LEXICAL FREQUENCY AND LISTENER'S RESPONSE TO PACKET LOSS IN TELEPHONE CONVERSATIONS

Thilo Michael[1], *Omnia Ibrahim*[2]

[1]*Quality and Usability Lab, Technische Universität Berlin*
[2]*Language Science and Technology, Saarland University*
thilo.michael@tu-berlin.de

Abstract: In today's Voice-over-IP (VoIP) telephony, packet loss is one of the most prominent degradations. Severely bursty packet loss can lead to multiple consecutive packets and thus important parts of the transmitted speech to be lost. In listening-only tests, the understandability of packet loss-affected speech can be modeled based on the available audio cues in the signal. However, in real conversation scenarios, not every unintelligible word is important for the continuation of the conversation. Thus, even utterances that are largely affected by packet loss might not lead to a request of retransmission of the information (e.g., "Could you please repeat that?") and thus disruption of the conversation flow. lexical frequency can be used as a tool to measure the importance of the information transmitted. In this paper, we analyzed a set of 84 packet loss degraded telephone conversations and investigated the ability of listeners to recover from missing words resulting from the packet loss as a function of their frequency. We found that the request for information retransmission appears more often for messages with less frequent words.

1 Introduction

Since the introduction of Voice-over-IP telephony (VoIP), packet loss has become one of the main degradation experienced by the users of these services. While the audible effects of packet loss have been studied and modeled in detail, the intelligibility of packet loss-affected speech has been studied mostly in listening-only scenarios, where only the amount of perceivable content is considered. However, in interactive conversational scenarios, the audible effects are only part of the quality judgment. Utterances that have been rendered unintelligible by packet loss may lead to requests for retransmission (e.g., "*Could you please repeat that*"), which alter the flow of the conversation. However, not every lost part of an utterance leads to a disruption of the conversational flow. For example, users may not notice a packet loss if it occurs during silence. Also, a word that has been rendered unintelligible by packet loss might not be strictly necessary for the continuation of the conversation. In this case, the continuation of the dialogue may be prioritized over the intelligibility of every word.

Recent research has shown that given the same severeness of packet loss, the amount of these conversation disruptions vary between different conversational scenarios [1]. While the frequency of these disruptions doesn't directly correlate with the overall quality judgment on the conversation, the increase in conversation disruption introduces more speaker turns and longer utterances into the conversation, which can lead to interactivity effects with other degradations.

While the decision to disrupt the conversation and ask for the retransmission of the lost information is based on whether the unintelligible part of the speech is important for the goal of the conversation, it is not trivial to predict whether a conversation disruption occurs based on the

missing parts of the utterance. One way to measure the importance of the information transmitted is the lexical unigram surprisal or word frequency [2]. Surprisal captures the intuition that highly predictable expressions carry less information than predictable ones. It is defined as the probability of a linguistic unit (e.g., phoneme, syllable, word, etc.) to occur in specific context and expresses the amount of information that is conveyed in terms of bits [3]. Equation 1 shows the formula for calculating surprisal S of a linguistic unit based on the probability P of the unit given its context.

$$S(unit_i) = -log_2 P(unit_i|Context) \tag{1}$$

In this paper, we investigate the connection between conversation disruptions (e.g., "*I did not understand.*" or "*Could you please repeat that?*") and the lexical unigram surprisal (i.e., based on the probability of the unit without context) of the packet loss-affected words in the preceding turn in a corpus of 84 telephone conversations with varying degrees of zero-insertion packet loss. Unigram surprisal is measured as the negative log probability of the missing word occurrence. We transcribed the conversations and added markers in the position of the utterances where packet loss occurs and also which turns resulted in a conversation disruption. Based on a German-language model, we calculated the word probability and the resulting unigram surprisal of missing words that resulted in a conversation disruption and missing words that did not.

2 Related Work

The subjective evaluation of conversation quality is standardized by the International Telecommunication Union (ITU-T) in the ITU-T Recommendation P.805 [4]. In this recommendation, the assessment of the conversational quality via conversation tests is standardized, in which two participants converse with each other over a simulated telephone line and talk about topics given by a conversation scenario. For example, the Short Conversation Test (SCT) defines multiple everyday telephone tasks (e.g., ordering a pizza or booking a hotel room), and the Random Number Verification test consists of blocks of numbers that have to be matched with the numbers given to the respective interlocutor. Recent work showed that different conversational scenarios (SCT and RNV conversations) have a different amount of conversation disruptions [1]. While packet loss may disrupt the course of the conversation, the main research on the effects of packet loss on perceived quality has mostly been performed in listening-only tests [5, 6, 7]. Models that predict the listening quality of packet loss-affected speech include ITU-T Recommendation P.862 (PESQ) [8], ITU-T Recommendation P.863 [9] and the multidimensional prediction model NISQA [10]. The only parametric model to predict *conversational* quality standardized by the ITU-T is the E-model [11]. The most recent version for super-wideband and fullband communication scenarios, the fullband E-model [12], calculates the codec-related degradation based on the packet loss probability Ppl of the connection and the packet loss robustness factor Bpl of the codec used. However, the model does not include parameters for the information density of the conversation and thus does not take into account how likely conversation disruptions are to affect the flow of the conversation.

