INTELLIGIBILITY IN TELEPHONE CONVERSATIONS WITH PACKET LOSS

Thilo Michael

Quality and Usability Lab, Technische Universität Berlin
thilo.michael@tu-berlin.de

Abstract: With the rise of mobile Voice over IP networks, packet loss becomes one of the most prominent degradations experienced by the users of those services. Depending on the probability and burst rate of the packet loss, portions of the speech become unintelligible as more and more information is lost during transmission. In real-world conversations, the intelligibility is not only dependent on the amount of lost signal but also influenced by the amount of contextual information present in the conversation and the importance of the lost part of speech. Additionally, misunderstanding due to lost packets may lead to changes in the conversation's flow and contents. This paper analyzes the intelligibility and misunderstandings in about 200 conversations of two distinct scenarios with three levels of bursty packet loss. We examine how packet-loss changes a conversation's structure in terms of parameters obtained by a parametric conversation analysis. We also discuss how the conversation's content differs between the two types of conversation.

1 Introduction

Mobile telephone services and over-the-top communication services today rely on the Voice over Internet Protocol (VoIP) to transmit speech in real-time. With this protocol, the speech is coded, split into chunks, and sent as packets through a routed network. Because the service providers do not control the quality of the whole network, routing errors may occur. This error may result in the late arrival of coded speech packets, which leads to a transmission delay, where the speech of an interlocutor arrives noticeably late and disturbs the flow of the conversation. Packets that do not arrive at the receiving end (or arrive too late to be useful for decoding the transmitted speech) result in packet loss. Depending on the codec used, the missing information is either replaced with silence or masked with an estimated signal through packet loss concealment (PLC). During packet loss bursts, where multiple consecutive packets are not arriving on time, the codec cannot ensure proper concealment, and the speech drops out, rendering it unintelligible.

The quality of service (QoS) for communication networks is usually assessed in listening scenarios, where participants listen to degraded speech samples and rate them on a 5-point absolute category rating (ACR) scale. These ratings are then averaged into a mean opinion score (MOS). With this method, the impact of packet loss on the perceived quality can be quantified, and the intelligibility of the degraded speech can be measured. Predictive models such as the Articulation Index (AI) and Speech Intelligibility Index (SII) can estimate the amount of usable audio cues present in the degraded signal which for a prediction on how likely the speech is intelligible [1]. Signal-based models like POLQA can estimate the MOS of speech samples based on the degraded and a reference speech signal [2].

However, in conversation scenarios, degradations like packet loss do not only impact the listening quality. Also, the speech intelligibility is influenced not only by the portion of speech

lost during transmission. Depending on which part word of an utterance was rendered unintelligible, the conversation may not be impacted at all (e.g., when the packet loss occurred during a silent part of the conversation). However, if the speech hinders the understanding of the utterance, the user may need to ask for the information to be retransmitted by the interlocutor, thus changing the course of the conversation. Additionally, the amount of contextual information available to the interlocutors influences the likelihood of misunderstandings. For example, in a casual conversation, not every word may be necessary to understand the meaning of a sentence, and interlocutors may reconstruct lost words given the conversation context. In contrast, while transmitting credit card information, every word contains information that needs to be understood, and a packet loss is more likely to disrupt the conversation.

While the effects of packet loss on the perceived quality and the intelligibility have been a focus of research for a long time, the impact of packet loss on the conversation structure and contents has not received much attention. With more insights into how packet loss affects a conversation's structure, conclusions on how it might interact with other degradations can be drawn. In this paper, we perform an initial analysis of the impact of packet loss on the intelligibility and thus the structure of a conversation. For that, we analyze recorded conversations of different conversational interactivity that are degraded by varying amount of bursty packet loss. We examine how packet-loss changes the structure of a conversation in terms of parameters obtained by a parametric conversation analysis (P-CA) and how the type of conversation influences the likelihood of a misunderstanding occurring.

