Speech recordings for systematic assessment of multi-party conferencings

Janto Skowronek, Alexander Raake, Katrin Hoeldtke, Matthias Geier
Deutsche Telekom Laboratories, Technische Universität Berlin, Berlin, Germany.

Summary
Towards an assessment method for multi-party audio conferencing systems, speech recordings have been made. The recordings were specifically designed to provide an inter-comparability between conversations with different numbers of interlocutors. This paper describes preparation, implementation, and validation of the recordings and closes with respective conclusions.

PACS no. 43.72.Kb, 43.71.Gv

1. Introduction

Users of current audio conferencing systems often report general dissatisfaction with the system and the experienced fatigue. While an integral approach to assess the Quality of Experience (QoE) of conferencing systems is still needed, several aspects have been addressed in the literature already. The user's ability to separate speakers in a conferencing situation has been investigated in the context of spatial audio reproduction [1, 2]. Others addressed the high cognitive load required in conferencing by comparing communicative aspects between face-to-face conversations and conversations via collaborative remote working systems [3, 4]. Furthermore, much research has been done in the International Telecommunications Union on the QoE assessment of speech communication systems. However, those assessment methods are designed for one-to-one conversations; there is no agreed method for the multi-party conferencing case available yet.

Towards such an assessment method, this paper presents an approach to obtain recordings of multi-party conference conversations that can serve as test stimuli. First, we developed scenarios that provide the content for the recordings (Sec. 2). Emphasis here was to obtain scenarios that remain comparable even if the numbers of speakers would be changed. Then, we prepared and conducted a recording session (Sec. 3) in which experienced speakers had conversations accordingly to the scenarios via a simulated conferencing system. After post-processing the recordings (Sec. 4), we analyzed them to validate the intended inter-comparability of the actual conversations (Sec. 5) and to draw some conclusions (Sec. 6).

2. Conferencing Scenarios

In QoE assessment of telecommunication systems, two test paradigms are typically used: conversation tests and listening-only tests. For that reason, we developed scenarios that can be used either directly in a conversation test or for making recordings, which in turn would serve as stimuli in a listening-only test. In the literature, test scenarios differ in the naturalness of the conversation by giving different degrees of freedom to the speakers. Scenarios with a high degree of freedom, e.g. free discussions [1], have a limited comparability due to the lack of a common conversation structure; scenarios with low degree of freedom, e.g. reading aloud numbers [5], suffer from a low naturalness. As a compromise, structured test scenarios have been developed for conversations with two [6] and three interlocutors [2].

Those instructions have been developed yet only for a fixed number of interlocutors in the same experiment. Since the number of interlocutors in a conferencing experiment appears to influence the observed effects [2], a systematic investigation would be beneficial to quantify such scaling effects. For that purpose, scenarios were required that remain comparable while the number of interlocutors is changed. In addition, the scenarios should increase the cognitive load with increasing number of interlocutors as it can be observed in real-life conferences. Therefore we modified the scenarios from [2]. Starting from a common underlying conversation structure, with each additional interlocutor we added a fixed amount of information and thus a fixed amount of complexity to the conversation.

The scenarios consist of four phases: welcome, problem solving, information exchange and farewell. During the welcome phase, the first interlocutor assumes the role as discussion leader, checks if everybody is present, asks for an introduction round, and mentions the reason for the conference call and the agenda items. During the problem solving phase, small problems are solved by the interlocutors, whereby each problem represents one agenda item and consists of four contributions: a demand, a constraint, a conflicting constraint, and a solution. In this phase, most of the intended scaling of information and complexity is realized. With each additional interlocutor, a new problem is introduced (increasing the amount of information) and each time the four contributions demand, constraint, conflict and solution are distributed across different interlocutors (increasing the amount of complexity). Figure 1 shows how we distributed with each new interlocutor the new information items (gray blocks) to the scenario. The figure also shows that a few blocks needed to be shifted (gray arrows) in order to better balance the amount of contributions per interlocutor.

