Influence of Packet Loss and Double-Talk on the Perceived Quality of Multi-Party Telephone Conferencing with Binaurally Presented Spatial Audio Reproduction

Maxim Spur[1], Dennis Guse[1], Janto Skowronek[2]
[1] *Quality and Usability Lab, Technische Universität Berlin, Deutschland*
[2] *Technische Universität Ilmenau, Deutschland*
Email: maxim.spur@tu-berlin.de, dennis.guse@tu-berlin.de, janto.skowronek@tu-ilmenau.de

Abstract

In this paper, we investigated how the perceived quality of a telephone conferencing system with spatial sound reproduction is affected by packet loss degradation and double-talk. We conducted a listening-only experiment investigating the impact of packet loss when two speakers talk either in sequence or concurrently.

Regarding packet loss, the results show that subjects are able to differentiate between different degrees of packet loss, i. e., larger amounts of loss results in a lower perceived quality. These findings are in line with packet loss on non-spatial single-channel telephony, and they suggest that quality assessment methods for conventional telephony are also applicable for spatial audio scenarios.

In addition, the results show that participants were able to discern which of the two speakers' connections were affected by packet loss. Furthermore, the presence of double-talk significantly improved the subjective quality ratings of the impaired conditions compared to sequential speech. Nevertheless, the impaired double-talk conditions were rated lower than the unimpaired conditions, showing that the degradation as such is still perceived.

This leads to the conclusion that double-talk makes degradations on individual connections less apparent, which suggests for future work that the benefit of spatialization in double-talk situations reduces the impact of impairments affecting individual speakers.

Introduction

Spatial audio reproduction, as achieved through e. g., convolving monaural signals with *head-related transfer functions* (HRTFs), has been shown to offer significant benefits to conferencing calls [1], i. a. by harnessing the *Cocktail Party Effect* [2]. Since most current systems for audio conferencing are based on *Voice over IP* (VoIP) technology, they are susceptible to common issues arising from using the underlying IP networks, such as *packet loss* (PL). If no countermeasures such as forward error correction or packet loss concealment are applied, this (random) PL leads to lost audio segments—usually 20 ms in length with each lost packet—and can result in choppy audio at best and a complete loss of spoken content at worst.

Depending on the degree of PL, which can be measured as frequency of the random PL events as well as the number of successive packets lost in one PL event (PL *burst length*), the user-perceived quality can drop significantly.

Prior work exists which investigates the effect of PL on the perceived quality within the context of monaural telephone conferences (e. g. [6]). This paper aims to complement this research by providing a first study assessing the perceived quality impact of PL within a spatial-only conferencing setup. In addition, the study also explores for the first time how the inclusion of double-talk (i. e., both speakers talking at the same time) influences listeners' perceived quality in this context.

The results from the experiment highlight listeners' sensitivity to slight increases in PL and their ability to correctly attribute PL to individual speakers. They further show that conditions with double-talk were consistently rated higher than their sequential speech equivalents.

Experimental Setup

To assess the impact of PL and double-talk on spatially represented audio conferences a listening-only test was conducted. It employed the established research methods for speech systems as outlined by ITU-T recommendations P.800 [11] as well as P.1301 [13] and also used in prior work (e. g., [6, 7]). This section details the design of the experiment and how it was conducted.

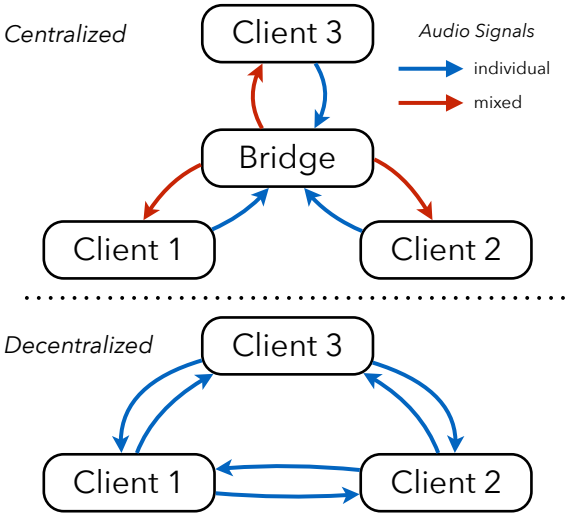

Figure 1: Centralized vs. decentralized conferencing topologies shown with three interlocutors/clients.