2 Related Work

The Internation Telecommunications Union (ITU-T) has standardized the subjective evaluation of conversation quality in ITU-T Recommendation P.805 [3]. The conversational quality assessment is done via standardized conversation tests, where two participants converse with each other over a simulated telephone line and talk about topics given by a conversation scenario. It has been shown that the impact and perception of some degradations are influenced by the interactivity of the conversation [4, 5]. In order to evaluate those dependencies, conversation tests with distinct levels of conversational interactivity (CI) have been standardized. One prominent conversation test with a high CI is the random number verification (RNV) test, where participants alternatingly exchange a list of numbers organized in 4 blocks [6]. An example of a conversation test with low CI is the short conversation test (SCT), where participants solve real-world tasks like ordering pizza or booking a flight [3].

Parametric Conversation Analysis (P-CA) is a framework to assess the structure of conversations programmatically [7]. With an independent voice activity detection of the two speakers, four conversation states can be derived: *M* ("mutual silence"), *D* ("double talk"), *A* ("speaker A") and *B* ("speaker B") [8]. 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 [9]. 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 [10].

The effects of packet loss on VoIP speech transmission have been studied and modeled it the E-model [11]. 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 [12]. 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 [13].

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 [12]. This behavior can be modeled with a two-state Hidden Markov Model [13].

The Speech Intelligibility Index (SII) and its predecessor Articulation Index (AI) are standardized measures that have a high correlation with the intelligibility of speech in a listening situation [1]. The SII itself is not a measure of how likely a spoken sentence is understood, but instead on how many audio cues are usable in a given setting [14]. The SII uses frequency-specific information of the speech levels, the "noise" levels, and their auditory thresholds, which is weighted by the importance of each frequency band in regards to speech understanding. The resulting index can be transformed into speech understanding scores with the help of transfer functions. These functions are specific to the material that is listened to, as unknown random syllables and previously known full sentences have different chances of being understood [14].

3 Experimental Data

The conversation data used for this paper is published in [15] and follows the ITU-T Recommendation P.805 for the subjective evaluation of conversational quality [3]. The experiment was conducted in German, the participants were located in separate soundproofed rooms, and they 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 zero-insertion packet loss levels of 0, 15, and 30 %, each with a burst-ratio of 4. We selected a high burst-ratio of 4 to incite misunderstandings and repetition of information. The participants carried out short conversation tests as well as random number verification tests for each of the three packet loss levels. The two participants in each experiment were recorded on different channels, and the degraded, as well as the clean speech, were stored for later analysis. In summary, the reference and degraded recordings of 200 conversations were used for the analysis.

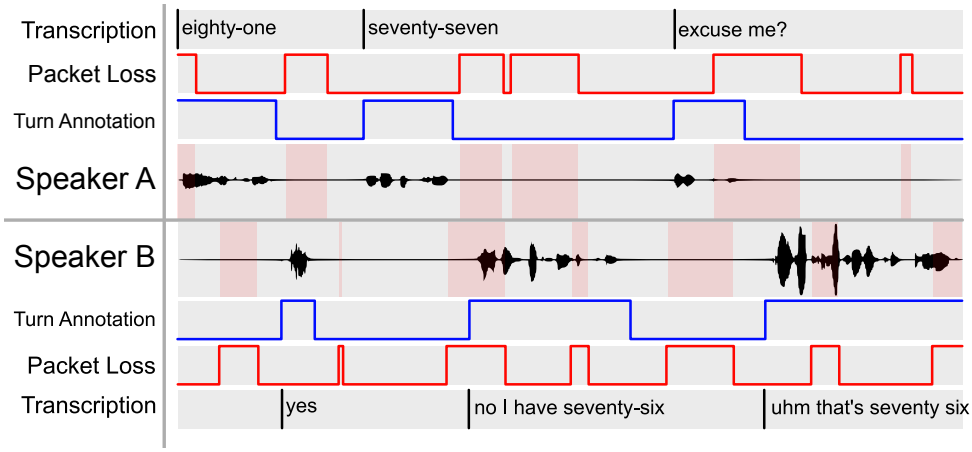

Figure 1 – Exemplary overview of information available for a part of a conversation. The speech of the interlocutors are recorded on separate channel, the turns are annotated in blue and the packet loss pattern is shown in red. The turns are transcribed[1] for each turn, but not force-aligned.

