Effects of Delay and Packet-Loss on the Conversational Quality

Thilo Michael[1], Sebastian Möller[1,2]

[1] *Quality and Usability Lab, Technische Universität Berlin, Germany, [firstname.lastname]@tu-berlin.de*

[2] *German Research Center for Artificial Intelligence (DFKI), Berlin, Germany*

Abstract

In current packet-switched telephone transmissions, packet-loss and delay are two of the most noticeable degradations affecting not only the conversational quality but also the structure of the conversation. The transmission delay may lead to interruptions and changes in the way the speaker take turns. At the same time, bursty packet-loss may cause important information to be requested again, which adds additional turns to the conversation. In this paper, we investigate the effects of the combination of three levels of transmission delay (0 ms, 800 ms, 1600 ms) and three levels of bursty packet-loss (0%, 25%, 50%) on the conversational quality, the interactivity of the conversation and the contents of the conversations. For this, we perform a conversation experiment using the Short Conversation Tests and Random Number Verification Tests, as described by ITU P.805. While the conversational quality is assessed with the Mean Opinion Score (MOS) of the conversational quality, four listening dimensions, and one interaction dimension, the interactivity is assessed with a conversational analysis based on the states of the conversation (double talk, mutual silence, speaker A, speaker B).

Introduction

Today's communication networks rely mostly on packet-switched Voice over IP (VoIP) transmission of speech signals, where the coded speech is split into packets that are then routed through a heterogeneous network. With codecs filling the frequency range up to fullband, two of the most noticeable degradations that occur in speech transmissions are packet-loss and delay. Both are accounted for in parametric and instrumental models that try to predict and asses the perceived quality of a transmission network. Because of the uncertainty of transmission paths in today's networks, high amounts of delay and lost packet are to be expected, especially in third-party services that have no control over the quality of the transmission network.

Conversational quality can be assessed with conversation tests, where two participants communicate with each other over a simulated telephone network. What kind of information is exchanged is dependent on the type of conversation test that is conducted. Two popular tests that have been standardized by the ITU are the Short Conversation Test (SCT) [11] and the Random Number Verification test (RNV) [12]. The SCT consists of real-world "role play" scenarios where participants book flights or order pizza. Resulting conversations are typically slower, and the information transmitted is oftentimes redundant and may be derived from context. During RNV tests, participants exchange and swiftly compare numbers. The resulting conversations are typically highly interactive, and the transmitted information is dense and can't be derived from context.

Both the impact of delay and packet-loss on conversations are long studied phenomena. Delay impacts the interactiveness of conversations. Thus, people judge the quality of a conversation not only based on the amount of transmission delay but also based on the interactiveness and thus type of conversation they are having [2]. During an SCT conversation, speaker alternations occur less frequent and turns are held longer. During an RNV conversation, the speaker alterations occur more frequently, and turns are very short. Because delay affects mostly speaker changes, the SCT test is rated consistently better than the RNV test when the transmission is degraded with high amounts of delay [2].

It stands to reason that the quality judgments of conversations with high amounts of packet-loss also changes depending on the type of the conversation. Here, the type of conversation may be defined as the redundancy of information transmitted or the amount of context a person can use to reconstruct an utterance. For example, in a scenario where two conversations with the same amount of packet-loss are rated, but in the second scenario, a credit card number is transmitted (and can't be inferred by context), the latter may have a lower quality rating.

In this paper, we analyze and compare the effects of delay and packet-loss on conversational quality. We conducted SCT and RNV tests with combinations of $0ms$, $800ms$, and $1600ms$ delay and 0%, 15%, and 30% bursty packet-loss. We recorded the overall conversation quality and analyzed conversations with parametric conversation analysis. We compare the two different conversation tests, the changes in the conversation structure, and look at interactivity effects between delay and packet-loss.

