VoIP-based Calibration of the DQX Model

Christos Tsiaras, Manuel Rösch, Burkhard Stiller
University of Zurich, Department of Informatics (IFI), Communication Systems Group (CSG)
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland
tsiaras@ifi.uzh.ch, manuel.roesch@uzh.ch, stiller@ifi.uzh.ch

Abstract—**In the Internet Protocol (IP) ecosystem, Quality-of-Experience (QoE) is important information needed by Service Providers (SP) to improve their services. However, end-user's satisfaction, which can be reflected by QoE metrics, cannot be easily measured like technical variables, such as bandwidth and latency. QoE can either be estimated through mathematical models or it can be measured through an experimental setup. In this work a Voice-over-Internet Protocol-based (VoIP) QoE measurement setup has been designed to capture end-user's QoE in VoIP services. The data measured during these experiments are used to define all necessary parameters of the Deterministic QoE model (DQX) in this VoIP scenario. Such a calibration of the model is essential to adapt it to the particular service and its technical and non-technical conditions in which it is used. Furthermore, those DQX results achieved are compared with those results of the IQX Hypothesis and the E-Model, being proposed by the ITU-T. Thus, it is finally shown that DQX can capture more accurately end-user's QoE in VoIP scenarios.**

Index Terms—**Quality-of-Experience (QoE), Voice-over-IP (VoIP), Mean Opinion Score (MOS), E-model, IQX Hypothesis, DQX model**

I. Introduction

Quality-of-Service (QoS) is defined application-specifically by a value threshold of technical variables such as latency, packet loss, and bandwidth. These values are well known for different technologies and services and they can be measured [11]. Furthermore, selected values of those variables are often used for marketing purposes, *e.g.*, Mobile Network Operators (MNO) and Internet Service Providers (ISP) advertise "high bandwidth" or "high performance". However, QoS variables are not explicitly linked to the end-user's satisfaction. It is naive to conclude that end-users' Quality-of-Experience (QoE) can be increased by adjusting one QoS variable, because the relationship between QoS variables and end-users' experience depends on the Type-of-Service (ToS). Large latency can serve as an example here, since latency has a higher negative effect on Voice-over-Internet Protocol (VoIP) services than on video streaming [5].

Therefore, this work here is focused on defining the deterministic relationship between QoS variables and QoE in the VoIP scenario. The four step QoE formalization methodology in this work reads as follows: (1) Define an experimental setup allowing for the emulation of various network connection performance settings on jitter, latency, packet loss, and bandwidth; (2) perform test VoIP calls in pre-defined experimental setups and collect QoE-related feedback from end-users in terms of Mean Opinion Scores (MOS) [7]; (3) use the feedback collected to determine through non-linear regression the Deterministic QoE (DQX) model parameters for different variables in this VoIP scenario; and (4) compare DQX [24] to the two QoE-predicting models, the exponential relationship connecting QoS parameters, called IQX Hypothesis [3], and the E-model [12] of the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T).

In support of the accurately and timely measurement of QoE for VoIP services a Web Real-Time Communications (WebRTC) VoIP client was newly designed and developed such that it collects directly all necessary user feedback from experimental VoIP calls under different network conditions in various scenarios. Those varying network conditions are emulated by the network emulation framework WANem [23], which utilizes a real network. Therefore, three computers were attached to each other through a switch via Local Area Network (LAN) cables. Using such an experimental architecture guarantees a fully controlled network emulation that is not influenced by external traffic.

This experimental setup served for the collection of more than 500 data points and was used to evaluate how accurate the previously mentioned QoE-predicting models [3][12] reflect these collected data points. In this experiment, it was shown that the DQX model is the most accurate model to capture QoE in given scenarios. Moreover, there exist by now two additional evaluations of the DQX model concerning (a) the influence factor of a variable that affects QoE and (b) the proposed equation that estimates QoE, when multiple variables are considered simultaneously.

The remainder of this work is structured as follows. The background and related work is discussed in Section II. Section III describes the design and utilization of the experimental setup. Section IV presents results collected and evaluates the outcome by comparing the DQX model to related work. Finally, Section V summarizes this work, draws conclusions and discusses critical thoughts required for future work.

II. Background and Related Work

While the MOS [7] determines a commonly agreed upon scheme to evaluate VoIP services from an end-user's perspective, different QoE models exist in the literature, capturing QoE for various services. The most well known QoE models are (1) the E-model [4] for VoIP

ISBN 978-3-901882-68-5

services and (2) the IQX Hypothesis [3] which is a generic exponential QoE model. DQX will be compared with those two models, to show that it captures in a more accurate way the end-users' satisfaction.

A. Mean Opinion Score (MOS)

To capture the end-user's experience in QoE experiments, the five-point opinion scale recommended by the ITU [7] was applied. This opinion scale is used in many QoE-related research and determines an excellent basis for comparing results. This scale defines scores from one to five, while each score defines a certain meaning. The ITU recommendation [8] assigns to each score an English word (cf. Table 1).

Table 1: MOS Levels of End-to End Perceived Quality

Score	Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

B. E-model

The E-Model is a transmission planning tool that can be used to predict QoE for a typical telephone user in an end-to-end (e2e) conversational scenario. The model takes a wide range of transmission variables into account and it can be used to assess the voice quality of wired and wireless services, based on circuit-switched and packet-switched technology [4].

The output of this model is — in contrast to other models — not in form of MOS values. The E-model uses the Transmission Rating Factor *R* as output, which can be transformed into MOS and, therefore, it becomes possible to compare the E-model to other models [12], too.

The E-Model uses mathematical algorithms based on the analysis of a large number of subjective tests with a wide range of transmission variables. These algorithms can transform transmission variables into "impairment factors". According to the E-model tutorial [4], five impairment factors are used to calculate the *R* value.

