It Takes Two to Tango - Assessing the Impact of Delay on Conversational Interactivity on Perceived Speech Quality

Sebastian Egger, Raimund Schatz, Stefan Scherer

1Telecommunications Research Center, Vienna, Austria
2Institute of Neural Information Processing, University of Ulm, Germany

egger@ftw.at, schatz@ftw.at, stefan.scherer@uni-ulm.de

Abstract

This paper analyzes the relationship between transmission delay, conversational interactivity and perceived quality of bidirectional speech. Our work is grounded on the results of subjective speech quality tests conducted in our lab and recent studies in this field. The test experiments do not only quantify the impact of network delay on speech quality as perceived by untrained subjects. They also assess the mutual influences between conversational interactivity (CI) and delay using three different conversation scenarios.

Our results show a clear positive correlation between the level of conversational interactivity and interlocutors’ delay sensitivity. Another key finding is that even in contexts of high interactivity, one-way delay values up to 400 ms did not have any significant impact on untrained participants’ perception of overall speech quality. Furthermore, we examine the surface structure of participants’ conversations across a wide range of delay conditions (up to 1600 ms). Our analysis demonstrates how additional metrics such as unintended interruption rate (UIR) can be successfully used to determine the surface structure and delay sensitivity of a conversation.

Index Terms: Conversational interactivity, Delay, QoE

1. Introduction

With the current evolution of telecommunication networks towards packet-based transmission and the related widespread introduction of voice over IP (VoIP) services, measurement and management of speech quality has become a substantial challenge for network operators. However, while the impact of degradations such as packet-loss, jitter, noise, and codec distortions on user-perceived voice quality have been extensively addressed by existing research, the impact of network delay has received much less attention. In particular, this concerns the influence of the actual structure and content of the communication on delay perception.

The aim of this paper is to address this issue by providing a reliable evaluation of how transmission delay affects Quality of Experience (QoE) of voice communications in different controlled conversational contexts. In the light of existing research and standardization efforts in the domain of voice QoE, we present our lab-study results on the delay sensitivity of untrained subjects and analyze the relationship between the impact of delay, interactivity and other conversational parameters. To this end, Section 2 reviews existing literature in the field of delay testing and extraction of conversational parameters, arguing for the necessity of pairwise lab experiments on delay and the introduction of new conversational metrics. Section 3 describes our experimental setup used for voice QoE assessment. The results of our study as well as the underlying conversational analysis are then presented in Section 4. Finally, Section 5 discusses our conclusions and recommendations as well as planned future work.

2. Theory and Related Work

2.1. Impact of Delay and Conversational Interactivity on Perceived Quality

The influence of delay in conversational situations has been investigated by several studies. Seminal work on the effect of delay on perceived quality was performed by Kitawaki [1] who assessed delay detectability thresholds of study participants and measured the perceived quality of the conversations. Although the detectability threshold of one-way delay1 of untrained subjects was up to 560 ms, the quality degradations perceived by the subjects were pronounced even for moderate transmission delays. In this respect, their results are in line with the recommendations from ITU-T G.114 [2]. G.114 states that voice communication systems should have a one-way delay of maximally 150 ms, with delays above 400 ms being totally unacceptable. However, more recent studies that assess delay influences on perceived speech quality, report a substantially lower quality impact of transmission delays up to 600 ms (cf., [3, 4, 5, 6]). In order to examine delay ranges where severe degradations should occur we decided to expose users to one-way delays up to 1600 ms. Together with the methods described in Section 2.2 we aimed for measuring features that identify the irritations caused by delay artifacts and the resulting perceived quality impairments.

2.2. Conversational Parameters and Delay

The analysis of interlocutor behavior and its alteration due to (disturbed) communication processes is discussed within the pragmatic aspects of human communication as laid out by [7]. Within this research tradition, conversation analysis (CA) [8] has exhaustively analyzed the process of the human communication. CA provides features for assessing interferences in the course of interaction which includes alterations of turn taking, feedback abilities, and so on. A subset of these features is related to nonverbal interaction cues of conversations also known as vocalic cues (cf., [9, 10]). In his work on conversational interactivity, Hammer [5] has used several of these vocalic cues to determine the conversational interactivity (CI) level of human communication. He utilized

1The delay reported in [1] is 1120 ms round trip delay. We have converted this value to one-way delay as the majority of the other publications in this area uses this delay notion.
double talk (DT) and mutual silence to characterize the conversations and identified the speaker alternation rate (SAR) as well as the active- and passive interruption rate (AIR and PIR respectively) to be valuable predictors of the CI level of a conversation. However, these measures do not differentiate between interruptions caused by the technical system’s delay and deliberate interruptions caused by the speaker. In order to overcome this differentiation problem we introduce the unintended interruption rate (UIR):

The UIR is based on the rate of passive interruptions that interlocutors experience during a conversation. However, it counts only those passive interruptions which were actually caused by delay, thereby excluding all occurrences of active interruptions2 that were deliberately caused by a speaker:

The advantage of the conversational parameter UIR is its ability to differentiate between disturbances introduced by the system’s transmission delays from interferences deliberately caused by one speaker interrupting the other. We will report on our findings on the relationship between delay and conversation parameters (including UIR) in Section 4.2.