Related Work

The ITU has standardized subjective evaluation of conversation quality in [11]. Recent research proposed to separate the analysis of the conversation into three phases: the *listening* phase, the *speaking* phase and the *interaction* phase [**?**]. These phases can be evaluated separately to better model the conversational quality [**?**]. Recent work has been analyzing the conversational quality in its different phases over multiple dimensions [13]. The assessment of conversational quality is done via standardized conversation tests like the Short Conversation Test (SCT) or the Random Number Verification test

(RNV) [12, 11].

Transmission delay affects the flow of a conversation due to the delayed arrival of turn-taking signals [4]. However, the degree to which turn-taking and the interactivity of a conversation is degraded additionally depends on the interactivity of the conversation itself [18, **?**]. Parametric Conversation Analysis (P-CA) is a framework to assess the structure of conversations programmatically [4]. With an independent voice activity detection of the two speakers, four conversation states can derived: M ("mutual silence"), D ("double talk"), A ("speaker A") and B ("speaker B") [15, 10]. Based on these four states, interactivity metrics like the speaker alternation rate (SAR), interruption rate (IR), as well as turn-taking information like gaps and overlaps between speaker turns, can be calculated [5, 16]. For delayed conversations, the unintended interruption rates (UIR) measures the number of interruptions that were caused by the delay and were not intended to be interrupting the interlocutor [2].

The effects of packet-loss on VoIP speech transmission have been studied and modeled it the E-model [1]. The effects of packet-loss can be defined by the percentage of packets lost over a given time frame, the length of speech contained in a single packet, the burstiness of the loss, and the codec that is used [6]. When a packet is lost or not transmitted in time, it usually gets replaced by silence. However, current codecs employ packet-loss concealment (PLC) where the lost packet is remodeled given the previous and sometimes next frames [17]. The burstiness of the signal is measured with the burst-ratio that is defined as

$$BurstR = \frac{Average\ length\ of\ observed\ bursts}{Average\ length\ of\ bursts\ with\ random\ loss}$$

in [6]. This behavior can be modeled with a two state Hidden Markov Model [17].

The E-model is the most popular parametric model used for transmission planning. For narrowband transmission has been standardized by the ITU [6] and has since received updates for wideband [7] and fullband [8]. In this model, the terms for impairments (e.g., Id for impairments due to delayed transmission and Ie, eff for codec related impairments like packet-loss) are subtracted from a maximal transmission rating Ro and are thus independent of each other. While the E-model for narrowband telecommunications scenarios accounts for different levels of interactivity, the wideband and fullband E-model do not [7, 8].

Experimental Setup

The experimental setup of our conversational test is based on the recommendation P.805 by the ITU [11]. The experiment was conducted in German, and the participants were located in separate soundproofed rooms and communicated through diotic headsets to simulate a telephone conversation. The mono speech signal was encoded with 16-bit PCM at 44.1 kHz. During the conversations, we introduced three different end-to-end echo-free delay levels of 0, 800, and 1600 ms as well as three packet-loss levels of 0, 15, and 20 %, each with a burst-ratio of 4. We selected these high levels for delay and packet-loss (in contrast to the levels recommended by [9, 6]) because we wanted to explicitly analyze the interactivity effects that occur during severely degraded conversations. Especially the high burst-ratio of 4 was used to incite misunderstandings and repetition of information. The combination of every packet-loss and delay condition results in 9 different overall conditions tested.

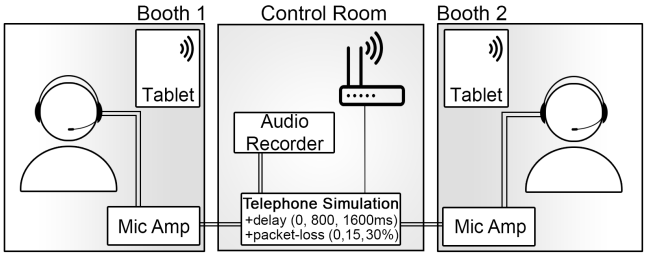

Figure 1: A schematic of the experimental setup. Two participants talking over a simulated telephone network and rating them on tablet devices.