After the problem solving phase, the last agenda item in each scenario consists of an information exchange round in which information items such as email addresses or telephone numbers are exchanged. Some scaling is realized here as well, since each additional interlocutor adds one information item to the conversation. After that phase, the discussion leader closes the conference with the farewell.

To realize scenarios based on this general structure, we compiled for each scenario and interlocutor instruction sheets that provide a) general information about the scenario, b) the roles the interlocutors are asked to assume and c) about the interlocutor's contributions in detail. Bearing a future listening-only test in mind, we realized 13 scenarios, four scenarios for three interlocutors and three scenarios for two, four and six interlocutors, respectively.

3. Recording session

The main technical constraint was to be able to record up to six speakers simultaneously while they are communicating via a simulated conferencing system. That required a facility in which all six speakers can be seated such that they can not see and hear each other directly. We had the opportunity to use a large size anechoic chamber (area = $120m^2$, lower frequency limit = $63Hz$) at the Technical University of Berlin. All speaker were placed close to the walls with a maximum possible distance between each other and they were facing the non-reflecting walls. In the middle of the room, we placed the recording setup and the operator. Figure 2 shows a schematic drawing of the arrangement of the speakers and the operator inside the room, while the photograph in Figure 3 gives a visual impression of the recording scene.

We recruited six male German-speaking volunteers with professional experience as speakers or actors. Following the recording plan in Table I, we assigned the speakers to the different roles in the scenarios by balancing a number of boundary conditions:

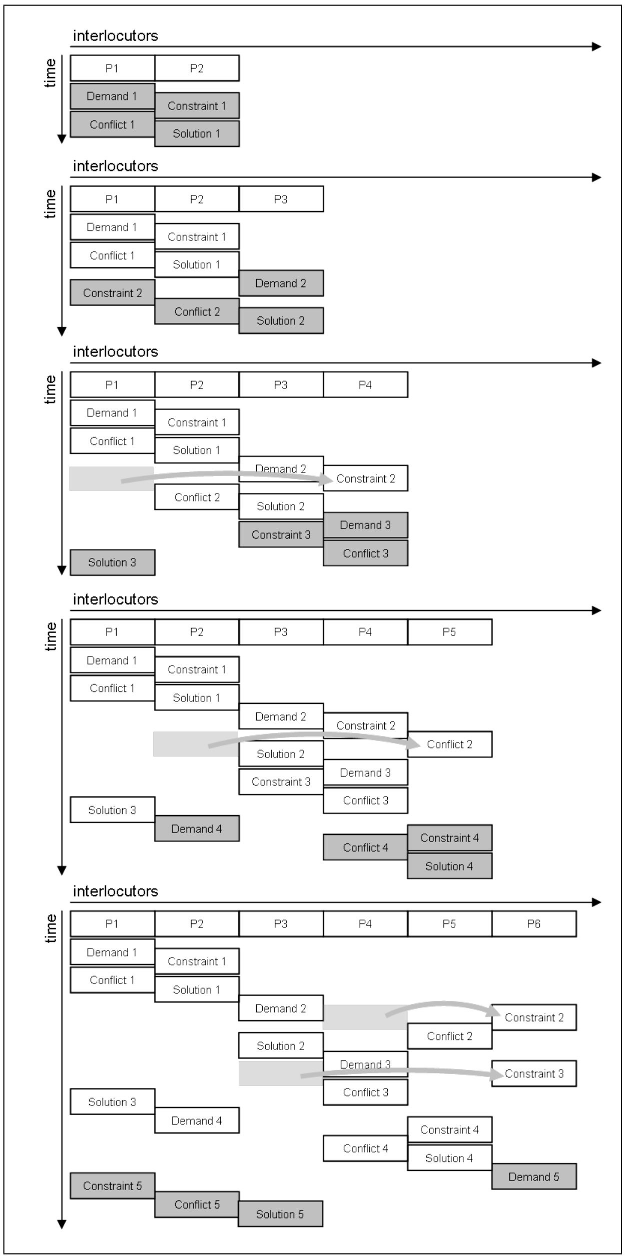

Figure 1. Scheme to scale the problem solving phase of a scenario with the number of interlocutors. With each interlocutor P1 to P6 a new problem solving phase is created by adding a new problem with the four contributions demand, constraint, conflict and solution (gray blocks). Some blocks are also shifted to better balance the amount of contributions per interlocutor (gray arrows).