- **Ro:** Expresses the basic signal-to-noise ratio, including various noise sources, such as circuit noise and room noise.
- **Is:** This term takes impairments into account that exist more or less simultaneously with the voice signal, such as, (a) too loud speech level, non-optimum Overall Loudness Rating (OLR), (b) non-optimum Side Tone Masking Rating (STMR), and (c) impairment caused by quantizing distortion.
- **Id:** This factor represents all impairments that are caused by too long absolute delay and potential echo effects on both talker's and listener's side.
- **Ie:** Equipment impairment factor represents impairments that are caused by the respective codec used and packet-loss.
- **A:** The advantage, or expectation factor, considers the advantage of service access. *E.g.*, a user in a region which is hard to provide connectivity, such as regions where a satellite link is needed, expects a lower quality, and therefore, tolerates more impairment.

Equation 1 considers all impairment factors to calculate the R value [12]:

$$R = Ro - Is - Id - Ie + A \qquad (1)$$

All impairment factors are calculated through algorithms that take several transmission variables as input. An overview over all variables being used for the calculation is illustrated in Figure 1, where a telephone connection and all impairment factors affecting the quality of the conversation according to the E-Model is illustrated.

As can be seen in Figure 1 the E-Model has a high complexity considering many different parameters. A detailed calculation of each impairment factor can be found in the ITU-T recommendation G.107 [12]. A question of this work is if the DQX model which has a comparatively low complexity can keep up with the E-Model.

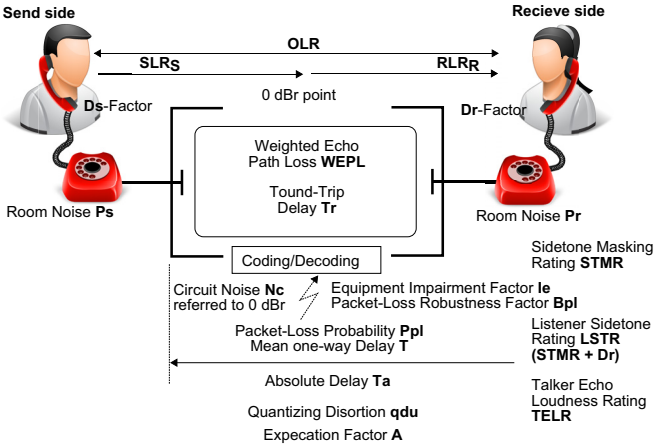

Fig. 1: Reference Connection of the E-Model [12]

C. IQX Hypothesis

The IQX hypothesis proposes a generic formula, which can predict QoE for specific QoS variables. While many formulas in this field are based on a logarithmic relationship between QoS and QoE, such as the ITU-T formula in [6] or the formula in [17], the IQX hypothesis applies an exponential approach. This approach is proven to be more accurate through non-linear regression and comparison of the correlation coefficient [3].

The exponential formula is derived from the idea that the change of QoE depends on the current level of QoE, given the same amount of change of the QoS value. When a linear dependence on the QoE level is assumed, the relationship can be written as a differential equation and this equation can be resolved to the final formula of this Hypothesis (cf. Equation 2) [3]

$$\frac{\partial QoE}{\partial QoS} \sim -(QoE - \gamma) \rightarrow QoE = \alpha \cdot e^{-(\beta \cdot QoS)} + \gamma \qquad (2)$$

Equation 2 contains the three parameters α, β, and γ which must be determined by experimental sessions

where non-linear regression is applied to the collected MOS. Such a regression was made in a VoIP scenario for packet loss and the resulting formula, including the found values for α, β and γ, is presented in Equation 3 and illustrated in Figure 2 [3].

$$QoE = 3,01 \cdot e^{-(4,473 \cdot p_{loss})} + 1,065 \quad (3)$$

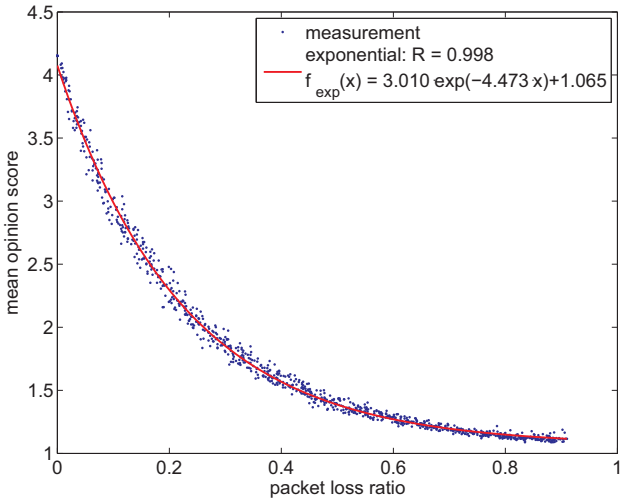

Fig. 2: QoE Mapping Function of Packet Loss Ratio in the IQX Hypothesis [3]

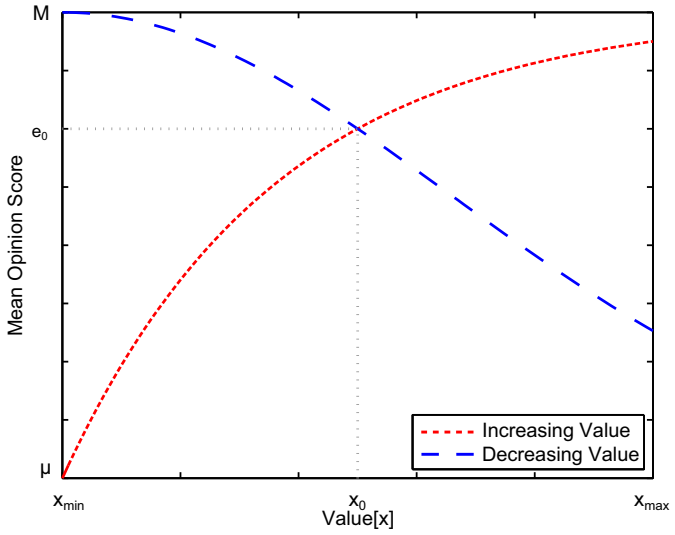

Fig. 3: IV and DV of the DQX Model [24]

D. Deterministic QoE Model (DQX)

The DQX model [24] is illustrated in Figure 3 through an exemplary plot of the two types of variables, which DQX considers affecting QoE. DQX, like the IQX Hypothesis, uses an exponential approach to link QoS variables and QoE. The difference between the two models is that DQX is deterministic and, furthermore, proposes a way to calculate QoE with multiple QoS variables as an input. This work here has its main focus on the DQX model.

