University of Tartu

School of Economics and Business Administration

Oliver Loper

RELATIONSHIP BETWEEN CALL DURATION AND PERCEIVED CALL QUALITY BASED ON SKYPE AUDIO CALL TELEMETRY FROM ANDROID DEVICES

Master Thesis

Supervisors: senior research fellow Oliver Lukason, Kuldar Kõiv

Tartu 2017

Recommended for defense ………………………………

Oliver Lukason

Accepted for defense "   "...................... 2017.

I have written the Master Thesis myself, independently. All of the other authors' texts, main viewpoints and all data from other resources have been referred to.

…………………………………..

Oliver Loper

ABSTRACT

Measuring the impact of a change is essential for quality control, work prioritization, and conducting experiments. The telecommunication industry has studied how Quality of Service based problem handling and coding techniques affect the perceived quality. That has been done mainly relying on simulations conducted in laboratory environments due to lacking the possibility of collecting immediate user feedback.

This thesis demonstrates the relation between call duration and subjective call quality which is statistically significant and continuous. Analysis shows how network degradation, device, device usage, and cultural impacts have significant impact to the relation and related factors separately.

The relation provides a practical workaround for the case of not having subjective quality ratings, and a faster method to collect required sample sizes for statistical analysis. The latter is very useful when the change is relatively small needing more samples to achieve statistical significance or affecting small groups resulting in slower data collection.

Keywords: call duration distribution, QoE, VoIP, MOS, QoS, telecommunication.

1. INTRODUCTION

This thesis is analyzing the impact of quality to user behavior in the context of telecommunication core service – calling. The basis is telemetry analysis with the purpose of relating subjective call quality affecting factors (device form factor, network degradations, device, country) and call duration. The relationship can be used to convert the call quality into revenue and call duration into measurable impact of a change made.

Voice over Internet Protocol (VoIP) enables calling over packet-switched network. The VoIP definition covers a wide range of solutions from on premises deployed central server based schemes to peer to peer applications hosted on mobile devices by end users.

Monitoring the quality of service (QoS) and quality of experience (QoE) is critical to detect existing problems and to measure the impact of planned improvements. QoS is needed to understand the technical parameters, but QoE based models are needed to understand the impact to users.

Improvements can be dynamic addressing a specific condition by adjusting the tradeoff within the technical constraints. Getting it evaluated in laboratory environment is possible if the system parameters are known, but it will not capture unwanted impact to other conditions if there is any. The later a problem is discovered the costlier it will be. Collecting direct user feedback from each call would be desired for an improvement.

The research gap is related to very few publications having the direct user ratings related to calling. There is no mechanism in traditional telecommunication to collect immediate feedback about the call quality. Only one paper (De Pessemier et al. 2015) was found to cover the relation between user satisfaction and call duration, but it was not the primary topic and the analysis method was inconclusive.

More generally, this thesis positions in the literature covering marketing, retailing and consumer relations. In this literature stream, the service quality, customer satisfaction, and customer value are historically amongst the core topics (Oh & Kim 2017: 2–3). In retailing research, it is a long-established fact, that higher perceived quality leads to increased purchases, also in the telecommunications area (Taylor & Baker 1994: 171). Still, due to limitations related to large scale data collection immediately after service consumption, the relations between perceived quality and single purchase have been so far relatively understudied.

The objective of this thesis is to show the relation between call duration and perceived call quality based on immediate consumer feedback, also to analyze the call quality affecting factors. The thesis is structured as follows. Section 2 provides a review of literature on service quality and customer satisfaction, standardized methods in assessing call quality, classification of QoE modelling, standardized objective call quality assessment methods bringing out the factors impacting QoE, concerns related with the objective methods about factors that also affect QoE without having been covered by the models, data driven quality assessment, and practicalities related to conducting controlled experiments. Section 3 describes Skype (from services and organization culture perspectives), provides an overview of the dataset available to the author together with limitations for revealing the business sensitive data, methods used in analysis, and description of variables used in section 4. Section 4 brings out the call duration distribution for all calls and subjectively rated calls, shows the relation between mean opinion score (MOS) and call duration, analyzes if the impacts brought out in literature review are statistically relevant and how much of call duration and MOS variation they explain, shows how these factors impact the relation demonstrated, summarizes the results on analysis, and finally brings out some practical implications. Section 5 provides a brief conclusion of the discussion and analysis offered in section 4.