Topology

Multi-party conferencing systems usually follow one of two designs: a centralized or decentralized topology (see Figure 1). The main difference between those manifests itself in the location where the mixing and further processing (e. g., spatialization) takes place. In the centralized topology, a central conferencing bridge connects all clients together. The bridge receives each client's outgoing audio signal and prepares an individual mixed signal for each client to receive. In the case of spatial audio, packet loss during transmission of a stereo signal could result in artifacts that diminish the advantageous spatial effect. This motivates the choice of the decentralized topology for this study, since only mono signals are sent in that case. Each client receives every other's outgoing audio and does the mixing and processing by itself. PL then only results in silent parts with no impact on the spatial effect.

Spatial Rendering and Packet Loss Insertion

A decentralized conferencing system with three participants (one listener as the test subject and two pre-recorded speakers) was simulated by playing the speakers' outgoing audio through a spatialization engine (the SoundScape Renderer [9]) and presenting the resulting binaural signal to the listener. The spatialization engine is set up so that both speakers are virtually positioned in front of the listener and separated from each other by 30 degrees, thereby creating a "left" and "right" speaker.

Since pre-recorded sentences were used to act as the speakers, PL could be simulated by muting the recorded speech for multiples of 20 ms (one lost audio frame per packet) for each packet loss event depending on the desired PL severity. For this experiment, three degrees of PL severity as measured in burst length (BL) were compared: 3, 6, or 9 packets lost in a row per PL event. The PL events were distributed randomly through parts of the recordings where speech was present at a frequency that resulted in 7.5, 15 or 22.5 % of spoken content lost, respectively.

Stimuli Presentation

The test conditions differed as follows: PL burst lengths of 3, 6 or 9 packets in a row, PL affecting the left, right or both speakers and double-talk occurring or not. Each simulated conversation was created from pre-recorded sentences and edited to be around 10 seconds long. Using permutations of the test parameters and repeating them with different spoken content, 96 unique stimuli (repeating each condition three times) were created. These stimuli were presented using a digital questionnaire tool (TheFragebogen, [10]) running on a tablet PC with the audio output through an Edirol UA-25EX sound card to a pair of Beyerdynamic DT 790 Pro headphones.

Scales

The questionnaire tool allowed repeated playback of the stimuli and provided the test subjects with scales to rate two aspects of each simulated conference after listening to it:

- Perceived quality of the whole system; overall impression of the connection
- Perceived quality of the individual connections (left/right speaker's connection)

Figure 2: Continuous 7-point scale with German labels as used in the study. Translated labels left-to-right with MOS value: extremely bad (0), bad (1), poor (2), fair (3), good (4), excellent (5) and ideal (6).

The first question used a discrete 5-point absolute category scale ranging from "excellent" to "bad" [11]; the second one used the 7-point continuous scale [12] shown in Figure 2 and extended the range to "extremely bad" and "ideal". These two different kinds of scales (with reversed orders of options) were chosen to further differentiate both types of questions and encourage the subjects to think separately about each question. Prior to the evaluation of the 96 stimuli, 16 stimuli were presented to the subjects during a training phase, showing the full range of variation of the conditions using the same setup. During the training phase and the subsequent experiment phase, the subjects were left alone in soundproof cabins according to ITU-T P.800 [11]. The experiment sessions lasted on average about 50 minutes. The experiment was conducted in Berlin, Germany in June 2015 with 25 native speaking subjects (13 female, 12 male), aging from 18 to 45 years old ($\mu = 29$).