The occurrences of packet loss in each conversation were extracted for further analysis. For this, we used the packet loss generation algorithm used in the experimental setup to reproduce the exact timecodes at which the speech signal drops out. We then used automatic voice

activity detection on the reference signal to segment each conversation into the turns based on Lunsford et al. [16]. We then automatically transcribed these turns with Google Automatic Speech Recognition. A short analysis of four conversations showed a word error rate (WER) of around 27 % and around 7 % of mostly short turns that could not be transcribed. For each turn in the conversation, we utilized the information about the locations of the packet loss to determine the percentage of lost speech for each turn. An exemplary conversation with this annotated data[1] is shown in Figure 1.

To analyze the effects of the packet loss on the intelligibility and thus on the contents and flow of a conversation, we used the transcriptions to find occurrences of "*conversation disruptions*". As a conversation disruption, we define every turn where a participant has to ask their interlocutor for the repetition of information, independent of what caused that misunderstanding to happen. Thus, only instances where a loss of speech signal is so significant that the conversation cannot be continued without intervention is counted as a *conversation disruption*. However, the real reason for a participant asking for a repetition of information cannot be inferred from the data alone. Also, misunderstandings unrelated to packet loss may be counted as a disruption.

To detect these disruptions automatically, we identified several key phrases in the data like "*excuse me?*" and "*I did not understand*". When then automatically marked each turn as a *conversation disruptions* that matched one of these key phrases.

4 Results and Discussion

From the annotated data, we extracted the state probabilities and sojourn times of the four conversational states, as well as the speaker alternation rate as part of the parametric conversation analysis and the *conversation disruptions* as part of the semantic analysis.

4.1 Parametric Conversation Analysis

To analyze the structure and turn-taking characteristics of the conversation, we performed a parametric conversation analysis to extract the state probabilities of the four states and calculate the speaker alternation rate as well as the number and length of turns.

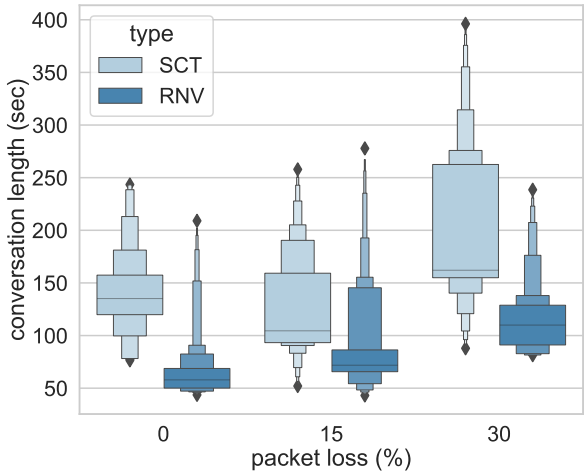

Figure 2 – Length of the conversations in seconds at 0, 15, and 30 % packet loss for SCT and RNV conversations.

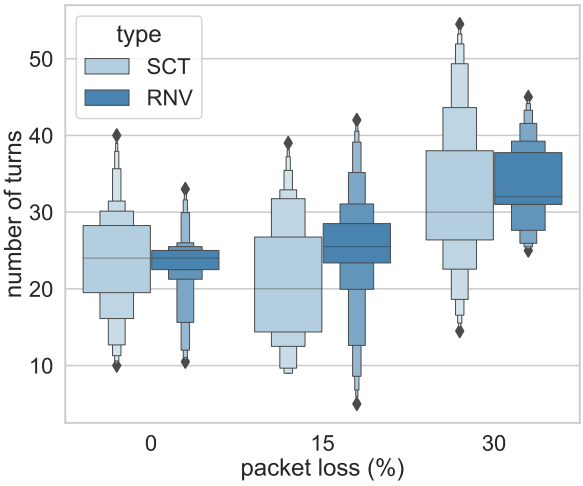

Figure 3 – Number of turns per conversation at 0, 15, and 30 % packet-loss for SCT and RNV conversations.