1. Different speakers should play the role as the discussion leader (role 1).
2. All speakers should be represented almost equally often and should take part in scenarios with different numbers of interlocutors. That means, not all two-, three- or four-interlocutor scenarios are spoken by the same two, three or four speakers.

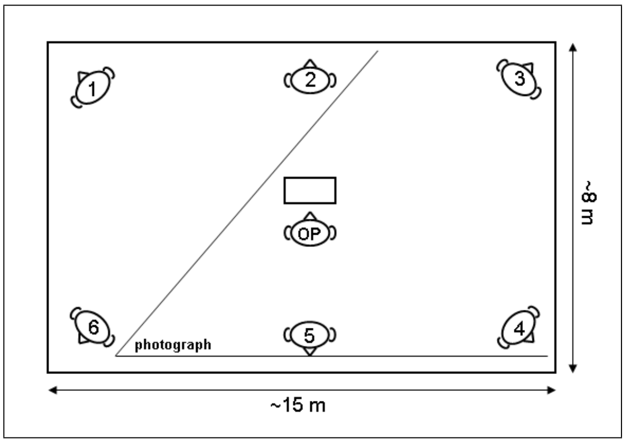

Figure 2. Recording room: positions and orientations of speakers 1 to 6 and the operator (OP) and the area covered by the photograph in Fig. 3.

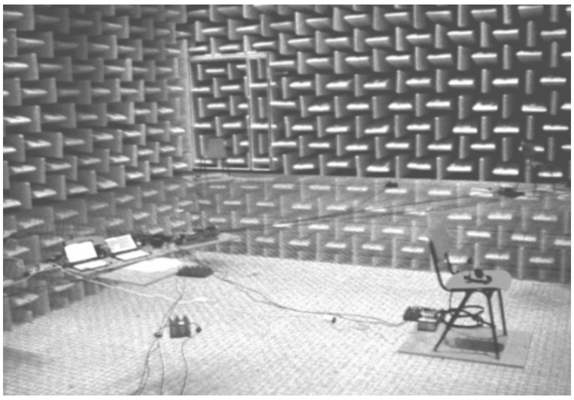

Figure 3. Recording room: in the front the seat of speaker 5, in the room's middle the recording setup, in the back barely visible seats 3 and 4.

3. The distribution of speakers across scenarios and the temporal order, in which the scenarios are recorded, should be aligned in order to minimize idle periods for speakers.
4. In order to further minimize crosstalk (microphones recording voices of other speakers) and background noises in scenarios with less than six speakers, the speakers should always take those seats with the maximum possible distance from each other as well as from the operator.

The speakers communicated via a simulated conferencing system. We aimed for a conversation situation that was as natural as possible while still using a speech transmission system and not having a face-to-face conversation. For that reason we used spatial audio reproduction via headphones as well as full bandwidth and low latency high-quality equipment. The spatial rendering of the audio signals was realized on a Linux laptop running the SoundScapeRenderer software [7], equipped with an RME HDSP card & Multiface II system and a Creamware Ultra A16 AD/DA converter. The SoundScapeRenderer convolved for each listener the five other speakers' input signals with a head-related transfer function (HRTF) that corresponds to five different angles, one per speaker: -60, -30, 0, +30, +60. Due to the rendering, each speaker could hear the conversation partners equally distributed on a semi-circle in front of him. A Windows laptop equipped with an RME HDSP card & Multiface II system and running Steinberg Cubase served as recording system. Both computers were connected via an ADAT link and using the low latency routing capabilities of the RME cards. The speakers used high-quality Sennheiser Aviation Headsets (HMD-410 & HME-46-3-6) connected to two RME Quadmic microphone amplifiers and six Millenium HA 4 headphone amplifiers. Figure 4 shows a schematic drawing of the recording setup.