For every service, there are diverse technical, such as latency or bandwidth, as well as non-technical variables, such as price, which affect QoE. The model distinguishes between two types of such variables. There are (a) Increasing Variables (IV), which increase the user's satisfaction with their growth, and there are (b) Decreasing Variables (DV), which do the opposite. For all these variables there exists a certain value at which the user is satisfied with the service. These values are called expected variable values and they are either defined in the Service Level Agreement (SLA) between the service provider and the customer, or by service-specific constraints. The expected variable value — see Figure 3 in the DQX formalization — is the x_0 value and the end-user's satisfaction corresponding to x_0 is defined as e_0.

The idea of the model is that there is a minimum user satisfaction μ and a maximum user satisfaction M. The IV curve begins in (x_{min}, μ) and crosses (x_0, e_0) towards (x_{max}, M). For the DV curve it is vice versa, QoE begins with the user satisfaction (x_{min}, M) and decreases through (x_0, e_0) towards (x_{max}, μ). The benefit of the DQX approach is that it adds logic behind the IQX hypothesis parameters α, β, and γ, which are defined in [3] through a non-linear regression.

Since the shape of the graph differs for each service, technology, and user-base, the model introduces the influence factor m. As illustrated in Figure 4, the m value leads to different shaped graphs. For small values the graph is flat and for high values of m it steepens [24]. Thus, m shows how fast QoE is affected by a given fluctuation of a variable.

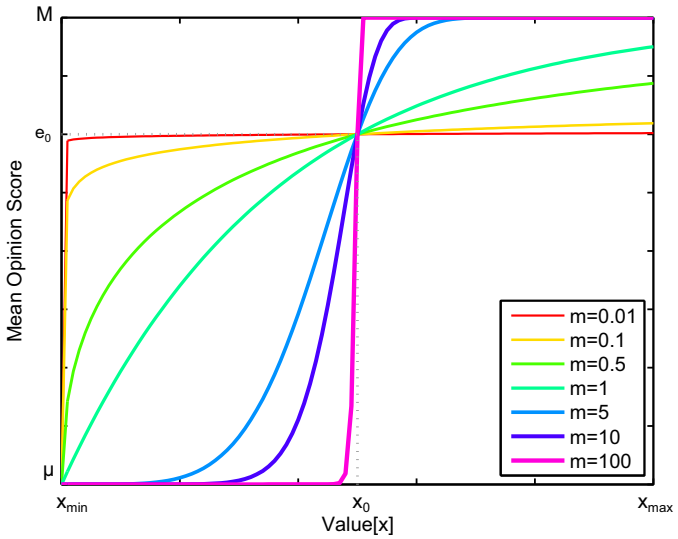

Fig. 4: IV Example with Different m Values in DQX Model

Moreover the model differs between the m value describing the QoE curve above and below x_0. Depending on the variable's value x, m is called m^+ if $x>x_0$ or m^- if $x<x_0$. For $x=x_0$, m does not influence the QoE curve at all [24]. In general m might be a function of the variable's value x ($m=f(x)$). However, in this work only two values of the influence factor have been considered. The m value used for the graph for x above x_0 is called m^+ and the m value used for the graph for x below x_0 is called m^-.

The equation for IV is defined in Equation 4. The parameter h stands for the difference between the maximum and the minimum QoE score ($h = M-\mu$), m is the

influence factor and λ is a coefficient that is defined through the expected variable value x_0. Equation 5 shows how λ can be derived from Equation 4 [24]. For DV there is an analog equation (cf. Equation 6), where the factor λ is defined like in the IV case [24].

$$e_i(x) = h \cdot (1 - e^{-(\lambda \cdot x^m)}) + \mu \tag{4}$$

$$e_i(x_0) = e_0 \Leftrightarrow \lambda = x_0^{-m} \ln\left(\frac{h}{h - e_0 + \mu}\right) \tag{5}$$

$$e_d(x) = h \cdot e^{-\lambda \cdot x^m} + \mu, \lambda = x_0^{-m} \ln\left(\frac{h}{e_0 - \mu}\right) \tag{6}$$

The DQX model also introduces an equation for multiple variables. The equation uses the single variable equations (Equation 4 and Equation 6) and combines them as seen at Equation 7 [24].

$$E(x) = \mu + h \cdot \prod_{k=1}^{N} \left[\frac{e_{i \vee d}(x_k) - \mu}{h}\right]^{w_k} \tag{7}$$

The idea of this Equation 7 is that even if only one variable outperforms the overall QoE score will still reflect it. Thus, for every variable a QoE score is calculated. All respective scores are weighted with the exponent w_k depending on the relevance of the particular variable and multiplied. Finally, the resulting percentage is applied to the rating scale.

Since the DQX model has no defined rating as an outcome, the maximum value M, the e_0 value, and the minimum value μ must be defined in advance and the output will be according to that. Respecting the ITU-T MOS scale, in this work here, the maximum selected is M=5, e_0 is 4, and the minimum is μ=1. Thus, the difference h between the two parameters is therefore, 4. Inserting these parameters into Equation 4, Equation 6, and Equation 7, results in the following formulas that produce MOS-compliant DQX score values [24].