[1] All transcriptions are translated into English for this paper and are originally in German.

Table 1 – Median length and turn count and turn length for the SCT and RNV conversation at 0, 15 and 30 % packet loss and increase of length and turn count relative to the 0 % packet loss condition.

Scenario	Packet Loss	Median length (s)	increase in %	Median turn count	increase in %	Median turn length (s)
SCT	0 %	135.15	-	24	-	1.25
	15 %	104.50	−22.68	20	−16.67	1.17
	30 %	162.05	19.90	30	25.00	1.29
RNV	0 %	57.90	-	25	-	0.63
	15 %	71.82	24.04	26	4.00	0.61
	30 %	109.95	89.90	32	23.08	0.75

Figure 2 shows the median and distribution of the conversation lengths for the SCT and RNV conversations at 0, 15, and 30 % packet loss. While the length of SCT conversation is similar for 0 and 15 % packet loss, the length increases for the 30 % packet loss condition. While RNV conversations generally produce shorter conversations, their length also increases with higher packet loss levels.

This increase in conversation length is also present in the number of turns per conversation (Figure 3), and expectedly, a drop in the number of turns is also present for SCT conversation at 15 % packet loss. While the data does not explain the drop in the number of turns and overall conversation length, the transcriptions seem to indicate that participants tend to leave out small-talk from the conversation when a degradation disrupts the conversation. However, this hypothesis needs to be investigated further.

The overall increase in conversation length and turn count for the higher packet-loss levels indicates that packet loss causes conversation disruptions, which need additional turns to fix and thus prolong the conversation. Table 1 shows the median values for length and turn count and also the increase relative to the 0 % packet loss condition. While the relative changes of the turn count seem to match the relative change in the conversation length, the meadian length of RNV conversations is almost twice as long (90 %) at the 30 % packet loss level. In contrast to that, the increase in turn count is only 23.08 %. Because the length of each turn is similar for each packet loss condition, the disproportionate increase in conversation length cannot be attributed solely to the increase in the length of the utterances.

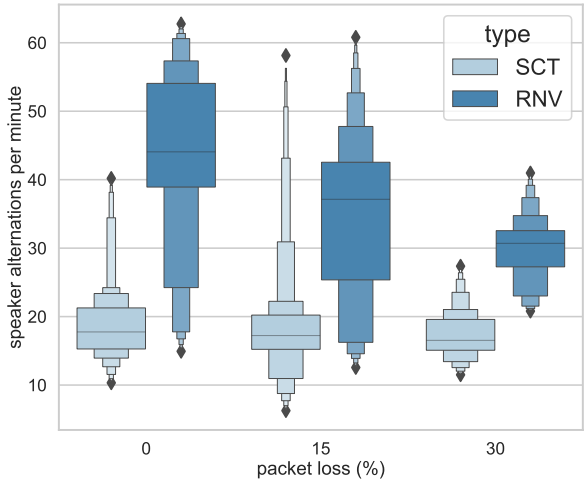

Figure 4 – Speaker alternations per minute for SCT and RNV conversations at 0, 15, and 30 % packet loss.

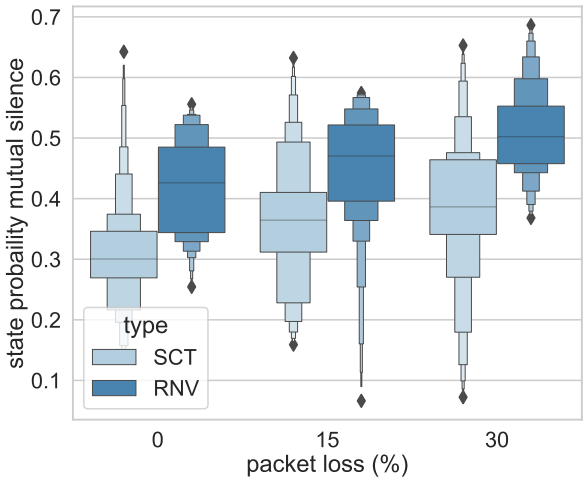

Figure 5 – State probability of mutual silence (M) for SCT and RNV conversations at 0, 15, and 30 % packet loss.