Table I. Recording plan: order of scenarios (No.), scenario topic, number of interlocutors (#IL), assignment of the six speakers A to F to the roles within the scenarios, and the speaker's physical position according to Fig. 2.

No.	Scenario	# IL	Role	Speaker	Recording
1	Conference	3	1	B	3
			2	C	4
			3	A	1
2	Meeting	3	1	A	1
			2	B	3
			3	C	4
3	CD	4	1	C	4
			2	D	6
			3	A	1
			4	B	3
4	Convention	2	1	D	1
			2	C	4
5	Interviews	6	1	A	1
			2	B	2
			3	C	3
			4	D	4
			5	E	5
			6	F	6
6	Internet	6	1	D	4
			2	E	5
			3	F	6
			4	A	1
			5	B	2
			6	C	3
7	Ice cream	6	1	B	2
			2	C	3
			3	D	4
			4	E	5
			5	F	6
			6	A	1
8	City festival	4	1	C	4
			2	D	6
			3	B	3
			4	A	1
9	Paintings	4	1	E	4
			2	F	6
			3	A	1
			4	B	3
10	Jubilee	3	1	D	1
			2	E	3
			3	F	4
11	Car	3	1	F	4
			2	E	3
			3	D	1
12	Project	2	1	E	1
			2	F	4
13	Movie	2	1	F	4
			2	E	1

We processed the recordings using the Audacity freeware sound editor

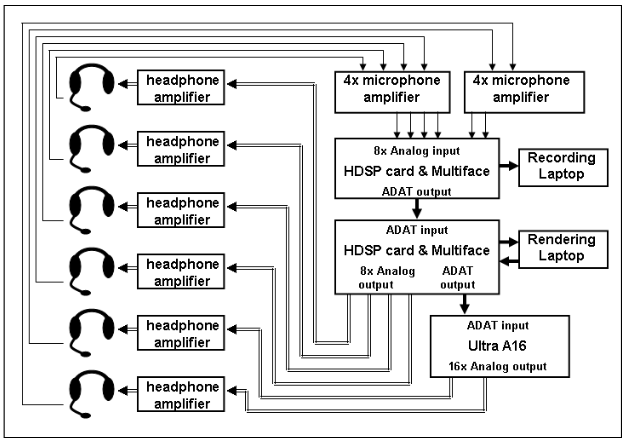

Figure 4. Recording setup: signal paths from the microphones through the recording and rendering laptops to the stereo headphones. Arrows: single-line = analog mono signal; double-line = analog stereo signal; bold single-line = multichannel digital signal.

- to adjust level differences of individual recording tracks (the speakers spoke differently loud or adjusted the microphones in different positions between the scenarios),
- to minimize background noise that got enhanced during the level adjustment using Audacity's adaptive multi-band noise gate (using moderate settings to avoid the introduction of audible artifacts),
- to remove any remaining audible microphone crosstalk (sometimes another voice was audible on a speaker's track when that speaker was silent),
- to remove other disturbing sounds such as rustling with the paper scripts or loud breathing into the microphone, and
- to reduce plopping sounds or extremely loud passages.

4. Editing the recordings

The recordings were supposed to resemble the intended structure as ideally as possible. However, the actual conversations partially deviated from the scenario scripts. Especially when speakers could not come to the intended solution of a problem, then the conversation turned into an open discussion phase, in which the speakers improvised such that they eventually came to a conclusion. Furthermore, the scenarios were two to three times longer than we expected from a pre-test with co-researchers from our laboratory. Apparently, the speakers could fill the scripts with much more details than we anticipated. The 13 recorded conversations had a total duration of about 4.5 hours. In order to achieve the desired high fit between theoretical structure and existing recordings and also to reduces recording lengths to feasible durations for a listening-only experiment, we edited the recordings.

