The Interplay between QoE, User Behavior and System Blocking in QoE Management

Tobias Hoßfeld, Luigi Atzori, Poul E. Heegaard, Lea Skorin-Kapov, Martin Varela

***Abstract*—In this position paper we highlight a shortcoming of current QoE management approaches that typically do not take into due account the resulting user behavior. As a result, a divergence is introduced between the predicted and the actual QoE, the later being affected by the reaction of the user to resource assignments. We believe the following two factors to be among those having the highest impact in this respect: the user (im)patience, and tolerance to low quality. To illustrate our claims, we model an example scenario where a user requests an online service, such as an online authentication service. The request is processed by a system with limited resources, which may also cause the request to be blocked or buffered, with a consequent impact on the QoE. Some aspects of aborting users, blocked users, and QoE of served users are investigated by means of a simple queueing system, $M/M/s/n+M$ which takes impatience into account. Insights from this theoretical study show that an increase in the user patience results in a decrease of the average QoE in the system, as the user may consume system resources without waiting to be finally served. Based on these findings, we argue the importance of incorporating these aspects of quality, often ignored in both QoE modeling and management, into any QoE management system that is expected to improve the provider's bottom line.**

***Index Terms*—QoE, QoE management, user impatience, abort probability, blocking probability,** $M/M/s/n+M$

I. Introduction

QoE-aware service management relies on the continuous monitoring and prediction of the quality perceived by the users; it aims at taking actions to fix potential quality degradations and at optimizing the usage of network and server/cloud resources. The success of these procedures depends on the reliability of the adopted quality models and the accuracy of the monitoring and availability of the parameters that such models demand. However, it is also important to understand the relations between QoE, user behavior, and engagement [1]. These latter two are related to the way the user interacts with the services and include both long-term actions, such as churning from a service, and short term actions, such as interrupting a service session or starting to do other activities concurrently. User behavior modeling is indeed of great importance for the providers' business as it affects the service usage, and thus the resources required and again the resulting quality.

Based on these considerations, in this position paper we highlight a shortcoming of current QoE-driven network and service management approaches that do not take into due account the user behavior. Indeed, it is a common practice to estimate QoE on the basis of adopted quality models that consider key influencing factors, such as network and application resources and parameters; however, the estimated QoE differs from the actual one as perceived by the served end users, as the actual QoE is also affected by the users' reactions to resource assignments. In this paper, we specifically focus on the influence of two factors. Firstly, user impatience determines how long a user is willing to wait to receive a service, after having initiated a service request. We highlight that there are cases where users being too patient can be detrimental for the overall quality of the user base, and that *late aborts* by the user can be problematic. Secondly, we consider the users' tolerance to low quality. Too much tolerance to low quality can in certain cases lower the quality for other users (who may or may not be so tolerant themselves).

User behavior, especially user impatience, has been the subject of several studies in different domains (marketing, web service developers) and communities (QoE, UX). On the one hand, it is relevant to know for how long users are willing to wait for different services before abandoning, as has for example been investigated in the case of web browsing [2]–[7]. On the other hand, the relation between waiting times, e.g., for web page downloads, and QoE is of interest and a research topic in the QoE community [8]–[14]. Similarly, the users' tolerance to service degradation has been analyzed for cases of video streaming services in terms of correlation between quality degradation and play time [15], [16], but usually not integrated into the QoE model or analysis.

In most of these works, user behavior has been inferred by analyzing real data traffic, with the aim being to determine the relationship between behavior and some service and traffic parameters so as to derive potential models. Conversely, limited efforts have been put towards including these models into the service management chain, with the aim to analyze the impact on the system's performance in terms of resource usage and achievable QoE, as we do in this work.

