Listening and conversational quality of spatial audio conferencing

Alexander Raake[1], Claudia Schlegel[1], Katrin Hoeldtke[1], Matthias Geier[1], and Jens Ahrens[1]

[1]*Deutsche Telekom Laboratories, TU Berlin, Berlin, Germany*

Correspondence should be addressed to Alexander Raake (`alexander.raake@telekom.de`)

ABSTRACT

We present the results of a listening and a conversation test on the quality of spatial and non-spatial audio conferences. To this aim, we have developed conversation test scenarios for audio conferences with three remote participants in order to carry out quality evaluation tests for audio-conferences that are comparable with similar scenarios for traditional one-to-one telephone conversation assessment. We have applied the test scenarios during a conversation test, to (i) validate the test scenarios, (ii) in a realistic usage context measure the advantages of spatial versus non-spatial audio conferencing, and in relation with the quality-impact due to the transmitted speech bandwidth, and (iii) provide recordings of conferences for later use in listening tests. In the conversation test, we have compared different bandwidths (narrowband/NB, 300-3400 Hz; wideband/WB, 50-7000 Hz; fullband/FB, 20-22000 Hz), spatial versus non-spatial headphone-based rendering, and channels with and without talker echo. In a subsequent listening test using recorded conferences, we have attempted to assess the quality of spatial and non-spatial audio-conferencing in a more detailed fashion, including aspects such as speaker identification and memory.

1. INTRODUCTION

Traditional teleconferencing often suffers from issues such as low intelligibility, limited ability of the participants to discern (in particular) unfamiliar interlocutors, to separate different speakers and to communicate over a long time without substantial fatigue. With VoIP, high-quality but low-bitrate codecs and an increasing processing power of user equipment, desktop and mobile conferencing is more and more ready to develop towards spatial audio and virtual speech chat rooms.

The advantage of a spatial auditory display of the interlocutors has been demonstrated in a listening context e.g. by [1, 2]. In [1], a listening test was conducted using fullband pre-recorded four-party conferences of 6 min duration each, presented using different spatial configurations of four loudspeakers: Non-spatial, i.e. play-back via one loudspeaker placed directly in front of the listener, and two spatial configurations, with a (15, 5, -5, -15)-configuration (“collocated”) and a (60, 20, -20, -60)-configuration (“scaled”), respectively. The tests consisted of a number of questionnaires: A memory test, where the participants were asked to indicate who of the four conferees made a particular statement (26 pre-transcribed statements per conference), and rate how sure they were about their choice; a focal assurance questionnaire, where for each conference the listeners had to outline the conferees’ opinions, and indicate their respective certainties; a post-conference questionnaire, including questions on conferee-identification difficulty, overall conference comprehension, the attention required to determine the conferees’ identity, the help due to the additional images, and the assistance due to the spatial location of the conferees. Both the “scaled” and “collocated” configurations showed significantly better performance according to almost all of the collected measures, with the “scaled” spatial configuration typically leading to the best results. The advantage of spatial configurations for focal assurance and speaker identification (recall) was explained with the hypothesis of a shared working memory load in this case: Profiting from using both the visuo-spatial sketch pad responsible for temporal retention of visual and spatial material, and the phonological loop, responsible for retaining verbal material and semantic meaning [1].

In [2], we have studied the performance of a downward-compatible tool for spatial conferencing. Here, narrowband, mono-channel, i.e. down-mixed multiparty conferences were split up into individual tracks corresponding to the different conferees using an automatic speaker classification algorithm, and subsequently rendered using the “SoundScape Renderer” [3]. In the perceptual evaluation, we compared the system output with a non-spatial output, and with the spatial presentation based on an ideally segregated stream, produced from the voice tracks prior to the down-mixing. In the context of this paper, only the “ideal spatial” and “non-spatial” cases are of relevance. Instead of pre-recorded conferences, we employed sequences of numbers read out by a given speaker, which were combined to longer sequences, thus containing several short passages of numbers from different speakers. One independent variable in the tests was the number of speakers per test-sequence, using 2, 3 and 4 speakers.

The task of the subjects was to identify the active speaker and speaker change points using a graphical user interface, with the goal of quantifying the number of correct identifications, substitutions and deletions. Further, the subjects were asked to provide judgments of the pleasantness of the audio reproduction (slider with “pleasant” and “unpleasant” at the extreme points), and the task difficulty (slider with “difficult” and “easy”). The results show that the number of change-point detection errors increases with the number of speakers N, with considerably lower rates for the spatial compared to the non-spatial case for each value of N. For $N = 2$ speakers, the audio representation has no impact on the anyways low number of errors; advantages due to spatial presentation can be observed for $N = 3$ and $N = 4$ speakers. Perceived task difficulty strongly increases with N, while the pleasantness ratings only slightly decrease with N. For the difficulty ratings, the advantage due to spatial presentation lies in the same relative scale-range as the impact due to increasing N; in turn, while pleasantness clearly declines between $N = 2$ and $N = 3$ speakers for the non-spatial case, for the spatial case it starts declining only when increasing N from 3 to 4.

The conversational situation has been studied to a far lesser extent, and related studies mainly focused on situations with two groups of conferees located at two remote locations. However, in a real-life context, many conferencing situations are characterized by the fact that none or only some of the individual conferees are spatially collocated. In this case, especially when using a uni-channel, narrowband conferencing system, the cognitive load for all or some of the participants may be high.

To assess the conversational speech quality of telephony services, it is necessary to involve the conversation partners in an appropriate conversation task using predefined conversation test scenarios. For classical two-person conversations, different types of conversation scenarios have been described in the literature (see [4] for a summary). The main shortcoming of many of these scenarios is that they reduce the naturalness of the assessment situation. Similarly, some of the existing multiparty communication scenarios represent unnatural tasks, and others employ free conversations about pre-defined topics [1, 5] that cannot easily be compared with each other.

