Sitting, Chatting, Waiting: Influence of Loading Times on Mobile Instant Messaging QoE

Anika Seufert*, Carina Baur*, Fabian Poignée*, Michael Seufert†, Tobias Hoßfeld*
* *University of Würzburg, Chair of Communication Networks*, Würzburg, Germany
{anika.seufert|carina.baur|fabian.poignee|tobias.hossfeld}@uni-wuerzburg.de
† *University of Augsburg, Chair of Networked Embedded Systems and Communication Systems*, Augsburg, Germany, michael.seufert@uni-a.de

***Abstract*—This paper explores the relationship between loading times and Quality of Experience (QoE) in Mobile Instant Messaging (MIM) applications. Using a web application that mimics MIM interfaces, we conducted a QoE study in which participants engaged with a virtual chat partner. We controlled image loading times during chatting and evaluated their impact on QoE, annoyance, and acceptance ratings. Although the results show no difference in the QoE ratings, they clearly show that longer delays lead to greater annoyance and lower user acceptance. These findings underscore the importance for MIM app providers to minimize loading times in order to increase user satisfaction and retention.**

***Index Terms*—Mobile Instant Messaging, Messaging QoE, Crowdsourced QoE Study, Image Loading Time**

I. Introduction

In the digital age, communication has evolved to become instantaneous with Mobile Instant Messaging (MIM) applications emerging as essential tools for interpersonal interactions. These apps offer users the convenience of real-time communication and enable the exchange of text messages, images, and multimedia content over long distances. However, despite the seamless nature of these interactions, there are often loading times – short pauses during the transmission or reception of messages – which can disrupt the flow of conversation and impact the overall Quality of Experience (QoE).

Motivated by the ubiquitous presence of MIM apps and the relatively low number of research papers in the field of QoE, this paper addresses the relationship between loading times and QoE. For this, we developed a web application for mobile instant messaging QoE studies, which emulates the look and feel of current MIM apps. In this app, the study participants chat with a virtual chat partner called Alice, who asks the participants to send messages and pictures and also sends pictures herself. When sending and receiving, we inserted artificial loading times, which users are made aware of using a familiar loading icon. We conducted a preliminary QoE study to test the feasibility of the methodology for collecting QoE ratings for MIM apps and to gain initial insights. In subsequent studies, a thorough investigation of the influence of loading times on QoE can then serve as a basis for future research efforts. These findings can be helpful in the design, development and optimization of MIM applications to increase user satisfaction and retention.

The remainder of this work is structured as follows. In Section II, we present background information and discuss related work. Section III outlines methodology of our study. The collected dataset and the evaluation of perceived QoE across various loading time patterns are detailed in Section IV. Lastly, Section V provides a summary of our findings and gives an outlook on future work in this domain.

II. Background and Related Work

For multimedia services and applications, the subjective end user experience is important for network and service providers alike. They are interested in quantifying the QoE of users. Guidelines for collecting QoE ratings via subjective studies are available [1]–[3], as well as for time- and resource-efficient crowdsourcing and remote QoE studies [4]–[8].

Although there is some research on evaluating the QoE of MIM applications, only few studies have been conducted recently. In [9], authors investigated WhatsApp traffic and protocols and found that 35% of file downloads were potentially badly perceived by users. They collected feedback on the duration of file transfers in a study with 50 participants. Results showed that users tolerated transfers of up to $20\,\mathrm{s}$ length with a good overall experience. More than $40\,\mathrm{s}$ download time lead to very bad quality. As part of an investigation on popular smartphone apps, the QoE of WhatsApp was investigated with respect to the downlink bandwidth in [10]. A $5\,\mathrm{MB}$ file was sent to the participant who rated the acceptability of the loading time. Results showed that the separation between good and bad experience was at the threshold of $2\,\mathrm{Mbps}$. In contrast to those works, our study investigates loading times in both uplink and downlink, incorporating text and media in interactive chat scenarios. Additionally, both studies were published in 2015, but user expectations and media file sizes have changed due to technological advancements since then. In a more recent study [11], application interaction and user expectations were collected over four weeks on 39 Android phone users and for different smartphone applications, including MIM. A machine learning model was developed using smartphone features to decide whether a service meets the user's expectation. While the network QoS was reported as an important feature, the experience of the participants has not been investigated. An approach on representative service based QoE monitoring for

instant messaging (IM) was proposed by [12]. Considering the subservices of audio, video, and text chat, they investigated the behavior of 8 persons during Skype audio and video calls respectively to understand which bandwidth levels would cause participants to switch to text chats. In contrast, this work focuses on the QoE during text chats and investigates the influence of perceived delays in uplink and downlink.

