Audio Engineering Society

Convention Paper

Presented at the 129th Convention
2010 November 4–7 San Francisco, CA, USA

Assessing Virtual Teleconferencing Rooms

Mansoor Hyder[1], Michael Haun[1], Olesja Weidmann[1], Christian Hoene[1]

[1] *Interactive Communication Systems (ICS), Wilhelm-Schickard-Institut für Informatik (WSI), University of Tuebingen, 72076, Tuebingen, Germany*

Correspondence should be addressed to Mansoor Hyder or Christian Hoene
(mansoor.hyder@uni-tuebingen.de, hoene@uni-tuebingen.de)

ABSTRACT

Spatial audio makes teleconferencing more natural, helps to locate and distinguish talkers in a virtual acoustic environment and to understand multiple talkers. This paper presents a study on how to design virtual acoustic environments used in 3D audio teleconferences to maxime loclization performance, easiness and subjective speech quality ratings. We conducted subjective listening-only tests considering different parameters describing the virtual acoustic environment, including acoustic room properties, virtual sitting arrangements, reflections of a conference table, number of concurrent talkers and different voice types of call participants. The experimental results help us to enhance the performance of our open-source, spatial audio teleconferencing solution named "3DTel" by enhancing the quality of its user experience in terms of naturalness and speech quality.

1. INTRODUCTION

Teleconferencing systems provide good means to communicate collaboratively with an added advantage of saving traveling costs and effort. But today's phone based conferencing tools display a lack in audio quality and naturalness of communication. More research efforts are required to put these audio conferencing solutions on new lines by providing users with a natural audio communication feeling. Incorporation of 3D audio is one way to improve the overall quality of audio conferencing solutions, but further optimizations are required. A 3D audio conferencing solution can be improved if the virtual acoustic environment, which is part of most 3D audio simulations, is chosen properly.

This paper describes a series of experiments and examines the effects that simulated acoustic room properties, virtual sitting arrangements, reflections of a conference table, number of concurrent talkers and voice characteristics have on the perception of

speech quality, locatability and speech intelligibility in a 3D teleconferencing system. Particularly, the tests conducted were designed to answer the following questions: To what extent are multiple talker localization performance and subjective speech quality ratings influenced by the size of the virtual conference room? What are the results when a conference table is simulated and what is the overall impact of changing the conference table size? What results are achieved when the number of simultaneous talker increases? Do different voice types have an influence on the easiness of locating simultaneous talkers? What are the results when there is an increase in talker position density?

The remainder of this paper is structured as follows: Section 2 lists related and ongoing research related to 3D audio, spatial audio teleconferencing systems and the quality assessment of such systems. Section 3 discusses the methodology, setup and performance of the listening-only tests presented in this paper by listing the utilized testing scenarios, procedures and terms. Afterwards, the results of these tests will be presented in detail in Section 4. Finally, the paper is concluded with a summary of the obtained results in Section 5.

2. RELATED WORK

Teleconferences suffer from many well known problems. For example, the listener performance in multi-talker scenarios decreases in terms of understanding speech, locating talkers and concentrating on a talker of choice as there is an increase in auditory scene complexity [4]. If binaural or even 3D audio is incorporated in teleconferencing systems, the quality of teleconferences can be increased [21, 3].

Multiple 3D audio teleconference systems have been implemented. In [6], Hughes presented a 3D audio teleconferencing system called Senate. Reynolds et al [18] presented a distribution model for headphone based spatialized audio conferences. Herre et al [5] described a combination of Spatial Audio Object Coding and Directional Audio Coding technologies to be used for interactive teleconferencing. Spors et al. presented the SoundRenderer Framework in [1] that can be used to render 3D audio for teleconferences [15]. In previous work [8], we described a 3D audio telephony and teleconferencing system called "3D Telephony". We have implemented the system using a 3D sound processing software called UniVerse [11]. The system enables the participants to have their calls placed in a virtual 3D environment.

Spatial audio teleconferencing systems under development are far from mass market usage as their quality of experience does not fulfill all user demands yet. Consequently, it is very important to measure the quality of existing systems to understand how to improve them. Kilgore et al [12, 14] presented experimental research to determine whether the combination of spatialization and simple visual representation of a voice's location helps recognizing completely unfamiliar voices. The test results evidently show that localization easiness benefits when spatial audio is coupled to a visual interface only with a large number of voices, as in this case with eight, but not with four voices. Versterinen [20] tested performance differences between 3D, monophonic and stereophonic audio conferences through subjective tests in her work "Audio Conferencing Enhancements". Results presented show that spatially mixed hemispherical audio produced the most pleasing listening experience of a multi-person conversation. Inkpen et al [10] explored the impact of spatialized audio and video on user-experience in multiway video conferences using a proprietary software. Their study didn't reveal any significant differences between mono audio and spatialized audio. The results of other studies [13, 21, 8] however showed positive influence of spatial audio. Because of these contradicting research results, we see it as an important task to improve spatial audio conferencing as different spatial teleconferencing systems might perform significantly different.

3. USER EXPERIMENTS

In order to enhance our 3D Telephony system we conducted formal listening-only tests to measure localization performance, localization easiness, spatial and overall speech quality of different virtual teleconferencing scenarios.