In order to assess conversational speech quality of teleconferencing in a realistic fashion and similar to the case of two-party telephony, we have developed two sets of 12 three-person conversation scenarios: A set of 12 scenarios representing business-type teleconferences, and another set of 12 scenarios representing family or spare-time conversations. These three-user conversation test scenarios (3CTs) were developed to study the perceived quality of different teleconferencing configurations. We have used these scenarios in a first conversation test for validation and to measure the quality-advantages of different technical conferencing characteristics: (1) Bandwidth — Fullband (0.02-22.1 kHz), wideband (0.05-7 kHz) and narrowband (0.3-3.4 kHz); (2) Reproduction — Spatial versus non-spatial; (3) Talker echo — effective in some of the narrowband and fullband conditions, with and without spatial rendering.

Here, it is particularly interesting to see whether wideband and spatial rendering actually lead to a higher preference in a conversational situation. The test results give first indications on how well the conversational conferencing quality judged by the users reflects the quality advantage found for wideband over narrowband in case of normal telephone dialogues [4, 6], and the listening advantages found for spatial over non-spatial audio conferencing [1].

In a subsequent listening test, we have extended the assessment and have included two ways of memory-performance assessment similar to [1]. For recording the conferences, the conversation test scenarios have been used with three male speakers carrying out conversations over clean transmission chains.

The paper is outlined as follows: Section 2 describes the set-up of the conversation scenarios, the conversation test, and the test results, Section 3 summarizes the listening tests, and Section 4 concludes with a discussion and outlook on future work.

2. CONVERSATION TESTS

2.1. Test Scenarios (3CTs)

The main advantage of conversation tests over listening tests is that they reflect the actual application of telephone or conferencing services in a more natural way (other advantages of conversation over listening tests are summarized e.g. in [7]). Their main limitation is that they are time-consuming and often involve test scenarios that do not represent telephone-typical conversations. In order to reduce some of the drawbacks of (dialogue-type) conversation tests, the SCTs (Short Conversation Test scenarios) developed by Möller [7] represent real-life telephone scenarios like ordering a pizza or reserving a plane ticket. They lead to natural but semi-structured, comparable and balanced conversations of approximately 2 to 3 minutes duration.

In recent work on multiparty conferencing assessment, free conversations on pre-defined (typically controversial) topics have been employed [1, 5]. In order to bridge the gap between the SCTs typically used in a two-party speech communication context and the multiparty conferencing assessment, we have developed a new set of conversation test scenarios for three interlocutors. Details of the test scenario development can be found in [8].

The layout of the scenarios loosely follows that of the two-person SCTs [7]. In the case of the 3CTs, each scenario is captured by two sheets of paper per interlocutor. The first sheet is identical for all participants, and briefly outlines the overall situation in which the conversation takes place, the actual topics to be discussed, and the roles and names of the participants. The second sheet is individual for the three interlocutors, and comprises a mix of pictograms that indicate the type and function of the information to follow, short instructions, and tabulated data. The participants have complementary information which are necessary to complete the conversation task. Example topics for the business scenarios are the planning of a meeting, selection of titles for a new music CD compilation, and the organization of an arts exhibition. Note that we have focused on the business scenarios in the remainder of this paper.

2.2. Test Conditions and Procedure

The conditions used in the conversation test are summarized in Table 1. Here, spatial presentation means a dichotic presentation of the two other participants' voices using static (i.e. non-headtracked) binaural room impulse responses (BRIRs) at ±30° azimuth. The BRIRs were recorded in an acoustically treated studio environment.

#	bandwidth	*TELR*	*T*	presentation
		[dB]	[ms]	
1	NB	65	0	diotic
2	NB	65	0	spatial
3	WB	65	0	diotic
4	WB	65	0	spatial
5	FB	65	0	diotic
6	FB	65	0	spatial
7	NB	35	100	diotic
8	NB	35	100	spatial
9	FB	35	100	diotic
10	FB	35	100	spatial

Table 1: Test conditions used in the conferencing test. *TELR* ≡ Talker Echo Loudness Rating, i.e. echo attenuation. *T* ≡ mean one-way delay (that is half the actual echo-delay).

Eight groups of three interlocutors took part in the test, yielding 24 judgements for each of the 10 test conditions. After each conversation, the subjects were asked to provide a quality rating on a 7-point continuous, absolute rating scale with the typical Absolute Category Rating scale labels [9, 10]. After the quality ratings subjects were asked for ratings of conversation effort on the so-called CR10-scale, a Category Ratio scale according to [11].

A 10x10 Greco-Latin Square design was used to ensure that each channel condition is combined with each test scenario only throughout the entire test. Since eight (and not ten) groups of three interlocutors took part in the test, eight out of the ten possible scenario–condition lists were employed.

For each group, the test was split into two sessions, in order to avoid subject fatigue. The first of the two test sessions was preceded by an initiation phase to familiarize the subjects with the test equipment and conditions. During this phase, the participants were asked to take a role in a section from Goethe's Faust and read it aloud in alternating turns. For each complete turn of the three subjects, one of six of the ten test conditions were used as the conferencing setting, to demonstrate the type of connections and quality range. In a second part of the initiation, the subjects carried out a training conference using one of the twelve scenarios (the same for all groups).

At the end of each test run, the subjects were asked to fill in a questionnaire, with questions e.g. on their experience with telecommunication services and in particular the use of audio conferences.

2.3. System Set-up

The conferencing system was implemented using a Linux-based audio server, with interconnections based on JACK audio [12], and BruteFIR as the convolution engine for the static BRIRs [13]. The system is a modified part of an earlier version of the SoundScape Renderer (SSR) [3], which was used for the listening test described in Section 3. The three participants of each run were seated in three independent and acoustically treated rooms available in the Usability Lab of T-Labs, conforming to [9]. High-quality open headsets of the type Sennheiser HMD 410-6 were used for sound playback.

For all conversations, the microphone signals of the interlocutors were recorded via the audio server using three independent audio tracks, one per speaker. These recordings had two goals: (1) Characterization of the conversation structure in terms of turns, utterance frequency and durations, overall scenario duration, etc. (2) Generation of a database for subsequent listening tests with a more analytical focus on aspects like memory and speaker recognition.

During the entire test, the call set-up was carried out by a test-supervisor. With the launch of a given test condition as well as with its termination, a sound sample was played out to the subjects indicating the call set-up and ending. The launch and termination of each call automatically started and ended the recordings.

