On test principles for a QoE evaluation using real services

overview on methodology and challenges for defining test principles

Albrecht Kurze
Chair Media Informatics
Technische Universität Chemnitz
Chemnitz, Germany
Albrecht.Kurze@cs.tu-chemnitz.de

Maximilian Eibl
Chair Media Informatics
Technische Universität Chemnitz
Chemnitz, Germany
eibl@cs.tu-chemnitz.de

***Abstract*— We report on our experiences from two user studies (lab experiments) with nearly 300 participants for QoE evaluation using real mobile services and devices in our WiFi network emulation testbed. We briefly introduce our principles for integrating real services in these studies: how we selected relevant services, how we investigated their testability and how we tested them with high efficiency.**

Keywords*— *QoE Evaluation, User Study, Empiricism, Mobile Services, Quality of Service (QoS), Quality of Experience (QoE), Mobile User Experience

I. INTRODUCTION

QoE evaluations mean empirical studies. For these studies it is essential on what model assumption they are based (black box vs. white box), how they are designed (lab experiment vs. field trial) and how the elements (e.g. network, devices, services) relevant for QoS and QoE are integrated (model, simulation, emulation, and real elements/systems). Classic QoE (lab) experiments use standardized test sets (e.g. [1] for web browsing) or emulated services itself (e.g. "Fakebook" in [2]). In conjunction with network emulation relevant influencing factors and elements (network, software, content, and remote server side) are under full control in some kind of sand boxed test environments. In speaking of test quality criteria this normally delivers results with high objectivity and reliability. But sometimes this suffers from low (external) validity in the meaning of applying such results to the real world.

In our studies we aimed at characterizing the QoS-QoE-relationship of relevant mobile services and at finding thresholds for necessary QoS parameter (mainly throughput and latency) for a "good enough" for these services. We decided to evaluate real services as this promised to generate directly usable results. Another benefit is that real services already bring along the client and server side software as well as appropriate content. We created a distributed network emulation testbed based on Linux (Openwrt) with tc+netem [3] on ordinary hardware WiFi-routers. We added several Web-APIs, a Web-GUI and coupled our system with Limesurvey (web survey software) for easy and highly automated test operation. This way we were able to use real services and devices (multiple form factors and platforms). We faced several challenges in evaluating real services in a setup like this. We learned a lot from all the glitches in our first study and successfully improved our test design and test procedure. On this base we formulated three test principles for real services.

II. TEST PRINCIPLES FOR REAL SERVICES

Test relevance: As it is impossible to test all (real) services available it is necessary to select some services. It is a viable way to concentrate on services with a high relevance. The definition of what is relevant might be based on several aspects and might be determined on several ways: from a theoretical (scientist's) perspective with relevant service classes and QoS-QoE-characteristics, or with high practical relevance from a users' point of view (popular, often and heavily used), or from mobile network operators' view (high load in networks, high demand, and high importance for customer satisfaction). We used public available data sources (e.g. AGOF Digital Facts, Sandvine Global Internet Phenomena, Ericsson Mobility Report, Cisco Visual Network Index (VNI) Forecast) and confirmed the services' popularities in our own user surveys in [4]. This way we selected six mobile services with high relevance now or in near feature: Google Drive (cloud storage), Facebook (social networks), Google Maps (LBS and navigation), MTV-Music (music streaming), Spiegel.de (popular german news site, mobile browsing), and Youtube (video streaming).

Testability: In [4] we have identified six distinct facets that define how testable a service is: reality, controllability, observability, usability, heterogeneity, and repeatability. We realized that each facet required intensive and time consuming preliminary investigations before starting our user study. Unfortunately, the facets are interlocked and cannot be easily separated. Therefore, we have identified the important relations between the facets that have to be kept in mind when selecting specific service elements (e.g. app or content), defining test scenarios, and analyzing the results.

Reality refers to the network situation as well the as the usage situation and all elements involved which influences all other facets. For the usage situation it should be as real as possible in all directly user observable aspects (real device, app and content). For the network it should be as real as possible in further effect (but not necessarily in cause). Differences between the different base technologies in the field (mobile networks, 3G/4G) and in the lab (network emulation with WiFi) and how the services and devices handle it have to be taken into account.

Controllability refers first of all to the QoS network parameters but also to the content used and anything else that causes QoE-relevant user observable effects. One key part of experiments is the active control over the independent variables.

We combined planned network condition assignments in our testbed with a passive control of external and random influence factors by monitoring network traffic and device outputs for later on-demand analysis to take into account unplanned variation.

Observability refers to all user observable factors that directly influence the QoE and especially to the QoS-dependent effects. There meet QoS-dependent performance, application-dependent presentation and user-dependent perception. We differentiate between three categories of user observable QoS-dependent stimuli: time and duration, effects and behavior (e.g. stalls), quality and quantity (mainly of content, e.g. HQ vs. HD).

Usability (in our meaning here) refers to the user interfaces of an app/website as part of human computer interaction as well as how usable a service is in a user study as a test scenario (with a specific task). Real services mean real apps – and this sometimes means bad usability. A bad usability might influence how QoS-dependent performance as well QoS-independent factors take effect in a user's perspective (observability).

Heterogeneity is caused by the test design, by different devices in the testbed (platform, form factor, performance etc.) and by introducing task variants necessary for task repetitions (within-subject variation) with relevant changes from pre- to post-conditions. The heterogeneity might cause distinctions in the other facets and might by this define service specific sub-groups in the result with different QoS-QoE-relationships.