To measure localization performance each test participant was presented a map with possible talker locations. Then, the actual location of each talker was compared to the location selected by the test participant. Localization easiness described the subjectively perceived effort required by test participants to localize a talker, while spatial quality

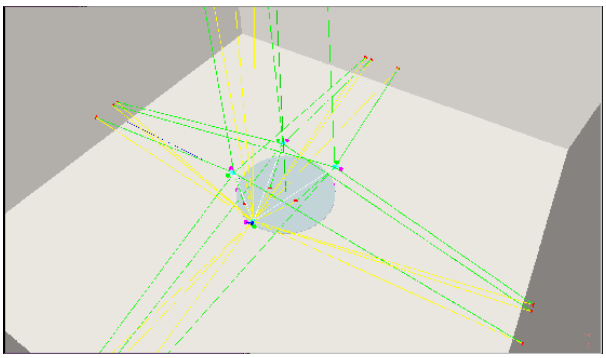


Fig. 1: Virtual room with three simultaneous talkers

described how well the participant could perceive that talkers were spatially separated, and overall speech quality referred to the perceived speech quality as compared to a real life conversation. Localization easiness, spatial and overall speech quality were measured using discrete $MOS-LQSW$ scores with the values 1 (bad), 2 (poor), 3 (fair), 4 (good) and 5 (excellent). The $MOS-LQSW$ values were named $MOS-LQSW_{LE}$ for localization easiness, $MOS-LQSW_{SQ}$ for spatial quality and $MOS-LQSW_{OQ}$ for overall speech quality.

During the tests the five parameters *voice type*, *number of concurrent talkers*, *table size*, *talker position density* and *room size* were modified. The influence of each parameter was evaluated by comparing a specially designed test setup consisting of a series of two tests to a given reference test.

User experiments were conducted with 31 paid subjects, 13 of them female and 18 of them male, according to ITU-T P.800 recommendations [16] as far as possible. All test participants were aged between 20 and 45 years with an average age of 27 years. 8 out of 31 participants showed earlier experiences with listening-only tests, and all subjects indicated a good to professional level of computer proficiency. The average time taken by the subjects to complete all tasks given in the tests was 62 minutes. Each subject participated in 11 different tests contained in 5 different setups and one reference test, thereby assessing quality and localization information on 71 audio samples. Thus, 2201 audio samples were collectively assessed during the tests all together.

All audio samples consisted of anechoic speech samples taken from the ITU-T Rec. P.50 Appendix 1 library [17]. They were prerecorded from and processed by the open-source 3D audio rendering engine Uni-Verse [11] at a sampling rate of 16 kHz. A screenshot of Uni-Verse's rendering engine is shown in Figure 1, and further details about the usage of the Uni-Verse framework can be obtained in [8]. The speech samples were recorded using three different male and three different female voices each speaking four sentences in American English. Table 1 lists all samples used during the experiments as well as their duration and corresponding file names in the ITU-T Rec. database. Using human speech samples as sound sources in the experiments has been thought of as a direct application to the problem of a multi-party teleconferencing system.

Sample	File name	Duration
Male 1	A_eng_m5.wav	14.62s
Male 2	A_eng_m7.wav	13.65s
Male 3	A_eng_m4.wav	11.11s
Female 1	A_eng_f4.wav	09.95s
Female 2	A_eng_f5.wav	13.14s
Female 3	A_eng_f7.wav	12.66s

Table 1: Listing of all used speech samples

All tests were conducted in a quiet listening room on a computer using a specially designed user interface as shown in Figure 2. Before the tests were conducted, each participant received an introduction into the testing environment and instructions about the tasks to be accomplished during the tests. Every test was preceded by a learning phase during which the participants were presented reference samples with their accompanying correct locations. In the training phase, all samples were presented in the same linear order to each participant and could be played up to three times using the provided play button, before moving on to the next sample by pressing the next button. To enable participants to distinguish the different talkers contained in each sample, each talker was represented by a number as well as its spoken text.

Each participant was asked a series of questions to be answered for each talker contained within each sample. First, the locations of all talkers were to be determined by selecting a location from a map

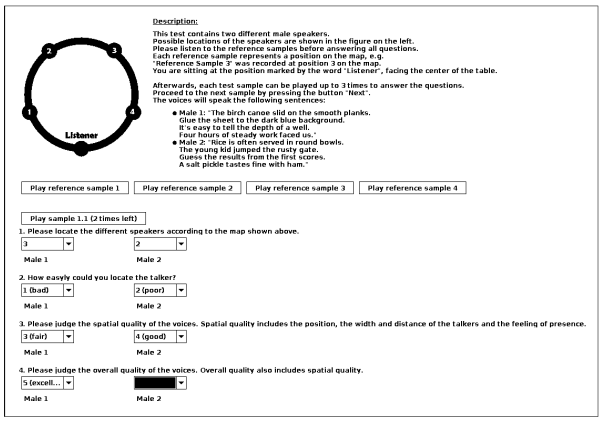
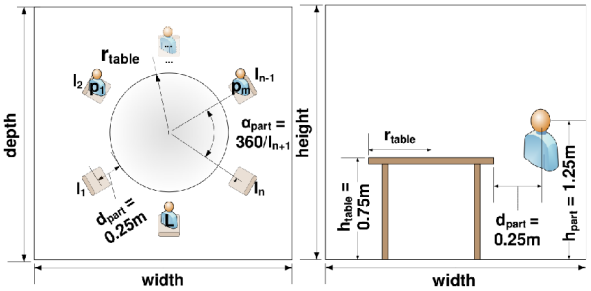


Fig. 2: Graphical user interface of the testing environment



(a) Top view (b) Frontal view

Fig. 3: The virtual conference room

of possible talker locations. Secondly, location easiness, spatial and overall speech quality had to be selected by using the previously described discrete MOS values MOS_{LE}, MOS_{SQ} and MOS_{OQ}.

3.1. Experimental Design

All tests were performed in cubic virtual conference rooms with varying dimensions. The walls of the rooms showed the typical acoustic properties of concrete. A schematic overview over the virtual test environment and all measured parameters is shown in Figure 3.

A round conference table showing the acoustic properties of wood was placed at the center of the room at a height of $h_{table} = 0.75m$ above the floor. The table had a variable radius of 2, 3 or 4 meters depending on the test.

Either 5, 7, or 9 participants were distributed equally around the table. All participants were placed at a distance of $d_{part} = 0.25m$ from the table and at a height of $h_{part} = 1.25m$ above the floor.