In what follows we will discuss several service scenarios in which user impatience (related to waiting times) or intolerance (related to poor service quality) drives users to abort, or abandon, a service, thus resulting in non-trivial effects on the success of QoE management approaches. We also consider the scenario where the user requests an online service, and we derive a system model with user aborts. We then focus on a more specific setting, namely an authentication service involving waiting times, to illustrate the impact that user impatience can have on overall QoE in a system. We use this illustrative scenario to argue that such user behavior is an outcome of QoE management mechanisms that are implemented in either

Tobias Hoßfeld (University of Würzburg, Germany), Luigi Atzori (University of Cagliari, Italy), Poul E. Heegaard (NTNU, Norway), Lea Skorin-Kapov (University of Zagreb, Croatia), Martin Varela (callstats.io, Finland)

the network or at the application level, and thus needs to be explicitly considered when benchmarking and evaluating the performance of such mechanisms. Moreover, we discuss potential ways in which knowledge regarding user impatience and tolerance to low quality can be exploited in the system to improve resource management and overall QoE for all users.

The remainder of this paper is organized as follows. Section II provides the background and rationale for considering user behavior in the context of QoE management. Section III introduces a general Markov model for a holistic analysis of QoE management with user behavior. Section IV presents numerical results and discusses the key findings. Finally, Section V provides a discussion of results and implications for future studies focusing on QoE management.

II. User behavior: the role in QoE management

A. User behavior modelling in QoE management

For the most part, QoE-driven network and service management is related to either allocating resources and/or configuring a service so as to maximize QoE subject to given constraints (e.g., resource availability, user device capabilities). When considering multiple users in the system, an additional objective may be to maximize QoE fairness among users [17]. As a result, corrective actions in the network and service configuration are introduced on the basis of the *estimated QoE*, which is obtained by applying appropriate quality models to the current network and services status. However, in the proposed frameworks in the literature, the contribution of the user behavior is often neglected (e.g., [18], [19]), notwithstanding the fact that it has a significant effect on the resource usage and consequently on the *actual QoE* as perceived by end users. Indeed, the later will be different from the estimated value, which however is considered when deciding on next corrective actions. Accordingly, we believe that for more effective QoE management, there is the need for closing the loop in the process by including user behavior modeling, see also [20].

Figure 1 shows the concept by depicting a generic QoE-driven service and network management scenario. Note that this can be applied to both short-term/in-session and medium/long-term management cases. When taking decisions on next corrective actions on the basis of the system status, appropriate QoE models (blue circle in the figure) are currently introduced in the management approaches, which can rely on simulations, heuristics or deterministic approaches. However, on the basis of the provided QoE, users may react in several ways, thus impacting the final resource allocation and *actual* provided quality: e.g., given a "QoE-driven" allocation of system resources, will certain users end up aborting the service due to long waiting times or poor quality? How will this affect their QoE, and also the QoE of other users in the system? This is in particular true in cases when users end up aborting a service prior to having actually consumed it. For that reason, additional user behavior modelling is needed, as shown in the figure with a green circle. A loop is created, which involves QoE estimation, modeling of users' reactions to a given resource allocation, and resulting input provided for

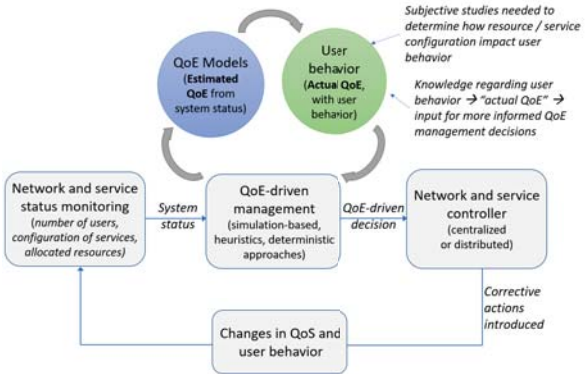


Fig. 1: Generic framework of QoE-driven management that highlights the importance of user behavior modeling.

consideration in the management approach. The methodology for including user behavior in the process needs to be further investigated and depends on the specific application. Different service scenarios imply different service states, and in turn invoke different types of user behavior that may be relevant to consider from a QoE management perspective.