The Speech Intelligibility Index (SII), a model of the intelligibility of speech, is standardized by the American National Standards Institute (ANSI) in [13]. The SII itself is not a measure of how likely an utterance is understood by the listener, but it rather measures how many audio cures are usable in a given setting [14]. However, with a transfer function, the SII can be transformed into a speech understanding score. The transfer functions are specific to the material that is listened to. For example, unknown random syllables and previously known full sentences have a different chance of being understood and thus need different transfer functions [14].

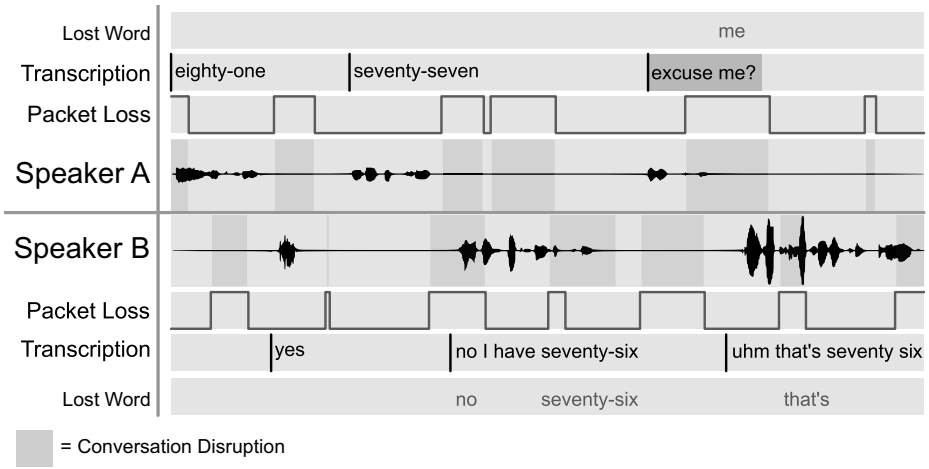

Figure 1 – Exemplary overview of an annotated conversation. The speech is recorded in separate channels, the packet loss pattern is shown in red, the conversation is transcribed[1] and force aligned. Conversation Disruptions are marked in blue and words where more than 50 % of phonemes were affected by packet loss are annotated in red.

Lexical surprisal is a measure to quantify the amount of information conveyed in a message [15]. It has been previously used to predict reading time [16] or as an indicator for sentence comprehension [17]. Recent work investigated the difference of syllable duration and intensity range of speech affected background noise [18].

3 Data Processing and Analysis

The conversational data used for the surprisal analysis in this paper is based on an experiment [19] following the ITU-T Recommendation P.805 for the subjective evaluation of conversational quality [4]. The participants of the conversation test 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, and the experiment was conducted in German. During the conversations, we used the telephone simulation to introduce three different zero-insertion packet loss levels of 0, 15, and 30 %, each with a burst-ratio of 4. We selected this high burst ratio to incite conversation disruptions. The participants carried out SCT as well as RNV conversation scenarios for each of the three packet loss levels. The speech of the two participants in each experiment was recorded on different audio channels, and the degraded, as well as the clean speech, were stored for later analysis. For the analysis, we used 84 of these conversations.

Figure 1 shows an exemplary overview of an annotated conversation[1]. We manually transcribed the conversations and force-aligned the transcription using the webMAUS force-alignment service [20]. Based on the transcriptions, we located conversation disruptions (i.e., turns where the participants indicated that they did not understand something in the previous turn) and annotated both the disrupting turn and the turn that caused the disruption.