2. LITERATURE REVIEW

2.1. Service quality and customer satisfaction

In general, this thesis positions into the field of customer satisfaction related empirical studies. This section provides a short overview of service quality and customer satisfaction: frequently quoted definitions and impact to consumer behavior in telecommunication industry.

By comparing relationship quality and transaction-specific quality concepts, Teas (1993:28–30) brought out that depending on the research perspective, definitions are differently used. Lewis & Booms (1983: 99) define service quality as "a measure of how well the service level delivered matches customer expectations. Delivering quality service means conforming to customer expectations on a consistent basis". Service quality centric research originating from Service Quality Model (Parasuraman et al. 1985: 44) uses the same definition.

Another central term related to this thesis is customer satisfaction as defined by Oliver (1999: 41): "fairly temporal postusage state for one-time consumption or a repeatedly experienced state for ongoing consumption that reflects how the product or service has fulfilled its purpose". In this thesis context, the term 'perceived call quality' is capturing the temporal post usage state as it is referring to consumer response to the request to rate the call quality immediately after the call has ended.

Taylor & Baker (1994: 170–172) analyzed telecommunication industry and found statistically significant interaction between the service quality and customer satisfaction, but recommended to conceptualize them as distinct constructs; they also brought out that both are positively impacting the purchase intensions. Gerpott et al. (2001:262) conducted an empirical study based on telecom sector in Germany concluding that the customer satisfaction is leading to customer loyalty and retention. This is aligned with studies from different industries (Mittal & Kamakura 2001: 137; Cooil et al. 2007: 77).

2.2. Typical scale for assessing call quality

This section provides an overview of standardized scale and methods for assessing call quality.

ITU-T Recommendation P.800: Methods for subjective determination of transmission quality defines Absolute Category Rating (ACR) listening quality scale from 5 to 1 as: excellent, good, fair, poor, and bad (International Telecommunication Union 1996: 18). The recommendation also defines Degradation Category Rating on a similar 5-point scale as: inaudible, audible but not annoying, slightly annoying, annoying, and very annoying (International Telecommunication Union 1996: 23).

These scales are used for subjective evaluation in ITU-T Recommendation P.911: Subjective audiovisual quality assessment methods for multimedia applications (International Telecommunication Union 1998: 5–6) and ITU-T Recommendation P.920: Interactive test methods for audiovisual communications (International Telecommunication Union 2000: 9).

Also, ACR is the output of ITU-T Recommendation P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications (International Telecommunication Union 2004: 5), ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs (International Telecommunication Union 2001: 4), and ITU-T Rec P.863: Perceptual objective listening quality assessment (International Telecommunication Union 2014: 4).

2.3. Modeling quality of experience

This section provides an overview of QoE model classifications, complexity of the technical challenges with internet calling, and concerns related to objective modeling based on standards.

2.3.1. QoE modelling

Tsolkas et al. (2017: 2-3) bring out the QoE modeling as subjective, objective, and hybrid on Figure 1. Subjective models can be achieved by 1) controlled experiments needing through design from selecting appropriate evaluators to preparing the environment and scripting the scenario, 2) real service evaluation where user feedback is asked directly during or after providing the service, 3) crowdsourcing where the experiment is conducted on anonymous online users who provide feedback to streamed or downloaded materials. Objective models are separated into 1) media-layer models depending on the use of reference signals, 2) packet-layer getting the input from packet headers and payload, 3) parametric models related to QoS. (Tsolkas et al. 2017: 2–3)