Increasing Variable:

$$e_i(x) = 4 \cdot (1 - e^{-(\lambda \cdot x^m)}) + 1, \lambda = x_0^{-m} \ln(4) \tag{8}$$

Decreasing Variable:

$$e_d(x) = 4 \cdot e^{-\lambda \cdot x^m} + 1, \lambda = x_0^{-m} \ln\left(\frac{4}{3}\right) \tag{9}$$

Multiple Variables:

$$E(x) = 1 + 4 \cdot \prod_{k=1}^{N} \left[\frac{e_{i \vee d}(x_k) - 1}{4}\right]^{w_k} \tag{10}$$

III. Experimental Setup

The experimental setup is based on the WANem framework for network emulation and a WebRTC messenger that was implemented during this work. Moreover the ITU recommendation P.800 [8] was considered for the experimental procedure.

A. Architecture (H/W and S/W)

A key element of the QoE evaluation architecture used in this work is a WebRTC messenger called "QoEssenger". WebRTC is World Wide Web Consortium (W3C) draft standard for real-time communication between browsers [26]. The goal of WebRTC is plug-in-free low-cost communication in real-time between any browser. And with communication not only audio and video communication is meant, but also the direct exchange of data. So with the help of WebRTC it is possible to create a Peer-to-Peer (P2P) connection from browser to browser and send audio, video and data over it. This WebRTC approach has been chosen because it is the new trend in the field of the VoIP communication, it is open source and there is no software necessary other than a browser on the client side [25].

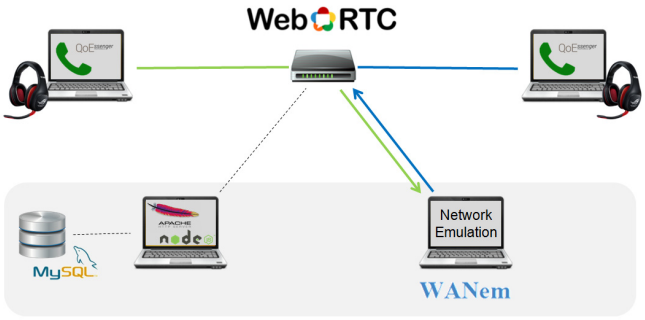

Fig. 5: Experimental Setup

Additionally, the WANem [23] framework for network emulation was used. This software, which is based on the Linux Operating System (OS) Knoppnix [18], is convincing because it makes use of the well-accepted open source network tool for Linux called NetEm [15][20]. Furthermore, WANem provides the possibility to build a UI on top of it, which facilitates the flexibility to create a control panel that meets with precision the demands of an QoE evaluation experiment, such as the easy setup of different scenarios with diverse latency, bandwidth, packet loss and jitter settings to be tested, as well as the automated collection of the user ratings.

The architecture using WANem is illustrated in Figure 5. There are four computers connected in a local LAN through a switch. Two computers run the QoEssenger, one computer runs an Apache [1] web server, a node.js [14] signaling server as well as a MySQL [1] data base and the last computer runs the WANem tool that can emulate the network. The WANem works as follows: The routing table of the two computers that run the QoEssenger is modified in such a manner, that all the packets are routed to the other peer through the computer that runs WANem. This computer is responsible for the network emulation. *E.g.*, if the packet loss is set to 50%, the WANem computer will drop every second packet that is routed through.Such architecture with LAN cables and a switch is necessary to guarantee a controlled network environment without the interferences that happen in a Wireless LAN (WLAN) network.

B. Experimental Procedure

The experimental procedure besides the hardware and software -related information, includes some important information concerning (a) the participants group (subjects) and (b) the procedure of the experiments.

The subjects of the experimental procedure of this work were 34 volunteers at the University of Zurich and the High School of Willisau in Switzerland. The volunteers were mainly computer science students between 20 and 25 years old. However, the overall age distribution range of the subjects was from 16 to 63 years old. A pair of two randomly selected subjects participated in several voice calls with different technical parameters. Each subject rated the quality of each call separately after the end of the call.

The goal of the overall experimental procedure was to affect as little as possible the QoE rating of each subject. Firstly, the number and the duration of the test calls defined carefully, so that the experiment about the human experience would not demand from a subject to actively participate for more than 30 minutes in voice calls. Otherwise, it was assumed that the subjects would become annoyed and/or bored and their answers would be influenced by emotions which would decrease the quality of the results. Thus, to avoid such situation, the total duration of each experimental session designed to not exceed one hour.

Having a fixed interview length influenced the decision concerning the number and the duration of the test calls. There is a trade-off between the number of measurements and the confidence of the results. If the test calls are longer, fewer experiments can be performed within a fixed time-frame. It was assumed that people are not able to have a free and balanced conversation of 45 second on their own since it does not seem to be enough time to develop a proper conversation, especially between strangers. Thus, the following method was used to support the conversation in the test calls: At the beginning of the experiment each participant got around 300 easy general knowledge questions [16][21] and the subjects had to ask and answer them alternately. This approach led to a fluent and balanced conversation without distracting the subjects from their evaluation task.

The decision about the procedure of the interview was as follows:

0-5 min Introduction, explanation of the experiment and rating system

5-25 min 16 Test Calls, around 45 seconds calling time + 15 seconds voting time each

25-30 min Question and Answers about the calling experience

IV. Results

The following evaluation is based on the MOS of 34 subjects, which produced in total more than 500 end-user's opinion score ratings at an overall calling time of approximately 6 hours. 80% of these ratings were collected in a single variable scenario, where only one variable was adjusted. The remainder of these ratings were mixed variable scenarios, where multiple variables were adjusted. The main focus of this work's QoE measurements was on the single variable scenarios, since DQX only demands knowledge of expected variable values and influence factors of individual variables to predict QoE considering multiple variables. The primary goal of this work was to calibrate DQX for the VoIP scenario. Thus, the data were collected for this purpose. The secondary goal of this work was to validate the DQX prediction performance in the multiple variables scenario. It was assumed that less but equally distributed set of data points would be sufficient to reveal the potential inaccuracy of the DQX model.