In the following, we review studies that have addressed the analysis of user behavior in the past, with a particular focus on scenarios where user impatience or intolerance drives users to abort or abandon a service.

B. User (im)patience resulting from waiting times

With a focus on objectively investigating real user behavior, studies have utilized network measurements and early interruptions of TCP connections. In [6], two months of real traffic traces on a campus access link is analyzed, containing more than 7,000 hosts. The measurement results indicate that "users mainly abort the transfer in the first 20 s." Similarly, in [7], a traffic trace is captured during the busy hour of a Broadband Access Server, between the access router and the first routers towards the Internet. For web traffic, the average time of interrupted streams was found to be approx. 4 s. In [10], the cancellation rate of web browsing users is found to be on average about 7 s.

Other work has focused on subjective studies (typically conducted in a lab setting) and investigated how long users are willing to wait for web pages to download before abandoning the web site. As a result, the precise maximum waiting time varies across different studies: about 10 s in [5], 28 s in [2], 8 s in [3], 2 s in [4].

A general overview on the relation between waiting times and QoE and the studies conducted so far is provided in [9] for different services. The study clearly differentiates between the case of the delay that occurs *before* service consumption (aka initial delay) and the one that happens *during* service consumption (e.g., stalling during video watching). Results show that the quality perception of the waiting time is strictly dependent on when the waiting time occurred. Moreover, logarithmic relationships between waiting times and QoE demonstrate the applicability of well-known principles such

as the Weber-Fechner law. Resulting models, however, do not consider user aborts.

We note that a service scenario in which a user becomes impatient and aborts before the actual service has started leads to likely user frustration, and the user not having in essence *perceived an experience* related to using the service. Hence, such a case is not captured by QoE models. In addition to a web browsing case, this may occur for example in the case of an online authentication service. After waiting a certain amount of time, the user aborts and tries again, or temporarily/permanently gives up.

C. User (in)tolerance to low service quality

While user aborts may result from impatience and long waiting times, such behavior may also result from low service quality. We can consider the case of an adaptive video streaming service, where a user consumes a low quality video stream for a certain amount of time, after which the user decides to abort the service. This phenomenon has been analyzed in [15], where the authors have studied a significant video data set that spans different content types which where used in video streaming services. The results show that there is a significant dependency between user engagement and quality, with a significant reduction in the play time as soon as the re-buffering ratio becomes significant. In [16], the authors have performed a similar analysis and went further by designing a user engagement prediction model that allows content providers to predict how long viewers remain in video sessions with specific video quality metrics. These results show that there is an immediate impact on the system (and potentially other users in the system), as the lower the quality, the higher the probability of users leaving the streaming session, and the greater the amount of resources that become available for the more tolerant users. Still, considering a more long term perspective, this phenomenon can also lead to user churn. Experimental analysis in this direction are more difficult to conduct, as they require longer observation. Moreover, it is more difficult to obtain data, as the operators are not willing to share data on the amount of users abandoning the services. The authors in [21] makes use of a Sigmoid function to link user satisfaction to the perceived quality; still, further empirical analysis is needed in this direction.

As another example, we consider a multiparty audiovisual conference call, where due to poor quality experienced by one user, all users decide to abort the call and switch to a different service or communication mode (e.g., switch from using Skype to WebRTC, or switch from a video call to an audio-only call). A network operator or service provider aware of these quality degradation issues may decide to take certain corrective actions, which should also rely on the understanding of the user tolerance, as it drives user reactions.

In Section III, we consider a scenario where the user requests an online service, and we derive a system model with user aborts. We then focus on a more specific setting, namely an authentication service involving waiting times, to illustrate the impact that user impatience can have on overall QoE in a system. We use this illustrative scenario in Section IV to posit that by understanding the impact that system performance will have on resulting user behavior, more informed QoE management decisions can be made.

