Quality Assessment of Spatial Audio Conferencing Systems from an End User Perspective

Janto Skowronek, Alexander Raake, Angelo De Silva, David Rogiers
Assessment of IP-based Aplications, Telekom Innovation Laboratories, Technische Universität Berlin
Email: janto.skowronek@telekom.de

Introduction

The idea of using spatial audio to improve the experience of teleconferencing applications has been investigated for a several years. Results were encouraging and first commercial products have actually been brought to the market. With the transition from research prototypes to running services, a number of questions arise concerning the quality evaluation of such systems from an end user perspective: To which extend will end users appreciate spatial audio conferencing beyond an initial "wow-effect" or "strangeness-effect"? Which technical parameters are crucial for the quality perception by the end users? To answer such questions, appropriate quality assessment methods are required. However, the goal of such methods is not simply to assess quality differences between non-spatial and spatial audio conference systems, but to be able to to distinguish between variants of spatial audio conferencing systems. To serve as a background for the development of such methods, this paper reviews existing literature, first focussing on studies that showed the added value of spatial audio, and then analyzing in more detail those studies that (attempted to) differentiate between spatial audio conferencing systems.

On the added value of spatial audio

The reviewed studies [1, 2, 3, 4, 5] applied different methods to assess the added value of spatial audio conferencing. Those methods differed on the one hand in the test paradigm: most studies applied listening-only tests; one study applied a conversation test. On the other hand, the methods differed in the set of measures used for assessment. Some studies used standardized ITU rating scales asking for quality, listening effort, or conversation effort. Some studies used additional rating scales asking for the perceived speaker separation effort or for the perceived cognitive load, and simple preference ranking was employed as well. Additional measures were not using direct ratings from the test participants, but were computed from the percentage of correct answers from test participants in a keyword spotting task or in a memory test. Furthermore, also a conversational analysis, measuring conversation duration and speaker state probabilities, has been investigated as well. Table 1 provides an overview about the applied test paradigms and used measures across the considered studies. Notice that some studies used a number of individual measures for similar aspects which we subsumed here under one measure, enabling us to focus on the essential aspects. This overview shows that in most cases the approaches revealed significant differences, i.e. confirmed the added value of spatial audio, even though some dependency on the actual study and measure can be observed.

On distinguishing spatial audio systems

The reviewed studies [2, 4, 6, 7, 8, 9] attempted to differentiate between variants of spatial audio conferencing systems. In terms of the applied assessment methods, sensitivity is apparently a key issue, because in only a few cases significant differences could be found between spatial audio conditions. Next to the tested aspects of spatial audio reproduction and the used measures, a number of factors (may) have influenced the sensitivity as well. First, it is known that listening-only test are usually more sensitive that conversation tests, probably a reason why the conversation test in [6] could not differentiate between the tested conditions. Second, the amount of training that subjects get before doing the actual rating can obviously influence sensitivity, and studies with more training revealed differences, studies with less training did not. A third aspect concerns any particular specifics of the stimuli or the stimuli context used in the tests. In terms of stimuli specifics, [7] comprised stimuli that were very complex (dynamic behavior of the spatial rendering), and [2] included additional aspects beyond the audio processing (i.e. personalization features). In terms of stimuli context, [4, 6, 9] comprised additional non-spatial conditions (audio bandwidths, speech shaped noise). Fourth, the characteristics of the speakers' voices can influence the difficulty for a listener to separate those voices. The similarity of voices is known to be relevant here, e.g. [4] found indications for that when comparing non-spatial and spatial conditions. However, also the familiarity with the speakers' voices appears to be important as the results of [6] suggest. Table 2 summarizes these findings in an overview.

Conclusions

Quality differences between spatial & non-spatial audio conferencing systems can be measured, while quality differences between different spatial audio conferencing systems are very difficult to measure. While the existing proposals are promising, future work should focus on improving the sensitivity of the available test methods. Based on this review, we recommend that those methods should describe precisely the measures to be used, the training of subjects, and the subject profiles in terms of voice characters and familiarity.

