Measuring 3D Audio Localization Performance and Speech Quality of Conferencing Calls for a Multiparty Communication System

MANSOOR HYDER*, GORDAN DAS MENGHWAR*, AND IMRAN ALI QURESHI**

RECEIVED ON 03.05.2012 ACCEPTED ON 21.06.2012

ABSTRACT

Communication systems which support 3D (Three Dimensional) audio offer a couple of advantages to the users/customers. Firstly, within the virtual acoustic environments all participants could easily be recognized through their placement/sitting positions. Secondly, all participants can turn their focus on any particular talker when multiple participants start talking at the same time by taking advantage of the natural listening tendency which is called the Cocktail Party Effect. On the other hand, 3D audio is known as a decreasing factor for overall speech quality because of the commencement of reverberations and echoes within the listening environment. In this article, we study the tradeoff between speech quality and human natural ability of localizing audio events/or talkers within our three dimensional audio supported telephony and teleconferencing solution. Further, we performed subjective user studies by incorporating two different HRTFs (Head Related Transfer Functions), different placements of the teleconferencing participants and different layouts of the virtual environments. Moreover, subjective user studies results for audio event localization and subjective speech quality are presented in this article. This subjective user study would help the research community to optimize the existing 3D audio systems and to design new 3D audio supported teleconferencing solutions based on the quality of experience requirements of the users/customers for agriculture personal in particular and for all potential users in general.

Key Words: **3D Audio, VoIP, Telephony, Teleconferencing, Virtual Reality.**

1. INTRODUCTION

Telephone is considered as one of the best inventions of the modern day world for communication with the people of all walks including businessmen, researchers and, students who often use this tool to communicate with their peers and market stack holders. Within the last decades; the number of user/customers for mobile and fixed lines phones has increased in many folds. On other side, there is hardly any improvement in the audio quality of phones. The chief limitation of the phones is extraneous noise, low speech intelligibility and poor audio

* Assistant Professor, Information Technology Centre, Sindh Agriculture University, Tando Jam.
** Assistant Professor, Department of Telecommunication Engineering, Mehran University of Engineering & Technology, Jamshoro.

quality, specifically in multiparty calls [1].

The major downside of the telephone and/or teleconferencing solutions is the lack of naturalness in the communication. The major theme of this research work is to develop telephone and teleconference solution that can support three dimensional audio, since our natural listening ability is also three dimensional.

Furthermore, we see it as an important task to investigate the difference between speech quality and localization performance of the participants of a conferencing call of our communication solution. In order to study this difference, we performed subjective user studies or listening-only tests by incorporating two different HRTFs, different placements of the teleconferencing participants and different layouts of the virtual environments. Subjective user studies were conducted by utilizing two different Head Related Transfer Functions, two different geometries of the room and different sitting positions and finally two different heights and head-sizes of the conferencing call participants.

The basic theme of this investigation is to study the four configurations; which are further described in the next sections, and judge them for their suitability for the users/ customers in telephony and teleconferencing purposes. Furthermore, we intend to address some specific queries such as: what is the performance of the participants of our conferencing call in locating their partners within virtual acoustic space? What is the performance of our telephony and teleconferencing solution in multi-talker situations specifically when more than one person starts talking at the same time?

2. RELATED WORK

Three dimensional sound supported systems were first initiated at NASA (National Aeronautics and Space Administration) research center. Further, a good understanding of three dimensional sound was achieved by Begault, [2]. Most of the research work in the area of three dimensional audio has been achieved by different research groups and individuals relating to conferencing solutions and audio localization. Hughes, [3] describes Senate which is a personal computer dependant client for audio and speech communication. Senate supports local audio files and has customizable graphical user interface for audio sources/streams to be played by the user.

Senate utilizes a server concentration system, which has client-server architecture similar to the centralized processing. It differs from our approach because it does the spatial rendering by centrally combining all audio signals from all clients. Afterwards, the audio signal is broadcasted to all the clients which then perform the spatial rendering . The server concentration scheme is suitable for use with new SAOC (Spatial Audio Object Coding) coding standard. The SAOC encoder is placed in the central server and the input 'objects' are the up streamed audio channels from each terminal. The down mixed signal, which is then down streamed to all the terminals, is decoded locally and conferees may configure their own audio experience as they wish. Local processing can include, in addition to the spatial rendering itself, the necessary control of the conferees own voice and echo control if necessary. The major downside of the Senate is a lack of user studies which supports author's claims.