During the experiment, we asked the participants to run through a sequence of short conversation tests and number verification tests. In every experiment, both conversation tests were carried out with all 9 delay and packet-loss conditions. At the beginning of the experiment, the participants carried out an SCT and an RNV scenario with delay or packet-loss to familiarize themselves with the setting. This results in 20 conversations carried out per experiment (2 practice conversations, 9 SCT conversations, and 9 RNV conversations). After each conversation both participants rated the overall quality as well as the 4 listening and 1 interactive dimensions on the extended continuous rating scale (ECS). The ECS was used instead of the traditional 5-point absolute category rating (ACR) scale recommended by [11], because it reduces scale-end effects and is more sensitive [14]. We expected scale-end related issues because of the unusually high delay and packet-loss levels in our experiment design. SCT and RNV scenarios were alternated, with every test beginning with the SCT scenario. However, the order of delay and packet-loss conditions, the caller and receiver roles, as well as the scenario of the conversation tests were randomly chosen for every test.

The setup of the experiment can be seen in Figure 1. The headsets in the two soundproof rooms were connected to the control room through direct audio connections. The ratings for the conversations were displayed and collected with "The Fragebogen" [3] on two Windows tablets that were connected to the experiment computer in the control room via WiFi. The participants used a stylus to fill out the conversation tests and to select the quality on the rating scales.

Results and Discussion

We recruited 58 participants without hearing impairments who were between 18 and 71 years old (mean 32, median 27.5, 28 female). The first language of all partic-

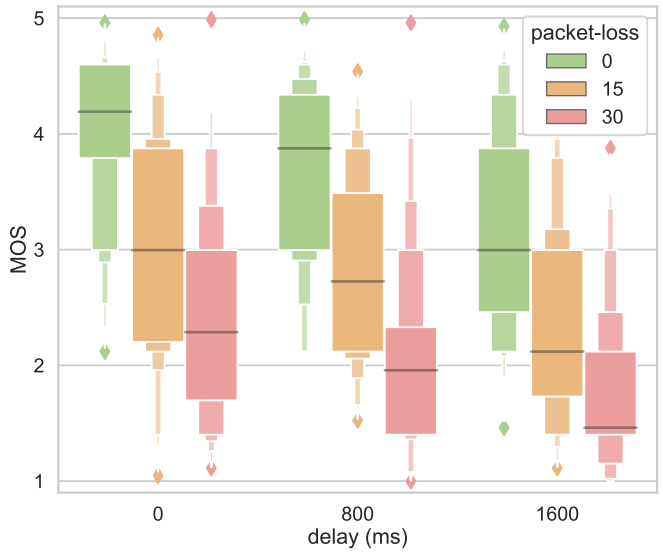

Figure 2: Conversation quality MOS of SCT conversations for 0, 800, and 1600 ms delay and 0, 15, and 30 % packet-loss.

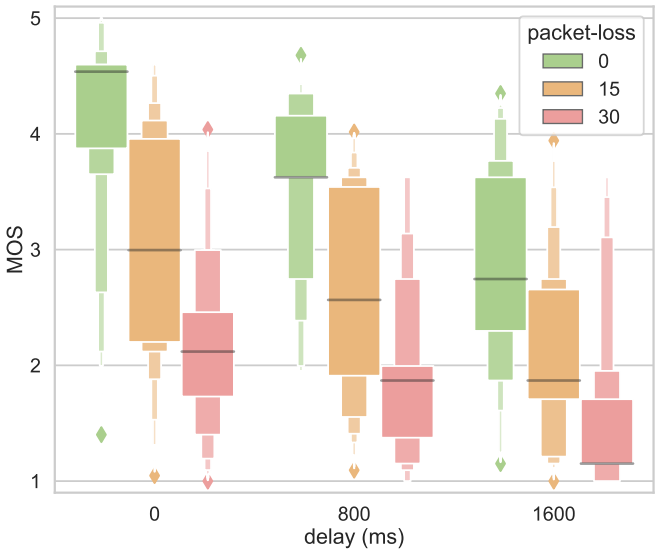

Figure 3: Conversation quality MOS of RNV conversations for 0, 800, and 1600 ms delay and 0, 15, and 30 % packet-loss.

ipants was German. Because three participants always chose roughly the same value on the rating scale for every condition, we removed judgments from participants where the variance of the ACR MOS was less than 0.35. Some quality ratings and conversation files had to be excluded due to technical failures during the test. This results in 938 subjective quality ratings and 534 recorded conversations.