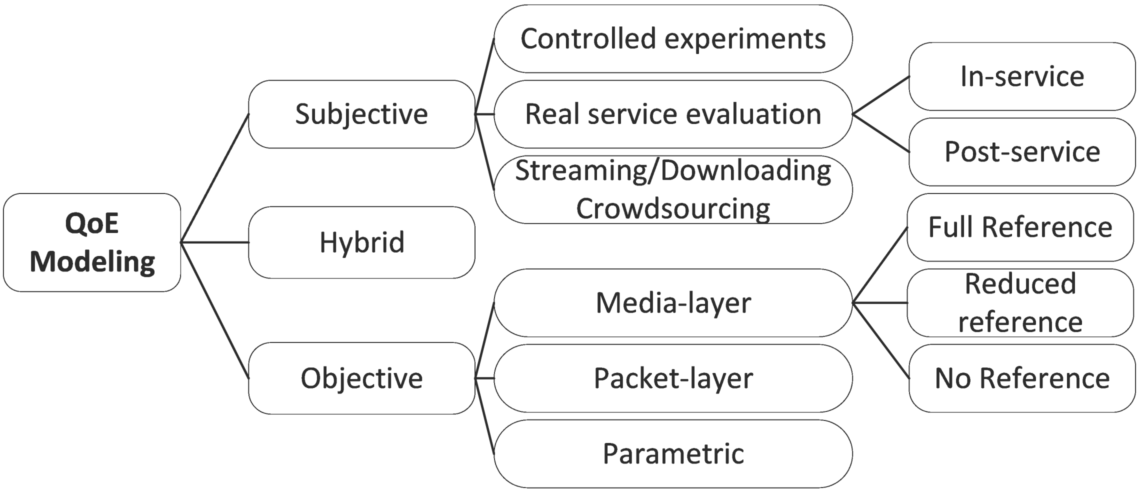

Figure 1: Classification of QoE modeling approaches (Tsolkas et al. 2017: 2)

It seems to be a wide agreement that subjective controlled experiments are time consuming and expensive (Takahashi et al. 2004: 28; Tsolkas et al. 2017: 3; Jelassi et al. 2012: 495).

The main target for the communication services is subjective quality meaning that the subjective quality assessment is the most reliable method. As the VoIP providers need implement solutions based on QoS information, there is a need to model the relationships. (Takahashi et al. 2004: 28–29)