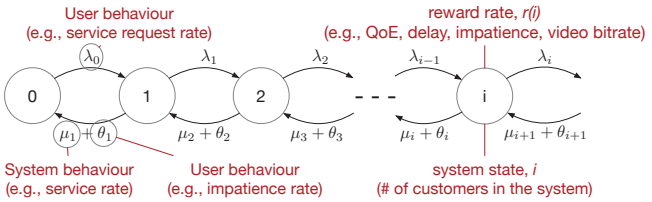


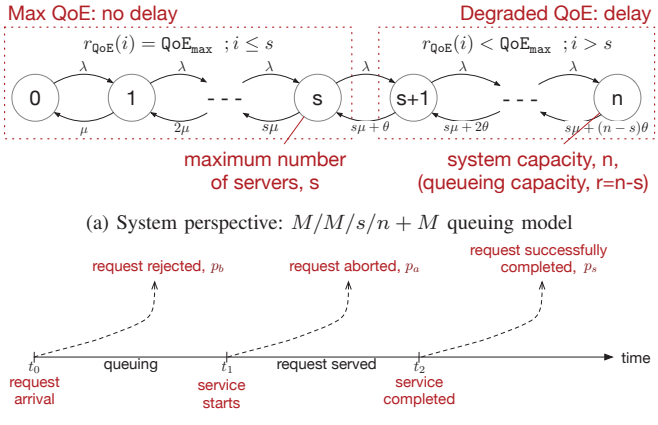
Fig. 2: Markov model of the system with requests rate λ_1, users impatience (abort) rate θ_i, and service rate μ_i.

III. General System Model with User Aborts

1) General case: We consider a system with shared or limited resources. For example, the users in the system are served immediately but share equally the server capacity (case A). Or, there are several servers in the system and a user is served individually if a server is available (case B). In both cases, the system state is reflected by the number i of users in the system, which determines the system behavior, i.e. the service rate μ_i a user obtains (A) which may also include the waiting time a user experiences (B). If the waiting queue is finite, the request will be rejected when the queue is full and the system blocks the user. However, a user may abort the service due to low quality or due to impatience regarding the experienced waiting times. If there are more users in the system, the users get a lower service rate or experiences higher waiting times. The abort rate θ_i as well as the request rate λ_i reflect the user behavior in the model and depend on the system state i. The arrival rate λ_i of users requesting the service may also depend on the actual system state, e.g. to reflect a finite number of users. If the request arrival process is Poisson, and service and impatience times are exponentially distributed, this is a Markov model. Figure 2 shows a general Markov model of such a system.

To each state we can assign *reward rates*, $r_M(i)$, which is the value of the metric, M, of interest in state i, like the video quality level or the waiting time of users. The expected reward is obtained by $E[M] = \sum_{\forall i} r_M(i)\pi_i$, where π_i is the steady state probability of state i. We can also assign rewards for QoE in each state, $r_{\texttt{QoE}}(i)$, which is the QoE a (tagged) user experiences when she arrives and finds the system in state i. This general Markov model allows to investigate jointly the interplay between system behavior and user behavior with respect to QoE for a QoE management system.

2) $M/M/s/n+M$ waiting and blocking system with impatient users: In the remainder of the paper, we consider a system with a finite number of servers s (maximum number of simultaneous requests in service), and a finite number of queuing positions r, which implies a finite system capacity $n = s + r$. The user gets served immediately, or will have to wait until a server is available. The user request will be



(a) System perspective: $M/M/s/n+M$ queuing model

(b) User behavior perspective

Fig. 3: Illustration of system and user behavior for the online authentication use case.

rejected when the queue is full. If the waiting time is too long, the user may decide for abandonment due to impatience. The user abort rate θ_i of all waiting users in state $i > s$ is then $\theta_i = (i-s)\theta$. The considered $M/M/s/n+M$ queuing system is a Markov model, known as Erlang-A model with $\lambda_i = \lambda, \forall i$, and illustrated in Figure 3(a).