In each test, one of the participants always represented the listener and was placed at a fixed position. To simulate the listener, a generic HRTF for five frequency bands was assumed due to good experiences obtained during our previous studies [9, 7, 8]. All other participants represented talkers whose positions and numbers were varied in the different setups, with at least 2 and at most 4 participants talking concurrently at the same time. Additionally the distribution of male and female talkers was varied to examine the influence of the different voice types on localization performance and subjective speech quality.

Beside a reference test setup, we tested five different setups varying one of the above mentioned parameter at a time as compared to the reference test. Table 2 lists all setups and their respective parameters. The setups were called *Voice Type, Number of Simultaneous Talkers, Listener-to-Sound Source Distance, Talker Position Density* and *Sound Source-to-Wall Distance* and are described in the following sections.

3.1.1. *Reference Test*

The reference test is based on processed speech signals with an average length of $14.38s$, simultaneously spoken by two male talkers from four possible locations distributed around the table. The virtual conference room has a size of $20 \times 20 \times 20m^3$, the radius of the table is set to $2m$. Sound source positions are labeled relatively to the listener location as *1-NearLeft*, *2-FarLeft*, *3-FarRight*, *4-NearRight*, the position of the listener is labeled *Listener*. The listener and all sound sources are facing the center of table. Within the reference test, six samples with different combinations of voice-to-position assignments were recorded. The total number of samples assessed for this test is 186.

3.1.2. *Voice Type*

The goal of this setup was to test the impact of relative and absolute differences in voice types, such as two concurrent male, female or mixed talkers. Therefore, the two tests within this setup,

Name	Room dimension	Participants	Table radius	Simultaneous talkers	Voice type
Reference	$20 \times 20 \times 20m^3$	5	$2m$	2	m/m
Voice Type	$20 \times 20 \times 20m^3$	5	$2m$	2	**f/f** **m/f**
Number of Simultaneous Talkers	$20 \times 20 \times 20m^3$	5	$2m$	**3** **4**	$m/m/m$ or $f/f/f$ $m/f/m/f$
Listener-to-Sound Source Distance	$20 \times 20 \times 20m^3$	5	**3m** **4m**	2	$m/m/m$
Talker Position Density	$20 \times 20 \times 20m^3$	**7** **9**	$2m$	2	m/m
Sound Source-to-Wall Distance	$\mathbf{15 \times 15 \times 15m^3}$ $\mathbf{10 \times 10 \times 10m^3}$	5	$2m$	2	m/m

Table 2: Test setups and parameters

Voice Type-1 and *Voice Type-2* utilize two simultaneous female talkers with an average signal length of $13.03s$, and two mixed talkers with an average signal length of $14.42s$ as opposed to the two male talkers used in the reference test.

It is assumable that the results attained by this setup will show better localization performance and localization easiness scores for mixed gender samples, since both voices can be distinguished more easily than with two voices of the same gender.

3.1.3. *Number of Simultaneous Talkers*

This setup was used to measure the changes in localization performance, easiness and speech quality when the number of concurrent talkers increases. Thus, the two tests *Number of Simultaneous Talkers-1* and *Number of Simultaneous Talkers-2* employ three and four simultaneous talkers respectively. To level the effect of mixed-gender voice types, half the samples recorded for *Number of Simultaneous Talkers-1* consisted of three male talkers, while the other half consisted of three female talkers. The average signal length for this test was $14.36s$. In *Number of Simultaneous Talkers-2,* all samples consisted of two female and two male talkers and the average signal length was $14.47s$.

The results of this setup are likely to show two opposing trends: while an increase in the number of simultaneous talkers is likely to confuse the test participants, *Number of Simultaneous Talkers-2* is expected to show better results in localization performance, since four out of four possible sound source locations are occupied in each presented sample. Thus, the error of misperceiving a sound source's location with an empty location on the presented map should be minimized. Although the localization performance will decrease for *Number of Simultaneous Talkers-1* and increase for *Number of Simultaneous Talkers-2* in relation to the results achieved in the reference test, it is safe to assume that an increase in the number of simultaneous talkers will lead to a decrease in subjective localization easiness and speech quality scores.

3.1.4. *Listener-to-Sound Source Distance*

The *Listener-to-Sound Source Distance* setup was used to measure the impact of the distance between the individual sound sources and the listener. Therefore, the tests *Listener-to-Sound Source Distance-1* and *Listener-to-Sound Source Distance-2* use virtual conference tables with radii of $3m$ and $4m$ respectively. The average stimuli lengths were $14.42s$ and $14.47s$.

It can be assumed that the localization performance as well as subjective localization easiness scores will enhance with an increase in table radius, since the the distance between the individual sound source

locations also increases with the size of the table. Thus, the different possible locations of the sound sources should be distinguishable more clearly.

3.1.5. *Sound Source Density*

This setup was designed to measure localization performance, easiness and speech quality when the number of possible sound source locations increases. Thus, the two tests *Sound Source Density-1* and *Sound Source Density-2* employ six and eight possible talker locations respectively. All locations were distributed equally around the virtual conference table as shown in Figure 3a. The average signal lengths for these tests were $14.41s$ and $14.43s$. While *Sound Source Density-1* features six possible sound source locations occupied by two male voices and a fixed listener position, *Sound Source Density-2* uses eight possible sound source locations as well as a fixed listener position. Sound source positions are labeled as *1-NearLeft*, *2-CenterLeft*, *3-FarLeft*, *4-FarRight*, *5-CenterRight* and *6-NearRight* in *Sound Source Density-1*, and as *1-NearLeft*, *2-NearCenterLeft*, *3-FarCenterLeft*, *4-FarLeft*, *5-FarRight*, *6-FarCenterRight*, *7-NearCenterRight* and *8-NearRight* in *Sound Source Density-2*. *Sound Source Density-1* employs seven different voice-to-position assignments, and ten different voice-to-position assignments were recorded for *Sound Source Density-2*. The total number of samples assessed was 217 and 310 respectively.

