Placing the Participants of a Spatial Audio Conference Call

Mansoor Hyder, Michael Haun, Christian Hoene

***Abstract*—In teleconferencing calls it is difficult to identify the current talker, especially if the person is not well known. In addition, if more than one person is talking at the same time, none of them can be understood easily. In real meetings, however, these limitations are not relevant because humans hear spatially. They take advantage of the "cocktail party effect" to distinguish speakers. In this publication, we assess a teleconferencing solution that we call 3D Telephony, which adds a virtual acoustic room simulation to IP based telephony, thereby achieving a spatial audio experience. We have conducted subjective listening tests of a 3D audio rendering engine to study the impact of the participant's locations in a virtual environment on sound quality, understandability and locatability. Our listening test results show some interesting findings. We identified placements, in which listeners find it very easy to locate virtual talkers and in which their success rate in locating two simultaneous virtual talkers is nearly perfect.**

***Index Terms*—3D telephony, teleconferencing, spatial audio**

I. Introduction

Teleconferencing systems are well-established and important tool for interpersonal communication. But it has one main limitation. If multiple persons speak at the same time, they can hardly be understood or identified unambiguously. This adds to the problem of poor quality, especially in multi-user scenarios [1]. In a classic phone the acoustic origin of a remote speaker is always the phone of the listener. It would be more natural if the callers communicate as if they are standing in front of or next to each other, and when they are moving, the virtual source of the audio signal follows the speaker's movements.

We propose a 3D Telephony System, which generates a virtual 3D acoustic environment in which each participant of a telephone call is placed at a unique position in the virtual world. The 3D Telephony System can be used for teleconferencing. Conference calls would be improved in regards to the listener's sensation of the call's quality because the participants could identify who is talking by locating the origin of the sound.

Our 3D Telephony System takes advantage of spatial hearing and 3D sound. Auditory events, such as the speech of a human, are distinct in terms of time and position [2]. Spatial hearing refers to the human ability to locate the origin of auditory events. 3D sound refers to a sound which makes a listener discern significant spatial cues for a sound source such as direction, distance and spaciousness. Therefore generating 3D sound means that one can place sound anywhere—left or right, up or down, near or far—at one's disposal in 3 dimensional space [3]–[6]. The technical requirements that are needed to implement a headphone-delivered 3D sound are well known [7]. In order to improve sound localization performance three factors need to be considered, which are an individualized head-related transfer functions (HRTFs) to describe how the acoustic waves propagate through each listener's head [5], [8], sound processing (auralisation) to simulate reverberations and reflections within a virtual surrounding individually, and head tracking systems to follow the movements of the speaker's and listener's heads.

In our study related with the virtual placement of participants, besides focusing on the sound quality, understandability and locatability of virtual participants, we also wanted to check for the occurance of any front/back or elevation localizing errors which are commonly seen in 3D audio systems when non-individualized HRTFs are used [9]. In addition to this, we also want to study the trade-off between sound source direction perception and distance perception. According to Shinn-Cunningham [10], reverberation degrades perception of the sound source direction, but enhances distance perception.

In Section II we present related work. Our implementation of the spatial audio telephone is described in Section III. To figure out how to place the participants of a conference call in the virtual room, we conduct seven formal listening tests (Section IV). After presenting experiments on placement of participants, we summarize our work and list open ended research questions as well as the future work.

II. Related Work

Systems that support 3D sound were first presented at the NASA Ames Research Center by Wenzel. A substantial understanding of 3D sound has also been achieved by Durand Begault [4].

Peter Hughes has presented a 3D audio teleconferencing system called Senate [11]. Senate has a SIP based interconnect, PC based audio client and is capable of playing both streamed speech and local sound files, has a GUI interface where a listener is able to place the sound sources - incoming audio streams, audio files in places of own preference. The possible options pointed out by the author regarding the network design included fully interconnected mesh, central server processing, distributed processing, server concentration and spatial audio object coding. Senate differs from our approach because it is not an open source entity and no evaluations of the usability of the system has been undertaken.

This work was supported by the DAAD, Germany, the HEC, Pakistan, and in the Nachwuchswissenschaftler program of the University of Tübingen.

All authors are with University of Tübingen, 72076 Tübingen, Germany, mansoor.hyder|michael.haun|hoene@uni-tuebingen.de

V. Sundareswaran et al. [12] have developed a 3D Audio Augmented Reality (3DAAR) wearable system that can be used to provide alerts and informational cues to a mobile user in a manner so as to appear to be emanating from specific locations in the user's environment. They performed experiments to verify the ability of users to localize the audio signal. On the basis of the results of the experiments they suggested that there is a potential way to improve auditory localization through a perceptual training process that involves synchronous presentation of visual and auditory stimuli.

Sound localization performance between virtual and real three-dimensional immersive sound field in virtual auditory display (VAD), which renders three dimensional auditory spaces, is compared by Dae-Gee Kang et al. [13]. They specifically examined distance errors only to evaluate the performance of the newly developed VAD, in which a listener can move freely.