A user experiences maximum QoE if the request is served without waiting delay, and we assign a reward rate $r_{\mathtt{QoE}}(i) = \mathtt{QoE}_{\max}$ to all states $i \leq s$. Otherwise, the user experiences waiting delays resulting in lower QoE as modelled by Eq. (1). When the user finds the system in state $i > s$ is, the reward rate is less than max: $r_{\mathtt{QoE}}(i) < \mathtt{QoE}_{\max}$. The number of requests in the system is reduced when a user aborts the service (because she is impatient), or the request is served.

Figure 3(b) illustrates this from a user's perspective. The request arrives at t_0 and at this moments it can go into the following possible two states: *blocked* (system is in state n and the blocking probability $p_b = \pi_n$), if there are no servers available but it is delayed in the queue for $t_1 - t_0$. From the waiting state, the request moves to the *served* state (system is in state $0 < i \leq s$) if not aborted. If the system is in a state $0 < i \leq s$ upon a request arrival, then the request is immediately served and $t_1 = t_0$. In case the waiting time is too long, the user might get impatient and decide to cancel the request. If the impatient user aborts the request before it is served, the number in the queue is reduced by one. The reward rate that reflects the probability of abortion in state i is $r_a(i) = \frac{(i-s)\theta}{(i-s)\theta+s\mu+\lambda}$; $s < i \leq n$, and hence the expected probability of a user aborting is $p_a = \sum_{i=s}^{n} r_a(i)\pi_i$. Then, the probability of being served is $p_s = 1 - p_a - p_b$.

IV. Use Case: Online Authentication Service

As a use case, we consider access to online services such as shopping carts, online banking, online authentication, web etc. We consider a scenario where the user requests an online authentication service, and may have to wait until the request is served due to limited resources. After a given waiting time, which depends on the current system status, the user request is served. However, during waiting, the user may decide for abandonment due to impatience.

A. QoE and User Behaviour Model

We are interested in analyzing the resulting QoE of the users that have to wait before being served, with particular attention to the role of the impatience in this process and at varying load of the system. For this we need a QoE model.

A mapping function $Q(t)$ between the waiting time t and QoE is provided in [9] and used in our analysis. We want to highlight that only successfully served users are assigned a QoE value, but not blocked or aborting users. The mapping function is bounded in the QoE domain to the range $[1;5]$.

$$Q(t) = -2.816 \log_{10}(t + 1.378) + 5 \tag{1}$$

In our model the random variable W represents the waiting time of a user which is analytically provided e.g. in [22]. Once we know the waiting time W (random variable, RV) of served customers, we can compute the QoE value (RV) through $Y = Q(W)$. Accordingly, the CDF of QoE values is obtained with the inverse mapping function $Q^{-}1(x) = t$.

$$Y(x) = P(Y \leq x) = P(W \leq Q^{-1}(x)) \tag{2}$$

The overall QoE $\mathcal{Q}$ is the expected value of Y over the QoE range $[L;H]$ which is $[1;5]$ in the numerical results.

$$\mathcal{Q} = \mathrm{E}[Y] = \int_L^H x \cdot y(x)dx \text{ with } y(x) = \frac{d}{dx}Y(x) \tag{3}$$

We can also directly compute the overall QoE using the reward rates as introduced in the general model. The reward rate is the expected QoE in state i and the probability that the system is in state i is π_i. Hence the overall QoE is

$$\mathcal{Q} = \sum_{i=0}^{n} r_{\mathtt{QoE}}(i)\pi_i \tag{4}$$

If the waiting time W of a user is larger than their patience, the user aborts. The average impatience time of a user is $1/\theta$.

B. Numerical Results

Our investigation has been conducted by varying the system parameters as follows.

- arrival rate $\lambda \in [0.5;15]\mathrm{s}^{-1}$
- number of servers $s \in [1;50]$
- mean patience threshold $\theta \in [1;50]\mathrm{s}$
- number of extra waiting spaces $r \in [1;50]$
- service rate $\mu = 250\,\mathrm{s}^{-1}$

The main effects are observed in terms of $(p_b, p_a, p_s, \mathcal{Q})$ when varying one of the following parameters (λ, s, θ, r) and computing the average results for all the others taking any value in their ranges.