Tabelle 1: Studies comparing non-spatial and spatial audio. Paradigm: LOT = listening-only test, CT = conversation test. Measures: see text. Table entries: $\surd$ = the study found for the measure significant differences between non-spatial and spatial audio conditions; $\times$ = the measure did not reveal significant differences; no symbol = the study did not apply the measure; $(\surd)^1$ = significant differences were only found for one of the two tasks that participants were asked to do during that test; $(\surd)^2$ = significant differences were only found in combination with other conditions (i.e. audio bandwidth) used in the test.

Study	Paradigm	Measure							
		Quality	Listening / Conversation Effort	Keyword Spotting Spotting	Memory Test	Preference Ranking	Conversation Analysis	Speaker Separation Effort	Cognitive Load
[1]	LOT				$\surd$	$\surd$		$\surd$	
[2]	LOT				$\times$	$\surd$		$\surd$	
[3]	LOT	$\times$	$(\surd)^1$	$(\surd)^1$					
[4]	CT	$\surd$	$\times$				$\times$		
	LOT	$\surd$			$\surd$				$\surd$
[5]	LOT	$\surd$			$(\surd)^2$			$\surd$	$\surd$

Tabelle 2: Studies comparing spatial audio conditions. Spatial audio conditions: keywords summarizing the tested conditions; Measures: applied measures (see Tab. 1); Sensitivity: $\surd$ = significant differences; $\times$ = no significant differences; Paradigm: LOT = listening-only test, CT = conversation test. Measures: see text. Training: keywords summarizing the amount of training of subjects; Stimuli Specifics, Stimuli Context & Familiarity of speakers: see text;

Study	Spatial Audio Conditions	Measures	Sensitivity	Influencing factors				
				Paradigm	Training	Stimuli Specifics	Stimuli Context	Familiarity of speakers
[2]	Rendering Techniques	Preference ranking Task performance	$\surd$ $\times$	LOT	on voices, not on conditions	Personalization feature		yes, via training
[4]	Head Tracking	Quality Task performance	$\times$ $\times$	LOT	listening demo & 1 call		Audio Bandwidths	No
[6]	Positions (Angles)	Quality Cognitive Load	$\times$ $\times$	CT	listening demo & 1 call		Audio Bandwidths	all speakers were friends / relatives
[7]	Dynamic changes of angles	Quality Cognitive Load	$\times$ $\times$	LOT	visual explanation & 1 call	Very complex stimuli		No
[8]	Channel separation HRTFs, Head tracking	Quality	$\surd$	LOT	Specific explanations & 2 examples per condition			No
[9]	Sound capture, HRTFs, Angles	Task performance	$\surd$	LOT	5-8 % of stimuli as training			No

Literatur

[1] Baldis, J., *Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop*, J. Beaudouin-Lafon, M., Jacob, R. J. K. (Eds.), Human Factors in Computing Systems Conference, 3, 166-173, 2001.

[2] Kilgore, R., Chignell, M., Smith, P., *Spatialized audioconferencing: what are the benefits?*, Conference of the Centre for Advanced Studies on Collaborative research, 135 - 144, 2003.

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

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

[5] Skowronek, J., Raake, A., *Investigating the effect of number of interlocutors on the quality of experience for multiparty audio conferencing*, Interspeech, 829-832, 2011.

[6] De Silva, A., *Development of a Spatial Audio Conferencing Simulation System and Assessment of the Effect of Spatialization on Quality of Experience*, Diploma Thesis, Assessment of IP-based Applications, TU of Berlin, 2013.

[7] Rogiers, D., *Dynamical Aspects of Spatial Audio in Multiparticipant Teleconferencing*, Master Thesis, Assessment of IP-based Applications, TU Berlin, 2013.

[8] Volk, T., *Quality of Experience - Evaluierung eines Telekonferenzsystems in der Entwicklungsphase*, Diploma Thesis, Lehrstuhl für Datenverarbeitung, TU München, 2013.

[9] Malfait, L., *Differentiating among spatial multi-party conferencing systems with a subjective listening-only task*, ITU-T Contribution COM 12 – C 0121 – E, Internation Telecommunication Union, Geneva, 2013.