3. Experimental Setup

Our investigation is based on a set of conversational quality tests conducted within the i:Lab of FTW. We recruited 17 pairs of untrained subjects who were located in separate rooms and communicated with each other using a VoIP connection. The conversational tests utilized VoIP clients on standard consumer grade laptops in conjunction with monaural headsets (to come close to normal telephone conversations). For further details on the technical test setup please refer to [11].

We exposed the participants to five different delay conditions as described in Table 1. In order to test different degrees of conversational interactivity we used three different scenarios as depicted in Table 2, resulting in a total of 15 test conditions. We chose the scenarios based on the recommendations in [12, 13] and according to their degree of interactivity they should introduce in the conversation as assessed in [5].

In order to acquaint the users with the setting and the VoIP system, we offered them a warmup session with minimal delay (100 ms). After completing the warm-up condition we followed up with the different delay conditions. Subsequent to each test condition (2 to 3 minutes, depending on the time needed to complete the task), participants were asked to rate the perceived quality of the system on a 5-point absolute category rating scale (MOS) ranging from 1.0 (bad) to 5.0 (excellent) as recommended by [14, 12].

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Task</th>
<th>CI Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCT</td>
<td>Ordering a Pizza</td>
<td>1</td>
</tr>
<tr>
<td>iSCT</td>
<td>Exchanging weather data</td>
<td>2</td>
</tr>
<tr>
<td>RNV</td>
<td>Random number verification</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2: Scenarios used within the VoIP user study

4. Results

We present our results in two parts: the first part addresses the impact of delay and conversational interactivity on perceived speech quality. Secondly, we discuss the extraction of certain conversational parameters from the audio signal to determine key qualities of a conversation, including conversational interactivity and delay sensitivity.

4.1. Perceived Speech Quality as f(Delay, CI)

As described in Section 3 we exposed untrained subjects to a wide range of delay conditions, letting them perform conversations at different levels of interactivity. The main goal was to test the following hypothesis:

The higher the conversational interactivity (CI), the higher the interlocutors’ sensitivity to delay impairments.

Figure 1 presents the mean opinion score (MOS), and the corresponding 90% confidence intervals, obtained for overall speech quality as a function of the three different conversational contexts and delay (ranging from 100 to 1600 ms). Our results confirm the above hypothesis: the higher the interactivity of the used conversation scenario, the stronger participants’ quality ratings decreased with rising delays. This correlation becomes most evident in Fig. 1 showing that for the very interactive RNV, one-way delays above 400 ms had a significantly higher impact on perceived voice quality than in the other scenarios: at the maximum delay setting (1600 ms), the mean of the quality ratings lies at 2.9 MOS (corresponding to a drop by 1.6 units) for RNV, compared to 3.7 MOS (a drop by only 0.5) for the low interactivity SCT scenario.

Our key finding is that high transmission delays degrade perceived speech quality less than expected. One-way delay values up to 400 ms did not significantly affect perceived speech quality, even during the very interactive RNV task. For normal conversations featuring medium CI (SCT and iSCT scenarios), even delay in the range of 800 ms did not significantly impact participants’ quality ratings. Our study results indicate a fairly high tolerance of human users to delay. This is in stark contrast to existing standardization and other research studies stating that mouth-to-ear delay values above the critical threshold

---

2 An active interruption occurs when an interlocutor starts to speak, while he still hears his counterpart talking. In contrast, a passive interruption occurs when a speaker becomes interrupted by the (delayed) arrival of a counterpart’s utterance.
of 400 ms (as suggested by G.114 [2]) or 250 ms (as suggested by Kitawaki and Itoh [1]) to be unacceptable. We see mainly two explanations for these discrepancies: firstly, we tested with naive subjects who in general tend to be less critical than trained experts. This difference in sensitivity is also confirmed by [1] who reported delay detectability thresholds of 100-700 ms for trained and 350-1100 ms for untrained subjects (depending on the task given). Secondly, we believe that changing usage habits and user perception lead to divergent study results over time. For example, recent influences such as the diffusion of packet-based telephony services such as Skype™ and GoogleTalk have educated a growing user base to cope with higher transmission delays, and this might have effectively lowered average quality expectations. Such developments are plausible explanations for the high sensitivity to delay artifacts reported by Kitawaki and Itoh’s subjects, since low-latency analog telephony systems were standard in 1991.