Raake, et. al. [4] have descroned personal computer based 3D sound reproduction system for audio reproduction. However, from their work it is not clear whether this solution is customizable or not.

3. RESEARCH METHODOLOGY

Our three dimensional audio supported teleconferencing solution has been implemented with three dimensional sound software engine called Uni-Verse [5] for binaural processing. Uni-Verse is an open source entity which is

based on Verse-Server. Verse-Server keeps and shares all geometric data with all other clients attached to the virtual environment in real time. Furthermore, three dimensional audio is commenced by accessing geometry which is available at verse-server. Firstly, an acoustical simulation is achieved by acoustical propagation then the possible paths of the audio propagation within virtual environment are calculated [6]. Secondly, an audio renderer that implements HRTFs calculates stereo audio signal to be available for headphones playback. The audio renderer is based on Pure Data, Puredata, [7]; which is a graphical programming environment for a real time applications.

Furthermore, we found it an essential task to study how Uni-Verse acoustic simulations could be parameterized to obtain a better audio event or talker-localization performance but not distracting speech quality at all.

3.1 Test Parameters

In current study, we opted for four different sets of virtual acoustic simulation parameters and utilized them for the judgment of five different listener/talker placement positions in two different virtual acoustic rooms. Moreover, in total we tested 20 different combinations of the said parameters.

Furthermore, the audio test samples (one male and one female) were taken from ITU-T BS-1387 database [8] and then were binaurally processed using Uni-Verse audio engine having sampling rate of 24 kHz. In this user study, nine adult participants (subjects) (6 male and three female) took part. They all had a normal hearing threshold.

In each setup one of the Uni-Verse parameters was changed at any given time and other parameters were kept the same to study the effect that any parameter might have on the user/customer perception could be recorded. All the parameters are listed in Table 1.

3.2 Room Dimensions

This user study was based on two different sized rooms of the dimensions such as: (HxWxL=20x20x40m^3) and (HxWxL=10x10x20m^3). These rooms were named A big and a small room (Table 1).

3.3 Head Related Transfer Functions

Two different HRTFs were utilized which were named HRTFs-1 and HRTFs-2 having 5- and 10-frequency bands respectively.

3.4 Head Size

Head size is a Uni-Verse parameter which is defined as the internal difference between two ears having default value of 0.17cm. We utilized Head-size at its default value.

3.5 Placement

In this user study we tested five placement positions of the participants (listeners/Talkers). These placements positions are named Horizontal-, Frontal-1-, Frontal-2-, Surround-1-, and Surround-2 placement which are also listed in the Table 2.

TABLE 1. SUMMARY OF TEST SETUPS

Setup Name	Room Size	Height	HRTF	Head Size
Defaul	Big	A	1	0.17
HRTF2	Big	A	2	
Small Room	Small	A	1	
Talker Standing	Big	B	1	

3.6 Height

In this user study participants height within the Uni-Verse virtual acoustic room was kept different and has been summarized in Table 2.

In this user study, all listening-only tests were done by following the ITU (International Telecommunication Union) recommendation P.800. All 9 subjects took part in 20 listening only tests. All in all 180 tests were done to analyze the data. Each subject was presented the audio samples and were asked to complete two tasks for each listening-only test individually. Before the start of the test brief introduction was given to the subjects and were also requested to do some listening for their familiarity with our solution.

3.7 Tasks

Firstly subjects were offered audio samples and secondly they were requested to locate the talker within virtual acoustic space with the help of a given map which contains potential positions from where a talker might be talking.

Also, subjects were requested to provide us a quality of experience score in terms of easiness from 5-1 discreet MOS scale where 5=excellent, 4=good, 3=fair, 2=poor, and 1=bad.