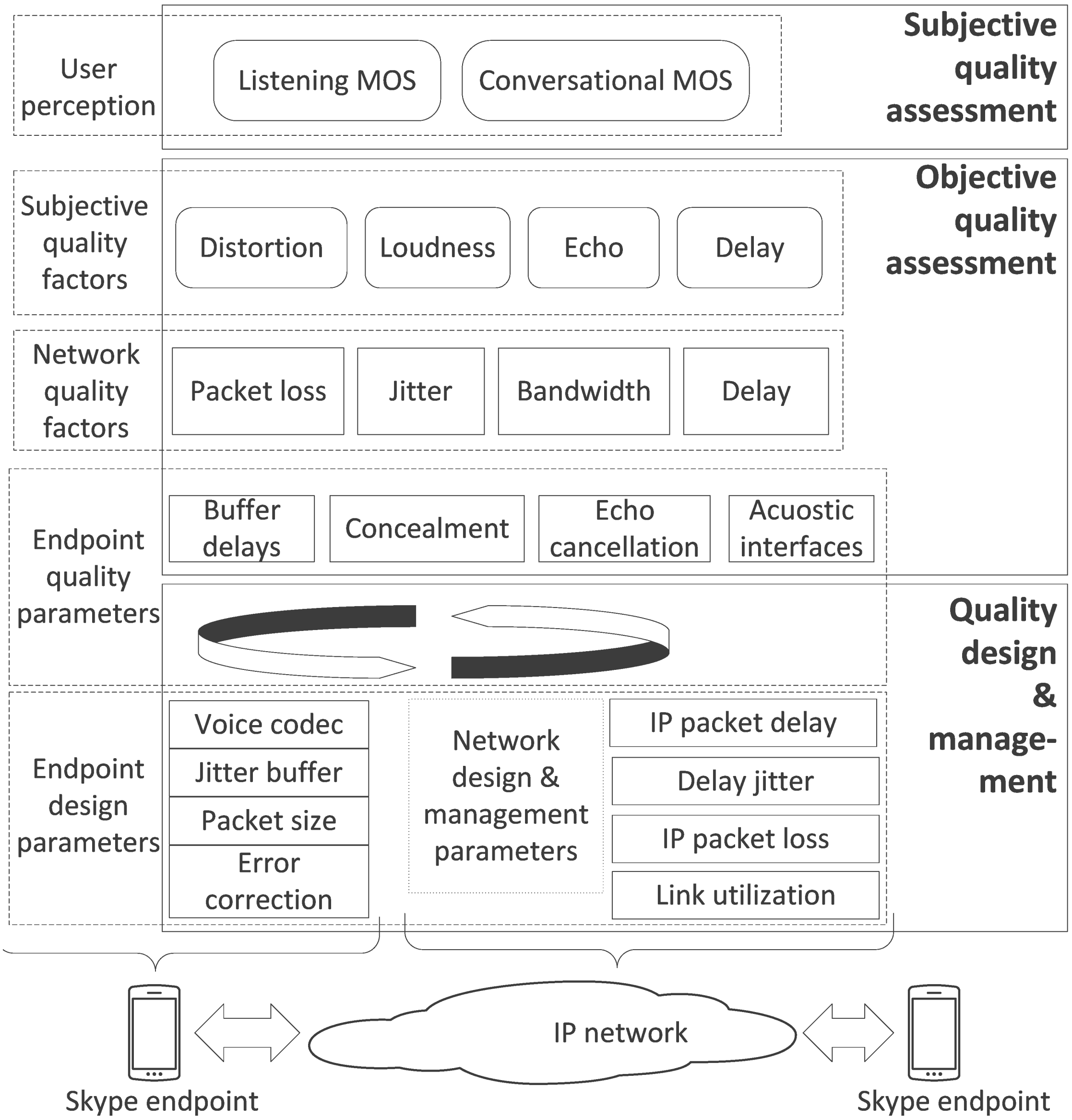

Figure 2: Factors that determine the quality of a VoIP call (modified based on Takahashi et al. 2004: 29)

Stankiewicz and Jajszczyk brought out the high level QoE provisioning to convergence requirements. From that classification, the user hosted VoIP is affected by anywhere requirement, anytime requirement, any user device requirement, any media and networking technology requirement, IP QoS support, by any operator requirement, and the impact of the network neutrality principle. The last one is especially interesting as it forbids internet service providers to prioritize traffic taken strictly, even in less strict concept one VoIP service provider traffic is forbidden to be prioritized over another. (Stankiewicz & Jajszczyk 2011: 1463–1469)

2.3.2. Concerns related to objective quality modeling

A major problem with ITU-T recommended objective speech evaluation tools is that the characterizations can be taken as parametric only due to the limitations of ITU-T methods P.563, P.862, and P.863 calling out in the applications that the impairments related to two-way interaction are not covered (International Telecommunication Union 2004: 3; International Telecommunication Union 2001: 2; International Telecommunication Union 2014: 1).

The shortcoming of PESQ not considering delay is addressed by adding E-model based roundtrip delay impact, but the authors bring out the shortcomings related to lacking interaction parameters and computation complexity of PESQ (Conway 2004: 2525).

The ITU-T G.107 and G.107.1 defined E-model is more generic providing a parametric approach that extends the coverage from speech to conversation. These models also consider user environment related parameters like background noise, device acoustic parameters like loudness loss, and talker echo (International Telecommunication Union 2015b: 1; International Telecommunication Union 2015a: 1). Applying the whole E-model in real-life solutions is questionable due to lacking parametrical characterization like noise levels (Falk & Chan 2009: 3). There are multiple simplifications and enhancements proposed to make it usable at least partially (Takahashi et al. 2004: 33; Jiang & Huang 2011: 499–500; Wuttidittachotti & Daengsi 2017: 8350) .

