial page asked participating subjects their age (defined as groups: 18-27, 28-40 and +40), gender (male or female) because we would like to investigate if age and gender cause a difference when using technology. No information that could identify particular participants was collected (in line with the University's ethics policy).

A web page (Fig. 7) was used to collect the participants' opinion on who was talking to whom, what they were talking about, and who was replying to whom. The time taken to identify the speakers-subject's match was recorded on a database when the answer was entered. After the 45 seconds, the participants were presented with another test page and another 6 video streams with 10 questions to answer in 45 seconds. Unanswered questions were recorded as 0 and time-stamped using the main page's submission timestamp. Abandoned tests, such as closing the browser, were not recorded, although statistics on these were maintained.

Once having answered a question, a participant could edit the answer but was not permitted to remove it. All questions answered can be updated by clicking on.

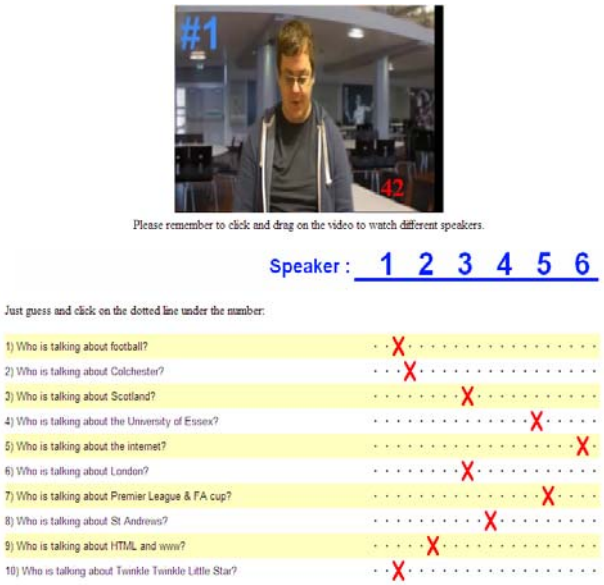


**Fig. 7: Test page presented to participants with the video and answer sheet.**

## III. DATA RESULTS

We advertised this experiment online through University and social professional networks (LinkedIn®. Groups) related to Immersive Media and Intelligent Environments, resulting in a diverse set of participants. Participants were tracked by the use of session IDs, page visits, page referral, date-time stamps

(for entering the page, clicking each row on answer sheets and submission). Empty rows where recorded as 0.

We had an initial response of 202 visiting the website, with only 49 experimental subjects undertaking the test.

Of the 202 visits, 51 were from my own University (Essex) where it had been internally advertised. Search engines and error pages were removed from the analysis.

The data is presented as tabular table (Figures 10-13) and graphs (Figures 8-10). Some important features to consider for further analysis are as follows:

From the 49 subjects who undertook the tests, 70% described themselves as male, and 30% female (see Fig. 8).

43% were on the age group of 18-27 years old, 22% were on the age group of 28-40 years old, and 33% above 40 years old of age (see Fig. 9).

Almost a 43% dropped out of the test (from different pages), with a 55% trying to answer some of the questions (see Fig. 10). More participants dropped out in test page 2 which introduced directional audio (omni directional audio has ended at test page 1).