2.4. Test Subjects

24 subjects participated in the conversation test. They were recruited from the employee's body of Deutsche Telekom Laboratories, and can all be considered as naive with respect to this type of tests, and as non-experts with regard to the employed conferencing technology. They were between 25 and 59 years old (mean 34.4 years), with 12 subjects female, 12 subjects male. 8 of the subjects had two conversation partners of equal sex (making the differentiation between them harder), and 16 subjects had two conversation partners of opposite sex. The average of the weekly usage time of conferencing services stated by the test 1 participants in the post-test questionnaire was 1.57 hours, and all subjects indicated some usage per week. Hence, the subjects can be considered as frequent users of teleconferencing systems. In the post-test questionnaire, 75% of the test subjects answered that they considered the scenarios to be reflecting real-life conferences.

All subjects were to their own account normal hearing. Subjects were tested for basic binaural hearing capability by presenting them a list of 20 numbers uttered by a male speaker, with a random presentation to the left or right ear. Subjects had to indicate from which side the heard each number. None of the subjects was excluded based on this test, with the criterion for exclusion being a threshold of more than 2 wrong answers.

2.5. Results

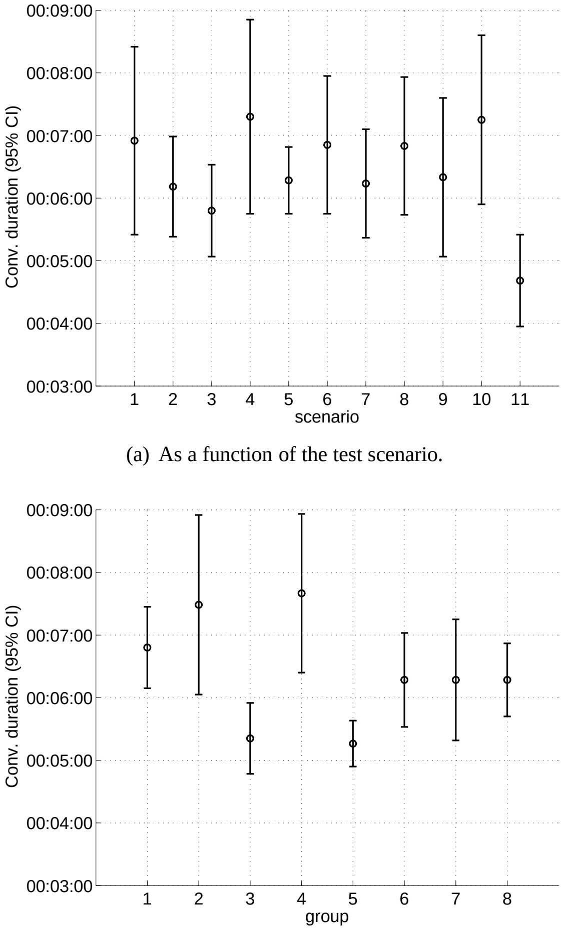
In the following, the test results are discussed from two perspectives: (1) The conversation recordings are analyzed instrumentally in terms of the conference durations and of additional parameters describing the conversational structures, in order to assess the variability induced by the different conversation scenarios, the different conferee groups, and the test conditions; (2) the actual conversation test results are analyzed.

2.5.1. Conversation Duration and Conversation Behavior

The recordings of the conversations have been evaluated for overall duration per scenario and per conferee group. Figure 1(b) (top) shows the conversation durations for the eleven different scenarios used in the test (means over groups and 95% Confidence Intervals, CIs).

As can be seen from the Figure, the average durations range between 5:50 to 7:20 minutes. One exception is scenario #11, which was the training scenario (planning of a meeting). This scenario is the only one that is significantly different from all others, with a mean duration of 4:41 min and a much smaller CI. A two-factorial ANalysis Of VAriance (ANOVA) was carried out using the group and the scenario as fixed factors. Both the scenario and the group were found to be statistically significant factors for conversation duration, with the group showing a larger impact (scenario: $F = 2.616,\ p < 0.05$; group: $F = 5.130,\ p < 0.005$).

A corresponding plot of the conversation durations as a function of the subject group is shown in Figure 1(b) (bottom). No statistically significant effect of the test



(a) As a function of the test scenario.

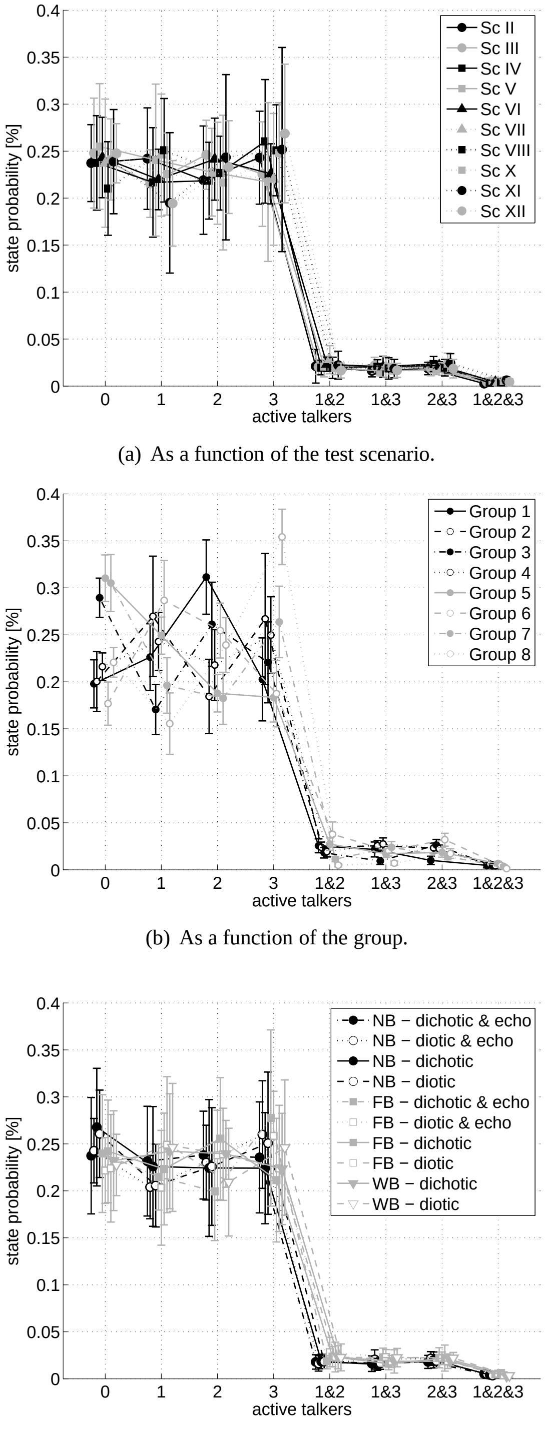
(b) As a function of the group.