Figure 2 and Figure 3 show the conversational quality ACR MOS for SCT conversations and RNV conversations respectively. As expected and described by previous research (e.g. [2]), the median MOS at 800 and 1600 ms is lower for RNV than for SCT conversations. However, in our results, we see a slightly higher overall sensitivity to delay than in similar studies. The difference in delay sensitivity for the different conversation types holds even for conversations with packet-loss. As expected, the MOS decreases for higher levels of packet-loss. Interestingly, the MOS for conversations with packet-loss are lower for SCT conversations than for RNV conversations. This seems to indicate that during severe amounts of packet-loss, the quality perception of a conversation is influenced by the type of conversation.

Looking at the MOS values for mixed packet-loss and delay conditions, it can be seen that the influence of both packet-loss and delay on the conversation is dependent on the severity of the other degradation. The additional communication necessary when information-carrying packets are lost is again affected by the delayed transmission. We believe that this results in interactions in the perception of packet-loss and delay. We believe that this difference is due to the high density of information that is being transmitted in RNV conversations, compared to SCT conversations.

Figure 4 shows the speaker alternation rate (SAR) of RNV and SCT conversations for the delay conditions. It can be seen that not only is the delay affecting the SAR,

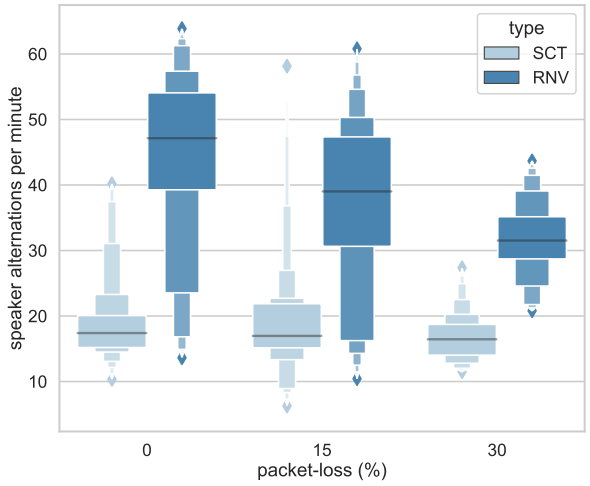

Figure 4: Speaker alternation rate of SCT and RNV conversations for 0, 15, and 30 % packet-loss and without delay.

but also that packet-loss has a substantial influence on it. The influence of packet-loss on the SAR is strong for RNV conversations while only being slightly noticeable for SCT conversations. This seems to confirm our hypothesis that packet-loss has a stronger influence on conversations with higher interactivity and less information context.

Because the RNV scenarios have a high information density transmitted by the participants (i.e., most of the time just a number is uttered), it is not possible to "repair" lost information from context when packet-loss occurs. Thus, the participants need additional communication to re-transmit the missing information. In contrast, the SCT scenarios have less information density (i.e., mostly full sentences with social conversation). When parts of a sentence are affected by packet-loss that can be assumed from context, the participants do not need to clarify and can carry on with the conversation. We believe that this

difference in information density in those two scenarios result in the difference in the conversational structure and the overall perceived quality.

