Measurements of Sound Localization Performance and Speech Quality in the Context of 3D Audio Conference Calls

M. Hyder, M. Haun, C. Hoene
University of Tübingen, Germany, Email: mansoor.hyder/michael.haun/hoene@uni-tuebingen.de

Introduction

Teleconferencing solutions supporting 3D audio have two advantages. First, the talkers can be identified easily by locating their position in the virtual acoustic environment. Second, if multiple persons talk at the same time, humans can focus on the talker of their choice by taking advantage of the cocktail party effect. However, 3D audio lowers the speech quality by decreasing loudness and adding reverberation and echoes. In this publication we study the trade off between speech quality and the ability of focusing and locating the talkers. We extend our initial work presented in [4], in which we study the 3D audio rendering engine by Raine Kajastila *et al*. [6]. We conduct subjective and objective listening-only tests using different head related transfer functions (HRTF), different positions of the conference call participants, and different geometries of the virtual environment. We present the statistical significant test results for sound localization, performance and the speech quality. This study helps us to design our teleconference solution that we call 3D Telephony.

It is important to study the relationship between speech quality and the ability of focusing and locating the talkers in 3D conference calls. We conducted listening only tests, applying different HRTF's, different algorithmic complexities, different geometries of the room and different positions of the conference call participants. We presented the results of 35 initial tests in an ITU-T workshop [4], for which we made user studies to check the best possible placement of listene

rs and talkers and their height in a virtual room for 3D teleconferencing. This time we have chosen to account some more parameters, i.e. different room sizes, different HRTF's, different heights of listeners and talkers and headsizes. In total, we conducted 180 more tests.

The goal of this study is, to see which configuration of the virtual room is the best out of the four described in the following section. We were particularly interested in answering the following questions.

- How well do listeners locate talkers in the virtual room?
- How well can listeners locate multiple simultaneous talkers?
- To what extent is the speech quality still fine, if more than one person is talking simultaneously?

Experimental Design

We are implementing a 3D teleconference solution using a 3D sound processing software that has been originally developed for game programming in the EU FP6 research project Uni-Verse[8]. Uni-Verse's open source 3D sound software implements "distributed interactive audio visual virtual reality system". The central component is called "Verse Server", which stores and shares 3D geometric data of a virtual environment. A real-time visual rendering application accesses the Verse server to display the current virtual environment. Also, 3D sound is achieved by accessing the 3D geometry stored in the Verse server. First, an acoustic simulation simulates the acoustic propagation in the virtual room [6]. The paths of sound propagation and virtual sound sources are calculated to figure out from where and when the sound waves arrive at the listeners head. Next, a sound renderer implementing an HRTF calculates stereo audio signals, which are played via the headphones. The sound renderer is built on Pure Data (PD), a real-time graphical programming environment developed by Puckette [7].

Using the Uni-Verse acoustic simulation, we needed to understand how it should be parameterized to achieve a good sound localization performance while remaining the speech quality. For that, we selected four sets of simulation parameters and used them for judging the five different placements of participants in the virtual room (in total 20 combination).

More precisely, we generate test signals using two reference samples. These two reference sound samples were taken from the database ITU BS.1387-1 [5], one of them is a male voice and other one is a female voice. Next, we calculated the impact of spatial sound on those samples using the Uni-Verse framework working at a sampling rate of 24kHz due to complexity reasons. The resulting samples were judged by nine subjects.

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. We have used two different HRTF's, two different room sizes, different heights of the listener and talker and kept the headsize constant to examine the speech quality and ability of focusing and locating talkers. The following parameters can be chosen for the acoustic simulation (Table 1).

Room dimensions: In our test experiments, we used two rooms. A *Big Room* having dimensions (HxWxL=20 x 20 x 40 m^3) and a *Small Room* having dimensions (HxWxL=10 x 10 x 20 m^3).

HRTF: We have used two HRTF's 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).

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 1: 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 2: Summary of listener and talker heights

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

Height: The placement of listeners and talkers in terms of height in the virtual room is summarized in Table 2. 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*.