The results of this setup are likely to show that localization performance decreases with an increasing density of possible sound source positions, since the distance between adjacent sound sources diminishes, thus reducing the difference in spatial information carried by the individual voices. Additionally, reflections produced by the conference table are more likely to be close to possible sound source locations, thereby confusing the participant's spatial perception.

3.1.6. *Sound Source-to-Wall Distance*

To determine the effect of room size and sound source-to-wall distance on all measured scores, this test uses two different rooms with dimensions of $15 \times 15 \times 15m^3$ and $10 \times 10 \times 10m^3$ for the tests *Sound Source-to-Wall Distance-1 and Sound Source-to-Wall Distance-2* respectively. The average lengths of the presented stimuli add up to $14.65s$ and $14.43s$ for the two tests.

The amount of reverberation and echo increases with the room size. According to Shinn-Cunningham [19], reverberation degrades perception of the sound source direction, but enhances distance perception. Hence it is assumable that the localization performance and easiness will increase with a decrease in room size.

4. RESULTS

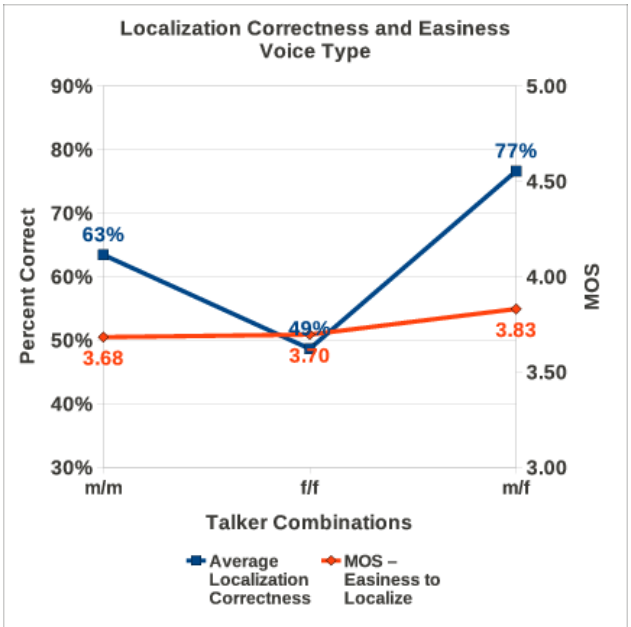
In normal listening situations we segregate the information masked by other simultaneous sounds by utilizing our natural ability to hear in three dimensions. We extract required sounds and/or information by taking advantage of the "cocktail party effect" [2]. It was of great concern for us to check to what extent our audio teleconferencing and telephony system "3D Telephony" helps users to solve the "cocktail party effect" problem and to what degree our solution is acceptable and comparable to natural human listening phenomena. A detailed analysis of the acquired experimental data is presented in the following sections.

4.1. *Reference Test*

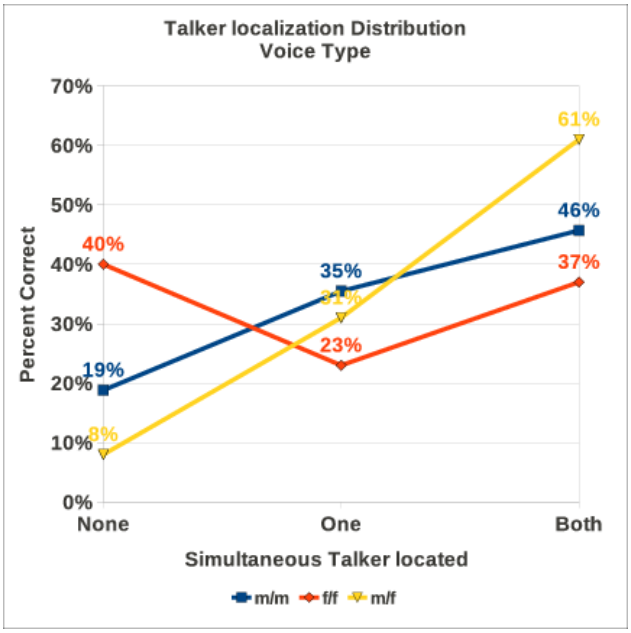
Since *Reference Test* was the foundational test and was designed in order to be used for the comparison to all other tests, the results obtained for this test are significant for the whole experimental process. The analysis of the reference test showed, that in 46% of all samples both talkers were located correctly, in 35% just one, and none of the talkers in the remaining 19%. Overall, 64% of all talkers were located correctly. Misperception occurred mostly between the *4-NearRight* talker location and *3-FarRight* location (30%) and between position *2-FarLeft* and *3-Far Right* (22%). MOS ratings on a 95% confidence interval (CI) were found to be 3.68 ± 0.11 ($MOS - LQSW_{LE}$ *on* 95% CI), 3.84 ± 0.10 ($MOS - LQSW_{SQ}$ *on* 95% CI) and 3.86 ± 0.10 ($MOS - LQSW_{OQ}$ *on* 95% CI).

4.2. *Voice Type*

For this setup, the localization correctness was yielding better results (overall 77%, both 61%, one 31%, none 8%) with mixed gender talkers (*Voice Type-2*) than with two male talkers (overall 64%, both 46%,



(a) Localization correctness vs. $MOS - LQSW_{LE}$ ratings



(b) Talker localization distribution

Fig. 4: *Voice Type*

one 35%, none 19%, as above) and two female talkers (*Voice Type-1,* with scores of overall 49%, both 37%, one 23% and none 40%) as shown in Fig. 4a and 4b. Misperception was observed to happen between similar positions as in the reference test. The MOS ratings for two female talkers and mixed gender talkers did not shown any statistical significant difference as compared to the reference values. The only exception is the $MOS - LQSW_{LE}$ *on* 95% CI rating for mixed gender talker, which was found to be 3.83 ± 0.12.