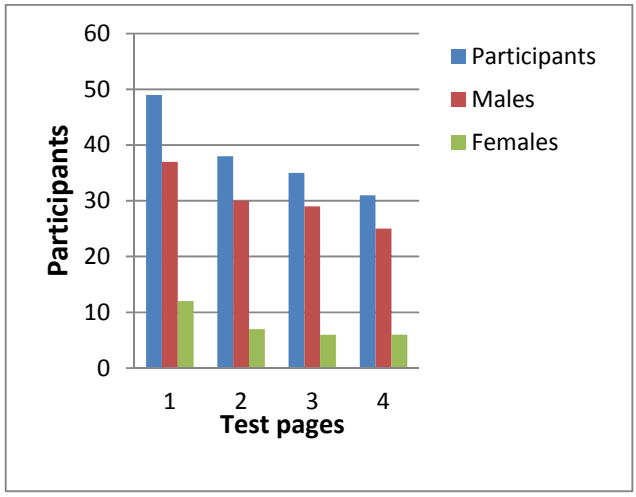


**Fig. 8: Participants by gender group.**

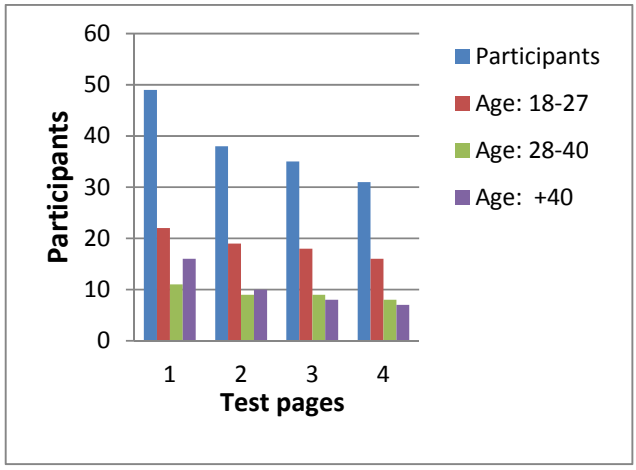


**Fig. 9: Participants by age group.**

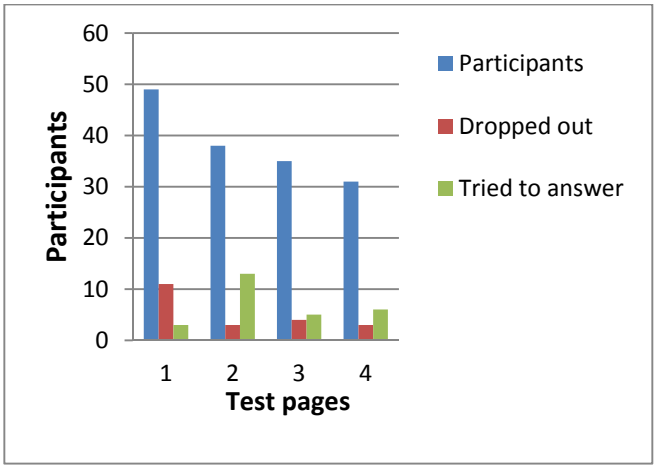


**Fig. 10: Participants dropping out of the test and participants trying to answer any question.**

**Set 1**

| Test page | TEST 1 | | | TEST 2 | |
|---|---|---|---|---|---|
| TOTAL | 49 | 100% | | 38 | 78% |
| Age: 18-27 | 22 | 45% | | 19 | 50% |
| Age: 28-40 | 11 | 22% | | 9 | 24% |
| Age: +40 | 16 | 33% | | 10 | 26% |
| Males | 37 | 76% | | 30 | 79% |
| Females | 12 | 24% | | 7 | 18% |
| Dropped out | 11 | 22% | | 3 | 8% |
| Tried to answer | 3 | 6% | | 13 | 34% |

**Set 2**

| Test page | TEST 3 | | | TEST 4 | |
|---|---|---|---|---|---|
| TOTAL | 35 | 71% | | 31 | 63% |
| Age: 18-27 | 18 | 51% | | 16 | 52% |
| Age: 28-40 | 9 | 26% | | 8 | 26% |
| Age: +40 | 8 | 23% | | 7 | 23% |
| Males | 29 | 83% | | 25 | 81% |
| Females | 6 | 17% | | 6 | 19% |
| Dropped out | 4 | 11% | | 3 | 10% |
| Tried to answer | 5 | 14% | | 6 | 19% |

**Fig. 11: Percentage at different pages and sets by groups.**

| Males | 37 | 76% |
|---|---|---|
| Age: 18-27 | 21 | 57% |
| Age: 28-40 | 6 | 16% |
| Age: +40 | 10 | 27% |

| Females | 12 | 24% |
|---|---|---|
| Age: 18-27 | 1 | 8% |
| Age: 28-40 | 5 | 42% |
| Age: +40 | 6 | 50% |

**Fig. 12: Initial participants by Gender and Age.**

| Dropping out at test page 1 | 11 | |
|---|---|---|
| Dropping out before ending page | 6 | 55% |
| Dropping out after ending page | 5 | 45% |

**Fig. 13: Relation dropping out of the test before and after ending test page 1.**

**Set 1**

| Test page | TEST 1 | |
|---|---|---|
| Tried to answer | 3 | 6% |
| Ratio Answers/Questions | 7 / 300 | 2% |
| Ratio Right/Wrong | 2 / 4 | 33% / 67% |
| | | |
| Test page | TEST 2 | |
| Tried to answer | 13 | 34% |
| Ratio Answers/Questions | 34 / 300 | 11% |
| Ratio Right/Wrong | 25 / 9 | 74% / 26% |
| **Set 2** | | |
| Test page | TEST 3 | |
| Tried to answer | 5 | 14% |
| Ratio Answers/Questions | 9 / 300 | 3% |
| Ratio Right/Wrong | 3 / 6 | 33% / 67% |
| | | |
| Test page | TEST 4 | |
| Tried to answer | 6 | 19% |
| Ratio Answers/Questions | 14 / 300 | 5% |
| Ratio Right/Wrong | 10 / 4 | 71% / 29% |

**Fig. 14: Ratio of questions answered by the total number of questions and proportion related to right and wrong answers.**

## IV. Data Analysis

From the number of 202 initial visits landing at the initial page, an introductory page were the experiment was explained and instructions given,, it seems that while the experiment initially drew attention to the public, either the lack of audio (compulsory), time or just because they were only curious made people to avoid entering finally the test, so actually a 25% of all visits (49 subjects) where positive by participating on the experiment and a 63% (31 participants) out of them managed to reach the end of a difficult experiment.