In his recent research [14], Raake has utilized a real time PC-based binaural sound reproduction system, which is based on a real-time convolution engine that is fed with the appropriate BRIRs derived from a user defined database [14].

Some experimental results related with phone motion-tracking to interact with mobile spatial audio content has been presented by Christina Dicke et al. [15]. The long term aim of the author is to explore how spatial audio can enhance multi-party conversations with mobile devices. The current experiments included user interaction with sound in 3D environment and in particular using phone as an input device through gesture tracking for navigation.

Vicky Hardman and Marcus Iken [16] have shown approaches to solve audio problems that invariably appear with multimedia conferences over shared networks, and which uses only general purpose hardware. They presented solutions to tackle the problems of audio such as gaps in the output stream and lack of hands free operations.

The imaginary product concept "CyPhone" was presented by [17] just as vision of the future. In their research, they have analyzed the sources of real time constraints in telepresence and augmented reality applications and have also depicted the general architecture and integration framework.

Human sound source localization is based on spectral filtering. Sound reflects and diffracts from our head, torso, shoulders and pinnae folds in a unique way for every angle. Those interactions combine at the entrance of the ear canal into a signal that has a different frequency response for each angle [18]. These frequency response variations are called HRTFs [19], which are required by the listener to localize the source of the sound. Digital sounds can be processed by these HRTFs to make left and right audio signals that will make the listener believe that sound emanates from the corresponding virtual source location [20].

This paper extends our previous work [21], in which five different placements of the virtual talkers and listeners were studied. User experiments were conducted with 9 subjects. In this follow on study our objective was to further study the occurance of any front/back or elevation localizing errors and also to study the trade-off between sound source direction perception and distance perception of proposed system by adding two more tests. We conducted our user experiments on 32 subjects, in addition of two more tests. As the number of subjects has increased substantially, the test results are more precise and their confidence interval is smaller.

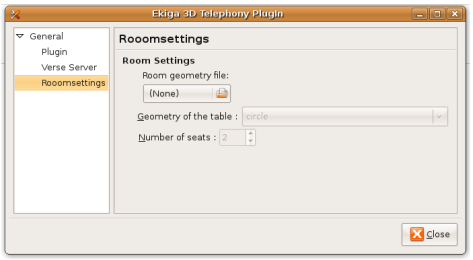


Fig. 1: Ekiga extended by a 3D sound options.

III. Implementation

Our implementation of the 3D telephone system extends Ekiga, a well known open-source Internet phone. Mostly, we based the prototype on software and components, which we briefly describe in the following sections.

A. 3D Sound Rendering and Ekiga Plug-in

We extend Ekiga by a 3D sound processing software, which has been originally developed for game programming in the EU FP6 research project Uni-Verse [22]. Uni-Verse's open source 3D sound software implements "a distributed interactive audio visual virtual reality system". A detailed explanation can be found in [23].

The central component is called "Verse Server", which stores and shares 3D geometric data of a virtual environment. Other applications can access and modify this data via the low-latency UDP-based Verse protocol. Also, the subscription based notification system allows an arbitrary number of applications to see all changes to the virtual environment immediately. The real-time visual rendering application uses the Verse server to display the current virtual environment. It can be thought of as a window to the virtual world that the Verse Server is hosting, an example is (rendered Figure 2). Changes in textures, modifications in the geometry and added objects can be seen in real-time by anyone using this tool. The sound renderer is built on Pure Data (PD), a real-time graphical programming environment developed by Miller Puckette [24].

We create a Verse plug-in for the Ekiga soft-phone which enables the users to have their calls placed in a virtual 3D environment (Figure 1). The plug-in is implemented as a Verse client forming the communication interface between the Ekiga Soft-phone and the Uni-Verse. Therefore, the plug-in offers several general configuration options, such as the server address, port and user credentials. It additionally provides a tool for uploading VML files containing the virtual geometry onto the Verse server and parses these files for virtual rooms. The plug-in chooses a suitable room for the call to take place in and computes the participant's positions according to the geometry of the seating plan and the maximum number of participants selected by the user. At the moment, two different

geometries of the virtual conference table, circular and rectangular are supported, but the implementation of arbitrary shapes is possible. The plug-in creates the necessary structures on the Verse server at initialization time and additional structures on every incoming call. It then intercepts Ekiga's output voice stream and redirects it to the Uni-Verse application over the Verse server resulting in the computation of spatial sound.

IV. Experiments on Placement of Participants

During assessing the 3D telephone system, the question arouse as how to place the participants in a virtual acoustic conference room, so that they can locate each other. In addition, the speech quality should not be impaired by reduced loudness, reverberations and echoes. Thus, we conducted subjective listening-only tests to study the impacts of virtual placement of participants on sound quality, understandability and locatability. In total seven different listening-only tests of 4 different setups were conducted while keeping in view the teleconferencing scenario.