The listening only tests have been done following the ITU-T Recommended P.800 recommendations as far as possible. Nine subjects (six male, three female) participated in listening only tests. Each subject participated in 20 tests comprising of 4 setups. In total there were 180 tests altogether.

Subjects were presented with two 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: How easily can you understand the talker? (When there is one talker and when there are more than one talkers?) Please score individual talker from 5 to 1 (5=excellent, 4=good, 3=fair, 2=poor, 1=bad).

Participants Placement and Listening-test Results

The five placements of participants and the respectively listening-test results are discussed in detail in the following.

In *Horizontal Placement* test, we placed the listener and talker at 1.8 meter height for *Default*, *HRTF-2* and *Small Room.* In *Talker Standing* listener is at 1 meter height and talker at 1.5 meter height. Listener is fixed at the center of the room and talkers or sound sources are moving left, right, front and back. We used one talker at a time. The orientation of the listener is facing talker at position-2. Listener and talker heights are summarized in Table 2. Layout for *Horizontal Placement* test can be seen in the Fig 1.

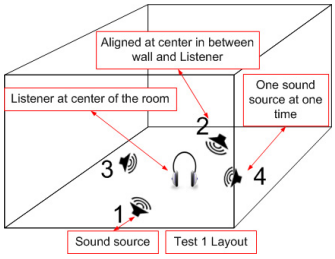


Figure 1: Layout: *Horizontal Placement*

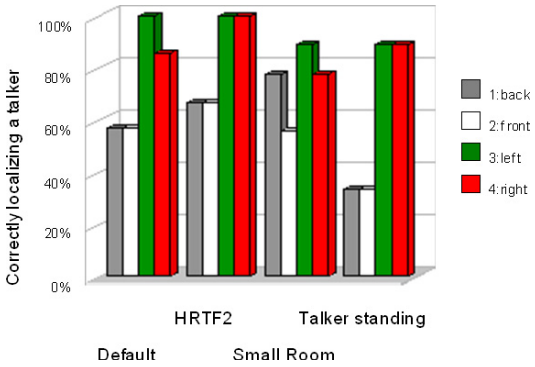
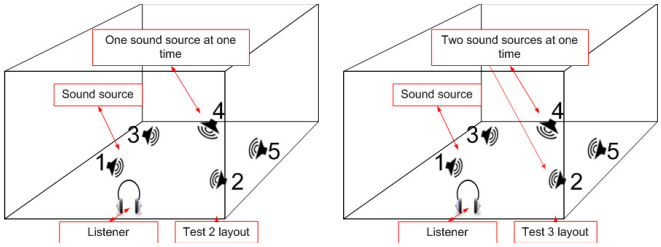


Figure 2: Results: *Horizontal Placement*

According to test results, the talkers were 74 percent correctly located in this test for all setups. Notably best located talker result was seen in *HRTF-2* which is 83 percent, and lowest located talker result was seen in *Talker Standing* which is 61 percent. It was seen that position left (3) and right (4) had better correctly located results than positions front (2) and back (1) (Fig 2). During this tests, the subjects pointed out that they had some difficulties to judge whether the sound was coming from front or back.

We used a normal sitting arrangement this time which we call *Frontal Placement-1*. This sitting arrangement is usually observed in normal meetings, by sitting around the table. We presented one talker at a time listener and talkers are at 1 meter height. Orientation of the listener is facing all the conference participants in front, being 1 and 3 are on the left side, 4 is in in front in the center of the room, and 2 and 5 are on right side of the listener. The placement of listener and talkers in terms of height is summarized again in the Table 2. The layout for the *Frontal Placement-1* can be seen in Fig. 3a.