Although the DQX model can be calibrated for diverse technical and non-technical variables, in this work, the focus is on four technical variables: jitter, latency, packet loss and bandwidth. The DQX model needs for every technical variable an expected variable value x_0 [24]. The WebRTC technology that was used in this work is relatively young and rich literature, that can be used to find possible expected variable values, does not yet exist. Thus, the x_0 values and references as of Table 2 were used for this evaluation.

Table 2: x_0 and References Used for the Evaluation

Variable	x_0	Reference
Latency	150 ms	ITU [5][9]
Jitter	100 ms	Cisco [2]
Packet Loss	5%	Opus Documentation [22]
Bandwidth	64 kbit/s	WebRTC Official Blog [27]

A. Single Variables

Here those results of scenarios are presented, where only one variable was affected. *E.g.*, in a scenario with 5% packet loss, the latency and jitter was set to 0 ms and the bandwidth was unlimited. Table 3 summarizes the influence factors (m) values found for each variable and the Goodness-of-Fit (GoF) for each m expressed as R^2 value.

Table 3: Results of the Single Variable Scenarios

	Latency	Packet loss	Jitter	Bandwidth
m^+ R^2	0,40 -0,65	0,09 0,85	1,06 0,96	4,53 0,75
m^- R^2	0,32 0,75	0,73 0,95	0,59 0,96	0,47 0,94

These high R^2 values show that the DQX model is able to capture QoE of end users quite accurately.

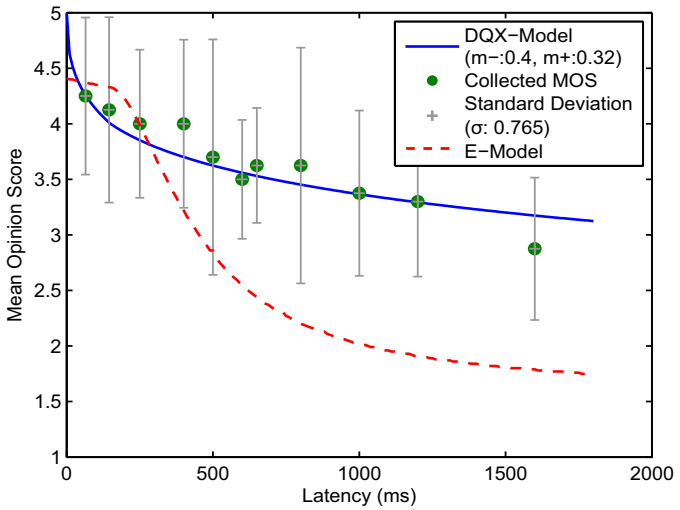

Fig. 6: DQX Model Fit and Comparison for Latency

The resulting graphs of the DQX model were also compared to the other two QoE-predicting models during this work. Figure 6 shows two plots which show the comparison between the DQX and the E-model (dashed line) QoE results as a function of latency. IQX Hypothesis is missing in Figure 6 because there are no IQX data available for latency.

It is noticeable in Figure 6 that the E-Model proposes lower MOS values than the fitted DQX model most of the time (there is not an equation proposed by IQX modeling MOS and latency). So for example for a latency of 1600ms the MOS for the E-Model is 1.79 and for the DQX-Model 2.875. The DQX QoE value is probably too high for such a high latency value. The reason for such high values could be the following. Latency is something that is not directly annoying, like a bad audio quality. It is something that gets more annoying the longer and faster a conversation becomes. Latency is not that disturbing in a short conversation with small talk characteristics. The conversations of the experiments had exactly these characteristics. Thus, it was assumed that the subjects did not report low MOS for high latency.

Since the collected MOS values seemed rather high, some extra experiments were performed with longer experimental calls in which only latency was tested. For these calls, three different conversational tasks proposed by the ITU-T were tested: (a) a travel office role play, (b) a random number verification task and (c) a contacts exchange task [11]. The results of these extra tests were unexpected. The subjects rated still high. For a test scenario with 1500ms latency the MOS was still 3.17. Therefore, it is further assumed that not only the duration and the type of conversational task are responsible for the unexpected outcome. It is possible that a cultural phenomenon leads to such results. As stated in [11] MOS can vary due to cultural differences. Except four subjects, all of them spoke Swiss German which is a rather slow language and therefore latency probably disturbs less. This hypothesis is supported by a test call between a Russian and an Italian participant held in English which seemed to be faster and more interactive than most of the native Swiss German speakers' conversations. However, a proof of this phenomenon could not be found in the literature and since the sample was not large enough it stays only a hypothesis.

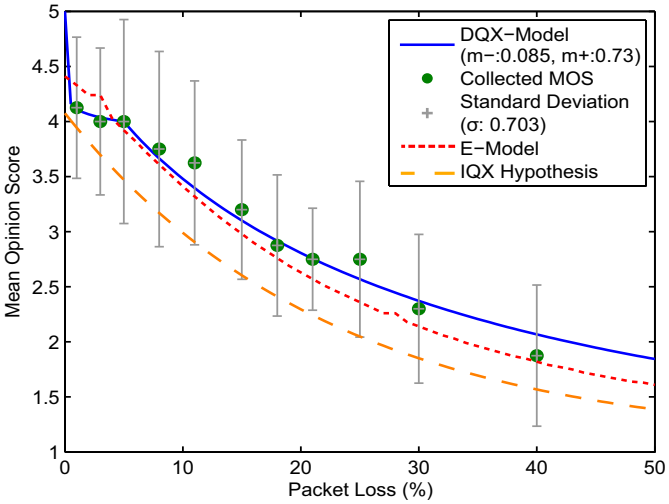

Fig. 7: DQX Model Fit and Comparison for Packet Loss

This experiment showed that E-model would be incapable to spot such behavior while DQX could predict the right MOS with high accuracy with the right influence factor selection. Thus, it is shown that E-model is not suitable for every VoIP scenario. The MOS depends on the service and the respective users. Therefore a model that allows flexible calibration, such as DQX, is needed to predict QoE accurately in diverse scenarios.