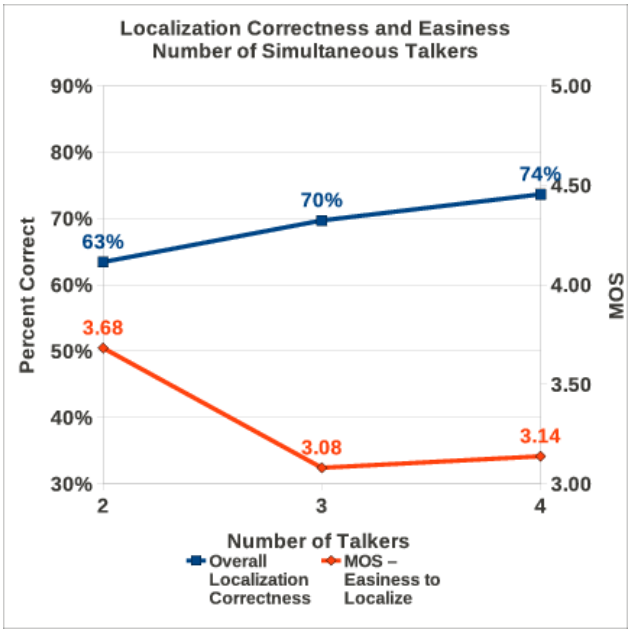
4.3. *Number of Simultaneous Talkers*

The results of this setup show an increase in localization correctness with an increasing number of concurrent talkers. In *Number-of-Simultaneous-Talkers-1*, an overall localizing correctness of 70% was observed (Fig. 5a). More precisely, in 51% of all samples, three simultaneous talkers were located correctly, in 17% of the cases only two out of three, in 24% of the cases one out of three, and in the remaining 9% none of the talkers was correctly located (see Fig 5b). The MOS values were all lower than the corresponding reference values: MOS_{LE} *on* 95% CI was 3.08 ± 0.13, MOS_{SQ} *on* 95% CI was 3.12 ± 0.11 and MOS_{OQ} *on* 95% CI was 3.19 ± 0.11.

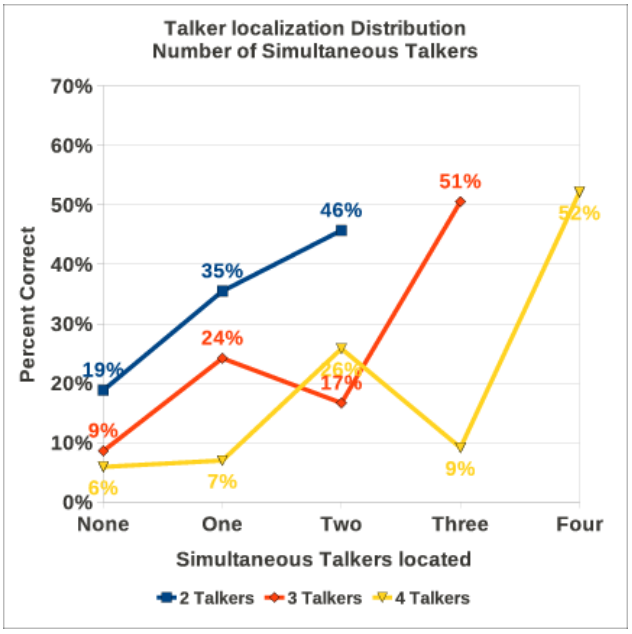
In *Number of Simultaneous Talkers-2*, overall 74% of all talkers were located rightly. All four talkers were located correctly in 52% of the cases, which is comparatively better than the results using three simultaneous talkers. Additionally, in 9% of the cases three talkers were located correctly, in 26% two talkers and in 7% of all presented talkers one out of four talkers was located correctly. Only at 6% of the time no talker could be located correctly. The MOS ratings were similar to the ratings found in *Number of Simultaneous Talkers-1*, only MOS_{LE} *on* 95% CI was slightly better at 3.14 ± 0.13.

4.4. *Listener-to-Sound Source Distance*

The results of *Listener-to-Sound Source Distance* show that a larger table leads to better localization performance. *Listener-to-Sound Source Distance-1* employed a table radius of $3m$. Here, 71% overall correctly located talkers were achieved as compared to 63% obtained in the reference test, as shown in