The packet loss patterns (red lines in Figure 1), that were introduced during the conversation experiment were generated with a two-state Markov model [6]. From there, we calculated the exact speech frames that were removed from the transmission and used this information to annotate for each phoneme if it was lost or not. Here, we only marked a phoneme as lost if all

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

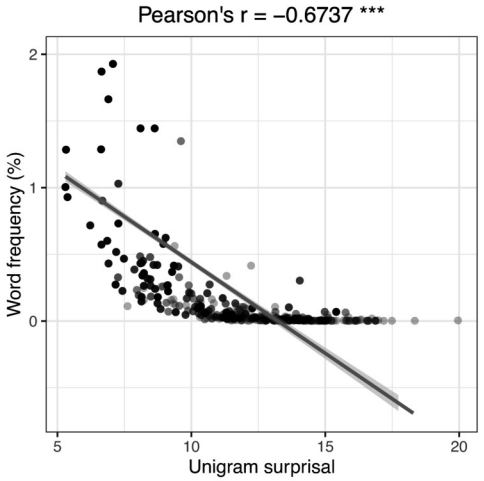

Figure 2 – Pearson's correlation coefficients between the word frequency and the unigram surprisal

of the corresponding speech signals were lost in the packet loss burst. With this information, we marked every word as lost in which more than 50 % of the phonemes were affected by packet loss (red words in Figure 1).

For the analysis, we extracted only words that were annotated as lost and divided them on whether the utterances in which they were said produced a conversation disruption in the next turn. To annotate the lexical frequency of each word, we used two sources for word frequency. The first source is a German frequency list[2],which has been generated in 2009 from TV and movie subtitles with a total of 25,399,099 words. The second source is the language model "de_core_ news_lg" provided by the spaCy library.[3] This language model consists of over 500,000 unique word vectors, each with the lexical frequency included. From the probabilities of each word, we then calculated the lexical unigram surprisal based on Equation 1, assuming no context.

Figure 2 shows the correlation between the unigram surprisal calculated from the language model and the respective word frequency extracted from the frequency list inside the corpus. The words that occur with less frequency tend to have a higher surprisal based on the language model used. The Pearson's correlation shows a strong negative correlation between the word frequency and the unigram surprisal.

For the statistical analysis, a generalized linear mixed model was used to evaluate the effects of unigram surprisal and packet loss type (15 % vs. 30 %) on listeners responses using the R lmer package [21]. We treated speakers and conversations as random effects.

4 Results & Discussion

Based on the fact that less frequent words in our corpus have a higher surprisal, we investigated the difference in word frequency and surprisal between packet loss-affected words that were part of an utterance that produced a conversation disruption and packet loss-affected words that didn't. Figure 3 shows that utterances that caused a conversation disruption (here "misunderstanding") have on average words with lower word frequency and higher unigram surprisal.

[2]https://en.wiktionary.org/wiki/User:Matthias_Buchmeier/German_frequency_list-1-5000
[3]https://github.com/explosion/spacy-models/releases/tag/de_core_news_lg-3.2.0

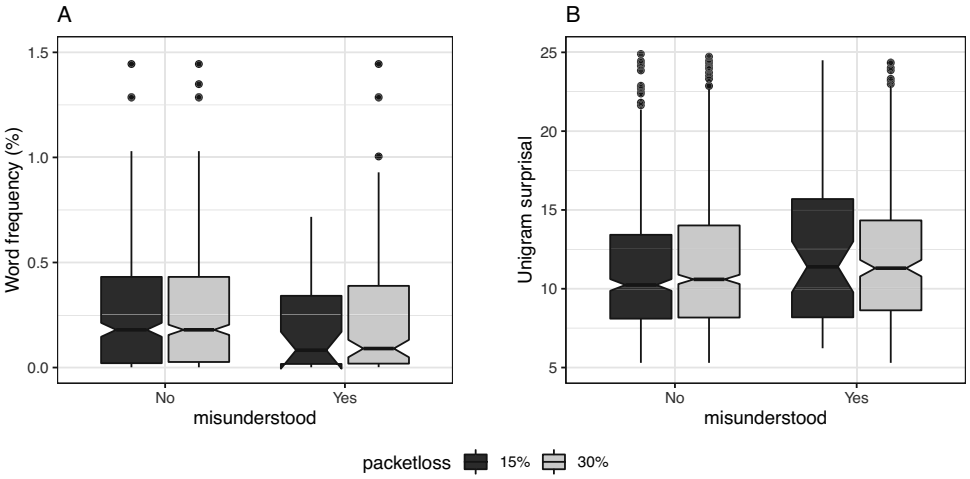

Figure 3 – Listeners' response as a function of word frequency (left) and unigram surprisal (right)

While for utterances that did not cause a misunderstanding, the word frequencies of conversations with 15 and 30 % packet loss are very similar, the word frequency for conversations with 15 % is lower than with 30 % for misunderstood turns. This might indicate that for the high level of 30 % packet loss, the information content of the missing words is less relevant for the decision on whether to disrupt the conversation because so little of the original speech is transmitted. Statistically, we found a significant effect of both surprisal and packet loss type on the listener's response but without interaction (Table 1). Missing words with higher unigram surprisal elicit conversation disruptions. This indicates that the listeners were able to recover from the packet loss when the missing words are highly frequent in the language.