In Figure 7 the MOS as a function of the packet loss is illustrated and compared to the E-Model and the IQX Hypothesis. The DQX model appears to capture QoE better than the E-model and the IQX hypothesis. In the IQX hypothesis work [3] the Internet Low Bitrate Codec (iLBC) was used during the measurements. For the E-Model calculation tool [4] the G.711 Codec [13] is assumed, and the codec in the experiments of this work was Opus [22]. The reason of having different codecs is that there was no control in codecs used in the related work and the use of the same codecs was not possible in WebRTC by the time of the experiments. However, Opus has advanced error correction mechanisms similar to the most advanced version of G.711 that is used in the E-Model and the one of the IQX hypothesis experiments, therefore the results presented here are comparable.

Another important result of this work is the analysis of the development of the influence factor values (m). During the analysis of the experiment's results, it was examined if the m, which needs to be determined empirically, remains constant, or if it is necessary to further adjust it during the evolution of a variable's value (x). This analysis has been conducted by fitting the DQX model through all two neighbored data points and the so determined m values were compared to each other. Figure 8 illustrates this comparison and shows that m should not be considered to be constant in every case. Just for the variable jitter the influence factor m values appear to be almost constant, since m oscillates around an almost horizontal line. Thus, as a result, it can be said that treating

m as a constant is only valid for small fluctuations of a variable, and further research towards the selection of *m* values must be found in future work.

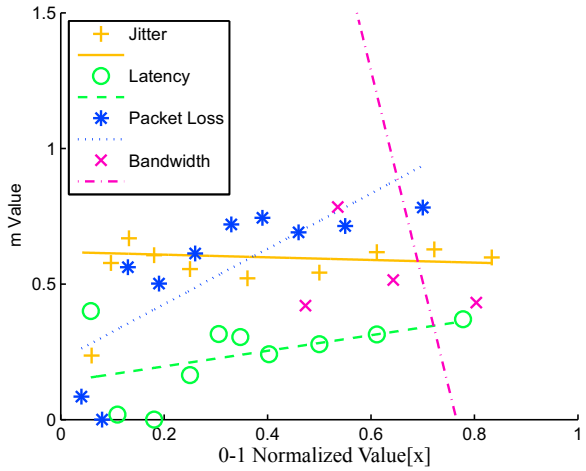

Fig. 8: Development of the *m* Values

B. Multiple Variables

This part evaluates the DQX model for scenarios where multiple variables were tested. Since the main focus was on single variable tests, there are not so many data points for these mixed scenarios. However, it was sufficient to run basic evaluations for different values of latency (*L*), packet loss (*PL*), jitter (*J*), and bandwidth (*B*).

Table 4: Collected MOS for Mixed Variables Compared to the Calculated MOS

L (ms)	*PL* (%)	*J* (ms)	*B* (ms)	*Collected MOS* (std. dev.)	*DQX*
600	10	0	0	**3,13 (0,64)**	**2,59**
500	7	0	0	**3,56 (0.73)**	**2,82**
500	10	0	0	**3,00 (0.67)**	**2,63**
500	10	0	60	**3,25 (0,46)**	**2,05**
400	0	0	75	**3,38 (0.74)**	**3,09**
400	7	0	0	**3,25 (0,71)**	**2,87**
400	20	0	75	**2,50 (0.93)**	**1,95**
250	10	0	0	**2,80 (0.63)**	**2,77**
0	7	0	64	**3,88 (0.64)**	**3,08**
0	7	0	98	**3,88 (0.64)**	**3,26**
0	10	0	60	**3,25 (0.46)**	**2,60**
0	12	0	98	**3,25 (0.71)**	**2,89**
0	0	300	63	**3,13 (0.83)**	**2,64**
0	12	400	0	**2,63 (0.74)**	**2,21**

In this work 14 multiple variables scenarios were tested and their results are summarized in Table 4. The first four columns indicate which variable values were tested. Column five and six contain the MOS collected from the subjects in experiments as well as the standard deviations of these collected ratings. The last column shows the MOS calculated by the Equation 10 of the DQX model using the parameters determined by the single variable scenarios.

Comparing these results in Table 4 it has to be noted that Equation 10 creates promising results, since the differences between calculated and collected MOS are small. The mean of all MOS differences is 0.53, which is small for an unadjusted calculation where all weights equal 1. Each variable's weight serves as another degree of freedom, allowing further calibration of DQX. However, there is not a sufficient amount of data points to make any significant statement in this work. Thus, in such cases the additional degree of freedom that DQX allows for could not be used. Considering these high standard deviations of those measurements, another thorough verification should be done in future to validate the accuracy of the DQX model in VoIP scenarios.

The result of this multiple variables scenario (latency and packet loss) is illustrated in Figure 9, where two variables are mixed. The 3D-curve is the calculated DQX model for the two variables and the large black bullets show the MOS collected. The size of these bullets has been chosen for visualization purposes and they should be ideally cut in half by the 3D-curve of the DQX model.