In the editing process, it was rather easy to delete the free discussion parts, because those parts were always a kind of detour that left the intended discussion thread and went back to it such that most of the intended contributions were made nevertheless. In addition, the natural but still controlled pronunciation styles of the speakers allowed us often to remove sentences that were unnecessary for the discussion without introducing unnatural intonations or speech rhythms. Removing "ehms", coughs or longer speech pauses was easily possible as well for the same reason. Eventually, the edited recordings had a total duration of about 97 minutes. A research colleague volunteered to listen to the edited recordings and to double-check the naturalness in terms of intonation, speech rhythm and conversation flow.

The requirement of editing the recordings was not problem introduced by the speakers' performance, but due to some flaws of our implementation of the scenarios: a solution was sometimes understood as another conflict, meaning the speakers tried to find another solution; speakers sometimes pursued a different goal than we had intended; it was not always clear to the speaker that he should not provide his solution before the others mentioned their constraints or conflicts. Before using the scenario instructions in conversation tests or in future recording sessions, those findings should be addressed, for instance by providing alternative solutions to the interlocutors or by providing reminders to wait with a solution until other opinions have been shared.

5. Validating the recordings

The first validation step was to analyze the conversation structures of the recordings. Listening to the recordings, we identified passages that represented correct, inserted, deleted or substituted contributions. Correct contributions (COR) were passages that correspond to intended contributions (INT) from the scripts. Deletions (DEL) were contributions from the scripts that were missing in the actual recordings. Substitutions (SUB) were passages that served as a replacement of a missing contribution, e.g. an alternative solution was found by the speakers. In addition, correct contributions were changed to substitutions if they occurred in a different order than intended by the scripts and if this order was critical. This accounted for the fact that the temporal order of certain contributions is critical for the conversational flow, e.g. a solution should not be given before the demand, while it is not critical for other contributions, e.g. a conflict can be mentioned before the constraint. Insertions (INS) were any additional contributions by the speakers. With this labeling of contributions we

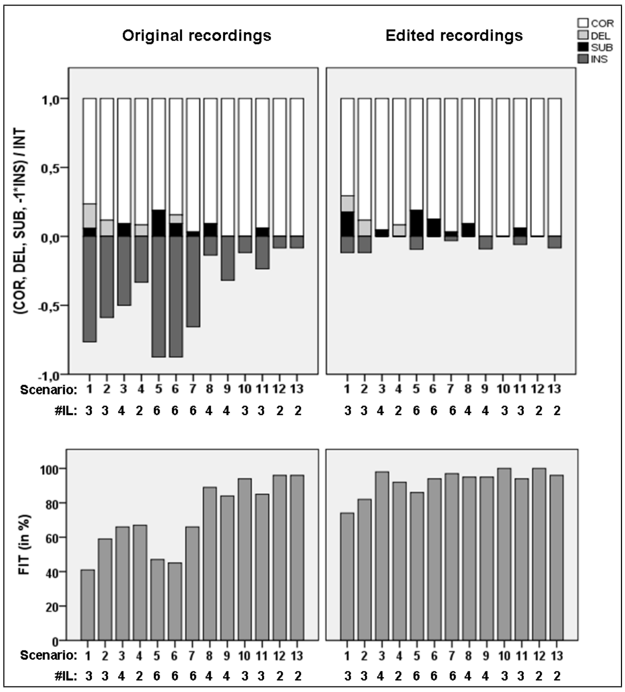

Figure 5. Structural analysis. Top panel: number of correct (COR), deleted (DEL), substituted (SUB) and inserted (INS) contributions relative to the number of intended (INT) contributions. INS are depicted on the negative axis for better interpretability. Bottom panel: computed FIT value in percent between actual recordings and intended structure from the scenario scripts. All panels: results for original recordings to the left, results for edited recordings to the right, plotted order of scenarios correspond with temporal order during recording day. Abscissa labels: scenario number & number of interlocutors (#IL) in that scenario.

also computed a measure for the fit between actual recording and scenario in percent:

$$FIT = \frac{COR + 0.5 \cdot SUB - 0.5 \cdot INS}{INT} \cdot 100 \quad (1)$$

Note that the chosen weights for SUB and INS mean that some deviation of the actual conversations from the scenario scripts is allowed without punishing those deviations too much. The top panel in Figure 5 shows the relative amount of the different contribution types per scenario recording. The relative amount is computed by total number of contributions divided by intended number of contributions and the insertions INS are plotted on the negative axis. The bottom panel in Figure 5 shows the corresponding FIT.