The QoS related degradations can be addressed by forward error correction to handle packet loss, buffering and packet concealment to handle jitter, and codec switching or complexity adjustment) to handle bandwidth (Ogunfunmi & Narasimha 2012: 44–48). These solutions come with tradeoffs to delay, computing resources, and bandwidth.

Adaptive jitter buffering in real networks is causing time alignment problems between the reference and degraded signal resulting in PESQ providing lower ratings compared to perceptually rated samples (Qiao et al. 2008: 4). The same was shown and solution proposed to address this gap was to use ViSQOL based model as it is shown to provide better prediction for VoIP related issues. This model still requires reference and degraded and signals as input (Hines et al. 2015: 17).

A research to model Skype SILK codec concluded that PESQ is not suitable due to being too conservative compared to the same perceptually rated audio samples. The paper recommends a Weber-Felcher’s Law of psychophysics instead to model the quality dependency of bitrate (Chen et al. 2012: 526).

There are also cultural concerns related to the objective modeling. Study conducted in Thailand based on perceptual evaluation claims the standardized E-model to be inaccurate for Thailand and suggests that the model should be customized based on countries that have their own culture and language (Daengsi & Wuttidittachotti 2013: 411). A very similar conclusion is pointed out also in later studies comparing Thai, British English, and American English (Wuttidittachotti & Daengsi 2016: 22) Another study using Chinese concluded that PESQ can be inaccurate for other languages than English (Zhang et al. 2015: 5).

2.4. Call duration relationships

This section provides an overview of publications where the call duration is analyzed in relation to technical parameters or user feedback.

A study was conducted in university campus in Taiwan to collect Skype call traces and to compare these against the measured network QoS. The article demonstrates strong and consistent relation of median call duration to network jitter and bandwidth - the results are plotted on Figure 3. (Chen et al. 2006: 403)

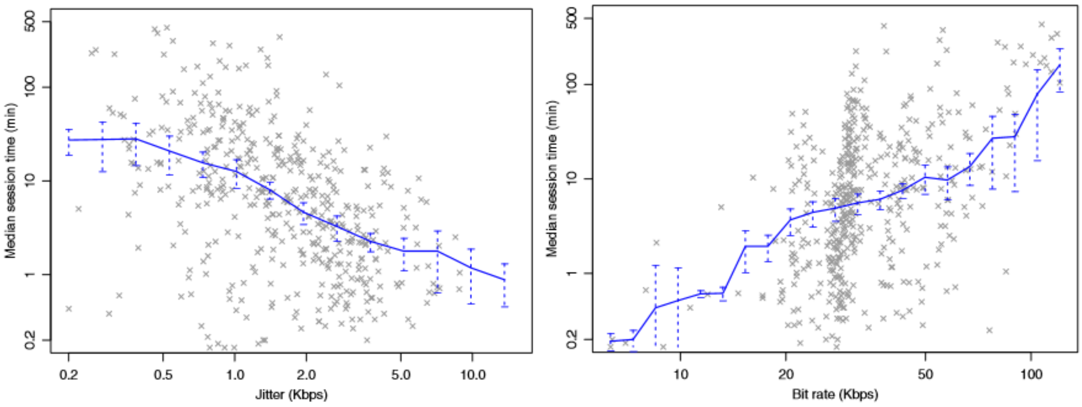

Figure 3: Median call duration relation to jitter and bandwidth (Chen et al. 2006: 403)

Skype speech codec SILK was introduced on Internet Engineering Task Force in section Real-time Applications and Infrastructure Area. The presented relation of codec impact to user ratings and call duration is shown on Figure 4. The relation between ratings and duration was not covered directly but it is visible on the plots.