Results

The quality ratings obtained from the experiment are expressed as *Mean Opinion Scores* (MOS), ranging from 1 to 5 or 0 to 6 for the two scales (see Figure 2). Figure 3 displays these for the examined conditions, separated by type of speech (sequential/double-talk), PL severity (burst lenght of 3, 6 or 9) and which connections are affected by PL. The individual connection quality ratings also show the MOS for the connections which were unaffected while the other one experienced a degree of PL. Examining these allows to investigate the noticeability of the degradation, i. e., were participants able to tell which connections were affected by PL or not affected.

To analyze the statistical significance of differences in MOS values with changing PL severity, the non-parametric Friedman test with post-hoc analysis was carried out on all groups (e. g., double-talk, overall quality, both sides affected: burst length of 3 vs. 6 vs. 9 or none). As seen in Figure 4, the MOS values change significantly with a change in PL severity; post-hoc analysis further reveals that most pairwise comparisons of conditions show significant change with increasing PL. Exceptions are most comparisons between burst lengths 6 and 9 as well as ratings of the unaffected connections when comparing any two burst lengths higher than zero.

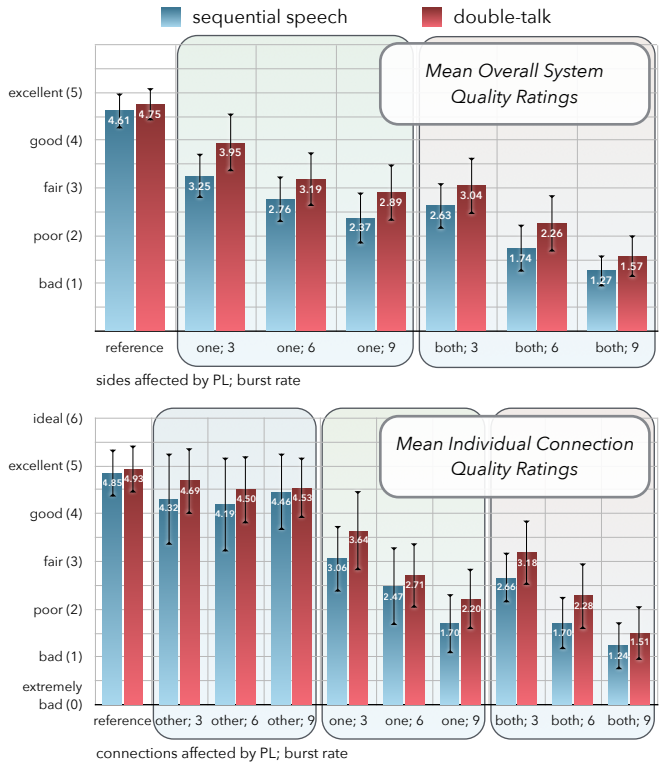

Figure 3: MOS values for all examined conditions, separated by type of speech, PL severity and affected connection; shown for both overall and individual connection quality.

This last point indicates that any degradation occurring on one connection impacts the quality judgements of the other unaffected one, while the severity of the PL does not have an effect.

Comparing conditions with PL on one connection to those with PL on both connections, the used non-parametric Wilcoxon signed-rank test showed significant differences for every burst length (p <.0001; effect sizes >0.5; both connections affected resulted in worse ratings).
The same test was used to compare conditions with sequential speech to those with double-talk. The results, along with medians of the analyzed data are displayed in Figure 5. Most comparisons show a significant improvement in MOS when double-talk is present for all PL conditions. The effect sizes are larger when both connections were affected by PL.

Discussion

The first finding from the study—decrease in MOS with an increase of PL severity—is in line with related work from the conventional non-spatial context, c.f. [5]. Another insight is the finding that the MOS of an individual connection's quality slightly decreases when the other connection is also affected by PL. This is also the case for unimpaired connections in which MOS decreases when there is PL on the other connection. This finding confirms for the spatial context that a mutual influence of the quality perception of individual connections can be observed [8].