Figure 4 shows that the increase in length stems from the fact that the RNV conversations

are slowed down by the increasing packet loss, while the SCT conversation stays at the same level of speaker alternation rate. Figure 5 shows that the lower speaker alternation rate is mainly caused by an increase in silence. This indicates that the high interactivity scenario (i.e., the rapid exchange of numbers) has an influence on how much packet loss impacts the conversation. One reason for the difference in the two scenarios might be the density of information in each utterance. While the sentences in SCT conversations tend to be longer, with relatively few words important for the understanding of the conversation, the utterances in RNV conversations mostly consist of only the information that needs to be transmitted in order to advance the conversation.

4.2 Conversation Disruptions

To analyze how conversation disruptions are influenced by packet loss, we calculated the number of disruptions per minute for every conversation.

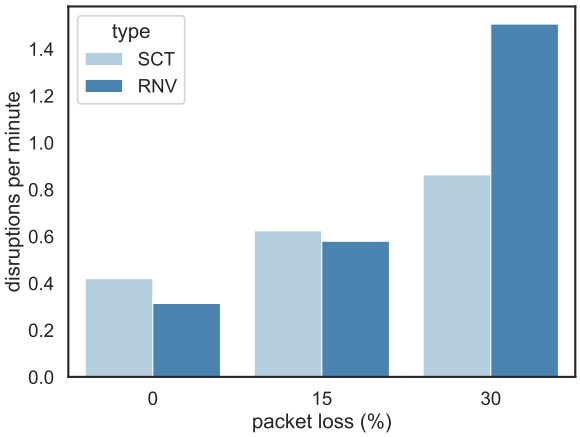

Figure 6 – Conversation disruptions per minute for SCT and RNV conversations at 0, 15, and 30 % packet loss.

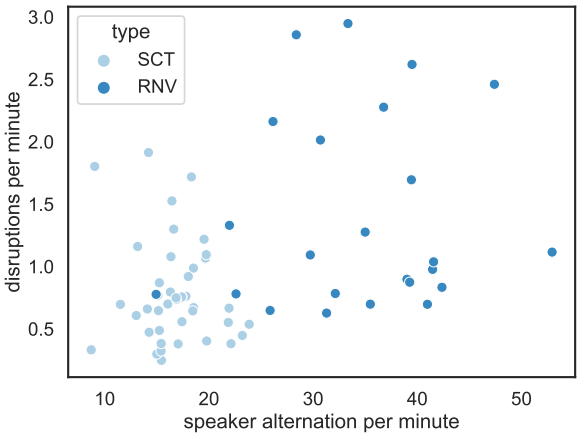

Figure 7 – Speaker alternation rate versus the discruptions per minute for SCT and RNV conversations.

Figure 6 shows the expected increase in conversation disruption with higher packet loss levels. While the disruptions with no packet loss are lower for RNV conversations than for SCT scenarios, the disruption rate is much higher at 30 % packet loss. Again, this may be explained by the information density present in the RNV conversations, paired with the high speaker alternation rate of this type of conversation. Figure 7 shows that the disruption rate has a weak positive correlation with the speaker alternation rate ($r = 0.38$, $p < 0.05$). This also suggests that the percentage and burst ratio of packet loss as well as the conversational interactivity influence the conversation disruptions.