Fig. 1: Conversation durations.

condition on the conversation duration could be observed (one-factorial ANOVA with condition as fixed factor).

In summary, it can be said that the subject group has a higher impact on the conversation duration than the particular scenario does. The very similar conversation durations for the 10 actual test scenarios indicate a good match with the scenario design goal. The mean duration is 6:25 min.

In addition to the conversation durations, we have analyzed the influence of the group, the scenarios and the test conditions on the conversation behaviour of the test subjects. To this aim, we have analyzed the recorded three-channel conferences according to an eight state Markov model, with three states representing single talk



(a) As a function of the test scenario.

(b) As a function of the group.

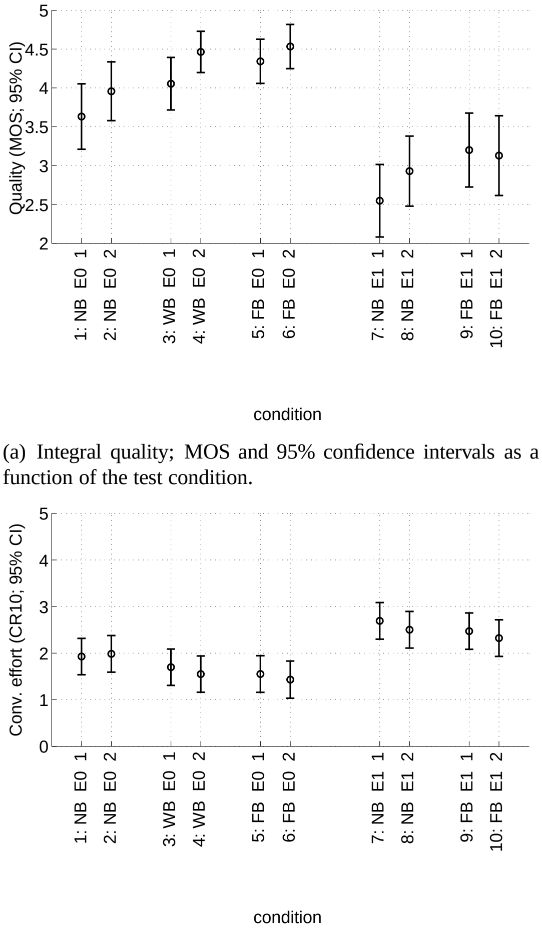
(c) As a function of the condition.

Fig. 2: Conversation state probabilities.

(talker 1, 2 or 3 talks), three states for the possible cases of double talk (1 & 2, 1 & 3, or 2 & 3 speak at the same time), and one state each for the cases mutual silence and “triple-talk” (see [14, 15] for a foundational analysis of two-party telephone conversations). The following steps were taken during the analysis: (1) We have down-sampled the recordings from 44.1 to 16kHz, (2) applied a 2-2.5 kHz bandpass-filter to exclude breathing noise captured by the headsets for some of the conferees, (3) applied a simple energy-related voice activity detection on the resulting signal, (4) omitted all talkspurts with less than 15 ms duration (see [15]), and (5) filled in all pauses during the active period of a given talker that were shorter than 200 ms (see [15]). The resulting speech-contours were used to calculate state probabilities. In Figure 2, the state probabilities are shown in terms of the impact due to the scenarios 2(a), due to the respective user group 2(b), and due to the test condition 2(c). It is shown that the state probabilities are approximately independent of the scenario and condition, while they strongly depend on the different user groups, indicating that some conferees are more active than others.

2.5.2. Quality and Conversation Effort

Based on a visual inspection of the quality ratings over conditions given by individual subjects, subjects showing no variation between conditions or high quality ratings for the echo conditions have been excluded from the final data analysis (2 out of 24 subjects). The resulting mean quality and conversation effort ratings are depicted in Figure 3.



(a) Integral quality; MOS and 95% confidence intervals as a function of the test condition.

(b) Conversation effort; mean and 95% confidence intervals as a function of the test condition.

Fig. 3: Test ratings. The x-axis-labels have the form ’N: XX YY P’, with: N≡ condition number as in Table 1; XX≡bandwidth; YY≡E0 for no talker echo, and YY≡E1 in case of talker echo; P≡1 for diotic, and P≡2 for dichotic (spatial) presentation.

The quality (“MOS”) and conversation effort ratings (“CR10”) have been positively tested for normal distribution per condition using a Kolmogorov-Smirnov test, and visual inspection based on Q-Q plots. Both the quality and conversation effort ratings were analyzed using a repeated-measures mixed linear models ANOVA [16] with the test condition as fixed factor. Condition proves to be a highly significant factor for quality ($F = 14.291, p < 0.001$) and for conversation effort ($F = 7.948, p < 0.001$). A subsequent marginal means analysis using a Bonferroni-adjustment of the confidence intervals to compensate for multiple comparisons revealed that 18 of the $10 \cdot (10-1)/2 = 45$ possible condition pairs are statistically significantly different from each other in terms of quality, and only 12 out of the 45 condition pairs in terms of conversation effort[1]. For

[1] When using a more conservative univariate general linear model

the conversation effort ratings, all of the 12 pairs contain one condition with echo disturbance and one condition without echo disturbance: The only discrimination possible from the CR10-ratings is that between echo and non-echo conditions.