We selected four sets of simulation parameters and used them for judging the seven different placements of participants in the virtual room (in total 22 combinations). We tested by changing one Uni-Verse parameter at a time in every setup and kept the other parameters the same to see the effect of every single changing parameter to study the impact of virtual placement of participants on sound quality, understandability and locatability. We have used two different HRTFs, two different room sizes, different heights of the listener and talker and kept the headsize constant The following parameters can be chosen for the acoustic simulation (Table I).

Room dimensions: In our test experiments, we used two rooms. A *Big Room* having dimensions (HxWxL=20 x 20 x 40 m³) and a *Small Room* having dimensions (HxWxL=10 x 10 x 20 m³).

HRTF: We have used two HRTFs in these tests, *HRTF-1* and *HRTF-2*. *HRTF-1* has 5 reverberations for 5 frequency bands and *HRTF-2* has 10 reverberations for 10 frequency bands.

Head size: We kept the head size to its default value which is 0.17 in all the setups, because we did not notice any difference by changing its value ranging between 0.1 to 0.3. (head size is a Uni-Verse UVSR parameter scalable from 0.1 to 0.3).

Placement: Seven different placements of the talkers and listeners were studied. We name these placements *Talkers in the Corners*, *Listener in the Corners, Horizontal Placement, Frontal Placement-1, Frontal Placement-2, Surround Placement-1* and *Surround Placement-2*. They are described further in the following sections.

Height: The placement of listeners and talkers in terms of height in the virtual room is summarized in Table II. We have used the same height parameters for *Default, HRTF-2* and *Small Room* which we call *Height-A* and for *Talker standing* we have used *Height-B*.

Setup Name	Room	Height	HRTF	Headsize
Default	Big Room	Height-A	HRTF1	0.17
HRTF-2	Big Room	Height-A	HRTF2	0.17
Small Room	Small Room	Height-A	HRTF1	0.17
Talker standing	Big Room	Height-B	HRTF1	0.17

TABLE I: Summary of test setup

	Height-A		Height-B	
Test	**Listener**	**Talker**	**Listener**	**Talker**
Horizontal Placement	1.8 m	1.8 m	1 m	1.5 m
Frontal Placement-1	1 m	1 m	1 m	1.5 m
Frontal Placement-2	1 m	1 m	1 m	1.5 m
Surround Placement-1	1.8 m	1 m	1 m	1.5 m
Surround Placement-2	1.8 m	1.8 m	1 m	1.5 m

TABLE II: Summary of listener and talker heights

A. Sample design

The samples were processed by the open-source 3D audio rendering engine Uni-Verse [23] (refer to Section III-A). The virtual rooms were based on the sample UVAS file "testscene_no_doors.vml", a big white room with dimensions of about 20x20x40 m³ (HxWxL) and a *Small Room* having dimensions (HxWxL=10 x 10 x 20 m³). The walls of the rooms had the typical acoustic properties of concrete.

The acoustic simulation uses the room geometry to find out how the sound emitted from the sound sources propagates through the geometric model. The sound propagation is simulated as beam using beam tracing technique. As a result, the acoustic simulation provides a set of real and virtual sound sources which the listeners would hear. The virtual sources are reflections of the real sound sources. For example, Fig. 2 displays acoustic simulations having one listener and two sound sources.

All test samples were produced with the UVAS open source program version using beamtracing, a maximal of 30 reflections per source, maximal order of reflection of 2, a maximal distance between listener and phantom source of 50 m and a maximal number of reflections of a source of 4. Further details on the acoustic simulation can be found in the publication of Min et al. [25].

Based on the results of the acoustic simulation, a sound renderer auralizes the direct sound and early reflections paths calculated by the room acoustic simulation module. The acoustic simulator transmits the listener, source and image source information including position, orientation, visibility and the URL of the sound source to the sound renderer. Then the sound renderer applies a minimum phase HRTF on the sound source. The HRTF is implemented by a 30 taps FIR filter. A detailed explanation of the used minimum-phase HRTF can be found in the paper by Savioja et al. [26]. The reverberation algorithm used in the implemented system was introduced by Vaananen et al. [27], which has been modified by Kajastila et al. [23]. Because the reverberation time (RT) is frequency dependant, the sound renderer uses 10 individual reverberators for 10 frequency bands and separate RTs for different frequency bands.

Further parameters used for the sample design, such as position of listeners and sound sources are given in the following test descriptions.