4. RARTICIPANTS PLACEMENT

The five placements of participants are discussed in detail in the following.

4.1 Horizontal Placement

In this test listener and talker were positioned at 1.8m height. Further, the height and the layout for horizontal placement can be observed in (Table 2 and Fig. 1).

4.2 Frontal Placement-1

Frontal Placement-1 was formulated by keeping in view our normal and natural sitting positions where each participant face each other. Subjects were presented only single talker in this tests. Further, the height and the layout for frontal placement-1 can be observed in (Table 2 and Fig. 2).

4.3 Frontal Placement-2

The frontal placment-2 test has basically the same layout as frontal-placement-1 having only one difference that is the introduction of two simultaneous talkers. This test was

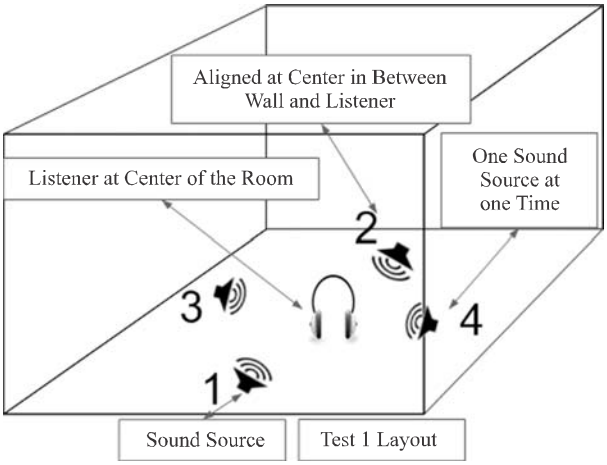

FIG. 1. HORIZONTAL PLACEMENT LAYOUT

TABLE 2. SUMMARY OF LISTENER AND TALKER HEIGHTS

Test	Height-A		Height-B	
	Listener (m)	Talker (m)	Listener (m)	Talker (m)
Horizontal Placement	1.8	1.8	1.0	1.5
Frontal Placement-1	1.0	1.0	1.0	1.5
Frontal Placement-2	1.0	1.0	1.0	1.5
Surround Placement-1	1.8	1.0	1.0	1.5
Surround Placement-2	1.8	1.8	1.0	1.5

meant to design to study the Cock Tail Party Effect phenomena [9-11]. Further, the height and layout for frontal placement-2 can be observed in (Table 2 and Fig. 3).

4.4 Surround Placement-1

In this test the person performing a role of the listener is placed at the center of the room. Further, the height and layout for surround placement-1 can be observed in (Table 2 and Fig. 4).

4.5 Surround Placement-2

In this test surround placement-1 setting was continued with only one difference that is the introduction of two simultaneous talkers at the same time. Further, the height and the layout for surround placement-2 can be observed in (Table 2 and Fig. 5).

5. SUBJECTIVE LISTENING ONLY TEST RESULTS

5.1 Horizontal Placement

74% localization results were observed in this test. 83% successful results were observed in HRTF2 setup. Additionally, it was evident from the results that left and right oriented positions were easier to localize than front and back positions (Fig. 6).