According to the results, *Frontal Placement-1* has 75 percent correctly located talkers in all setups, which is the best ratio among all tests. Among all the setups, *Default* yielded 97 percent correctly located talkers which is the best result, and *Small Room* yielded 51 percent correctly located talkers,



(a) *Frontal Placement-1* **(b)** *Frontal Placement-2*

Figure 3: Placement of talkers and listener

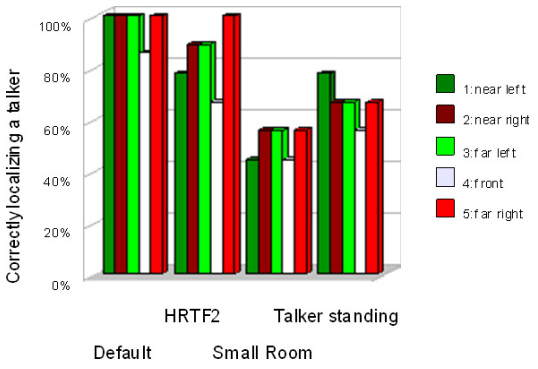


Figure 4: Results: *Frontal Placement-1*

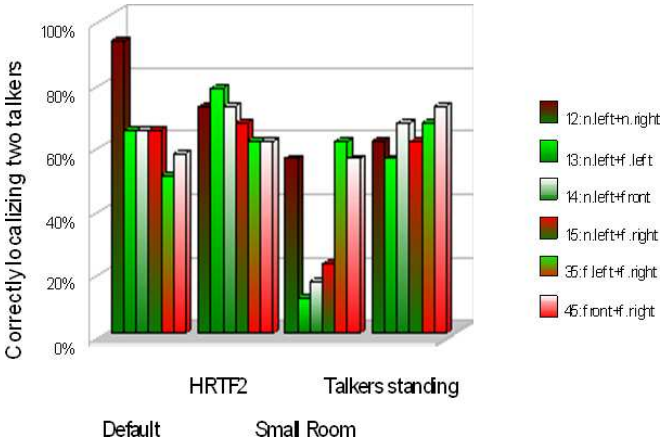


Figure 5: Results: *Frontal Placement-2*

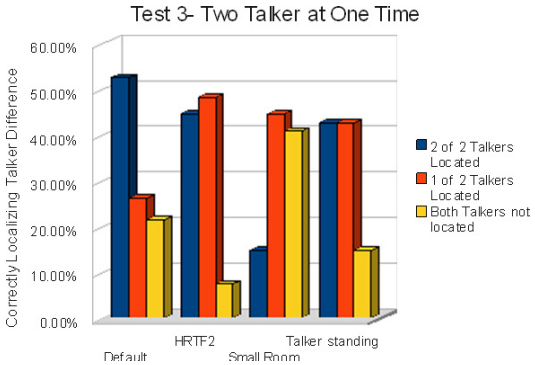
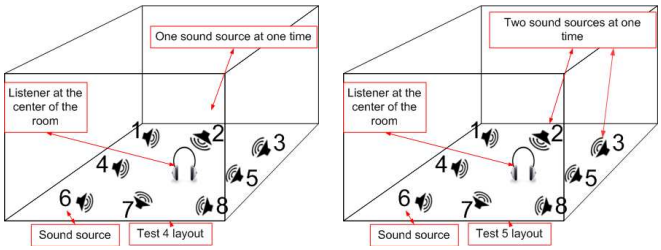


Figure 6: Results: *Frontal Placement-2*



(a) Layout: *Surround Placement-1* **(b)** Layout surround placement-2

Figure 7: Placement of participants

which is the lowest result. By looking at Fig. 4, we can conclude that both, *Small Room* and *Talker Standing*, produced less correctly located talker results than *Default* and *HRTF-2*. Based on the evaluation of results for the initial two tests, we can notice the effectiveness of *Default* and *HRTF-2* setups.

The layout for *Frontal Placement-2* can be seen in Fig. 3b. This is a similar layout like we presented in *Frontal Placement-1,* but having the difference that two talkers are talking at the same time. Testing the natural ability of listeners/subjects to concentrate and to understand one talker when there are many talkers, like the "cocktail party effect" [1, 2, 3], by using 3D teleconferencing solution was the main focus point.