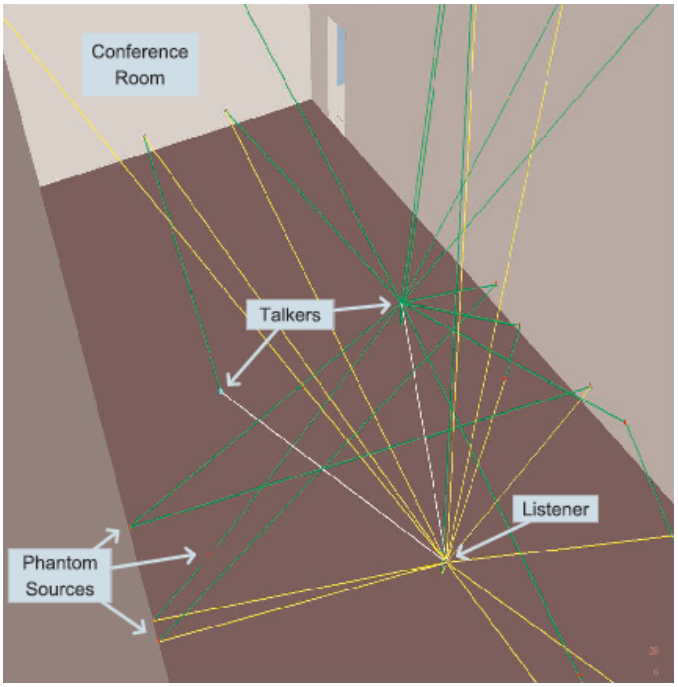


Fig. 2: Acoustic simulations having one listener and two sound sources. The white lines show the direct beam between sound source and listeners. The yellow lines are due to phantom sound sources (plotted as red points). The green lines are reflection of real sound sources.

B. User Experiements

User experiments with 32 normal-hearing subjects (29 male, 3 female) were conducted to find out the sound quality, understandability and locatability of the virtual talkers in the implemented system.

The listening only tests have been done following the (ITU-T) Recommended P.800 recommendations as far as possible. Two reference sound samples were taken from the database ITU BS.1387-1 [28] of which one of them is a male voice and other one is a female voice. The tests were conducted with pre-recorded samples played in random order to the listeners.

The technical equipment was based on an ordinary computer running Linux operating system. The computer had a 2.60 GHz Intel Pentium 4 processor and 1 GB physical memory. We used M-Audio Delta 44 external sound card and Sennheiser HD 280pro headphones.

Subjects were presented with three tasks in the same order for every individual test. Prior to each setup tests, subjects were asked to familiarize themselves with the given technology. Following were the tasks presented to each subject.

Task-1: Please locate the talker with the help of a map which describes possible locations of the talker.

Task-2: Can you understand and concentrate when there is one or more than one talkers?

Task-3: Please score the talker's sound quality from 5 to 1 (5=excellent, 4=good, 3=fair, 2=poor, 1=bad) (When there is one talker and when there are more than one talkers?)

They were also given the general layout of the room and the possible locations of the sound sources including the case of moving listener or moving sources.

C. Test 1: Talkers in the corners

For the case of *Talkers in the corners*, we placed the listener at the center of the room at ground level and placed the talkers in all eight corners of the room to study the impact of moving sound sources. The listener is facing the wall appointed by the corners 5, 6, 7 and 8. We wanted to see whether the subjects can locate the sound sources correctly by describing the orientation of the sound and their judgment about the quality of the sound. The layout for the room can be seen in Figure 3a.

Results indicate that it is very difficult for subjects to correctly locate the virtual talkers in this test. Possible factors that might cause the listeners to falsely locate the talkers could be the listener's position which is at the center of the room at ground level and as we know it is not a normal listening position and secondly the use of non-individualized HRTFs which results in two particular kinds of localization errors, front/back confusions and elevation errors [9] commonly seen with 3D audio systems. Localization errors were found and subjects also pointed out the difficulty to distinguish between front/back and up/down positions while seeming quite sure about its orientation. Subjects correctly located 27 percent of the talkers placed at the corners.

MOS-LQS value (95% Confidence Interval) was 3.85±0.76.

D. Test 2: Listener in the corners

Listener in the corners has quiet similar kind of exposure that is of *Talkers in the corners*, except the talker is fixed at the center of the room at ground level while the listener's position changes to all eight corners of the room and the orientation of listener remains facing the wall depicted by the corners 5,6,7 and 8. Layout for Test 2 can be seen in Fig. 3b. Though this is also not a normal way to hear the sound but we wanted to try with every possibility that can help us to see the impact of 3D sound in virtual room.