Figure 4 shows both the blocking probability p_b and the abort probability p_a. These numerical results demonstrate that higher patience of users leads to higher effective system load, as the users may abort later. Therefore, aborting users may

waste more resources when being more patient and aborting later. As a consequence, the higher effective system load leads to higher blocking probabilities. Clearly, the higher patience of the users is reflected in lower aborting probability. If the system provides more waiting spaces (r increases), users are less often blocked and at the same time they may abort frequently due to an increase in the average waiting time. As expected, higher arrival rates result in higher system load and hence higher blocking probabilities. This increase impacts on the abort probability with an initial increase to a certain point after which the abort probability decreases, as more users get blocked in that case. Finally, more servers help to serve more users and reduce also the abort probability as the waiting time is clearly reduced.

To analyze the performance of the system in terms of percentage of users successfully served, it is important to consider both the blocking and the abandon probabilities. The main effect plot in Figure 5 shows that the patience threshold of users and the number of extra waiting spaces only have a tiny impact on successful service completion. The number of arrivals and servers determine the traffic intensity in the system and are the main effects on the success probability. In this figure, the dashed lines show the results of the same system but without waiting spaces and hence no user aborts (the blocking probability is given by the well-known Erlang loss formula). We see that the system behavior is quite similar in terms of probability p_s, showing that having waiting spaces may provide an increase of almost 5%.

The overall QoE in the system shows an interesting behavior. Again, higher load and server utilization decrease the overall QoE. An operator may improve QoE and success probability by investing in more servers, which result however in larger costs for the operator. It needs to be evaluated what is the customer-lifetime value compared to the cost investments [23] to find an optimal operation point in terms of number of servers depending on the arrival rate.

While user patience does not influence the success probability, it strongly affects the overall QoE. If users are more impatient, they will likely abandon, thus leading to better QoE for other users. In the same way, more waiting spaces are not helpful, as they lead to larger waiting times and nevertheless aborting users, i.e. a waste of resources. It is therefore tempting to conclude that for improving QoE, only a small waiting space or even no waiting space is required, as the waiting space has a tiny effect on the success probability only. However, it remains unclear how aborting or blocking affects the overall experience of customers and their churn behavior.

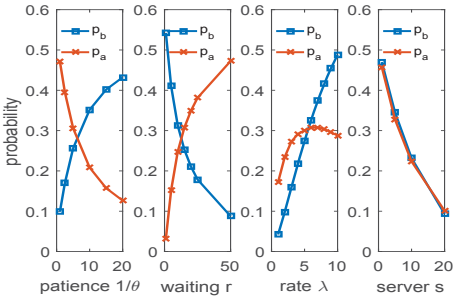


Fig. 4: Main effects on abort probability p_a and blocking prob. p_b.

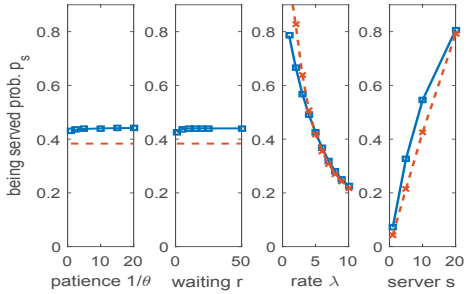


Fig. 5: Main effects on being served probability p_s. The dashed line compares the success probability $p_s = 1-p_b-p_a$ with an $M/M/s/s$ loss system without waiting space.