According to the test results, 59 percent of the subjects located the talkers correctly in all setups. (Fig 5.) Among setups, *HRTF-2* yielded the best combined male and female talkers correctly located result which is 69 percent and *Small Room* yielded the lowest correctness result of 37 percent. *Default* and *Talker Standing* yielded 66 percent and 64 percent correctly located talkers result respectively.

It was important for us to know the results, when both talkers were correctly located during *Frontal Placement-2* test. (Fig. 6.) Test results revealed that 39 percent both talkers were located correctly. On the other hand, one out of two talkers was 40 percent and none of two talkers was 21 percent correctly located. Among the setups, *Default* yielded 52 percent result by localizing both talkers correctly which is the best performance and as worst setup remained *Small Room* which yielded 15 percent results. *HRTF-2* and *Talker Standing* setups yielded 44 percent and 43 percent results by locating both talkers correctly. We observed that the correctly located talkers percentage has decreased in *Frontal Placement-2* compared to *Frontal Placement-1.*

The layout for *Surround Placement-1* can be seen in Fig. 7a. In order to find out the best possible placement to be used in 3D Teleconferencing solution, we used different placements of listener and talkers this time. In *Surround Placement-1*, the listener is placed at the center of the room while 8 talkers are placed surround the listener in a way that talker 1, 4 and 6 are placed on the left side. Talker 3, 5 and 8 are placed at the right side. Talker 2 is placed in front and 7 is placed in the back of the listener.

According to the test results, 43 percent of the subjects correctly located the talkers in *Surround Placement-1* in all setups. Among setups, *HRTF-2* and *Talker Standing* yielded 47 percent and 37 percent, the best and poorest correctly located results respectively. (Fig 8.) *Default* and *Small Room* yielded 44 percent and 43 percent correctly located talkers respectively.

In *Surround Placement-2* we continued with *Surround Placement-1* layout but with one change, instead of one talker we presented two simultaneous talkers to the subjects. *Surround Placement-2* layout can be seen in Fig 7b. For a description about the placement of talkers and listener refer to the previous test.

According to the test results, 59 percent of the subjects located the talkers correctly in all setups. Among setups, *HRTF-2* and *Small Room* yielded 46 percent and 33 percent, the best and the poorest located talkers results respectively. *Default* and *Talker Standing* setups yielded 34 percent and 36 percent located talkers results respectively. When it comes to locate both or one of the two or none of the two talkers correctly,