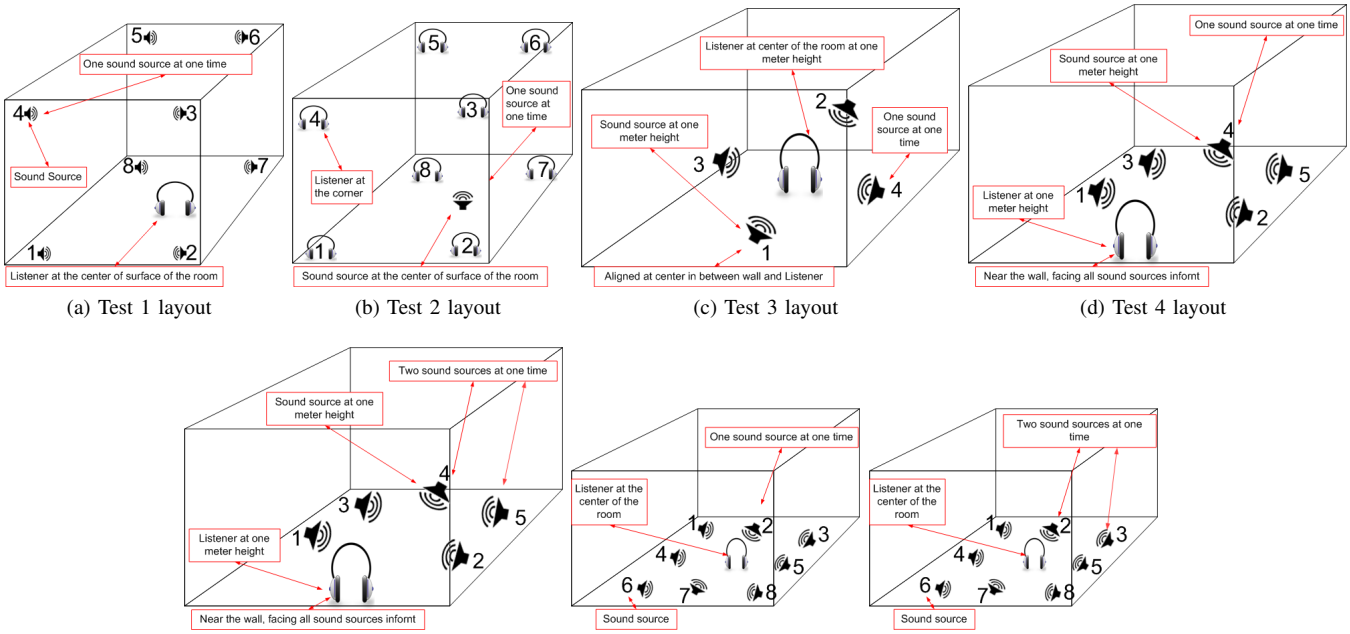
This test yielded 25% correctly located listeners. Most of the time subjects seemed confused in front/back and up/down position.

MOS-LQS value (95% CI) was 3.68±0.79.

E. Test 3: Horizontal placement

Virtual placement of participants and listener/talker heights for *Horizontal Placement* can be seen in the Fig. 3c and in Table II. Listener is fixed at the center of the room and sound sources are moving left, right and front, back of the listener. The orientation of listener is facing position 2. In this test, we tried to depict the normal meeting arrangement but at the same time we wanted to check that to what extent our current setup helps to reduce front/back confusion [9], which is normal when non-individualized HRTFs are used.

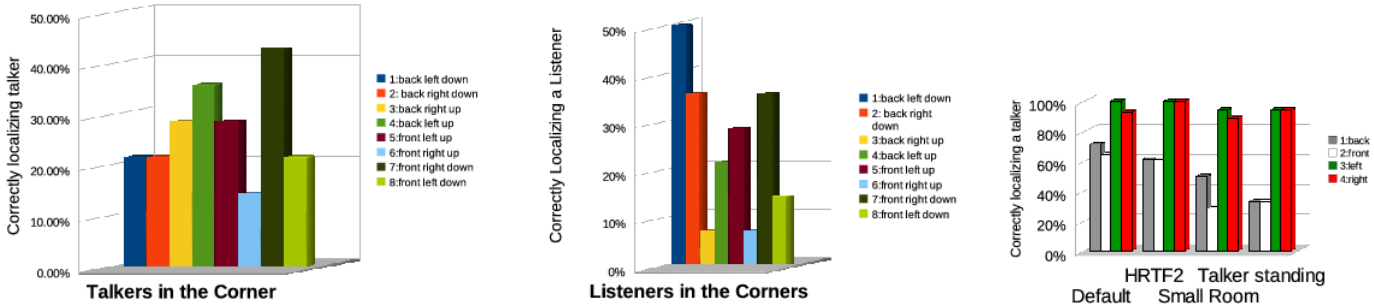
According to the test results, the best localizing talkers result is achieved by *Default* and lowest results are produced by *Talkers standing*. Importantly 73 percent subjects correctly located talkers and encouraged us to carry on with this kind of



(a) Test 1 layout (b) Test 2 layout (c) Test 3 layout (d) Test 4 layout

(e) Test 5 layout (f) Test 6 layout (g) Test 7 layout

Fig. 3: Placement of participants in the listening-only tests



(a) Talkers in the corners results (b) Listener in the corners results (c) Horizontal placement results

virtual placement of participants. On the other hand, 50 percent front/back confusion among the subjects was observed during the test where as left/right orientation had a 96 percent success rate.

MOS-LQS value (95% CI) was 4.25±0.63.