III. Study Description

To evaluate the influence of image loading time on the perceived experience, we conducted a QoE study on mobile devices, following the best practices for crowdsourcing-based QoE studies [6]–[8]. The study was conducted remotely and delivered as a web application with a user interface tailored for optimal performance on mobile devices and accessible via a web browser. It comprised three main components:

1. Study Preparation and Demographic Questionnaire: Before the study began, all images were first loaded into the local browser cache to avoid unexpected delays in remote execution and thus ensure accurate measurement results. Participants' devices were checked to confirm the use of a mobile device by asking about touchscreen functionality and screen size. In cases where the study was not accessed via a mobile device, a QR code with the study URL was displayed for participants to scan with a mobile device in order to participate in the study. Participants then provided basic demographic information such as gender, age, country of residence, and frequency of use of mobile instant messaging services.

2. Study Tasks and QoE Questionnaire: The main part of the experiment consisted of four different study tasks, in each of which the participants were asked to test and evaluate a MIM application. Each task consisted of short interactions on everyday topics with a virtual chat partner named Alice. In the first and third tasks, participants were asked to send a picture by choosing one from a predefined selection, while in the second and fourth tasks Alice sent pictures to the participant. During the picture exchange, participants experienced randomly selected loading times of $0.2\,\mathrm{s}$, $0.5\,\mathrm{s}$, $1\,\mathrm{s}$, $5\,\mathrm{s}$, $10\,\mathrm{s}$ or $25\,\mathrm{s}$. These waiting times were derived from the WQL model of [13], as a broad spectrum of QoE ratings (from 5 for $0.2\,\mathrm{s}$ down to 2.5 for $25\,\mathrm{s}$) can be expected if the results from Web QoE studies are transferable to our study. After each task, participants completed a QoE questionnaire about the task they had just completed. This questionnaire included questions about the content of the conversation, how they would rate the experience they had (*"How would you rate your experience with the messaging app?"* (excellent, good, fair, poor, bad)), whether a waiting time was perceived or not and how disturbing it was (*"How disturbing was this waiting time?"* (extremely annoying, very annoying, moderately annoying, slightly annoying, not at all annoying)), and about the acceptance of the application's performance (*"Would you use a messaging app showing this performance?"* (yes, no)).

3. Context Questionnaire: At the end of the experiment, participants were asked several contextual questions aimed at understanding their habits regarding the use of MIM apps. In addition, the questionnaire included the repetition of some questions from the original demographic questionnaire to be able to check the consistency of the responses.

Filtering of unreliable participants: To ensure the integrity of the data and minimize the impact of unreliable participants on the subsequent analysis, rigorous assessments of participant reliability were conducted by analyzing the collected data. First, participants who ignored the instructions we clearly gave them in the study description, e.g., who reloaded the study website, were excluded from the analysis to maintain consistency of data collection. Subsequently, questions on consistency and content-related questions were used to identify and exclude unreliable employees. Participants who could not detect a delay despite a loading time of 10 seconds or more were also excluded from the analysis. Finally, the results of employees who rated their experience worse in low-loading time conditions than in high-loading time conditions were also excluded, resulting in a data set with high intra-rater reliability, as suggested by [14].

IV. Results

The study was conducted on the crowdsourcing platform Microworkers. The participants received a reward of USD 0.75 with an estimated time to complete the task of less than 10 minutes. No further restrictions, like country or skill filters, were applied to limit the workers' access to the task. Overall, a total of 415 workers took part in the study. After filtering, 161 workers remain (51 female, 110 male), resulting in 644 ratings. On average, 46 ratings were collected for each of the 14 conditions, with a minimum of 23 ratings per condition.

Before analyzing the ratings in detail, we first want to know if there is a significant difference in the perceived waiting time for receiving and sending images. To investigate this, we randomly sampled an equal number of ratings from all image sending conditions without repetition, as well as from all image receiving conditions. These will be considered separately later. We then applied a Mann-Whitney U test to these two groups to see if there is a difference between them. The test shows that there is no significant difference between the ratings for receiving and for sending images, since $p > 0.05$ for both for QoE and for annoyance. For this reason, we do not differentiate further between receiving and sending images in the following evaluation.