Since the effect of spatial separation is particularly useful when conversing with two interlocutors of equal sex and thus similar voice characteristics (see e.g. [17]), we have analyzed the conversation results by comparing the quality ratings of subjects with conversation partners of equal sex.

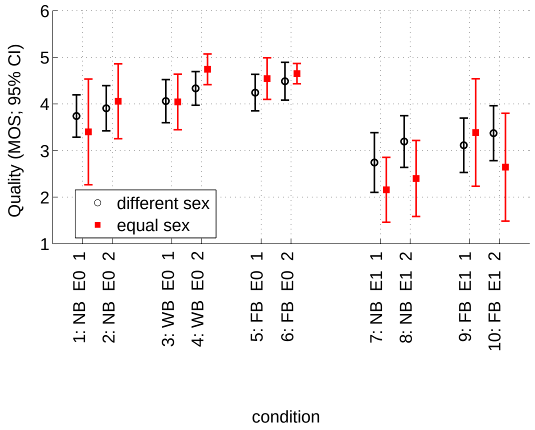


Fig. 4: Quality ratings depending on whether the two conversation partners of a subject were of equal or different sex: Mean and 95% confidence intervals as a function of the test condition. The x-axis-labels have the form 'N: XX YY P', with: N≡ condition number as in Table 1; XX≡bandwidth; YY≡E0 for no talker echo, and YY≡E1 in case of talker echo; P≡1 for diotic, and P≡2 for dichotic (spatial) presentation.

As can be seen from the results, the dynamic of the ratings of subjects with equal-sex-interlocutors is higher than that for subjects with conversation partners of opposite sex. Obviously, the good conditions (spatial, higher bandwidth) are more appreciated and the bad conditions (low bandwidth, echo, non-spatial) less well perceived than in the case of different-sex conversation partners. Note that due to the limited number of subjects with conversation partners of equal sex, the confidence intervals are larger in this case.

In a repeated-measures mixed linear models ANOVA without grouping and considering only the echo-free conditions, the fixed factors bandwidth (three levels) and reproduction (two levels) were found to both be significant (bandwidth: $F = 7.488, p < 0.005$; reproduction: $F = 5.555, p < 0.05$). A marginal means analysis with Bonferroni correction of the confidence intervals showed that narrowband can clearly be distinguished from wideband and fullband, but not wideband from fullband. The distinction between spatial and diotic presentation is statistically significant. The similar F-values for reproduction type (diotic versus spatial) and bandwidth imply that they are more or less equally important for quality. When comparing only the narrowband and fullband case, the repeated-measures mixed linear models ANOVA yields a statistically significant effect of bandwidth ($F = 14.898, p < 0.001$), but not of the employed reproduction ($F = 2.392, p = 0.126$). When comparing narrowband with wideband, both bandwidth and reproduction are found to be significant effects, with a slightly higher impact due to bandwidth (bandwidth: $F = 7.888, p < 0.01$; reproduction: $F = 5.336, p < 0.05$). When using a univariate general linear model analysis with bandwidth and reproduction as fixed, and the test subject as random factor, we directly observe a significant effect of bandwidth ($F = 14.831, p < 0.001$), and only a close-to significant effect of reproduction ($F = 3.597, p = 0.059$).

3. LISTENING TEST

To increase the sensitivity of the test method, we have conducted a listening-only test using pre-recorded conferences. In earlier studies, it was shown that conversation tests are more realistic, but also less critical than listening tests [18]. We tried to achieve a further increase of sensitivity by: (1) Employing a fixed group of interlocutors for generating recordings to be used in the listening test, and hence reduce the variation in conversation style and duration. (2) Employing both conditions with and without head-tracking to investigate whether more accurate spatial cues increase recognizability of speakers and yield better memorization of utterances. Background: No head-tracking was used in our conversation tests, and thus the set-up was comparable with that employed by [19]. Instead, in [1] and [5] real loudspeakers were used. Hence, in these cases head movements of the test subjects automatically translate into dynamic spatial cues. (3) Using echo-free conditions in order to avoid a potential compression of the rating scale.

analysis with condition as fixed and the test subject as random factor, a subsequent Bonferroni Post-Hoc analysis yields 21 significantly different condition pairs in terms of quality.

3.1. Scenario Recordings

An informal listening to the recordings from the conversation test revealed that there were sometimes problems related with breathing noise captured by the microphone, which is especially audible in case of fullband. Also, as discussed earlier, there are dependencies of conference duration and conversation style on user group and scenario. Consequently, we could not identify a set of satisfactory recordings that was covering all scenarios stemming from one group, or all scenarios with each stemming from a different group.

We thus decided to re-record the business conferences with one group of users recruited from our Lab. We used three male users in order to yield similar voice characteristics. They were all three experienced and frequent audio conferencing users. The set-up was the same as during the conversation test, with the following differences: Each interlocutor briefly introduced himself prior to each conference, indicating his name, affiliation and function. No degradations were used during the conversation recordings, and a static spatial presentation was employed. The conversation partners were instructed to talk as if they were carrying out an actual conference call.

The recordings were transcribed and annotated with regard to individual utterances to be later used in a memory test.

3.2. Test Conditions and Procedure

The conditions used in the listening test are summarized in Table 2. Besides non-spatial, diotic presentation, two variants of spatial presentation of the three voices were employed: Static (i.e. non-headtracked) or dynamic binaural synthesis, in both cases based on the same BRIRs as used in the conversation test. For the spatial case, the relative positions of the speakers were chosen at 0,° and ±30 ° azimuth.

Directly after each trial, the subjects were asked to judge the integral quality of the conversation they listened to on the previously used 7-point continuous scale (*MOS*).

In a first recall phase, the subjects were then asked to write down statements and arguments the three conversation partners had made during the conversation (free recall, resulting in the mean number of correctly recalled items per condition $FREm$).

In a second recall phase, the subjects were asked to indicate which of the talkers had uttered a certain statement, with 4 options (“A”, “B”, “C”, “don’t know”). 24 statements of this type were provided to the subjects on paper for each recorded conference. The answers result in mean correct, incorrect and not assigned statements per condition and subject (*CORm*, *FALm*, *NASm*).