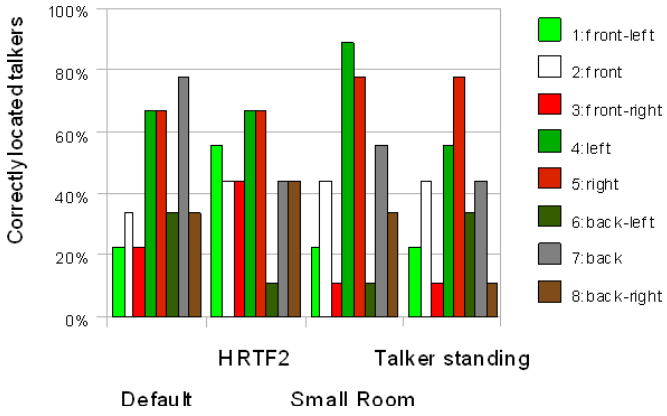


Figure 8: Results: *Surround Placement-1*

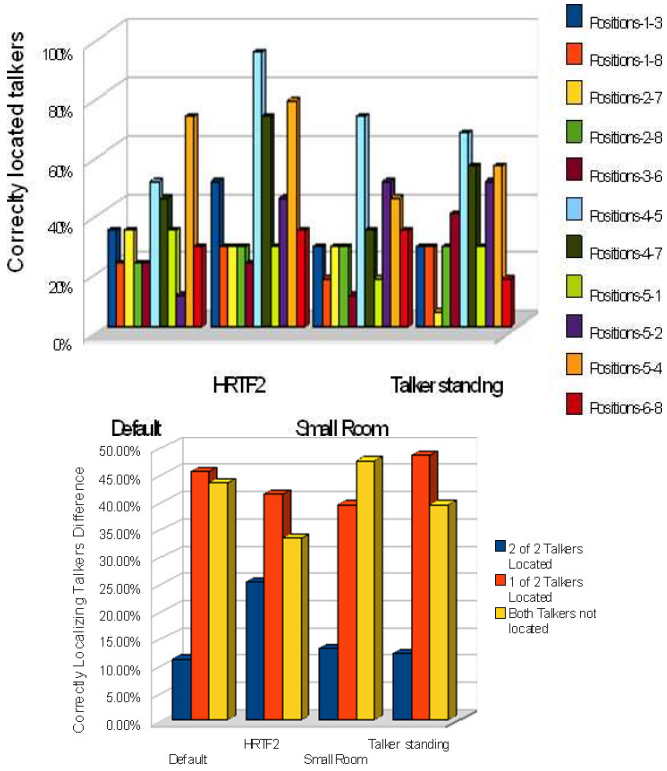


Figure 9: Results: *Surround Placement-2*

HRTF-2 yielded better results, which is quiet evident in Fig 9. We observed a very low result for both talkers correctly located in *Surround Placement-2*. Among setups, *HRTF-2* and *Small Room* yielded 25 percent and 13 percent localizing both talkers correctly. It was noticed that the subjects had difficulties whether the sound was coming from front or back but all the time they seemed very clear about the orientation of the sound.

Summary

According to the test results *HRTF-2* and *Default* proved to be the best among all the setups. *Small Room* produced the lowest located talkers results. It is quiet clear now that neither *Talker Standing* nor listeners standing is the suitable placement, because the listeners had difficulties in locating the talkers. Best located performance was observed in *Frontal Placement-1*, five talkers sitting in front of the listener. It is evident from the results that 1 meter height for both listener

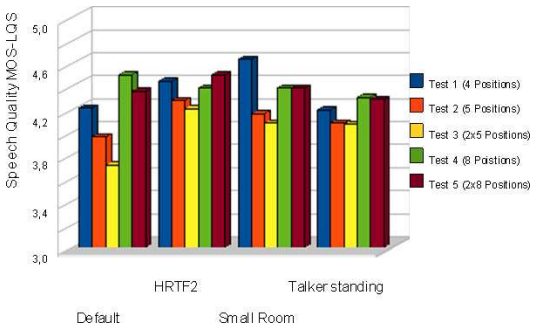


Figure 10: Speech qualitiy (MOS-LQS)

and talkers using 3D Teleconferencing setup, is a good height. The Subjects declared that they had difficulties to understand whether the sound was coming from front or back, although they seemed sure about the orientation of the talkers.

The natural frontal position is working fine with the 3D Teleconferencing solution. Further, we will conduct more subjective and objective listening only tests with *Frontal Placement* because it produced good results. We can do this by increasing the number of participants and also by increasing number of talkers simultaneously. In addition, we will add headphones that include a head tracking unit to overcome the front/back confusion.

References

[1] 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*, 43(11):932–941, 1995.

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

[3] M.A. Ericson and R.L. McKinley. *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter The intelligibility of multiple talkers separated spatially in noise, pages 701–724. Erlbaum, Mahwah, NJ, r. h. gilkey and t. r. anderson edition, 1997.

[4] Mansoor Hyder and Christian Hoene. 3D Telephony. *From Speech to Audio: bandwidth extension, binaural perception", in Lannion France*, Sep 10-12 2008.

[5] ITU-R. Method for objective measurements of perceived audio quality. Recommendation BS.1387, November 2001.

[6] Raine Kajastila, Samuel Siltanen, Peter Lunden, Tapio Lokki, and Lauri Savioja. A distributed real-time virtual acoustic rendering system for dynamic geometries. *Audio Engineering Society Convention Paper Presented at the 122nd Convention Vienna, Austria*, May 5-8 2007.

[7] Miller Puckette. Pure data webpage. http://puredata.info/, 2008.

[8] Uni-Verse consortium. Uni-verse webpage. http://www.uni-verse.org/, March 2007.