Comparing the relative amount of contributions only between the original recordings, one can see that many insertions were present in some but not all scenarios. Accordingly the conversation structure of the original recordings differed substantially, which can be also seen in the different FIT values, ranging from 41 to 96 %. If one neglects the results for the six person scenarios (No. 5, 6 & 7), a learning curve for the speakers is visible: recordings made towards the end of the session day achieved higher fits and lower insertions than those made in the beginning of the session day. Interestingly, this learning effect appears to be superseded with a non-linear effect due to the number of interlocutors: for six interlocutors, the fits are rather low and many insertions are made; for less than six interlocutors, this effect is not visible. The results for the edited recordings verify our expectation that the FIT as well as the similarity between the edited recordings could be improved compared to the original recordings. However, the intended structure could not always be perfectly achieved; maintaining natural conversations was the limiting factor during the editing process.

In a second validation step we conducted an conversational analysis similar to [8]. First, we computed state probabilities for the different speaker states that occur during the conversations: Silence, Single-Talk (one speaker is talking), Double-Talk (two speaker are talking simultaneously), etc. The top panel of Fig. 6 shows a stacked bar plot of those state probabilities for the original and the edited recordings, whereby states with more than one speaker are summarized to a Multi-Talk state for visibility reasons. One can see that the state probabilities for Silence and for Multi-Talk are slightly smaller for the edited recordings than for the original recordings. That confirms that we reduced the speech pauses during the editing process, but it also shows, that we lost some degree of interactivity because apparently we deleted some of the Multi-Talk situations. Second, we computed the speaker alternation rate per minute, quantifying on average how often a speaker change occurred per minute. As the middle panel in Fig. 6 shows, the speaker alternation rate is enhanced for most of the scenarios; especially when removing details and individual sentences during the editing process, we shortened in particular passages that were monologues, which in turn explains the observed increase of speaker alternation. But we did not achieve any improvement in terms of equalizing the speaker alternation rates between the scenarios: while the three- and four-interlocutors scenarios have slightly more similar alternation rates after the editing than before, the two- and six-interlocutors scenarios deviate strongly. Third, we computed for each individual interlocutor the state probability that he is speaking (Single- and Multi-Talk) representing the proportion of each speaker during the conversations. Comparing the plots at the bottom panel in Fig. 6 between edited and original recordings, we do not observe any strong or surprising effects. The dominant role of the discussion leader (person 1) remains and there is a slight overall increase of the probability values that can be explained by the reduction of speech pauses and the corresponding silence state probabilities.

To summarize the results, the editing process slightly improved the structural similarity of the con-