F. Test 4: Frontal placement-1

Front/back confusion and left/right accuracy of subjects in *Horizontal placement* made us to think and make some changes in the virtual placement of participants for the next test. In *Frontal placement-1*, we tried with normal sitting positions during office meetings, omitted talker's back position and increased one more participant to the virtual meeting, layout of which can be seen in the Fig. 3d. Summary of listener/talker heights can be seen in Table II. Five talkers were then placed in such an order that talker 1 (near left) and 3 (far left) were on the left side of the listener, talker 2 (near right) and 5 (far right) are on the right side of the listener and talker 4 (front) is in front of listener in the center of the room while only one talker was talking at a time.

Horizontal placement results raised an unanswered question to be addressed in *Frontal placement-1*, that is to check specifically the effectiveness of left/right orientation success in talker distance perception scenario. According to Shinn-Cunningham [10], reverberation degrades perception of the sound source direction but enhances distance perception. Sound source direction is no more a problem according to *Horizontal placement* results. IV-E.

We had some interesting findings in the results, 76 percent of the subjects correctly localized virtual talkers in *Frontal placement-1*, which is the best result among all tests. *Default* remained once again at the top by yielding the best localizing talker results that is 93 percent, confirming [10]. These results revealed the effectiveness of *Frontal placement-1* and *Default*.

MOS-LQS value (95% CI) was 4.07±0.68.

G. Test 5: Frontal placement-2

In *Frontal placement-2*, we continued with the same virtual placement of talkers with same issues to be addressed that we had during *Frontal placement-1*, except that we introduced

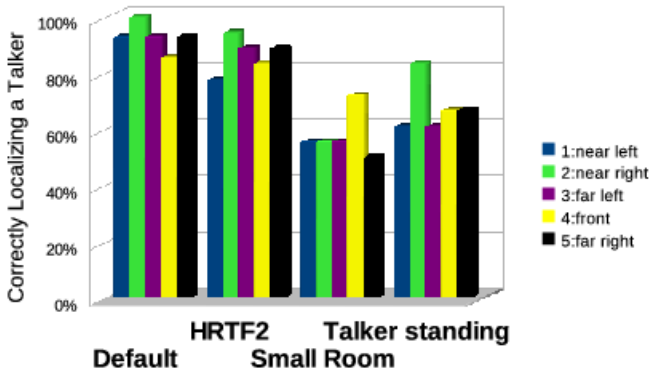


Fig. 4: Frontal placement-1 results

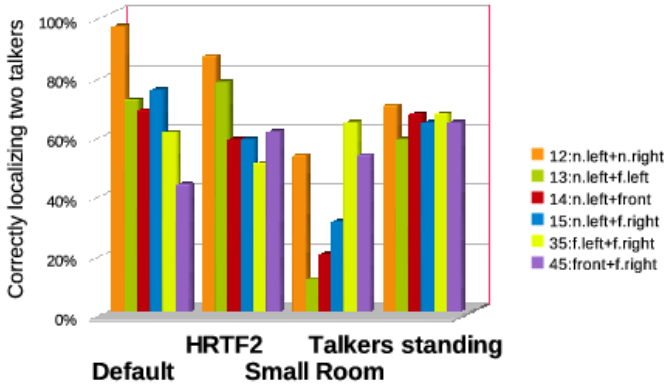


Fig. 5: Frontal placement-2 results

two talkers simultaneously. Human have natural ability to concentrate and to listen to a particular sound when there are many sounds like a cocktail party effect [29]. *Frontal placement-1* result success raised interesting question whether it would be easier for subjects in *Frontal Placement-2* to concentrate on single talker when two talkers are talking simultaneously and locate each virtual talker correctly, since the ability to recognize a talker is better for 3D presentation in presence of two or more talkers and even the time required to recognize a person is also shorter [30]. We introduced one male and one female talker in this test.

It was observed in the results that 59 percent virtual talkers were correctly localized in *Frontal placement-2*. *Default* remained once again at top by correctly localizing 69 percent of the talkers. *Small Room* had the lowest localizing talkers result which is equal to 38 percent.

Interestingly both talkers were 39 percent correctly located while one out of two talkers had 41 percent correctly localizing result. It was also observed in *Frontal placement-2* that 20 percent time none of the talkers were correctly located.

MOS-LQS value (95% CI) was 3.935±0.68.