5 Conclusion

In this paper, we performed a preliminary analysis of conversational data with three levels of packet loss and a high burst ratio. We defined a set of misunderstandings that lead to retransmission of information as *conversation disruptions* and showed that these disruptions increase with higher packet loss levels and are impacting conversations differently, depending on the interactivity of the conversation. We show that these additional interactions caused by the packet loss lead to more turns and more lengthy conversations. However, we also show that the increase in conversation disruptions alone cannot account for the large increase in conversation length of the more interactive RNV conversations. Additionally, a change in the participants' interactive behavior occurred with higher packet loss, but only for the RNV conversations. We believe that

the more pronounced increase in conversation disruptions in RNV conversations compared to SCT scenarios stem from the fact that the density of transmitted information is higher for RNV conversations. When participants transmit the numbers over the telephone line with the goal of being fast, the turns tend only to contain the number. When a packet loss burst happens during that time, a misunderstanding is much more likely. In contrast, the full sentences in SCT conversation have a much lower information density. Thus a packet loss burst has to occur during the transmission of words that are crucial to understanding the intention of the interlocutor.

These results give insight into how packet loss and resulting misunderstandings affect a conversation in realistic conversational scenarios and thus the quality perception. The increase in length due to the conversation disruptions may lead to new disruptions. Interestingly, because transmission delay reduces the interactivity of a conversation as well, the combination of these two degradations might change the quality perception more than the sum of the quality impairment of each degradation on its own.

In future work, we plan to transcribe the data manually to increase the quality of the conversation disruption annotations. With these new annotations, we will analyze the utterances that caused the disruptions (the preceding tuns) to understand how much of the speech signal was lost and also which words were affected during misunderstood turns. We will also investigate how the degradations delay and packet loss affect each other in a conversational scenario and how this affects the perceived quality.

References

[1] AMERICAN NATIONAL STANDARDS INSTITUTE: *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America, 1997.

[2] ITU-T RECOMMENDATION P.863: *Perceptual objective listening quality assessment*. International Telecommunication Union, 2014.

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

[4] RAAKE, A., K. SCHOENENBERG, J. SKOWRONEK, and S. EGGER: *Predicting speech quality based on interactivity and delay*. In *Proceedings of INTERSPEECH*, pp. 1384–1388. 2013.

[5] EGGER, S., R. SCHATZ, K. SCHOENENBERG, A. RAAKE, and G. KUBIN: *Same but different? — Using speech signal features for comparing conversational VoIP quality studies*. In *IEEE International Conference on Communications (ICC)*, pp. 1320–1324. IEEE, 2012.

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

[7] HAMMER, F.: *Quality Aspects of Packet-Based Interactive Speech Communication*. Forschungszentrum Telekommunikation Wien, 2006.

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

[9] REICHL, P. and F. HAMMER: *Hot discussion or frosty dialogue? Towards a temperature metric for conversational interactivity*. In *Eighth International Conference on Spoken Language Processing*. 2004.

[10] EGGER, S., 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*, pp. 1321–1324. ISCA, 2010.

[11] DING, L. and R. A. GOUBRAN: *Speech quality prediction in voip using the extended e-model. GLOBECOM '03. IEEE Global Telecommunications Conference*, pp. 3974–3978, 2003.

[12] ITU-T RECOMMANDATION G.107: *The E-model: a computational model for use in transmission planning*. International Telecommunication Union, Geneva, 2015. URL `http://handle.itu.int/11.1002/1000/12505`.

[13] RAAKE, A.: *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), pp. 1957–1968, 2006.

[14] HORNSBY, B.: *The Speech Intelligibility Index: What is it and what's it good for? The Hearing Journal*, 57(10), pp. 10–17, 2004.

[15] MICHAEL, T. and S. MÖLLER: *Effects of Delay and Packet-Loss on the Conversational Quality. Fortschritte der Akustik-DAGA*, pp. 945–948, 2020.

[16] LUNSFORD, R., P. A. HEEMAN, and E. RENNIE: *Measuring turn-taking offsets in human-human dialogues*. In *Proceedings of INTERSPEECH*, pp. 2895–2899. 2016.