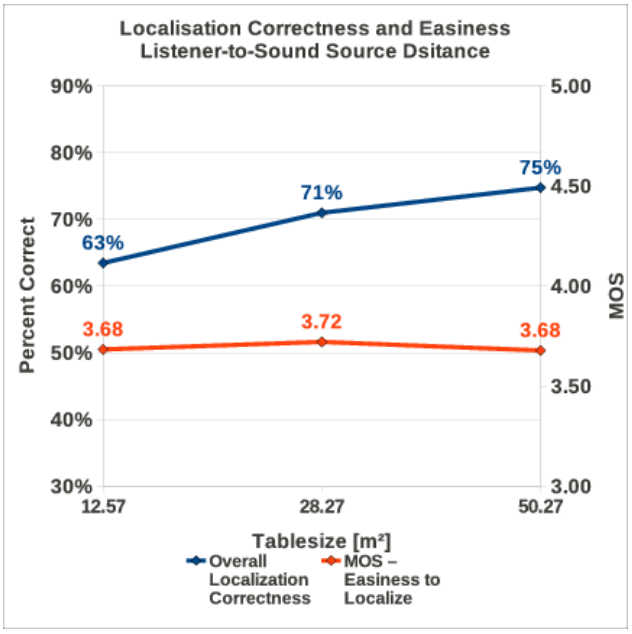


(a) Talker localization vs. MOS_{LE} ratings

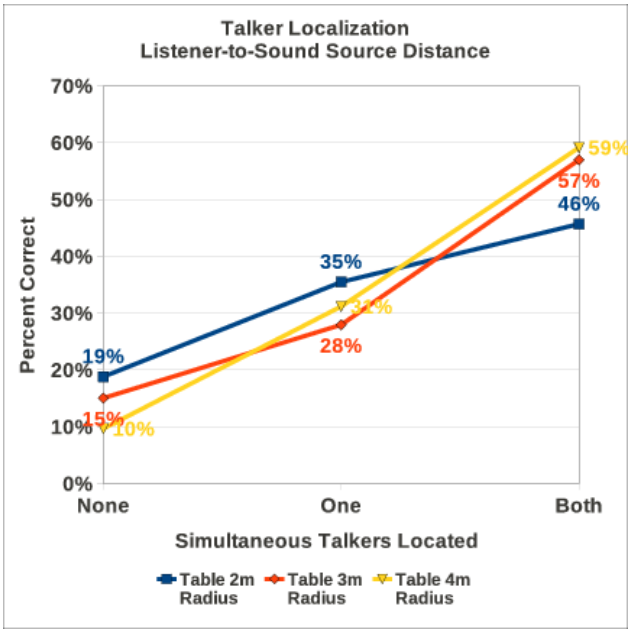


(b) Talker localization distribution

Fig. 5: *Number of Simultaneous Talkers*



(a) Localization correctness vs. MOS_{LE} ratings



(b) Talker localization results distribution

Fig. 6: *Listener to Sound Source Distance*

Fig. 6a. In 57% of the cases, both talkers were located correctly, one of two in 28%, and none in 15% of all cases (Fig. 6b).

Misperception occurred in a matter similar to the reference test, while all MOS scores were slightly higher at 3.72 ± 0.10 (MOS_{LE} *on* 95% CI), 3.68 ± 0.09 (MOS_{SQ} *on* 95% CI) and 3.75 ± 0.09 (MOS_{OQ} *on* 95% CI).

Using a radius of $4m$ for the virtual conference table in *Listener-to-Sound Source Distance-2* yielded 75% overall correctly located talkers, while in 59% both talkers were located correctly, in 31% only one of two and in 10% none of the talkers were located correctly. All MOS scores for this test were found to be within the confidence interval of *Listener-to-Sound Source Distance-1*.

4.5. *Sound Source Density*

In *Sound Source Density-1*, six possible talker locations were used. An overall correctness for talker localization of 47% was found (see Fig. 7a), while both talkers could be located in 28%, one talker in 39% and no talkers in 34% of all cases as shown in Fig. 7b. Misperception occurred mainly between *4-FarRight* and *5-CenterRight* (48%) as well as between *1-NearLeft* and *2-CenterLeft* (47%). MOS scores were slightly lower than in the reference test with values of 3.34 ± 0.10 (MOS_{LE} *on* 95% CI), 3.56 ± 0.09 (MOS_{SQ} *on* 95% CI) and 3.62 ± 0.09 (MOS_{OQ} *on* 95% CI).

In *Sound Source Density-2*, each talker could be placed on one of eight possible locations. Here, only 37% overall talker localization correctness was achieved. In 17% of all cases, both talkers were located correctly, in 41% only one and in 42% none of the talkers were located correctly. Misperception occurred between *5-FarRight* and *6-FarCenterRight* (44%) and between *1-NearLeft* and *2-NearCenterLeft (42%).* Again, MOS ratings were found to be of 3.15 ± 0.09 (MOS_{LE} *on* 95% CI), 3.48 ± 0.08 (MOS_{SQ} *on* 95% CI) and 3.53 ± 0.08 (MOS_{OQ} *on* 95% CI), which is slightly lower than those obtained in the reference test.

4.6. *Sound Source-to-Wall Distance*

In *Sound Source-to-Wall Distance-1* which was conducted in a medium sized room of $15x15x15m^3$ with a volume of $3375m^3$, an overall localization correctness of 72% could be achieved as depicted in Fig. 8a. Both talkers could be located in 57% of all cases, one out of two in 30% and in 13% none of the talkers was located correctly as shown in Fig. 8b. Location misperception was found to be near equal to the values found in the reference test. All MOS ratings were found to be slightly lower but within the confidence interval of the ratings achieved in the reference test.

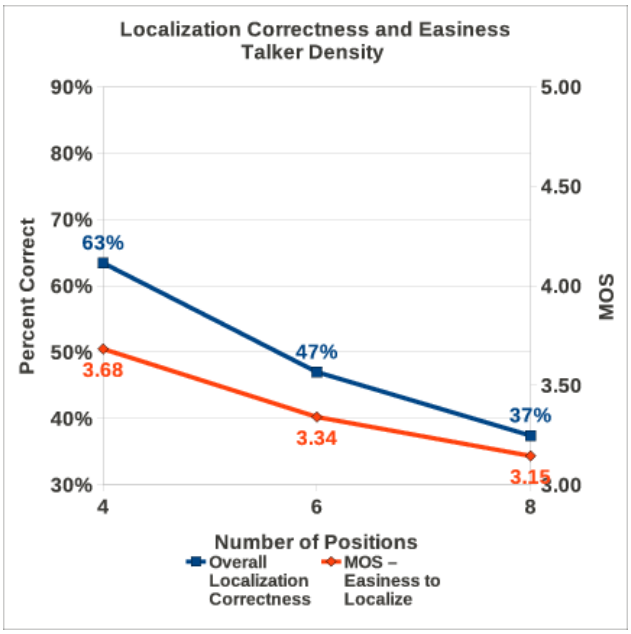
Sound Source-to-Wall Distance-2 with the dimensions of $10x10x10m^3$ and a volume of $1000m^3$ also exhibited a correctly located talker ratio of 72%, while both talkers could be located in 58% of the cases, one talker in 30% and none of the talkers in 13%. Again, misperception was found to be similar to the reference test, and MOS ratings were near equal to *Sound Source-to-Wall Distance-1* and the reference test.