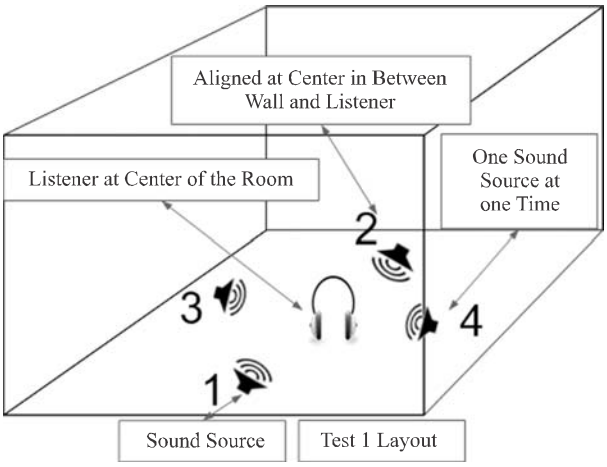

FIG. 2. FRONTAL PLACEMENT LAYOUT

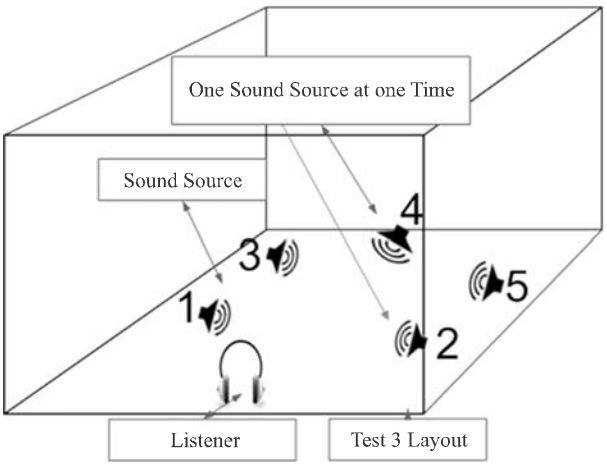

FIG. 3. FRONTAL PLACEMENT-2 LAYOUT

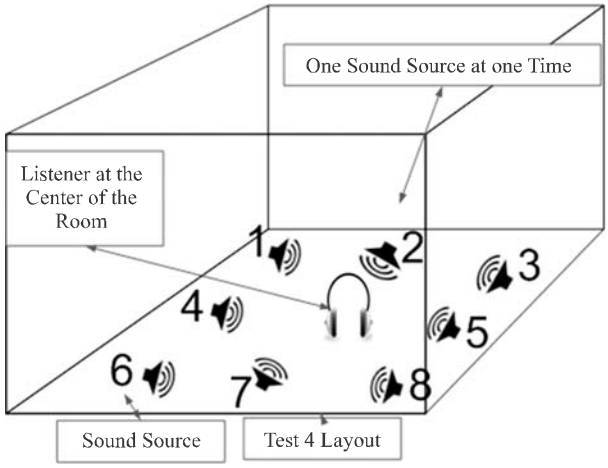

FIG. 4. SURROUND PLACEMENT-1 LAYOUT

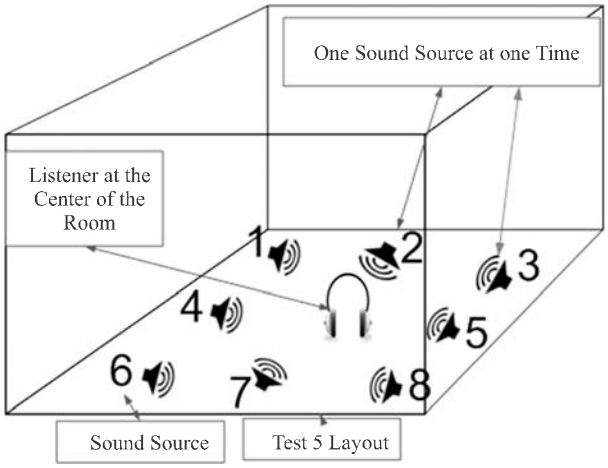

FIG. 5. SURROUND PLACEMENT-2 LAYOUT

5.2 Frontal Placement-1

75% localization results were observed in this test. Also, 97% localization was achieved by a Default setup (Fig. 7).

5.3 Frontal Placement-2

59% localization results were observed in this test. 69% successful results were achieved by HRTF2 setup (Fig. 8).

5.4 Surround Placement-1

43% localization results were observed in this test. 47% successful results were achieved by HRTF2 setup (Fig. 9).