Conclusion

In this paper, we investigated the effects of high levels of delay and packet-loss on a conversation, and it's perceived quality. We confirmed the difference in delay sensitivity between SCT and RNV scenarios, but we also found a slight effect of conversational interactivity on packet-loss sensitivity. We argue that this difference in sensitivity stems not only from the interactivity of the conversation but also from the information density present in the conversation. Additionally, we use the parametric conversation analysis to show that packet-loss affects the structure of the conversation as well. While packet-loss seems to not affect the SAR in SCT conversation, it strongly affects RNV conversations. Again, we believe that this is the difference in information density of the two conversation types results in the varying sensitivity to changes in the SAR.

In future work, we plan to repeat our studies with other types of conversations to investigate our information density hypothesis further. We also plan to annotate parts of the recorded conversation to investigate how the contents of the conversations are changing during high levels of delay and packet-loss. Further, we want to model our findings with a parametric model like the E-model to investigate the possible interactions of the two degradations.

Acknowledgments

This work was financially supported by the German Research Foundation DFG (grant number MO 1038/23-1).

References

[1] L. Ding and R. A. Goubran. Speech quality prediction in voip using the extended e-model. *GLOBECOM '03. IEEE Global Telecommunications Conference (IEEE Cat. No.03CH37489)*, 00(C):3974–3978, 2003.

[2] S. Egger, R. Schatz, and S. Scherer. It takes two to tango-assessing the impact of delay on conversational interactivity on perceived speech quality. In *Eleventh Annual Conference of the International Speech Communication Association*, pages 1321–1324. ISCA, 2010.

[3] D. Guse, H. R. Orefice, G. Reimers, and O. Hohlfeld. Thefragebogen: A web browser-based questionnaire framework for scientific research. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019.

[4] F. Hammer. *Quality Aspects of Packet-Based Interactive Speech Communication*. Forschungszentrum Telekommunikation Wien, 2006.

[5] F. Hammer, P. Reichl, and A. Raake. The well-tempered conversation: interactivity, delay and perceptual VoIP quality. In *IEEE International Conference on Communications*, volume 1, pages 244–249. Institute of Electrical and Electronics Engineers (IEEE), 2005.

[6] ITU-T Recommandation G.107. *The E-model: a computational model for use in transmission planning*. International Telecommunication Union, Geneva, 2011.

[7] ITU-T Recommandation G.107.1. *Wideband E-model*. International Telecommunication Union, Geneva, 2015.

[8] ITU-T Recommandation G.107.2. *Fullband E-model*. International Telecommunication Union, Geneva, 2019.

[9] ITU-T Recommandation G.114. *One-way transmission time*. International Telecommunication Union, 2003.

[10] ITU-T Recommendation P.59. *Artificial Conversational Speech*. International Telecommunication Union, 1993.

[11] ITU-T Recommendation P.805. *Subjective Evaluation of Conversational Quality*. International Telecommunication Union, Geneva, 2007.

[12] N. Kitawaki and K. Itoh. Pure delay effects on speech quality in telecommunications. *IEEE Journal on selected Areas in Communications*, 9(4):586–593, 1991.

[13] F. Köster. *Multidimensional Analysis of Conversational Telephone Speech*. Springer, 2017.

[14] Köster, Friedemann and Guse, Dennis and Wältermann, Marcel and Möller, Sebastian. Comparison between the discrete acr scale and an extended continuous scale for the quality assessment of transmitted speech. *Fortschritte der Akustik-DAGA*, 2015.

[15] H. Lee and C. Un. A study of on-off characteristics of conversational speech. *IEEE Transactions on Communications*, 34(6):630–637, 1986.

[16] R. Lunsford, P. A. Heeman, and E. Rennie. Measuring turn-taking offsets in human-human dialogues. In *Proceedings of INTERSPEECH*, pages 2895–2899, 2016.

[17] A. Raake. Short-and long-term packet loss behavior: towards speech quality prediction for arbitrary loss distributions. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1957–1968, 2006.

[18] A. Raake, K. Schoenenberg, J. Skowronek, and S. Egger. Predicting speech quality based on interactivity and delay. In *Proceedings of INTERSPEECH*, pages 1384–1388, 2013.