Both explanations are also supported by the fact that recent VoIP studies on delay [4, 5, 6] involving non-expert subjects confirm our findings. Gueguin et al. [4] found that delay in free conversations has almost no impact for values below 400 ms, with ratings decreasing only slightly between 400 and 600 ms. Similarly, the studies of Hammer [5] and Brauer [6] found delays around 600 ms still to be acceptable for the users. In conjunction with our findings, these results suggest that in the domain of VoIP services, the thresholds proposed by widely accepted standards and recommendations such as [2, 16] should be reconsidered and updated according to the results of contemporary studies.

4.2. Conversational Parameters as (Delay)

Our first analysis addresses the behavior of the conversational parameters as used by [5] for different delay conditions. From Fig. 2 one can see that for the highly interactive RNV scenario, both speaker alternation rate (SAR) and mutual silence (MS) do change with increasing delays. However, for the other two scenarios only a slight delay influence is visible. (Note, that delay dependency of the DT parameter behaved similar to SAR, thus the related plot was omitted due to space constraints.)

Similarly, active and passive interruption rates (AIR, PIR) as depicted in Fig. 3 are delay-invariant for SCT and iSCT. The diagrams demonstrate that SAR, AIR and PIR discriminate very well between the highly-interactive RNV scenario and the normally interactive SCT and iSCT scenarios. However, they also indicate that SCT and iSCT did not differ in terms of their degree of interactivity. Our observations made during the tests also confirm that in these scenarios, users conversed in a highly structured manner, because they rather waited for answers or confirming utterances rather than choosing to interrupt the opposite speaker. Due to the resulting well-defined answer-response structure without necessity for abrupt interruptions, the aforementioned conversational parameters did not change significantly under delay influence. For further studies focusing on conversational parameter extraction, we therefore recommend substituting the SCT scenario with e.g. free conversations.

Furthermore, we found the aforementioned measures to be inappropriate for capturing the delay influence on conversations of normal interactivity, due to their delay-invariance for SCT and iSCT. For this reason, we computed the unintended interruption rate (UIR) that we introduced in section 2.2. Fig. 4 overleaf illustrates the conversational parameter UIR as a function of delay. At 100 ms one-way delay, approximately one unintended interruption occurs per minute. Up to 800 ms, the rate of unintended interruptions grows steadily, but then remains constant for the 1600 ms condition. This behavior indicates that interlocutors happen to more frequently interrupt the opposite speaker unintentionally when delay increases. The saturation of the UIR for the two highest delay conditions can be explained by the fact that conversations become more structured.
and less interactive for higher delay as indicated by decreasing SAR, AIR and PIR and rising MS.

To summarize, our results demonstrate that in the context of delay analysis, using the UIR has the following advantages: independency of the scenario used, computability from plain talk spurs and good correlation with disturbances caused by transmission delay.

5. Conclusion

In this paper we reported on the impact of a wide range of delays on speech QoE across three different echo-free conversational contexts. In line with other recent studies [4, 5, 6], our results demonstrate that untrained subjects are considerably more tolerant (= 600-800 ms threshold) to delay artifacts than hitherto suggested by existing standards. We therefore recommend a reassessment and update of widely used standards and recommendations such as G.114 regarding delay influences. Furthermore, we found a pronounced effect of conversational interactivity on delay sensitivity, particularly in the context of the highly interactive RNV task. However, our results show only a small influence of delay on participant’s actual conversational behavior during conversations with normal interactivity, as reflected by the invariance of parameters such as SAR, DT and MS. Therefore, we introduced the UIR metric which can be extracted from talk-spurs and in contrast to aforementioned traditional conversational parameters, directly maps to the level of impairment that transmission delay causes to the conversation, independent of the conversation task at hand.

As concerns future work, we plan to repeat our studies with a wider range of tasks, including low-interactivity and free conversation contexts in order to generalize our findings regarding delay sensitivity and applicability of the UIR metric. Furthermore, we will perform a more detailed analysis of recorded test conversations, with a focus on participants’ adaption to delay and resulting transient effects.

6. Acknowledgements

This research has been performed within the project U-0 at the Telecommunications Research Center Vienna (FTW) and has been funded by the Austrian Government and the City of Vienna within the competence center program COMET. The authors would like to thank K. Manisch and O. Golderos for contributions to the user studies, and their colleagues from the Area U for further valuable input and discussions.

7. References