Test	Result			sequential speech, PL on one side	sequential speech, PL on both sides	double-talk, PL on one side	double-talk, PL on both sides
Friedman rank sum test	p			2.21E-15	6.01E-16	3.01E-14	3.01E-14
	$\chi^2(3)$			71.3	74.0	66.0	73.0
Multiple comparisons (post-hoc)	critical difference			*24.1*	*24.1*	*24.1*	*24.1*
	observed difference	ref	3	**26.0**	**25.0**	**24.5**	**26.0**
		ref	6	**53.0**	**52.0**	**55.0**	**50.5**
		ref	9	**71.0**	**73.0**	**66.5**	**73.5**
overall system quality		3	6	**27.0**	**27.0**	**30.5**	**24.5**
		3	9	**45.0**	**48.0**	**42.0**	**47.5**
		6	9	18.0	21.0	11.5	23.0

Test	Result			sequential speech, PL on other side	sequential speech, PL on one side	sequential speech PL on both sides	double-talk, PL on other side	double-talk, PL on one side	double-talk, PL on both sides
Friedman rank sum test	p			3.94E-05	1.99E-15	6.40E-16	2.77E-07	1.19E-15	6.40E-16
	$\chi^2(3)$			23.1	71.5	73.8	33.3	72.6	73.8
Multiple comparisons (post-hoc)	critical difference			*24.1*	*24.1*	*24.1*	*24.1*	*24.1*	*24.1*
	observed difference	ref	3	**29.0**	**27.0**	**25.0**	**26.0**	**25.0**	**26.0**
		ref	6	**41.0**	**49.0**	**51.0**	**49.0**	**52.5**	**49.0**
		ref	9	**30.0**	**74.0**	**74.0**	**39.0**	**72.5**	**75.0**
individual connection quality		3	6	12.0	22.0	**26.0**	23.0	**27.5**	23.0
		3	9	1.0	**47.0**	**49.0**	13.0	**47.5**	**49.0**
		6	9	11.0	**25.0**	23.0	10.0	20.0	**26.0**

Figure 4: Results of Friedman tests with post-hoc analysis of both overall system and individual connection quality judgements when comparing different PL severities (burst lengths). Significant differences highlighted in bold type and yellow background.

	overall system quality				*individual connection quality*			
impaired side, burst length	medians of data subsets		p	effect size	medians of data subsets		p	effect size
	sequential speech	double-talk			sequential speech	double-talk		
reference (none)	4.75	4.83	**7.2E-03**	**-0.38**	4.83	4.93	**2.7E-02**	**-0.31**
one side, 3	3.33	4.00	**2.7E-05**	**-0.59**	3.02	3.50	**1.3E-03**	**-0.45**
one side, 6	2.83	3.17	**9.1E-04**	**-0.47**	2.25	2.80	1.3E-01	-0.22
one side, 9	2.33	2.83	**3.8E-05**	**-0.58**	1.67	2.27	**7.7E-04**	**-0.48**
both sides, 3	2.67	3.17	**3.8E-04**	**-0.50**	2.68	3.27	**5.8E-05**	**-0.57**
both sides, 6	1.67	2.33	**8.6E-05**	**-0.56**	1.77	2.32	**1.8E-07**	**-0.74**
both sides, 9	1.17	1.50	**5.3E-05**	**-0.57**	1.19	1.57	**2.5E-04**	**-0.52**

Figure 5: Results of Wilcoxon signed-ranks tests of both overall system and individual connection quality judgements when comparing sequential speech to double-talk. Significant differences highlighted in bold type and yellow background; medium to high effect sizes in bold type and color.