H. Test 6: Surround placement-1

After better results obtained from *Frontal placement-1* and *Frontal placement-2*, we thought to make some changes in sitting arrangement of participants by adding three more

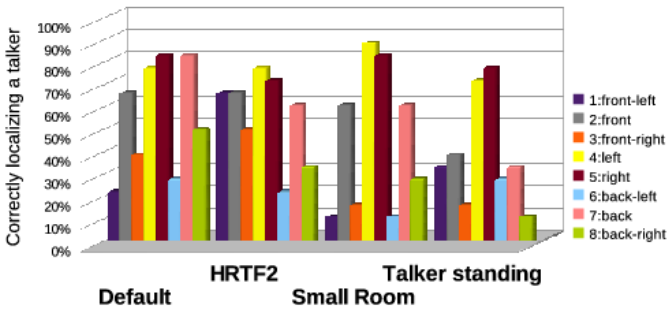


Fig. 6: Surround placement-1 results

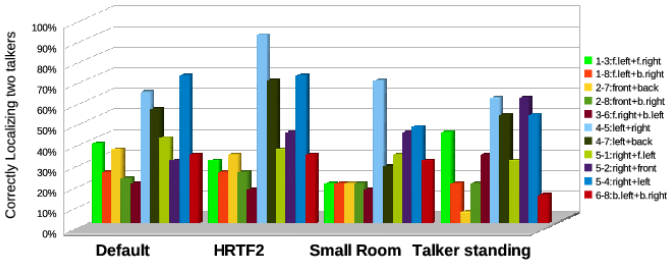


Fig. 7: Surround placement-2 results

virtual participants at the back of the listener by placing them to the back left, back right and back to see upto what extent our current setup helps to reduce front/back confusion [9].

Listener comes at the center of the room and talkers 1,4 and 6 are on the left of the listener. Talkers 3,5 and 8 are at the right of the listener. Talker 2 is in front and 7 is at the back of the listener. *Surround Placement-1* can be seen in the layout 3d whereas listener/talker heights can be seen in Table II.

In *Surround Placement-1*, 49 percent virtual talkers were correctly localized. Test results show that *Default* and *HRTF-2* had 56 percent correctly localizing a talker result while *Talker standing* produced 39 percent result and remained the lowest correctly localizing a talker setup for *Surround Placement*-1. Front/back confusion was once again observed and caused the overal result to reduce.

MOS-LQS value (95% CI) was 4.16±0.6.

I. Test 7: Surround placement-2

In *Surround Placement-2*, we introduced two talkers talking at the same time and kept the layout same as it was observed in *Surround Placement-1*.

In *Surround Placement-2*, 38 percent talker were correctly localized. Test results revealed that *HRTF-2* produced 43 percent correctly localizing a talker result and *Default* yielded 40 percent result and remained at the second best position. *Small room* had lowest result which is 32 percent.

Interestingly, subject were 17 percent sure about the location of both talkers and one out of two talkers had 42 percent success result. 41 percent time subjects were not sure about the location of any talker in *Surround Placement-2*.

MOS-LQS value (95% CI) was 4.12±0.61.

V. Summary

The quality of conference calls can be significantly enhanced if the telephones do not reproduce the speech in mono but instead use stereo headphone and spatial audio rendering. Then, one can identify participants by locating them and one can listen to one specific talker even if multiple talkers speak at the same time.

Listening-only tests using normal stereo headphones have shown that listeners can locate the origin of sounds and the position of talkers quite well. At the same time, the speech quality is only slightly reduced by adding reverberations echoes, and HRTF related filters. No subject complained about the lack of an understandability to understand the talkers or of any extra efforts required on user behalf to concentrate on talkers during user tests.

The test results have revealed that the performance of sound locating test is good when sound is placed at the same height with the listener and poor when it is vertically placed down or up in the direction of the listener. In our listening-only tests subjects seemed quiet sure about the sound orientation. The speech quality remained very good throughout all the tests and there were no impairments even with two echoes and reverberations. Same is true with two sound sources at a time, each of the sound source could be clearly heard and distinguished during the tests.

The *Default* setup employing an HRTF consisting of five reverberations for five frequency bands has produced better results among *HRTF-2* consisting of 10 reverberations for 10 frequency bands, *Small Room* and *Talker Standing* setups.

We believe that a 3D telephony system should not only be useful in conference calls but can also be beneficial for many other communication systems related to games like Doom, virtual environment aka Second Life and even for multimodal human computer interfaces.

References

[1] N. Yankelovich, J. Kaplan, J. Provino, M. Wessler, and J. M. DiMicco, "Improving audio conferencing: are two ears better than one?" in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work (CSCW '06)*. New York, NY, USA: ACM, 2006, pp. 333–342.

[2] J. Blauert and J. S. Allen, *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[3] H. C. Lee, H. B. Kim, M. J. Lee, P. M. Kim, S. W. Suh, and K. H. Kim, "Development of 3D sound generation system for multimedia application," in *Proceedings of 3rd Asia Pacific Conference on Computer Human Interaction*, Jul. 1998, pp. 120–123.

[4] D. R. Begault, *3-D sound for virtual reality and multimedia*. San Diego, CA, USA: Academic Press Professional, Inc., 1994.

[5] H.-J. Kim, D.-G. Jee, M.-H. Park, B.-S. Yoon, and S.-I. Choi, "The real-time implementation of 3D sound system using DSP," in *IEEE 60th Vehicular Technology Conference (VTC2004)*, vol. 7, Sep. 2004, pp. 4798–4800.

[6] C. Low and L. Babarit, "Distributed 3D audio rendering," *Computer Networks and ISDN Systems*, vol. 30, pp. 407–415, 1998.