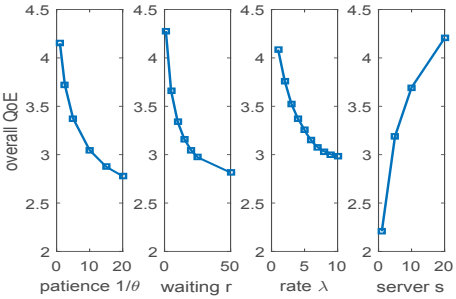


Fig. 6: Main effects on overall QoE $\mathcal{Q}$. Only succesfully served users are considered and their waiting time is mapped to QoE.

V. Discussions and Conclusions

Typical QoE management mechanisms rely on QoE models to drive resource allocation or service configuration decisions. However, for the most part they neglect actual user behavior in the system and the impact that this behavior has on overall QoE. We have discussed how user behavior modeling is a key aspect to be considered in the QoE management loop, and have focused on (im)patience with respect to waiting times or (in)tolerance to low quality. We further focused on the impact of user impatience on the performance of a system with a limited number of servers and waiting spaces. The major result from the conducted analysis is that user impatience has a great impact on QoE in the system, but a limited impact on the probability for a user to be served. Indeed, as users become more patient, the system becomes overcrowded with more users waiting in the system but still receiving the same probability to be finally served. Accordingly, the served users have to wait longer than in the case of a more impatient population, consequently resulting in decreased QoE.

From a QoE management perspective, different approaches can be followed by the service provider to react to this scenario. The most straightforward would be to increase the server resources so as to keep the system capable of serving the target percentage of users while keeping the waiting time lower than a desired threshold. However, this may not be feasible for

economic reasons; in this case, a decision could be to reduce the waiting spaces blocking the users before they wait too much. Another choice would be to keep users informed about the expected waiting times so that they can decide whether to wait or leave immediately. This has a positive effect on the QoE for the finally served users, as they will be waiting less. But what about the overall service utility, which should also consider the impact of the blocked users? In this case, the blocked or abandoning customers could be rewarded by the provider with additional free services (e.g., when retrying the customer won't pay) or could be informed about the best times when to retry and thus be guaranteed immediate service.

However, the crucial point from a QoE perspective is how users perceive blocking and how users perceive aborts. *Is it better if a user is blocked or if a user aborts?* Let us consider a visit to a restaurant. If the restaurant is occupied and you cannot wait, you will simply go to another place or come back in one hour. In this case you may not be disappointed, unless either the other restaurants are full as well or there are no other restaurants of choice nearby. Some other restaurants may have waiting spaces at the bar where you can wait until a table is free. In this situation, the restaurant may provide information about the expected waiting time. In a similar way, information about the expected service quality or expected waiting times may be provided to users. Future subjective studies are need to investigate the implications that informing users of expected waiting times will have on their behavior. Moreover, subjective studies need to be designed to investigate the relation between user behavior and QoE, and how blocking or aborting affect the QoE, also on a longer time-scale to cope with user churn. An important issue to also consider is the service context, as selecting another service provider may not be possible for a given service, or may be quite complicated.

To summarize, we believe that all these possible solutions need to be investigated after a more in-depth study about the QoE of waiting users, informed waiting users, blocked informed users, and rewarded blocked users. Additionally, different context aspects should be considered, with particular reference to task-driven aspects and the presence of alternative competing services.

REFERENCES

[1] W. Robitza, S. Schönfellner, and A. Raake, "A theoretical approach to the formation of quality of experience and user behavior in multimedia services," in *5th ISCA/DEGA Workshop PQS 2016*, 2016, pp. 39–43.

[2] B. Tedeschi, "Seeking ways to cut the web-page wait," *New York Times*, vol. 6, pp. 14–99, 1999.

[3] A. Web, "Is your website too big," *Accounting Web*, vol. 19, 2000.

[4] F. F.-H. Nah, "A study on tolerable waiting time: How long are web users willing to wait?" *Behaviour & Information Technology*, vol. 23, no. 3, pp. 153–163, 2004.

[5] J. Nielsen, "Top ten web design mistakes of 2005," *Posting in Jakob Nielsen's Alertbox. Retrieved February*, vol. 14, p. 2007, 2005.