Repeatability means to create comparable conditions and results that can be aggregated across multiple test runs and over the whole study at all. Therefore, the same QoS parameter settings (controllability) should cause the same or at least very comparable stimuli (observability). External factors might change over different time scales and will have a random effect caused by the named elements of reality outside of active control. Internal influence factors effecting repeatability are caused by the test design itself, often in conjunction with heterogeneity.

Test efficiency: One basic principle of empirical studies is to select the right input variables, to vary them meaningfully and to measure the right output variables. We narrowed the parameter space and focused on the most influential dimensions for the selected services. Instead of performing a one-factor-at-time input variation we defined and utilized mobile tech-oriented (typical) parameter combinations (tuples of max. download, max. upload, and min. latency, fig. 1). We estimated the range and stepping between lower and upper boundaries top down on model assumptions from expected (and wanted) typical QoE assessment over directly influencing QoE stimuli back into the QoS parameter space (described as "QoE Engineering Process" in [5]). We selected satisfaction as relevant QoE feature (more subjective/objective measures in [6]) and measured it with subjective self-assessments on a 5-step scale. For each service we collected assessments of satisfaction overall as well as satisfaction of two or three service specific sub-features (e.g. download time, video playback behavior). We wanted to see the change from slightly bad assessments (unsatisfied) to full overall satisfaction. Therefore, we selected five suitable QoS parameter variations for each service centered on the point of highest probability of uncertainty (where we expected a good mixture of assessments among the participants).

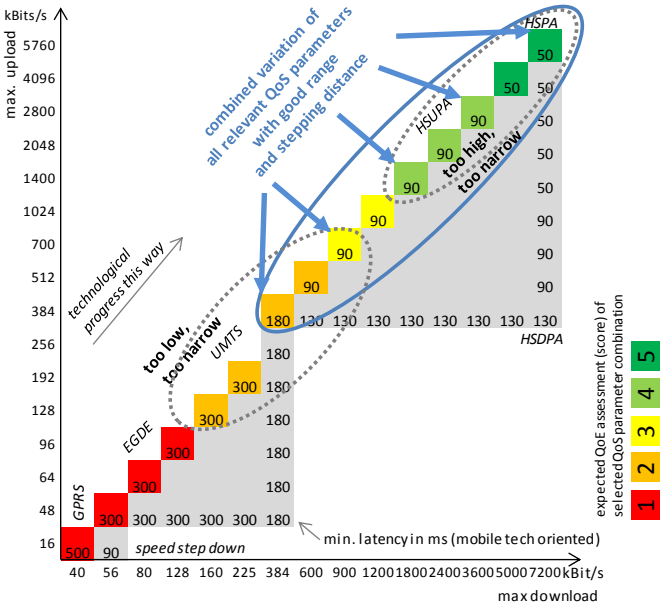

Fig. 1. Planning test efficiency, adapted from [4]
axes and cell values: QoS network performance parameters of max. download (x-axis), max. upload (y-axis), and min. latency (cell value), with viable assumptions on parameter range and stepping distances derived from different tech specification and network measurements
filled triangle shaped areas: typical QoS parameter combinations lie within this boundaries (depending on mobile tech, load, reception)
colored cells *(on upper boundary)*: expected typical QoE assessment of this QoS parameter combination (example for a specific service)
diagonal bubbles: selections of five QoS parameter combinations each
reading example: {3600, 2048, 90} is a typical combination for HSUPA networks and is expected to lead to a typical (mean) opinion score of 4 for a service XYZ (example)

III. Summary and Acknowledgment

We introduced three principles for integrating real services in QoE evaluations which we have learned from two user studies. The work is based on a project funded by industry from telecommunications sector performed in interdisciplinary cooperation of the chairs of Prof. Eibl, Prof. Krems, and Prof. Heinkel. Further explanations and details can be found in [4].

References

[1] "ETSI TR 102 505 V1.3.1: Speech and multimedia Transmission Quality (STQ); Development of a Reference Web page," European Telecommunications Standards Institute (ETSI), Technical Report, Nov. 2012.

[2] S. Egger, "Interactive Content for Subjective Studies on Web Browsing QoE: A Kepler derivative," presented at the ETSI Workshop on Selected Items on Telecommunication Quality Matters, Wien, 28-Nov-2012.

[3] S. Hemminger, "Network Emulation with NetEm," in *linux.conf.au*, Canberra, Australia, 2005.

[4] A. Kurze, "Modellierung des QoS-QoE-Zusammenhangs für mobile Dienste und empirische Bestimmung in einem Netzemulations-Testbed," PhD thesis, Technische Universität Chemnitz, Universitätsverlag Chemnitz, ISBN 978-3-944640-60-0, http://nbn-resolving.de/urn:nbn:de:bsz:ch1-qucosa-195066, 2016 (in press).

[5] Digital Subscriber Line Forum, "TR-126: Triple-play Services, Quality of Experience (QoE) Requirements," Digital Subscriber Line Forum, Architecture & Transport Working Group, DSL Forum Technical Report TR-126, Dec. 2006.

[6] P. Brooks and B. Hestnes, "User Measures of Quality of Experience: Why Being Objective and Quantitative Is Important," *IEEE Netw.*, vol. 24, no. 2, pp. 8 –13, Apr. 2010.