[7] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *JAES*, vol. 49, no. 10, pp. 904–916, Oct. 2001.

[8] C.-J. Tan and W.-S. Gan, "User-defined spectral manipulation of HRTF for improved localisation in 3D sound systems," *Electronics Letters*, vol. 34, no. 25, pp. 2387–2389, Dec. 1998.

[9] E. Wenzel, M. Arruda, D. Kistler, and F. Wightman, "Localization using nonindividualized head-related transfer functions," *Journal-acoustical Society of America*, vol. 94, pp. 111–111, 1993.

[10] B. Shinn-cunningham, "Learning reverberation: Considerations for spatial auditory displays," in *Proceedings of the ICAD*, 2000, pp. 126–134.

[11] P. Hughes, "Spatial audio conferencing," in *ITU-T Workshop: "From Speech to Audio:bandwidth extension,binaural perception"*, Lannion France, 2008.

[12] V. Sundareswaran, K. Wang, S. Chen, R. Behringer, J. McGee, C. Tam, and P. Zahorik, "3D audio augmented reality: implementation and experiments," in *Proceedings of the Second IEEE and ACM International Symposium on Mixed and Augmented Reality*, Oct. 2003, pp. 296–297.

[13] D.-G. Kang, Y. Iwaya, R. Miyauchi, and Y. Suzuki, "Comparison of sound localization performance between virtual and real three-dimensional immersive sound field," *Acoustical Science and Technology*, vol. 30, no. 3, pp. 216–219, 2009.

[14] A. Raake, S. Spors, J. Ahrens, and J. Ajmera, "Concept and evaluation of a downward-compatible system for spatial teleconferencing using automatic speaker clustering," in *8th Annual Conference of the International Speech Communication Association*, Aug. 2007, pp. 1693–1696.

[15] C. Dicke, S. Deo, M. Billinghurst, N. Adams, and J. Lehikoinen, "Experiments in mobile spatial audio-conferencing: key-based and gesture-based interaction," in *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*. ACM New York, NY, USA, 2008, pp. 91–100.

[16] V. Hardman and M. Iken, "Enhanced reality audio in interactive networked environments," in *In Framework for Interactive Virtual Environments (FIVE)*, 1996.

[17] P. Pulli, T. Pyssysalo, J.-P. Metsavainio, and O. Komulainen, "CyPhone-experimenting mobile real-time telepresence," in *Proceedings of 10th Euromicro Workshop on Real-Time Systems*, 1998, pp. 10–17.

[18] D. R. Begault, *3-D sound for virtual reality and multimedia*. San Diego, CA, USA: Academic Press Professional, Inc., 1994.

[19] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," in *AES Convention: 107,*, Sep. 1999.

[20] N.-M. Cheung, S. Trautmann, and A. Horner, "Head-related transfer function modeling in 3-D sound systems with genetic algorithms," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 6, pp. 3529–3532 vol.6, May 1998.

[21] M. Hyder, M. Haun, and C. Hoene, "Measurements of sound localization performance and speech quality in the context of 3D audio conference calls," in *Internation Conference on Acoustics*. Rotterdam, Netherlands: NAG/DAGA, Mar. 2009.

[22] Uni-Verse consortium, "Uni-verse webpage," http://www.uni-verse.org/, Mar. 2007.

[23] R. Kajastila, S. Siltanen, P. Lunden, T. Lokki, and L. Savioja, "A distributed real-time virtual acoustic rendering system for dynamic geometries," in *122nd Convention of the Audio Engineering Society (AES)*, Vienna, Austria, May 2007.

[24] M. Puckette, "Pure data webpage," 2008. [Online]. Available: http://puredata.info/

[25] P. Min and T. Funkhouser, "Priority-driven acoustic modeling for virtual environments," in *Computer Graphics Forum*, vol. 19, no. 3. Blackwell Publishers Ltd, 2000, pp. 179–188.

[26] L. Savioja, J. Huopaniemi, T. Lokki, and R. Vaananen, "Creating interactive virtual acoustic environments," *Journal of the Audio Engineering Society (AES)*, vol. 47, no. 9, pp. 675–705, 1999.

[27] R. Vaananen, V. Valimaki, J. Huopaniemi, and M. Karjalainen, "Efficient and parametric reverberator for room acoustics modeling," in *ICMC 97*, 1997, pp. 200–203.

[28] ITU-R, "Method for objective measurements of perceived audio quality," Recommendation BS.1387, Nov. 2001.

[29] K. Crispien and T. Ehrenberg, "Evaluation of the "Cocktail Party Effect" for Multiple Speech Stimuli within a Spatial Auditory Display," *Journal of the Audio Engineering Society*, vol. 43, no. 11, pp. 932–941, 1995.

[30] R. Drullman and A. Bronkhorst, "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *The Journal of the Acoustical Society of America*, vol. 107, p. 2224, 2000.