Table 1 – Glmer model for unigram surprisal: glmer(misunderstood ~ surprisal+ packetloss + (1 | Speaker)+ (1 | convID), family = binomial). Number of observations =1918. The reference level for packet loss is 15 %.

	Estimate	**Std. error**	***t***	**Pr (>|t|)**
Intercept	-3.517	0.00104	-3369.35	<0.0001 ***
surprisal	0.032	0.00104	31.31	**<0.0001*** **
packetloss 30%	1.437	0.00104	1377.25	**<0.0001*** **

5 Conclusion

In this paper, we presented the first analysis on the lexical frequency and unigram surprisal of words affected by packet loss. We showed that missing words in utterances that produced a conversation disruption (i.e., have been misunderstood by the listener) have a lower frequency and higher surprisal than missing words in utterances that did not produce a disruption. These results show that the understandability of packet loss-affected speech depends not only on the amount of speech signal that is lost but also on the amount of information contained in the missing speech.

In future work we plan on investigating whether lexical surprisal with more linguistic context (n-gram surprisal) can be a better predictor for conversation disruptions. We also plan to model understandability and conversation disruption based on the n-gram surprisal of the

missing words. While the current paper focus on listener's response to packet loss, one future direction is to investigate the speaker's strategies to recover from the packet loss by linguistically and phonetically comparing the original utterance to the repeated/clearer version.

Acknowledgements

This joint work was financially supported by the German Federal Ministry of Education and Research through Software Campus grant 01IS17052 (QUASIKO), and by the the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

[1] MICHAEL, T.: *Intelligibility in Telephone Conversations with Packet Loss*. In *32. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, pp. 311–318. TUDpress, Dresden, 2021.

[2] ERNESTUS, M.: *Acoustic reduction and the roles of abstractions and exemplars in speech processing*. *Lingua*, 142, pp. 27–41, 2014.

[3] SHANNON, C. E.: *A mathematical theory of communication*. *Bell Syst. Tech. J.*, 27(3), pp. 379–423, 1948. URL `http://dblp.uni-trier.de/db/journals/bstj/bstj27.html#Shannon48`.

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

[5] CLARK, A. D., P. D. F. IEE ET AL.: *Modeling the effects of burst packet loss and recency on subjective voice quality*. In *Proceedings of IP Telephony Workshop*. Citeseer, 2001.

[6] 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.

[7] DING, L. 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), pp. 3974–3978, 2003. doi:10.1109/GLOCOM.2003.1258975. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1258975`.

[8] ITU-T RECOMMANDATION P.862: *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. International Telecommunication Union, 2001.

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

[10] MITTAG, G., B. NADERI, A. CHEHADI, and S. MÖLLER: *Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets*. *Interspeech 2021*, 2021. doi:10.21437/interspeech.2021-299. URL `http://dx.doi.org/10.21437/Interspeech.2021-299`.

[11] 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`.

[12] ITU-T RECOMMANDATION G.107.2: *Fullband E-model*. International Telecommunication Union, Geneva, 2019.

[13] INSTITUTE, A. N. S.: *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America, 1997.

[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] HALE, J.: *Information-theoretical complexity metrics. Language and Linguistics Compass*, 10(9), pp. 397–412, 2016.

[16] MONSALVE, I. F., S. L. FRANK, and G. VIGLIOCCO: *Lexical surprisal as a general predictor of reading time*. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 398–408. 2012.

[17] LEVY, R.: *Memory and surprisal in human sentence comprehension.* 2013.

[18] IBRAHIM, O., I. YUEN, M. VAN OS, B. ANDREEVA, and B. MÖBIUS: *The effect of Lombard speech modifications in different information density contexts*. In *32. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, pp. 185–191. TUDpress, Dresden, 2021.

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

[20] KISLER, T., U. REICHEL, and F. SCHIEL: *Multilingual processing of speech via web services*. *Computer Speech & Language*, 45, pp. 326–347, 2017.

[21] BATES, D., M. MÄCHLER, B. BOLKER, and S. WALKER: *Fitting linear mixed-effects models using lme4*. *Journal of Statistical Software*, 67(1), pp. 1–48, 2015. doi:10.18637/jss.v067.i01.