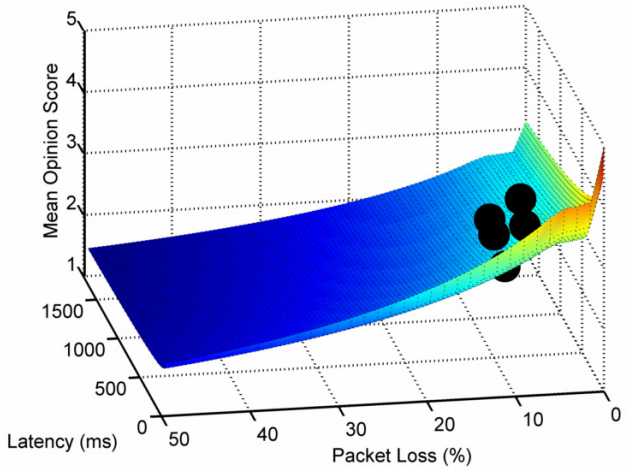

Fig. 9: 3D-Graph of the DQX Model for Multiple Variables

C. Further Calibration of the DQX Model

During this work two further calibrations of the DQX model for the VoIP scenario have been performed. The first one defines an adjusted MOS scale. As with the E-Model under normal conditions a MOS of 4.41 is expected [14], similar findings were made during experiments here. With all subjects an uninfluenced scenario was performed to observe the maximum possible MOS. The result is a MOS of 4.432, being very close to the one from the E-Model. The next assumption made now is that there does not exist any MOS higher than 4.432 and, therefore, the new scale is from 1 to 4.432.

The second calibration which can be done due to the DQX model is an adjustment of the x_0 parameter. Such an adaptation of x_0 is a contradiction to the idea of the

DQX model, because this parameter should be determined before experiments according to the SLA or service characteristics. However, one could argue that often x_0 values are not precisely defined and the reality might vary from proposed values in the literature. Therefore, as a second, further calibration x_0 values can be determined like the m values through a non-linear least square regression.

When these two further calibrations are applied to the DQX model, the GOF improves for all variables in single variable scenarios and this improvement of the GOF results in more accurate MOS calculations in the multiple variables scenario. This means that the mean difference between the MOS collected and the MOS calculated drops from 0.53 (c.f. Subsection B) to 0.21. Such difference is low regarding the fact that there is no calibration of these weight factors.

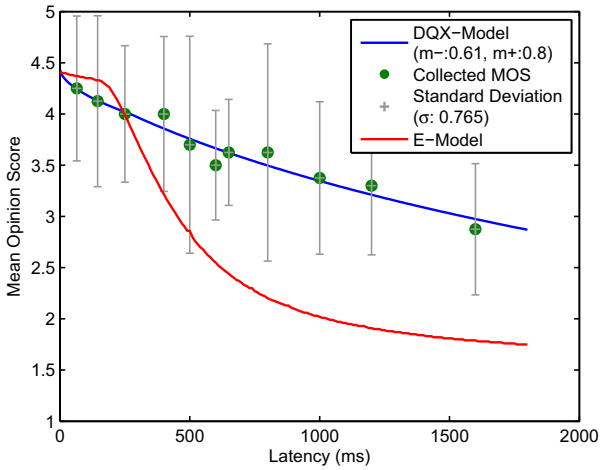

Fig. 10: Adjusted DQX Model Fit for Latency

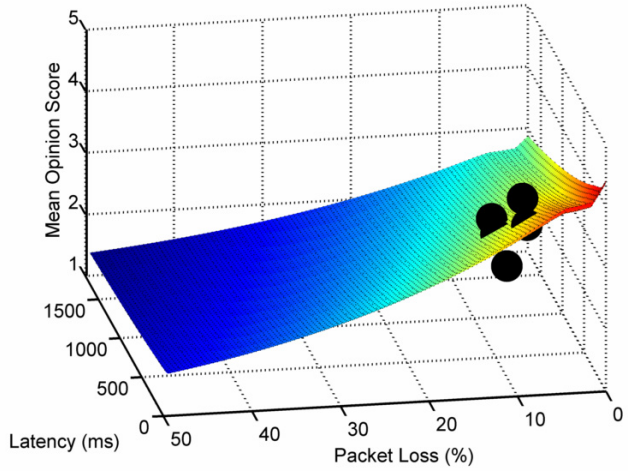

Fig. 11: 3D-Graph of the Adjusted DQX Model

The improvement through this calibration is also illustrated in Figure 10, where the DQX model is further calibrated for the variable delay. Compared to Figure 6, which shows the uncalibrated version of the model, the graph of the DQX model is now closer to the data points collected. Thus, this calibration of the minimum and maximum MOS and the expected variable's value (x_0) leads to a significant increment of the GOF R^2 value, compared to the unadjusted DQX model.

Figure 11 presents the adjusted DQX model 3D-curve in the two variables scenario (latency and packet loss). It can be seen that the 3D-curve cuts through more large black bullets than in Figure 9. Thus, it estimates better the MOS collected.

V. Summary, Conclusion, and Future Work

This paper designed and implemented a QoE measurement experiment setup, which is able to save and replay a sequence of different network scenarios emulations. Moreover, this setup provided the possibility to save user ratings and perform an application-specific analysis with adjustable variables being emulated and encompassing jitter, latency, packet loss, and bandwidth. The VoIP messenger — developed based on the WebRTC technology — collected over 500 data points in experiments with a total of 34 subjects.

The data collected was used to calculate DQX results for each scenario and respective MOS results were used to define those parameters of the DQX model for VoIP services. The evaluation performed three steps for each variable: (1) the DQX model was fitted through the MOS collected and the resulting GOF and the value of the influence factors m were analyzed; (2) the resulting DQX model was compared to the ITU-T E-Model and the IQX hypothesis; (3) these variables were evaluated in a mixed scenario with other variables.

It was shown that the DQX model reaches a high GoF. Moreover, an outcome of the analysis of m values is that they are not constant and further research is required in this area, to determine a model with a non-constant m value. Additionally, this work showed that the formula for multiple variables of the DQX model produces promising results, specifically for the set of measurements with mixed variables, which were performed. The DQX formula adopted for mixed values is also promising, especially if it is used with further calibration techniques concerning the MOS co-domain and the appropriate x_0 selection. All these findings lead to the conclusion that DQX is a highly adaptable and precise model, which outperforms all other state of the art models. Having provided the influence factor m for different services, the DQX model becomes a powerful and useful tool for service providers to predict and improve their services in terms of QoE.