To gain insight into the perceived experience of sending or receiving images, we next examine the distribution of QoE ratings per loading time, as shown in Figure 1. In accordance with the ACR rating scale [15], a rating of category 1 corresponds to a poor experience, while category 5 represents an excellent experience. Despite varying loading times, the QoE ratings mainly range between 3 and 5, i.e., between fair and excellent experience, with a good or better ratio (GoB) of 85.54% for a loading time of $25\,\mathrm{s}$ and a maximum of 96.12% or a loading time of $1\,\mathrm{s}$. Looking at the Mean Opinion Scores (MOSs) and the 95% confidence intervals (CIs), shown as white intervals on the right y-axis, loading times of $0.2\,\mathrm{s}$ to $1\,\mathrm{s}$ result in excellent MOS values ($0.2\,\mathrm{s}$: 4.61, $0.5\,\mathrm{s}$: 4.55 and

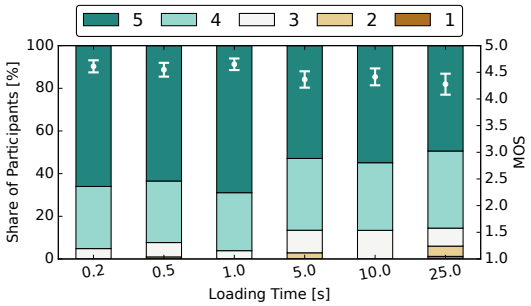

Fig. 1: Distribution of the QoE ratings per loading time.

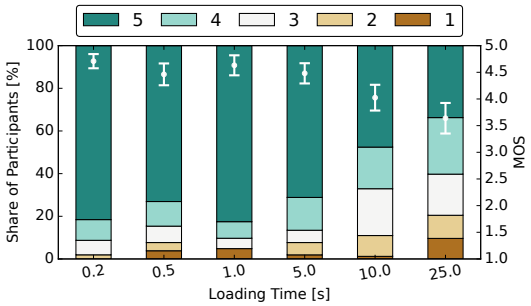

Fig. 2: Distribution of the annoyance ratings per loading time.

1 s: 4.65). Even at loading times of 5 s or more, the MOS scores indicate good perceived QoE (5 s: 4.37, 10 s: 4.41, and 25 s: 4.28). In particular, contrary to what might be expected, there is no recognizable downward trend. Furthermore, this does not match the results from [13] for web browsing or the results form [16] for various application types, in which the authors found a clear influence of loading time on QoE. To determine whether there are still notable differences in the QoE distributions, we perform pairwise Mann-Whitney U tests for all loading time conditions, applying the Šidák correction to address the multiple comparisons issue. Except for loading times of 1 s and 25 s, with a p-value < 0.04, p-values above 0.1 show no significant differences between the QoE distributions for all other loading time conditions. However, since the QoE ratings for 25 s do not differ significantly from the ratings of other loading times, no clear conclusion can be drawn. This could indicate that the loading time only plays a subordinate role and that users might also take other factors into account when evaluating their app experience.

Given the minimal differences in QoE ratings, we wanted to find out whether this was due to workers considering factors other than just loading times in their ratings, or whether the loading time actually had little impact on their experience. To investigate this further, we examined the annoyance ratings per loading time condition and plotted their distribution in Figure 2. Again, the MOS values and 95% CIS are shown as white intervals on the right y-axis. Compared to the QoE distributions, a clearer trend towards higher annoyance ratings can be observed here for longer waiting times. While the MOS values for a loading time of up to 5 s are still very high (0.2 s: 4.71, 0.5 s: 4.46, 1 s: 4.63, and 5 s: 4.48), a deterioration can be seen at 10 s (4.02) and 25 s (3.64). For a more thorough investigation, we again performed pairwise Mann-Whitney U tests with a Šidák correction for the annoyance ratings of all loading time conditions. The tests revealed that the ratings for loading times of 10 s and 25 s were significantly different from all others, although they were not significantly different from each other. For all other conditions, no significant difference was found. Thus, it can be concluded that loading times of 10 s or more are perceived as significantly more annoying. Nevertheless, compared to the results of [13,16] for other application types, the QoE deterioration is much weaker and should therefore continue to be investigated as a separate use case in further research.

Finally, we examine the impact on the providers of MIM apps. In addition to the ratings of QoE and annoyance, the study participants indicated whether they would continue to use the application despite the loading times that occurred. A high acceptance rate is observed for loading times between 0.2 s and 5 s, with rates between 92.23% and 93.27% of participants willing to tolerate the app's performance. However, if the loading time exceeds 5 s, the acceptance rate drops significantly. With a loading time of 10 s, the acceptance rate drops to 84.15% and with a loading time of 25 s to just 83.13%. Thus, increased loading times could result in a potential customer loss of around 10%. This trend is consistent with the earlier results of the annoyance ratings. However, the discrepancy between the acceptance rates and the QoE ratings raises questions that require further investigation. Future studies should look at this aspect in more detail. However, considering the conclusions that MIM app providers should draw from these results, the most important recommendation is to minimize such waiting times as much as possible. Since they cannot be completely avoided, app providers should explore strategies to reduce or hide the waiting time perceived by users. Techniques such as interlacing, discussed in [17], could be used to effectively address this challenge.