A large number of females took part on the test but mostly over 40s with an insignificant number (just 8%) of the youngest group of 18-27s, and equally a large number of people over 40s, a 33 %, participated in the experiment that surprisingly outnumbered younger group of 28-40s (see Fig. 12).

The first test page (test page 1) with omni directional audio was the page where more participants abandoned the test. This probably due to the confusion created at this stage where initially all audio from all speakers were at same level and talking at the same time. All participants that decided to move on to the next page, test page 2, managed to finalize the following test, and proceeded to new block of questions.

An unexpected feature of the data was that males were faster at answering questions, and were more precise with matching questions to video images. Females answered fewer questions, responded slower but were equal with precision when matching questions to video images.

More people tried to answer test page 2 (34%), the directional audio, than test page 1 (6%), the omni audio. The same pattern was repeated for test page 3 and 4. It's not clear why this is but possible reasons include the participant gaining in confidence or perhaps being more interested in the directional audio.

Apart from the initial page of test page 1, women did not drop out of the experiment at the same rate as men. Most females were within the age group +40. Another unexpected feature of the data is that as female age increased there was greater participation in these experiments, whereas the reverse is true for males (see Fig. 11). We have no explanation for this and it merits further investigation.

From the statistics of 202 people visiting the site but only 49 people participating, it is clear that the first initial page acted as a course grained filter, eliminating those web surfers who were curious about the advertisement but who, when were given more information, proved to have little interest or time for the project. As such, the dropout rates on the remaining pages were much smaller.

Thus test page 1 showed the largest number of participants abandoning the test, most participants abandoned the test before the 45 seconds was over, with almost similar number after the page were reloaded at the end of the 45 seconds (see Fig. 13). At test 1 for those who remained they were encouraged to answer questions and engage in the game, immersing themselves into the experiment.

Some of this behavior might be explained by Zicherman who analyzed gamification [8] and drew a flow zone between the boredom area (where the player lose all interest), and the anxiety area (where the player will probably shutdown the system), and this is clearly one of the cases we revealed in this experiment.

In general terms the engagement with questions increased as a participant moved between pages within a given set. Quite why this happened is interesting to consider. In part this may we suggest this may be due to the participants being given a balanced visual and auditory environment where they were able to control the information that flowed towards increased their motivation and engagement.

This view is supported by Self-Determination Theory (SDT), Deci Koestner and Ryan, which has determined that the more control someone has in choosing what to do, the better the chance the person will be internally motivated to do it [12]. Someone who wants to do something because it is fun is more likely to success than someone who is doing something for a reward or to "learn something." When someone is taken into a playful space then the flow of learning will come natural from the person. Many of those examples

can be found in real-world including museums, libraries, zoos, and botanical gardens [13] where the gaming and learning has been cleverly attached to the spatial environment. Many of these leisure settings employ game elements to help users find personal connections with the non-game setting, making out of it an enjoyable experience that the users will repeat, if the element of playful exploration is there. If the opportunity are created for individuals to explore their environments then they will engage in discovering the meaningful side of it, by engaging, reflecting and participating, allowing them to be transformed by the system what clearly is an opportunity for natural learning [14].

Looking at our data of engagement patterns it seems to follow this trend with more fun and flexible pages getting higher responses. Superimposed on this motivational model is, of course the level of cognitive load and it is clear that questions in Set 2 were more difficult than the questions at Set 1, needing deeper concentration to answer in the same 45 seconds as allocated to simpler tests.

In terms of the conclusions relating to omni versus directional audio that this work has set out to explore, our data revealed that firstly there is a significant advantage in using directional audio to facilitate recognition in noisy environments as it provides more clues for identifying which speaker was talking helping us to concentrate on the subject and the given task (to answer questions). Figure 11, where both sets had increased the number of participants trying to answer questions, support this assertion.

Secondly, by further breaking down the information regarding the number of answers, our findings confirmed that using directional audio improved the ability of participants to answer questions correctly (see Figure 14).

For example, for test 1, with 7 answers out of 300 possible answers, 33% were right and 67% were wrong while the figures for test 2 revealed 74% right against 26% wrong, which, in our view, confirms that the directional method adds a precision that greatly improves the performance of listeners using these systems.

In addition, for test 3 and test 4, in Set 2, the task increased in difficulty, so the number of respondent who engaged in responding was lower, but the the number of correct answer stayed proportionally in favour of directional audio and similar to those of Set 1.