#	bandwidth	presentation	head-tracking
1	NB	diotic	-
2	WB	diotic	-
3	FB	diotic	-
4	NB	spatial	-
5	WB	spatial	-
6	FB	spatial	-
7	NB	spatial	yes
8	WB	spatial	yes
9	FB	spatial	yes

Table 2: Listening test conditions.

After the two recall phases, the subjects were asked to judge their ability to recognize the interlocutors (*REC*), the intelligibility during the conference (INT), the attention required to recognize the conversation partners (ATT), and the usefulness of the spatial presentation (*USP*). These judgments were placed after the recall phase in order not to reduce the recall-performance. In turn, this has the disadvantage that the recall-test may influence the subsequent judgments.

A 9x9 Greco-Latin Square design was used, to ensure that each channel condition is combined with each test scenario at most three times over the 24 subjects participating in the test, with three different Greco-Latin squares there are $3 \cdot 9 = 27$ available playlists. Since 24 (and not 27) subjects took part in the test, three possible presentation orders of the third Greco-Latin Square were omitted.

As a first condition, subjects listened to a training condition (#9: FB, spatial, with head-tracking). The $9+1=10$ test runs per subject were seperated into two sessions held on two different days in order to avoid subject fatigue. Both test sessions were preceded by an initiation phase to familiarize the subjects with the test equipment and conditions. During this phase, the participants could listen to a continuously played conference and change between 5 of the 9 test conditions at will, so that spatial audio and the employed bandwidths could be listened and get used to. The training condition (#9) was presented as the first condition after the initiation in the first session only.

Similarly to the conversation test, at the end of each completed test run the subjects were asked to fill in a questionnaire, with questions e.g. on their experience with telecommunication services and in particular the use of audio conferences.

Note that prior to the main test, we conducted an informal pre-test in order to test the employed paradigm of asking speaker-identification/recall questions. In this test, 4 subjects recruited from the T-Labs staff were asked to listen to three conferences processed by conditions #7-9, and were asked to undergo the normal test described above. The results were promising so that we continued with the main test.

3.3. System Set-up

The conferencing system was implemented using the SoundScape Renderer (SSR) [3]. The same high-quality open headsets as in the conversation test were used for sound playback (Sennheiser HMD 410-6). They were equipped with a Polhemus FASTRAK sensor for providing head-tracking information.

The test set-up was fully automatic and playlist-based, with one list per subject and session. For each conference, the three-channel audio file and condition information was specified in the list. All scales and questionnaires were provided on paper.

3.4. Test Subjects

24 subjects participated in the listening test. The paid, naive subjects were mainly recruited from the university campus of TU Berlin. They were between 21 and 41 years old (mean 26.6 years; 13 female, 11 male). The average of the weekly participation in audio conferences indicated by the subjects in the post-test questionnaire was 0.17 hours (with 19 subjects indicating no participation in audio conferences at all per week), so that the user group can be considered as very unexperienced with audio conferences. This is a clear difference to the conversation test subjects, but was intentional due to the assumed higher sensitivity of the listening test, and its intended goal of assessing the advantages of spatial audio for naive subjects unexperienced with conferencing.

All subjects were to their own account normal hearing. One third of the subjects had taken part in an audiometric screening test some months earlier in the course of a different speech quality test, and at the time were normal hearing. As in case of the conversation test, subjects were tested for basic binaural hearing capability by presenting them a list of 20 numbers uttered by a male speaker, with a random presentation to the left or right ear. Subjects had to indicate the active ear. One subject was excluded based on this test, with the criterion for exclusion being a threshold of more than 2 wrong answers.

3.5. Results

At first, we have evaluated the correspondence of the ratings obtained from each subject per conversation with the average across all subjects, to validate the subject performance. Here, we found 3 subjects with a substantial root mean squared deviation from the general mean (RMSD), who have been excluded from the subsequent analysis. After removal of the respective subjects, all ratings were analyzed for normal distribution per condition using the Kolmogorov-Smirnov test. The different ratings were normally distributed for most of the conditions, and the numbers of non-normal conditions out of the 9 tested ones are given in brackets in the following list: Quality/MOS (0), speaker recognition/REC (0), intelligibility/INT (1), required attention/ATT (3), and usefulness of spatial reproduction/USP (4).

A repeated-measures linear mixed models ANOVA with condition as fixed factor revealed a significant effect due to condition for all ratings. The results are given in Table 3, as well as the number of condition-pairs that can be differentiated based on a subsequent marginal-means analysis including Bonferroni adjustment of the CIs.

The results indicate that all ratings permit the distinction of at least 8 of all 36 condition-pairs, and that the speaker recognition rating REC and usefulness rating USP appear to be most discriminative. The quality ratings MOS are less discriminative than expected. Here, it has to be noted that for some conditions the USP-ratings are not normally distributed so that the results need to be considered with some caution.

In order to investigate the impact of bandwidth versus reproduction, and the additional use of head-tracking, we have carried out a series of repeated-measures linear mixed models ANOVAs for all of the 5 ratings, with the test subject as repetition and the bandwidth (NB, WB, FB), reproduction (non-spatial, spatial), and head-tracking (yes, no) as fixed factors. Note that in this analysis head-tracking was not found to be significant for any rating.

For the quality ratings MOS, both bandwidth and reproduction were significant factors (bandwidth: $F = 8.468$, $p < 0.001$; reproduction: $F = 30.426$, $p <$

	MOS	REC	INT	ATT	USP
F	9.699	11.155	8.358	7.371	23.911
$p <$	0.001	0.001	0.001	0.001	0.001
N_d	10	16	8	11	18

Table 3: Results of repeated-measures mixed models ANOVA for test subject as repetition and condition as fixed factor (F-value and significance level p), and number N_d of condition-pairs out of $9 \cdot 8/2 = 36$ possible combinations that could be distinguished based on a subsequent marginal means analysis for the different ratings.