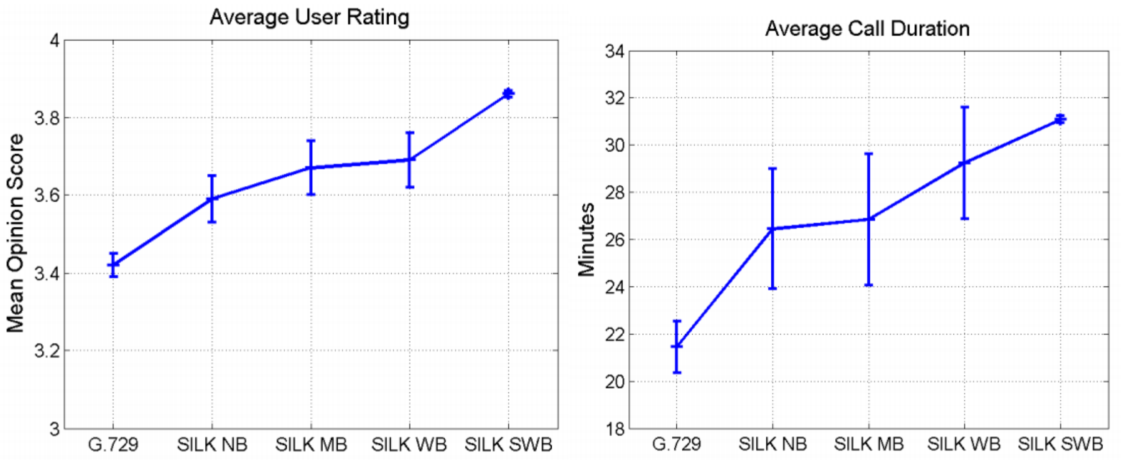

Figure 4: Skype SILK codec complexity modes compared to ITU-T G.729 codec (Vos et al. 2010: 2-3)

Analysis of VikingTalk VoIP service telemetry observed call duration relation between poor calls (rated as 1 or 2) and good calls (rated between 3 and 5), confirmed using T-test. However, the linear relation was not observed (De Pessemier et al. 2015: 5889).

2.5. Data driven Quality of Experience and A/B test

This section provides an overview of practical use of the relation and considerations related.

Based on a case with VoIP Chatterjee proposes 7 generic guidelines for modeling, debugging, and tuning QoE: 1) Using customer feedback to find patterns, 2) Designing QoE metrics to be measured, 3) Developing tools and processes to collect analyze, 4) Designing experiments to prove patterns, 5) Identifying the set of critical variables, 6) Designing experiments to identify bottlenecks, 7) Modeling the QoE for isolated variables (Chatterjee 2010: 1050).

“A/B testing is a common pattern for gradient-based, data-driven optimization of user experience“ (Nolting & von Seggern 2016: 277). A/B tests are controlled randomized field tests that provide a method to evaluate an idea. The basic way is to expose users randomly with 2 variants: A (control) as existing solution and B (treatment) as proposed idea.

Randomization related biases can occur. To validate the randomization, it is also highly recommended to conduct the A/A test always in parallel with other experiments to ensure that the users are split correctly, acquired data matches with records, and A/A results are statistically insignificant (Kohavi & Longbotham 2009: 174). The test is helpful for validation in case of experiments where only a specific subset fulfilling a certain condition needs to be exposed to the experiment (Kohavi & Longbotham 2010: 32) . The successful A/A tests can be used to find the variability of measured parameter to compute the minimum sample size needed (Kohavi et al. 2007: 175).

There is a threat that the treatment has an unwanted effect. This risk can be mitigated by applying the treatment to a small set of population and gradually increasing it if there are no severe problems detected, but to maximize the statistical power and enable drawing conclusions faster the eventual rate should be 50: 50 (Kohavi et al. 2007: 963–965).