So the benefits of using a directional method against an omni one are that it will increase not only the number of participants engaging in a meeting and the flexibility, but increases the number of correct answers (precision) in a two-way communication with consequent improvements in achieving a given collaborative task.

Finally there is an additional gain to efficiency arising from transmitting more than one conversation or audio information on a given multiparty situation.

Thus, in summary the data confirmed our hypothesis that by adding directional audio into online telepresence group meetings results in an improvement in the performance over and above systems using omni directional audio.

## V. CONCLUSIONS

In this paper, we have described a novel telepresence directional system, for intelligent immersive environments, based on selectable directional audio and video. We explained how the system can be adopted on brainstorming and collaborative meetings where multiple speakers at the same event are required to have concurrent one-to-one conversations. The experiment has shown that the proposed telepresence directional system provides us with the ability to attend multiple conversations at the same meeting and segregate sound sources that are out of our interest.

In particular we have presented a model that enables achieve a number of speakers to interact independently one-to-one or one-to-many, without interfering with each group's activity.

Concerning the experimental data is seems younger females at the range 18-27s, didn't show as much interest in new technology at this experiment, contrary to males on same age group who accounted as the largest group. Both males and females over 40s were attracted to new technologies probably on search of novelty or solutions.

Concerning the findings of our participants on directional audio, our experiment results highlighted that the audio aspect of video streaming is as important as or more important than the stream itself. Creating audio confusion to the participants generated the largest number of abandonments during the experiment. Segregating audio provided an added interest to the test, to our advantage.

Finally, further tests need to be done with live streaming instead of recorded video and more complex tasks with longer time to measure continuity in the performance and engagement.

## ACKNOWLEDGMENT

We would like to thanks all 6 speakers that volunteered to participate and be video recorded for the experiment, colleagues and friends at the University of Essex. Likewise we want to thank the 49 experiment subjects that entered the test and provided useful information.

## REFERENCES

[1] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions". Acta Acustica united with Acustica, 86(1), 117-128.

[2] R. Vertegaal, I. Weevers, C. Sohn and C. Cheung , "Gaze-2: Conveying eye contact in group video conferencing using eye-controlled camera detection", Proc. CHI 2003, ACM Press, 2003.

[3] A. Ohte, O. Tsuzuki, K. Mori, "A practical spherical-mirror omni directional camera", ROSE 2005 – IEEE International Workshop on Robotic and Sensors Environments, Ottawa, ON, Canada, Sept 2005.

[4] V. Grassi, J. Okamoto, "Development of an omni directional vision system", J. Braz. Soc. Mech. Sci. & Eng. vol.28 no.1 Rio de Janeiro Jan./Mar. 2006.

[5] V. Nalwa, "Outwardly pointing cameras", FullView, Stratford

[6] Lin Z., Dexiang D., Xi C., Yunlu Z. "A self-adaptive and real-time panoramic video mosaicing system", Journal of Computers, Vol.7, No.1, January 2012.

[7] Yuan H., Wang B., Zhang J., Li H., "A novel method for geometric correction of multi-cameras in panoramic video system", 2010 International Conference on Measuring Technology and Mechatronics Automation.

[8] A. Majumder, W. B. Seales, M. Gopi and H. Fuchs, "Immersive teleconferencing: a new algorithm to generate seamless panoramic video imagery", Proc. 1999 Proceedings of the seventh ACM international conference on Multimedia (Part 1), pp. 169-178.

[9] K. von Kriegstein, O. Dogan, M. Gruter, A.L. Giraud, C.A.Kell , T. Gruter, A. Kleinschmidt, S.J. Kiebel , "Simulation of talking faces in the human brain improves auditory speech recognition", Proc Natl Acad Sci U S A 105:6747– 675 , 2008.

[10] K. von Kriegstein, D. R. R. Smith, R.D. Patterson,, S. J. Kiebel and T. D. Griffiths, "How the Human Brain Recognizes Speech in the Context of Changing Speakers" , The Journal of Neuroscience, January 13, 2010 • 30(2):629–638 • 629.

[11] G. Zicherman, "Gamification by design: Implementing game mechanics in web and mobile apps. ", New York, NY: O' Reilly Media.

[12] Deci, E. L., Koestner, R., & Ryan, R. M. "Extrinsic rewards and intrinsic motivation in education: Reconsidered once again.", Review of Educational Research., 71, 1-27. 2001.

[13] J. Packer, "Learning for fun: The unique contribution of educational leisure experiences", Curator 2006 vo.l 49 Issue 3, 329-344.

[14] Kolb, A., & Kolb, D. A , "Learning to play, playing to learn: A case study of a ludic learning space", Journal of Organizational Change Management, 23(1): 26-50, 201.