0.001). Based on a subsequent marginal-means analysis for *MOS*, NB conditions could be separated from WB and FB, but not WB from FB, and the two reproduction types could clearly be distinguished. For the speaker recognition ratings *REC*, only the reproduction was found to be a significant factor, while bandwidth was not significant (reproduction: $F = 58.627,\ p < 0.001$). In case of the intelligibility ratings INT, again both bandwidth and reproduction were significant factors (bandwidth: $F = 6.955,\ p < 0.005$, again without discrimination between WB and FB; reproduction: $F = 40.975,\ p < 0.001$). For the required attention ATT and usefulness of spatial presentation rating USP, only reproduction was significant (ATT: $F = 36.876,\ p < 0.001$; USP: $F = 161.031,\ p < 0.001$).

When using the same analysis only for the cases of spatial reproduction, with bandwidth and tracking as fixed factors, we find a significant effect of bandwidth for *MOS*, and INT, and a close-to significant effect due to head-tracking only for the usefulness of spatial distribution of speakers USP.

The discrimination power of the two recall-phases is extremely low: The mean numbers of correctly remembered topics per conversation $FREm$, i.e. for the free recall test, range from 8.75 (#1: NB, non-spatial) to 11 (#6: FB, spatial, no head-tracking). The number of correctly recalled items *CORm* (out of 24) ranges from 14.7 (#1: NB, non-spatial) to 18.95 (#8: WB, spatial, head-tracking).

When applying a repeated-measures linear mixed models ANOVA to the number of correctly recalled items *CORm*, with the condition as fixed factor, condition is found to be significant ($F = 2.749, p < 0.05$). A univariate general linear model analysis with the subject as random and the condition as fixed factor indicates that both are equally decisive for *CORm* (condition: $F = 3.989, p < 0.001$; subject: $F = 4.083, p < 0.001$). The behavior of other measures *FALm*, *NASm*, and $FREm$ is very similar, and does not enable the substantial differentiation power shown e.g. by [1].

4. CONCLUSION AND OUTLOOK

We have presented a conversation test that clearly highlights the usefulness of an extended audio bandwidth and spatial reproduction in an actual audio conferencing context: The participants are able to notice and appreciate the advantage in terms of bandwidth and spatial presentation in spite of the substantial distraction due to the conversation task.

Obviously, no advantage due to head-tracking can be observed from the ratings collected in our listening test. An interesting observation can be made from the comparison between the listening and conversation test results, when it comes to the quality impact of bandwidth and spatial versus non-spatial reproduction: In the conversation test, the bandwidth was equally important as or more important than the reproduction, while in the listening test, reproduction was clearly more important for all of the collected ratings. A possible reason for this effect may lie in the higher engagement when participating in an actual conversation, where the bandwidth may be more noticeable and beneficial.

However, the main reason for this observation is thought to be the number of talkers a given subject is faced with: In the conversation test, each subject has 2 interlocutors, and in the listening test, each subject listens to 3 talkers. Consequently, the spatial separation becomes increasingly useful. In [2] we had shown that in an NB context the perceptual advantage due to spatial separation scales considerably with the number of talkers to be distinguished (speaker identification errors, task difficulty, pleasantness, see Section 1). The scaling with the number of interlocutors appears to be an important aspect also for the free and guided recall assessments undertaken as part of the listening test: Contrary to our expectation and findings e.g. by Baldis [1], the number of correctly remembered items was quite independent of the bandwidth and presentation type. In the listening test of [1], 4 talkers were used, and in our listening test only 3. Interestingly, in [1], the maximum percentage of correctly recalled items is 58.7% (FB, spatial), and the minimum 38.3% (FB, non-spatial), while the maximum observed in our test was 79.0% (WB, spatial, head-tracking), and the minimum 61.3% (NB, non-spatial), with comparable overall numbers of items (26 in [1], 24 in our test). The minimum value found in our test, which is comparable with the maximum value from [1], indicates that even for the worst condition our memory assessment may not have been as demanding as in a 4-talker case.

A clear weak-point of our listening test is that we have used subjects with little to no experience with audio conferences. This limits the comparability with the conversation test. In future work, we will conduct an additional listening test with experienced conferencing users. Here, it is planned to explicitly study the effect of scaling of the results with the number of conferees. In general, we aim for an assessment that more fully covers the range from more classical and wide-spread low-quality conferences (based on conference bridges with down-mixed single-channel transmission) to high-quality and spatial-audio conferencing. To make the work practically useful, the results will be fed into the new study activity on conferencing and tele-meeting assessment recently launched by ITU-T Study Group 12.

5. REFERENCES

[1] Jessica J. Baldis. Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In *CHI*, pages 166–173, 2001.

[2] Alexander Raake, Sascha Spors, Jens Ahrens, and Jitendra Ajmera. Concept and evaluation of a downward-compatible system for spatial teleconferencing using automatic speaker clustering. *In: Proc. 10th Int. Conf. on Spoken Language Processing (Interspeech 2007 – ICSLP), BE-Antwerp*, 2007.

[3] Matthias Geier, Jens Ahrens, and Sascha Spors. The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods. *In: Pro. 124th AES Convention*, May 17 - 20, NL–Amsterdam, 2008.

[4] Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. John Wiley & Sons Ltd, Chichester, West Sussex, UK, 2006.

[5] Kori Inkpen, Rajesh Hegde, Mary Czerwinski, and Zhengyou Zhang. Exploring spatialized audio & video for distributed conversations. In *Proc. 2010 ACM Conference on Computer supported cooperative work (CSCW)*, pages 95–98, 2010.

[6] Sebastian Möller, Alexander Raake, Nobuhiko Kitawaki, Akira Takahashi, and Marcel Wältermann. Impairment factor framework for wideband speech codecs. *IEEE Trans. Audio Speech and Language*, 14(6):1969–1976, 2006.

[7] Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic Publishers, USA–Boston, 2000.

[8] Alexander Raake and Claudia Schlegel. Auditory assessment of conversational speech quality of traditional and spatialized teleconferences. *In: Proc. 8th ITG Conference Speech Communication, to appear, DE-Aachen*, 2008.

[9] ITU-T Rec. P.800. *Methods for Subjective Determination of Transmission Quality*. International Telecommunication Union, CH–Geneva, June 1996.