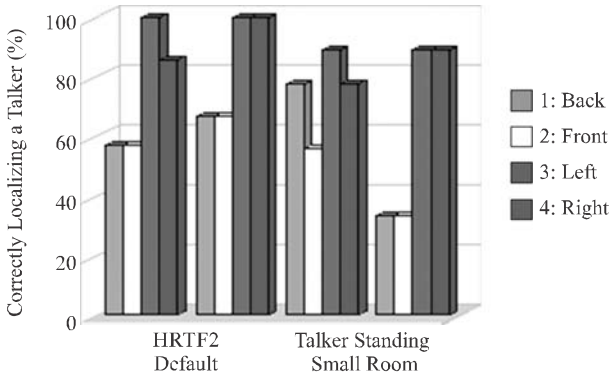

FIG. 6. HORIZONTAL PLACEMENT RESULTS

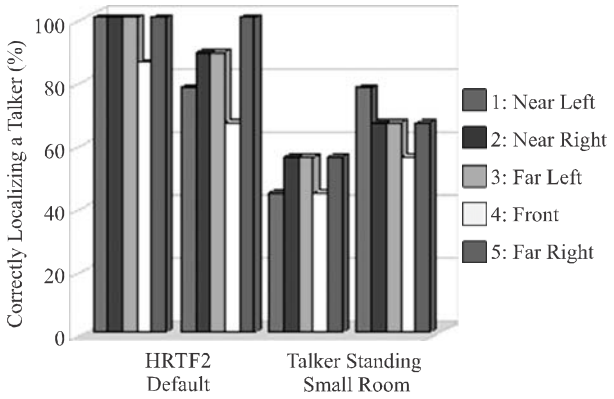

FIG. 7. FRONTAL PLACEMENT-1 RESULTS

5.5 Surround Placement-2

59% localization results were observed in this test. 46% successful results were achieved by HRTF2 setup (Fig. 10).

6. RESULTS AND DISCUSSION

Setups HRTF2 alongside Default produced over all better localization results as compared to the Small Room and

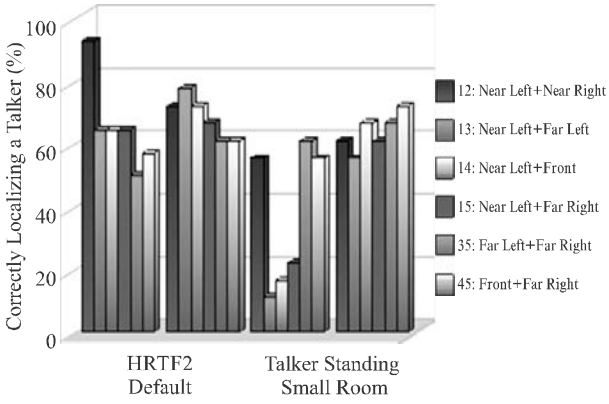

FIG. 8. FRONTAL PLACEMENT-2 RESULTS

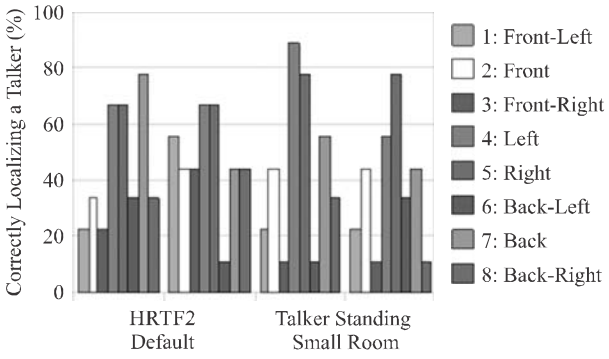

FIG. 9. SURROUND PLACEMENT-1 RESULTS

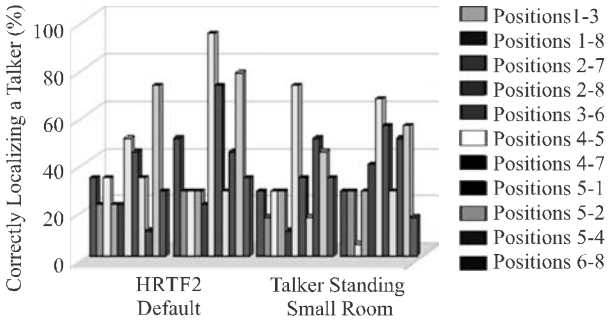

FIG. 10. SURROUND PLACEMENT-2 RESULTS

Talker Standing. The lowest localization performance was observed in Small Room. Additionally, it is evident from the results that standing positions for either a talker or a listener are not suitable for good localization performance and for speech quality.