This effect depends only on the presence of PL in the other connection but not its severity, as comparing different burst lengths of PL on the other connection yields no significant differences (see the individual connection comparisons of the "other sides" in Figure 4).
Since there is no effect between the severity of PL on the other connection and the drop in MOS of the rated connection, a confusion of both connections (either in hearing which of the two is affected by PL or in selecting the wrong rating scale) can be ruled out. Visualization of every test participant's individual rating data found no one who consistently rated. e. g., the worse connection as being better, although isolated instances of confusion can contribute to lower or higher means.

Conversely, double-talk has been found to increase MOS across almost all conditions. Here it could be argued again that confusions between the affected and unaffected connections are a factor in the MOS increase (balancing out the MOS decrease in the "other connection affected" ratings discussed above).

On the other hand, the MOS improvements are even greater in the overall system quality ratings, which are not affected by confusions of the two connections. Two effects could be responsible here:

- It may be harder to notice signal degradations with two people talking at the same time.
- The (desirable) spatial effect may be more perceptible with two separate sound sources being active at the same time.

Conclusion and Future Work

In this paper, a listening-only test was conducted to study the perceived quality of spatial conferencing systems in the presence of packet loss as well as double-talk. The results with regard to packet loss only are in line with prior work on the non-spatial case. We were able to confirm these findings for the case of spatial audio reproduction. In addition, it could be shown that the number of connections affected by PL as well as the presence of double-talk significantly influence a listener's perception of the system quality and also the quality of individual connections.

Future work is necessary to further substantiate the insights into how much of an effect spatial audio reproduction has on the perception of quality under the conditions investigated in this experiment. This could be done by a direct comparison of spatial vs. non-spatial reproduction. The understandings of how signal impairments on individual connections between interlocutors affect quality ratings gathered here should be further deepened by investigating a greater number of interlocutors and different kinds of signal degradation.

References

[1] Kilgore, R., Chignell, M. and Smith, P.: Spatialized Audioconferencing: What Are the Benefits? CASCON '03 Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research, IBM Press, 2003

[2] Arons, B.: A Review of The Cocktail Party Effect. Journal of the American Voice I/O Society, vol 12 pp 35–50; 1992

[3] Le Callet, P,. Möller, S. and Perkis, A.:Eds., Qualinet white paper on definitions of quality of experience Version 1.2, Qualinet, 2013

[4] Möller, Sebastian. Assessment and Prediction of Speech Quality in Telecommunications. Kluwer Academic publishers, 2000.

[5] Raake, Alexander. Speech Quality of VoIP – Assessment and Prediction. Chichester, West Sussex, UK: Wiley, 2006.

[6] Skowronek, J., Herlinghaus, J. and Raake, A.: Quality Assessment of Asymmetric Multiparty Telephone Conferences: a Systematic Method from Technical Degradations to Perceived Impairments. Proc. 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), pp 2604–2608, Lyon , 2013

[7] Skowronek, J. and Raake, A.: Assessment of cognitive load, speech communication quality and quality of experience for spatial and non-spatial audio conferencing calls. Speech Communication, vol 66 pp 154–175; 2015

[8] Skowronek, Janto, Anne Weigel, and Alexander Raake. "Quality of Multiparty Telephone Conferences from the Perspective of a Passive Listener". In: Fortschritte der Akustik - 41. Jahrestagung für Akustik (DAGA). Nürnberg, Germany, Mar. 2015.

[9] Geier, M. and Spors, S.: Spatial audio with the soundscape renderer. In 27th Tonmeistertagung—VDT International Convention, 2012

[10] TheFragebogen `http://thefragebogen.de`

[11] ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality, International Telecommunication Union, Geneva, 1996.

[12] ITU-T Recommendation P.851, Subjective quality evaluation of telephone services based on spoken dialogue systems, International Telecommunication, Geneva, 2003.

[13] ITU-T. Recommendation P.1301 - Subjective quality evaluation of audio and audiovisual telemeetings. International Standard. Geneva, Switzerland: International Telecommunication Union, 2012.