V. Conclusion

To sum up, our study investigated the impact of loading times on the QoE in MIM apps. For this, we designed a web application that mimics MIM interfaces and conducted a crowdsourcing QoE study, which showed that users generally perceive good QoE despite varying load times. Still, longer loading times lead to increased annoyance and lower acceptance rates. In particular, loading times of more than 5 s significantly affect user satisfaction and drop acceptance rates by around 10%. These results highlight how important it is for MIM app providers to minimize loading times in order to increase user satisfaction and retention. In future work, we will look at the nuanced factors that influence user perception and behavior in relation to loading times. Considering these insights can help MIM app providers optimize their platforms to meet user expectations and ensure long-term success in the competitive digital landscape.

ACKNOWLEDGEMENT

This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grants HO 4770/7-1 and SE 3163/1-1, project number: 442413406, as well as SE 3163/3-1, project number: 500105691. The authors alone are responsible for the content.

REFERENCES

[1] International Telecommunication Union, "ITU-T Recommendation G.1011: Reference Guide to Quality of Experience Assessment Methodologies," 2015.

[2] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives," *Multimedia Systems*, vol. 22, no. 2, 2016.

[3] M. Seufert, "Statistical Methods and Models based on Quality of Experience Distributions," *Quality and User Experience*, vol. 6, no. 1, 2021.

[4] M. Hirth, T. Hoßfeld, and P. Tran-Gia, "Anatomy of a Crowdsourcing Platform - Using the Example of Microworkers.com," in *Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. IEEE, 2011, pp. 322–329.

[5] B. Gardlo, S. Egger, M. Seufert, and R. Schatz, "Crowdsourcing 2.0: Enhancing Execution Speed and Reliability of Web-based QoE Testing," in *International Conference on Communications (ICC)*, 2014.

[6] S. Egger-Lampl, J. Redi, T. Hoßfeld, M. Hirth, S. Möller, B. Naderi, C. Keimel, and D. Saupe, "Crowdsourcing Quality of Experience Experiments," in *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Springer, 2017.

[7] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "Best Practices and Recommendations for Crowdsourced QoE – Lessons learned from the Qualinet Task Force "Crowdsourcing"," COST Action IC1003 European Network on Quality of Experience in Multimedia Systems and Services (QUALINET), White Paper, 2014.

[8] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, and M. Allahbakhsh, "Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions," *ACM Computing Surveys*, vol. 51, no. 1, 2018.

[9] P. Fiadino, M. Schiavone, and P. Casas, "Vivisecting WhatsApp in Cellular Networks: Servers, Flows, and Quality of Experience," in *Seventh International Workshop on Traffic Monitoring and Analysis (TMA)*. Springer, 2015, pp. 49–63.

[10] P. Casas, R. Schatz, F. Wamser, M. Seufert, and R. Irmer, "Exploring QoE in Cellular Networks: How much Bandwidth do you need for Popular Smartphone Apps?" in *Fifth Workshop on All Things Cellular: Operations, Applications and Challenges*, 2015, pp. 13–18.

[11] A. De Masi and K. Wac, "Towards Accurate Models for Predicting Smartphone Applications' QoE with Data from a Living Lab Study," *Quality and User Experience*, vol. 5, no. 1, p. 10, 2020.

[12] X. Xin, W. Wang, A. Huang, and H. Shan, "Representative Service Based Quality of Experience Modeling for Instant Messaging Service," in *25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*. IEEE, 2014, pp. 2018–2023.

[13] T. Hoßfeld, F. Metzger, and D. Rossi, "Speed Index: Relating the Industrial Standard for User Perceived Web Performance to Web QoE," in *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–6.

[14] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2013.

[15] International Telecommunication Union, "ITU-T Recommendation P.910: Subjective Video Quality Assessment Methods for Multimedia Applications," 2008.

[16] H. Z. Jahromi, D. T. Delaney, and A. Hines, "Beyond First Ompressions: Estimating Quality of Experience for Interactive Web Applications," *IEEE Access*, vol. 8, pp. 47 741–47 755, 2020.

[17] A. Seufert, S. Schröder, and M. Seufert, "Delivering User Experience over Networks: Towards a Quality of Experience Centered Design Cycle for Improved Design of Networked Applications," *SN Computer Science*, vol. 2, no. 6, 2021.