[10] Markus Bodden and Ute Jekosch. *Entwicklung und Durchführung von Tests mit Versuchspersonen zur Verifizierung von Modellen zur Berechnung der Sprachübertragungsqualität*. Final report to a project funded by Deutsche Telekom AG (unpublished), Institut für Kommunikationsakustik, Ruhr–Universität, D–Bochum, 1996.

[11] Gunnar Borg. A category rating scale with ratio properties for intermodal and interindividual comparisons. *In: Psychophysical Judgement and the Process of Perception (H.-G. Geissler and P. Petzold, eds.)*, pages 25–34, VEB Deutscher Verlag der Wissenschaften, D–Berlin, 1982.

[12] Paul Davis. http://jackaudio.org/, 2007.

[13] Anders Torger. Luleå Academic Computer Society, http://www.ludd.luth.se/ torger/brutefir.html, 2006.

[14] P.T. Brady. A technique for investigating on-off patterns of speech. *Bell System Technical Journal*, 44(1):1–22, 1965.

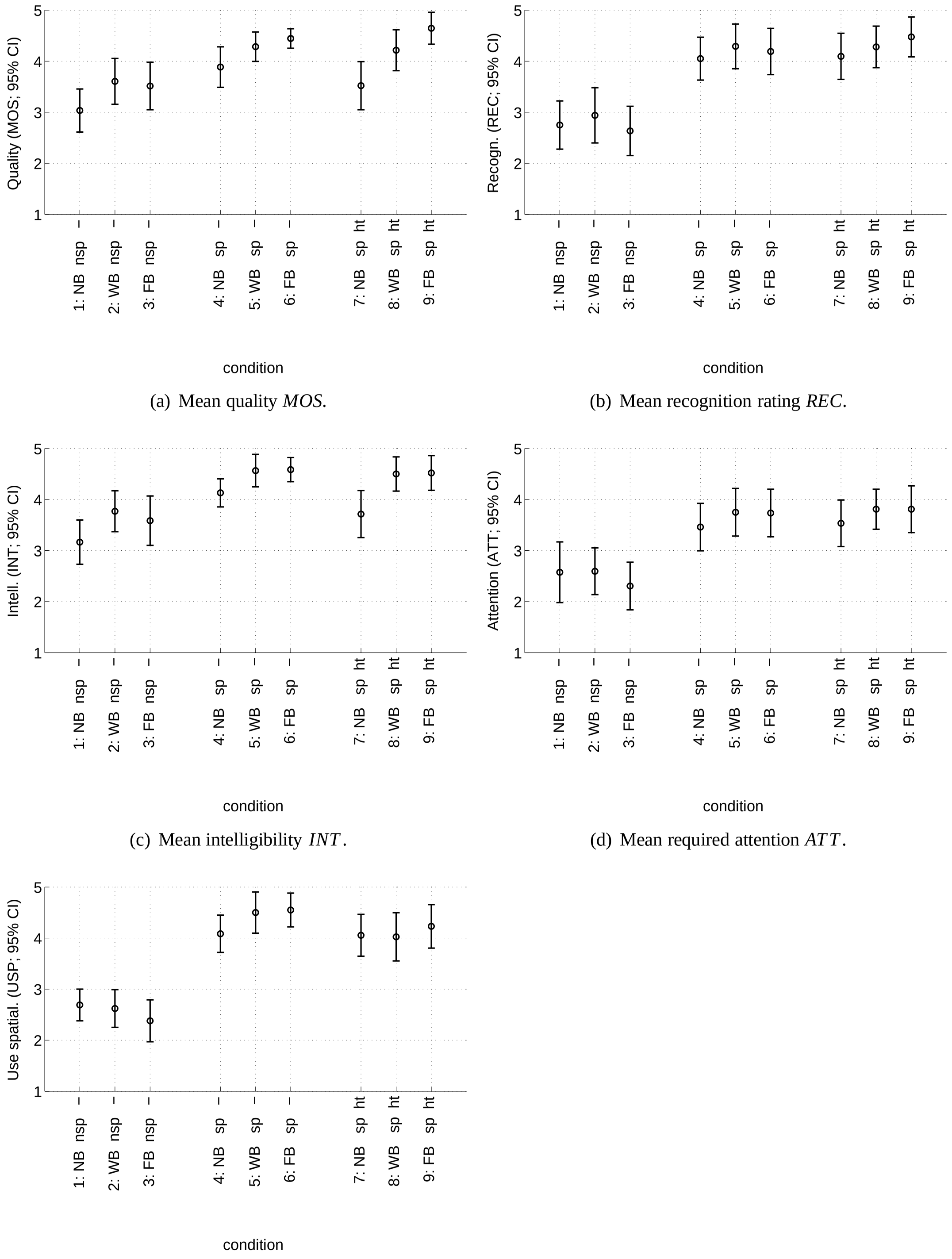
[15] P.T. Brady. A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal*, 47(1):73–91, 1968.

[16] Hugo Quené and Huub van den Bergh. On multilevel modeling of data from repeated measures designs: A tutorial. *Speech Communication*, 43(1-2):103–121, 2004.

[17] C. J. Darwin and R. W. Hukin. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J. Acoust. Soc. Am.*, 107(2):5, 2000.

[18] Laetitia Gros and Noel Chateau. The impact of listening and conversational situations on speech perceived quality for time-varying impairments. *In: Proc. MESAQIN 2002 (J. Holub and R. Smid, eds.)*, pages 17–19, Czech Technical University, CZ–Prague, 2002.

[19] Ryan Kilgore, Mark Chignell, and Paul Smith. Spatialized audioconferencing: What are the benefits? In *Proc. Centre for Advanced Studies Conf. on Collaborative Research*, pages 135–144, 2003.



(a) Mean quality *MOS*.

(b) Mean recognition rating *REC*.

(c) Mean intelligibility *INT*.

(d) Mean required attention *ATT*.

(e) Mean usefulness of spatial reprod. *USP*.

Fig. 5: Test ratings over conditions and 95% CIs. The x-axis-labels have the form ’N: XX PPP HH’, with: N≡ condition number as in Table 1; XX≡bandwidth; PPP≡nsp for non-spatial, and PPP≡ sp for spatial presentation; HH≡ - for no head-tracking, and HH≡ht for head-tracking.