However, in general it has to be stated with respect to the experiments performed that a sample of 34 subjects may not be fully ideal to generate representative data. Moreover the large cluster in the subject's age and gender distribution may not be favorable. The largest part of these subjects were men between 20 and 25 years, since the experiments took place in the Department of Informatics and the majority of students showed a technical background.

The noise-cancellation capable headset used in the experiment was unfamiliar to wear and the strong noise

attenuation impacted the subjects as they could not hear their own voice. On one hand, such a headset was important to guarantee that people can focus on the audio and that they are not disturbed by environmental noise. On the other hand, it is always a bias, whenever people feel uncomfortable during an experiment. Thus, a final recommendation on which headset is ideal for QoE measurements in VoIP services cannot be given.

Additionally, the measurement setup can be extended with further adjustable variables, like packet corruption, reordering, and duplication, since such an adjustment is already foreseen by the framework implemented. Additionally, QoE experiments of other IP-based services with different variables and variable values are foreseen. Next steps in the context of the DQX model will cover an analysis of different services, such as video streaming, Internet browsing, multi-player gaming, and finally in services where non technical variables, such as the price of a service, can also affect QoE. The assumptions made on influence factors m are planned to be confirmed in a larger-scale experiment and the formula for mixed variables will be evaluated in more detail with different weight factors.

Acknowledgments

This work was supported partially by the SmartenIT and the FLAMINGO projects, funded by the EU FP7 Program under Contract No. FP7-2012-ICT-317846 and No. FP7-2012-ICT-318488, respectively. Special thanks go to the Communication Systems Group (CSG), all participants of the experimental sessions, and Tobias Hoßfeld for helping to improve the camera ready version of this work.

References

[1] Apache Friends, XAMPP Installers and Downloads for Apache Friends, URL: https://www.apachefriends.org/de/index.html, Visited in July 2014.

[2] Cisco, Quality of Service for Voice over IP, URL: http://www.cisco.com/c/en/us/td/docs/ios/solutions_docs/qos_solutions/QoSVoIP/QoSVoIP.html, Visited in July 2014.

[3] M. Fiedler, T. Hossfeld, P. Tran-Gia, A Generic Quantitative Relationship between Quality of Experience and Quality of Service, IEEE Network, Vol. 24, No. 2, pp. 36-41, March/April 2010.

[4] ITU-T, E-Model Tutorial, URL: http://www.itu.int/ITU-T/studygroups/com12/emodelv1/tut.htm, Visited in June 2014.

[5] ITU-T, End-user multimedia QoS categories, ITU-T Recommendation G.1010, November 2001.

[6] ITU-T, Estimating End-to-End Performance in IP Networks for Data Application, ITU-T Recommendation G.1030, November 2005.

[7] ITU-T, Mean Opinion Score (MOS) terminology, ITU-T Recommendation P.800.1, November 2006.

[8] ITU-T, Methods for Subjective Determination of Transmission Quality, ITU-T Recommendation P.800, June 1998.

[9] ITU-T, One-way transmission time, ITU-T Recommendation G.114, Mai 2003.

[10] ITU-T, R Value Calculation, URL: https://www.itu.int/ITU-T/studygroups/com12/emodelv1/calcul.php, Visited in July 2014.

[11] ITU-T, Subjective Evaluation of Conversational Quality, ITU-T Recommendation P.805, October 2007.

[12] ITU-T, The E-Model, a computational model for use in transmission planning, ITU-T Recommendation G.107, March 2003.

[13] ITU-T, Transmission impairments due to speech, ITU-T Recommendation G.113, July 2014.

[14] Joyent, Node.js - About, URL: http://nodejs.org/about/, Visited in July 2014.

[15] A. Jurgelionis, J. Laulajainen, M. Hirvonen, A.I. Wang, An Empirical Study of NetEm Network Emulation Functionalities.

[16] Keloo Network, 10000 general knowledge questions and answers, URL: http://www.keloo.ro/doc/10000_intrebari.pdf, Visited in July 2014.

[17] S. Khirman, P. Henriksen, Relationship between Quality-of-Service and Quality-of-Experience for Public Internet Service, 3rd Passive Active Measurement Workshop, March 2002.

[18] K. Knopper, Knoppnix - Live Linux File-system on CD, URL: http://www.knopper.net/knoppix/index-en.html, Visited in July 2014.

[19] T. Krenn, Linux Netzwerk Analyse mit mtr, URL: http://www.thomas-krenn.com/de/wiki/Linux_Netzwerk_Analyse_mit_mtr, Visited in July 2014.

[20] Linux Foundation, netem | The Linux Foundation, URL: http://www.linuxfoundation.org/collaborate/workgroups/networking/netem, Visited in July 2014.

[21] Novascola, Quizfragen und Antworten, URL: http://www.novascola.ch/quizfragen-und-antworten, Visited in July 2014.

[22] Opus, Voice Coding with Opus, URL: http://www.opus-codec.org/presentations/opus_voice_aes135.pdf, Visited in July 2014.

[23] TATA, WANem: The Wide Area Network Emulator, URL: http://wanem.sourceforge.net/, Visited in June 2014.

[24] C. Tsiaras, B. Stiller, A Deterministic QoE Formalization of User Satisfaction Demands (DQX), 39th IEEE Conference on Local Computer Networks (LCN), Sep. 8-11, 2014, Edmonton, Canada.

[25] J. Uberti, S. Dutton, WebRTC Plugin-free realtime communication, URL: http://io13webrtc.appspot.com/, Visited in July 2014.

[26] W3C, WebRTC 1.0: Real-time Communication Between Browsers, June 2014, URL: http://dev.w3.org/2011/webrtc/editor/webrtc.html, Visited in June 2014.

[27] WebRTC, Issue 2025: Opus in low bandwidth conditions, URL: https://code.google.com/p/webrtc/issues/detail?id=2025, Visited in July 2014.