Within talker placement positions, Frontal Placement-1 produced most successful localization results and better speech perception scores. It was also found that 1meter height is the most suitable for the participants within virtual acoustic environment.

Test participants found it very difficult to locate front and back talkers however they did not have any difficulties in locating right or left talkers. Additionally, it was found that frontal position produced most successful results with our three dimensional audio supported telephony and teleconferencing solution. We plan to conduct more user studies with frontal placement by increasing the number of test participants and also with more simultaneous talkers. Furthermore, we will incorporate head-tracking to the headphones to avoid front and back confusion which was commonly found during listening-only tests.

7. CONCLUSIONS

A good score/percentage in localization refers the performance of subjects in successfully locating concurrent talkers. In test results, 40% or above successfully localizing results are considered better localization scores. The usefulness of our solution can also be seen from the fact that most of the times subjects achieved 40% or above successfully localization results. Since frontal placement achieved better localization results which implies that frontal placement resembles the normal meeting scenario where all participants of the meeting face each other. In other words, frontal placement is near to the natural listening environment that we observe in everyday meeting. Normally, we do not sit in meeting by showing our back to the meeting members. This might be the reason that frontal placement produced better localization scores as compared to the other sitting arrangements. Furthermore, to optimize our solution to support more participants (at least from 5-9 persons), we need to conduct more tests with three or four simultaneous talkers. These tests are important to further study the "Cocktail Party Effect" using our solution. Cocktail Party Effect is nothing but the ability of human to concentrate on one talker in presence of other simultaneous talkers or background noises. However to conduct more tests with more simultaneous talkers require extensive efforts to optimize our solution and further work is being carried out in this regard.

ACKNOWLEDGEMENTS

Authors of this article are very thankful to Dr. Ing. Christian Hoene and Michael Haun, for their valuable feedback, help and suggestions. Authors are also thankful to the administration of Sindh Agriculture University, Tandojam and Mehran University of Engineering & Technology, Jamshoro, Pakistan, for their support.

REFERENCES

[1] Yankelovich, N., Jonathan K., Joe, P., Wessler, M., and Joan, M.D., "Improving Audio Conferencing: Are Two Ears Better Than One?", ACM, pp. 333-342, 2006.

[2] Begault, D.R., "3-D Sound for Virtual Reality and Multimedia", Academic Press Professional, Inc. 1994.

[3] Hughes, P., "Spatial Audio Conferencing", ITU-T International Workshop from Speech to Audio: Bandwidth Extension, Binaural Perception, Lannion France, 2008.

[4] Raake, A., Spors, S., Ahrens, J., and Ajmera, J., "Concept and Evaluation of a Downward-Compatible System for Spatial Teleconferencing Using Automatic Speaker Clustering", Eighth Annual Conference of the International Speech Communication Association, pp. 1693-1696, 2007.

[5] Uni-Verse, Uni-Verse Consortium, Uni-Verse Webpage, http://www.uni-verse.org/, 2007

[6] Kajastila, R., Siltanen, S., Lunden, P., Lokki, T., and Lauri S., "A Distributed Real-Time Virtual Acoustic Rendering System for Dynamic Geometries", 122nd Convention on Audio Engineering Society, Vienna, Austria, 2007.

[7] Puredata, M.P., "Pure Data" Webpage, http://puredata.info/, 2008

[8] ITU-BS-1387, ITU-R, "Method for Objective Measurements of Perceived Audio Auality", 2001.

[9] Crispien, K., and Ehrenberg, T., "Evaluation of the Cocktail Party Effect for Multiple Speech Stimuli within a Spatial Auditory Display", Journal of the Audio Engineering Society, pp. 932-941, 1995.

[10] Drullman, R., and Bronkhorst, A.W., "Multichannel Speech Intelligibility and Talker Recognition Using Monaural, Binaural, and Three-Dimensional Auditory Presentation", The Journal of the Acoustical Society of America, pp. 2224, 2000.

[11] Ericson, M.A., and McKinley, R.L., "Binaural and Spatial Hearing in Real and Virtual Environments", Lawrence Erlbaum, pp. 701-724, Mahwah, NJ, 1997.