[6] D. Rossi, M. Mellia, and C. Casetti, "User patience and the web: A hands-on investigation," in *IEEE GLOBECOM'03*, IEEE, vol. 7, 2003, pp. 4163–4168.

[7] D. Collange, M. Hajji, J. Shaikh, M. Fiedler, and P. Arlos, "User impatience and network performance," in *Next Generation Internet (NGI), 2012 8th EURO-NGI Conference on*, IEEE, 2012, pp. 141–148.

[8] S. Egger, P. Reichl, T. Hoßfeld, and R. Schatz, ""time is bandwidth"? narrowing the gap between subjective time perception and quality of experience," in *ICC 2012*, IEEE, 2012, pp. 1325–1330.

[9] T. Hoßfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, "Initial delay vs. interruptions: Between the devil and the deep blue sea," in *2012 4th QoMEX*, IEEE, 2012, pp. 1–6.

[10] S. Khirman and P. Henriksen, "Relationship between quality-of-service and quality-of-experience for public internet service," in *In Proc. of the 3rd Workshop on Passive and Active Measurement*, vol. 1, 2002.

[11] D. Strohmeier, M. Mikkola, and A. Raake, "The importance of task completion times for modeling web-qoe of consecutive web page requests," in *2013 5th QoMEX*, IEEE, pp. 38–39.

[12] D. Strohmeier, S. Egger, A. Raake, T. Hoßfeld, and R. Schatz, "Web browsing," in *Quality of experience*, Springer, 2014, pp. 329–338.

[13] T. Hoßfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler, "The memory effect and its implications on web qoe modeling," in *Teletraffic Congress (ITC), 2011 23rd International*, IEEE, 2011, pp. 103–110.

[14] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, 2010.

[15] F. Dobrian, A. Awan, D. Joseph, J. Ganjam A. Zhan, V. Sekar, I. Stoica, and H. Zhang, "Understanding the impact of video quality on user engagement," *Communications of the ACM*, vol. 56, no. 3, 2013.

[16] Z. Chen, L. Cui, Y. Jiang, and Z. Wang, "Understanding viewing engagement and video quality in a large-scale mobile video system," in *IEEE Symposium on Computers and Communications*, IEEE, 2017.

[17] T. Hoßfeld, P. E. Heegaard, L. Skorin-Kapov, and M. Varela, "No silver bullet: Qoe metrics, qoe fairness, and user diversity in the context of qoe management," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2017, pp. 1–6.

[18] N. Bouten, S. J. Latral Famaey, W. Van Leekwijck, and F. De Turck, "In-network quality optimization for adaptive video streaming services," *IEEE Trans. on Multimedia*, vol. 16, no. 8, pp. 2281–2293, 2014.

[19] E. Liotou, D. Tsolkas, N. Passas, and L. Merakos, "Quality of experience management in mobile cellular networks: Key issues and design challenges," *IEEE Communications Magazine*, vol. 53, no. 7, pp. 145–153, 2015.

[20] P. Reichl, S. Egger, S. Möller, K. Kilkki, M. Fiedler, T. Hossfeld, C. Tsiaras, and A. Asrese, "Towards a comprehensive framework for qoe and user behavior modelling," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, IEEE, 2015, pp. 1–6.

[21] A. Ahmad, A. Floris, and L. Atzori, "Qoe-centric service delivery: A collaborative approach among otts and isps," *Computer Networks*, vol. 110, pp. 168–179, 2016.

[22] H. Takagi, "Waiting time in the m/m/m/$(m + c)$ queue with impatient customers," *International Journal of Pure and Applied Mathematics*, vol. 90, no. 4, p. 519, 2014.

[23] A. Ahmad, A. Floris, and L. Atzori, "Ott-isp joint service management: A customer lifetime value based approach," in *Integrated Network and Service Management (IM), 2017 IFIP/IEEE Symposium on*, IEEE, 2017, pp. 1017–1022.