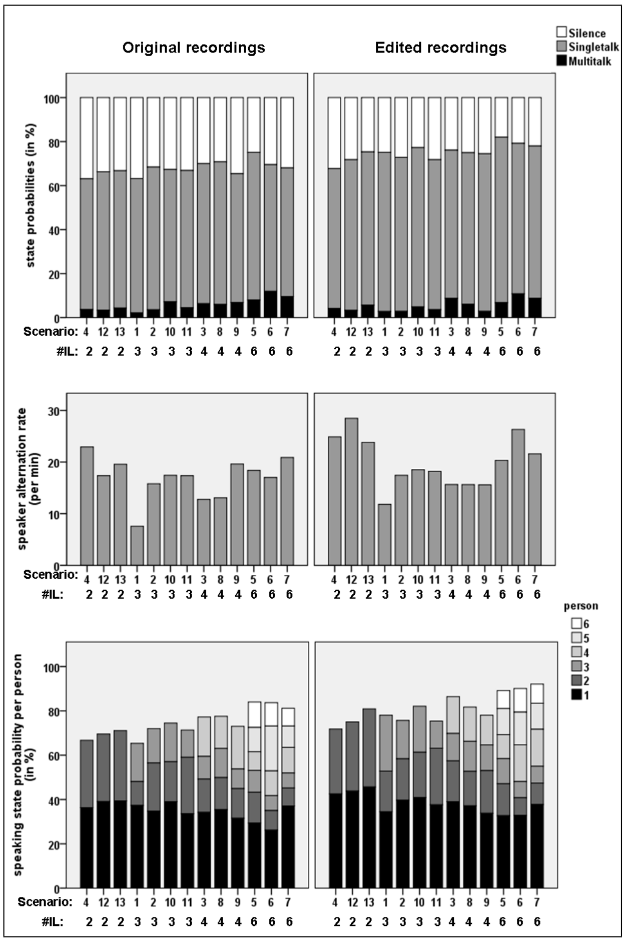

Figure 6. Conversational analysis. Top panel: state probabilities for Silence, Single-Talk, and Mutli-Talk periods. Middle panel: Speaker alternation rate per minute. Bottom panel: State probability for each interlocutors that he is speaking. All panels: results for original recordings to the left, results for edited recordings to the right, scenarios are ordered according to number of interlocutors per scenario for better interpretability. Abscissa labels: scenario number & number of interlocutors (#IL) in that scenario.

versations without introducing any severe changes in the conversational aspects.

6. Conclusions

Going through the process from scenario development to the final edit of the recordings revealed that individual details can have a strong impact on the outcome. The process could be improved for similar recordings in the future at several points, such as: create fallback solutions in the scenario scripts to avoid that speakers need to improvise, double check the exact headset microphone positions before each recording-take to minimize loudness variations and breathing noise, record extra scenarios to allow the speakers some learning time.

Despite the room for improvement, the validation showed that this approach enabled us to make recordings with a high structural and also a rather good conversational similarity. Thus the present recordings can be especially used in listening-only tests in which the number of interlocutors plays an important role. A series of such listening-only tests is currently ongoing from which we expect further insights on the applicability of the present recordings for the QoE assessment of audio conferencing systems.

Acknowledgement

We thank the six speakers for their excellent work that made these recordings possible.

References

[1] J. J. Baldis, “Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences”, Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference, pp.166-173, 2001.

[2] A. Raake, C. Schlegel, K. Hoeldtke, M. Geier, J. Ahrens, “Listening and conversational quality of spatial audio conferencing”, AES 40th International Conference, Tokyo, 2010.

[3] S. R. Fussell, N. I. Benimoff, “Social and Cognitive Processes in Interpersonal Communication: Imlications for advanced telecommunicatinos technologies”, Human Factors, Vol. 37, pp. 228-250, 1995.

[4] G. M. Olson, J. S. Olson, “Distance Matters”, Human-Computer Interaction, Vol. 15, pp. 139-178, 2000.

[5] N. Yankelovich, J. Kaplan, J. Provino, M. Wessler, J. M. DiMicco, “Improving Audio Conferencing: Are Two Ears Better than One?”, Proceedings of CSCW 2006, ACM Press, 2006.

[6] S. Möller, “Assessment and Prediction of Speech Quality in Telecommunications”, Kluwer Academic Publishers, USAŨBoston, 2000.

[7] M. Geier, J. Ahrens, S. Spors, “The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods”, Proceedings of 124th AES Convention, May 17-20, NLŨAmsterdam, 2008.

[8] K. Hoeldtke, A. Raake, “Conversation analysis of multi-party conferencing and its relation to perceived quality”, accepted for presentation on IEEE International Conference on Communications ICC 2001, 5-9 June 2011, Kyoto, Japan.