5. SUMMARY

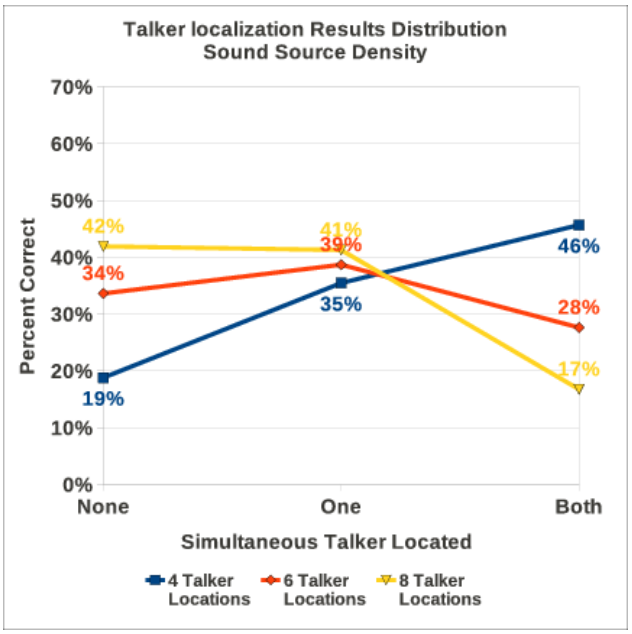
As shown by the results listed in Section 4, each of the measured parameters has a substantial influence on talker localization performance.

Results of the *Voice Type* setup clearly state that participants were able to locate two simultaneous talkers more often when the presented stimuli were of different genders as was previously assumed, and that two male talkers were easier to locate than two female talkers. While the first finding can be explained by the fact that it is much easier to distinguish two different voices if their pitches differ greatly. A possible explanation is that the male voices showed greater differences in voice pitch and hence were easier to differentiate than the female voices. But since the subjective location easiness ratings do not show any significant differences between the reference test and *Voice Type-1/2*, one can assume that the reasons were not that obvious. Another explanation can be give by the fact, that the experiments were performed by more male than female participants. Both tests achieved subjective MOS quality ratings at an acceptable level.

It could also be shown that an increasing number of participants leads to higher localization correctness ratios, which partly contradicts the preliminary assumptions made in Section 3.1.3. Although this result seems counter-intuitive, one has to keep in

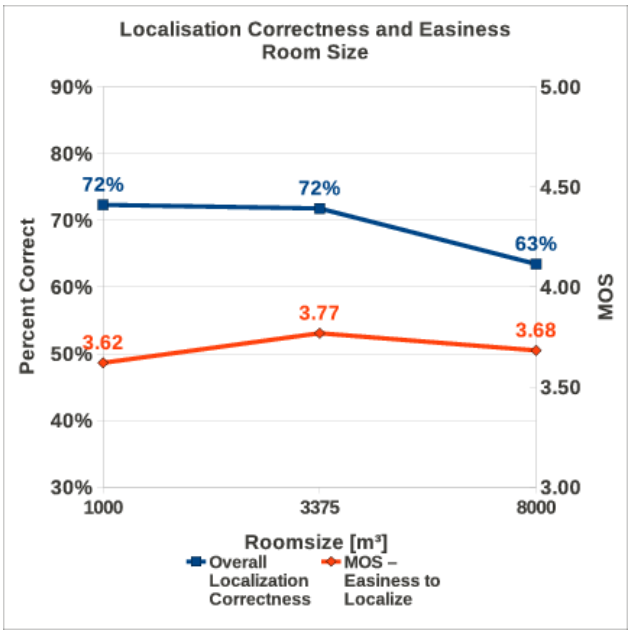


(a) Localization correctness vs. MOS_{LE} ratings

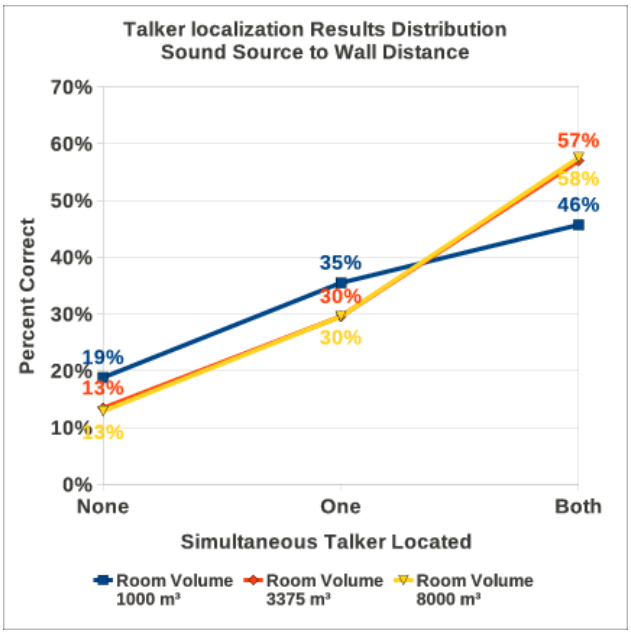


(b) Talker localization results distribution

Fig. 7: *Sound Source Density*



(a) Localization correctness vs. MOS_{LE} ratings



(b) Localization correctness results distribution

Fig. 8: *Sound Source-to-Wall Distance*

mind that the number of possible talker locations was kept constant while the number of concurrent talkers increased, and hence the talker-to-location ratio increased with the number of concurrent talkers. Therefore, participants were able to directly compare all concurrent talkers and the error of misperceiving a talker location with an empty location was minimized. Subjects reported that spatial separation of all simultaneous talkers helped them to determine the corresponding locations to a good extent, although echoes and reverberations for three simultaneous talkers made it difficult to absorb the situation for a longer period, thereby resulting in significantly lower $MOS - LQSW_{LE}$ ratings for 3 and 4 simultaneous talkers.

It was assumed that as the size of conference table would increase, better localization rates could be observed, since the increase in size is accompanied by greater distances between different talker locations. This assumption was verified by the results obtained in *Listener-to-Sound Source Distance-1/2*. While subjects found it similar easy to judge the talkers' correct locations in all three tests, their performance improved from 63% using a table with a radius of $2m$ in the reference test, to 72% and 75% when using tables with radii of $3m$ and $4m$ respectively.

While in *Number of Simultaneous Talkers* the talker-to-location ratio increased with the number of concurrent talkers, this ratio significantly decreased in *Sound Source Density-1/2* as compared to the reference test. Additionally, the distance between two adjacent talker locations decreased since the table radius was kept constant. Hence, the hypothesis of a significant decrease in talker localization performance could be verified by the results presented in Section 4.5. These results state that the density of possible talker locations distributed around a conference table has a significant impact on talker location performances, as the ratio decreases from 63% in the reference test to as low as 47% and 37% for six and eight possible talker locations. Additionally, *Number of Simultaneous Talkers-2* yielded the lowest $MOS - LQSW_{LE}$ ratings of all tests, while subjective speech quality ratings were found to be only slightly lower than those of the reference test.

Results of *Sound Source-to-Wall Distance* state that with a decrease in room size and volume localization performance increases. This verifies the prediction made in Section 3.1.6 and can be explained by the increasing number of echoes and reverberation in larger rooms, which according to Shinn-Cunningham [19] enhance the distance perception but degrades localization performance. While localization performance increases, test subjects found it equally easy to judge the talkers' locations for the reference, *Sound Source-to-Wall Distance-1* and *Sound Source-to-Wall Distance-2* tests, and the perceived spatial and overall speech quality showed no statistically significant differences between these three tests.

Aside from influences on localization performance and easiness, subjects were found to misperceive the talker locations. These misperceptions occur more often, as the density of possible talker location increases, and can be explained due to the phantom sources created by reflections on the virtual conference table. When the density of talker location increases, a phantom source is more likely to be close to one of the possible talker locations and hence might be confused with that position. Therefore, misperception mostly occurred between two adjacent positions either on the near left (between positions l_1 and l_2 as shown in Fig. 3a) or near right side (positions l_n and l_{n-1}) of the listener.

6. REFERENCES

[1] Jens Ahrens, Matthias Geier, and Sascha Spors. The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods. In *Audio Engineering Society Convention 124*, 5 2008.

[2] Barry Arons. A Review of the Cocktail Party Effect. *Journal of the American Voice I/O Society*, 12:35–50, 1992.

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

[4] D.S. Brungart, B.D. Simpson, C. Bundesen, S. Kyllingsbaek, AM Burton, and AM Megreya. Cocktail party listening in a dynamic multitalker environment. *Perception and Psychophysics*, 69(1):79, 2007.

[5] J. Herre, C. Falch, D. Mahne, G. del Galdo, M. Kallinger, and O. Thiergart. Interactive teleconferencing combining spatial audio object coding and dirac technolog. *Audio Engineering Society, 128th Convention*, May 2010.

[6] Peter Hughes. Spatial audio conferencing. Lannion France, 2008.

[7] Mansoor Hyder, Michael Haun, and Christian Hoene. Measurements of sound localization performance and speech quality in the context of 3D audio conference calls. In *Internation Conference on Acoustics*, Rotterdam, Netherlands, March 2009. NAG/DAGA.

[8] Mansoor Hyder, Michael Haun, and Christian Hoene. Placing the participants of a spatial audio conference call. Las Vegas, USA, January 2010.

[9] Mansoor Hyder and Christian Hoene. 3D telephony. ITU-T Workshop 'From Speech to Audio: bandwidth extension, binaural perception', September 2008.

[10] K. Inkpen, R. Hegde, M. Czerwinski, and Z. Zhang. Exploring spatialized audio & video for distributed conversations. pages 95–98. ACM, 2010.

[11] Raine Kajastila, Samuel Siltanen, Peter Lunden, Tapio Lokki, and Lauri 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.

[12] R. Kilgore and M. Chignell. Listening to Unfamiliar Voices in Spatial Audio: Does Visualization of Spatial Position Enhance Voice Identification. *Human Factors in Telecommunication*, 2006.

[13] R. Kilgore, M. Chignell, and P. Smith. Spatialized Audioconferencing: what are the benefits? In *Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research*, page 144. IBM Press, 2003.

[14] R.M. Kilgore. Simple Displays of Talker Location Improve Voice Identification Performance in Multitalker, Spatialized Audio Environments. *Human Factors*, 51(2):224, 2009.

[15] Claudia Raake, Alexander; Schlegel. Auditory assessment of conversational speech quality of traditional and spatialized teleconferences. In *Beiträge der 8. ITG-Fachtagung Sprachkommunikation 2008 (ITG-FB 211)*, Aachen, Germany, October 2008.

[16] I. Rec. P. 800: Methods for subjective determination of transmission quality. *International Telecommunication Union*, 1996.

[17] ITU-T. Rec. P. 50: Appendix 1, Test signals. *International Telecommunication Union*, 1998.

[18] CJ Reynolds, MJ Reed, and PJ Hughes. Decentralized headphone based spatialized audio conferencing for low power devices. In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, pages 778–781. Institute of Electrical and Electronics Engineers Inc., The, 2009.

[19] B. Shinn-cunningham. Learning reverberation: Considerations for spatial auditory displays. In *Proceedings of the ICAD*, pages 126–134, 2000.

[20] L. Vesterinen. Audio Conferencing Enhancements. 2006.

[21] Nicole Yankelovich, Jonathan Kaplan, Joe Provino, Mike Wessler, and Joan Morris 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)*, pages 333–342, New York, NY, USA, 2006. ACM.