The requirements for data driven QoE are brought out as following: 1) Measurable, 2) Informative, and 3) Business fitting. That is based on the possibilities to collect the data about QoE metrics from real usage and relate the measurable metrics to user feedback. (Chen et al. 2015: 1157)

For controlled experiments it is recommended to agree the evaluation criteria before it is conducted, that is to ensure it being business fitting (Kohavi et al. 2007: 966). From a VoIP service QoE point of view it would be preferable to relate it to the user feedback. However, the VikingTalk study showed that only 23.8% of calls received the user rating (De Pessemier et al. 2015: 5879). This means reduced number of samples, resulting in slower experiments and potentially exposing the users to a bad treatment for longer period than necessary.

The VikingTalk study also concluded that QoE wise the platform and device have statistically significant impact (De Pessemier et al. 2015: 5882–5884). When targeting the experiment to a respective subset then it would magnify the problem with sample size.

3. DATASET AND METHODS

3.1. Dataset

Skype as a service is mainly known as popular user hosted calling application. There are consumer and business versions[1] covering all popular platforms, including browsers. These applications have a graphic user interface asking regularly (after the call) to rate the call quality. Skype also offers solutions like Skype Connect[2] that run on SIP enabled PBX and do not have the graphic interface for collecting this feedback. If Skype is integrated into hardware like TVs[3] and IP phones[4], then it might also be so that collecting the ratings is not feasible due to lack of control over user interface.

From organization perspective Skype has adopted ‘data-driven culture’. All prototype solutions, release candidates and improvements are tested to the reasonable extent internally, then A/B tested on actual user base following the industry practices. Decisions are made based on the data. (Kohavi et al. 2007: 966)

Skype collects many different parameters from each call. This data covers a variety of QoE parameters. That allows slicing the user base for studies and experiments based on a specified set of parameters that can be related to QoS, internet service provider, platform, device, video, and acoustic interfaces.

For this thesis purpose, audio call duration, local ratings, and remote ratings are extracted to analyze the relation between them. Additionally extracted: estimated QoS impact to audio, device model, device form factor, and country as some parameters of interest as brought out in articles referenced above to check if and how they impact the relation.

In total the initial dataset contains 324,558,870 established calls. Out of these 3,913,685 were rated locally on the Android device and 2,992,001 rated from the other end of the call. The ratings were given after the call on the 5-point scale.

The scope of study is limited on audio calls on devices running Android operating systems. Android is selected as the platform as the devices are typically used without

[1] https://www.skype.com/en/business/
[2] https://www.skype.com/en/features/skype-connect/
[3] https://www.skype.com/en/download-skype/skype-for-tv/
[4] http://partnersolutions.skypeforbusiness.com/solutionscatalog/ip-phones

attached accessories as input/output devices. This allows to analyze data without additional complex mapping to estimate the attached device type.

Device form factor	Amount of all calls
Handset	30.5%
Headphones	33.7%
Speakerphone	34.3%
Other	1.5%

Table 1: Device form factor popularity

The other modes are mainly Bluetooth accessories like headsets or speakerphones. The usage is so low that these are left aside for this study.

Skype to Skype audio calls are selected as the QoS requirements are lower than for video and the analysis are more straightforward because it allows to assume two-way communication whereas with video there is a need to look the impact of one-way video, video being transmitted only during certain parts of the call, video frame rate and resolution capping due to QoS, and video caps or stopping due to QoS issues.

Due to the business sensitive nature of the dataset, averaged rating values and call counts are not brought out although they are available in the dataset. Data is acquired during the composition of thesis, but the exact period cannot be revealed.

3.2. Methods

3.2.1. Data binning

Statistical data binning is grouping a continuous variable into a definite number of bins. This method is commonly used to observe data distribution.

Hogg (2008: 5–6) discusses the considerations when choosing the binning to make histograms. In this thesis, the binning is used on call duration:

- In section 4.1 Call duration distributions the binning is done using equal width steps. That is to give an overview of the distribution and bring out the exponential nature.
- Afterwards equal width step binning is used on the logarithmic scale to handle the call duration distribution by balancing the call counts per bin. An additional benefit of using logarithmic scale is that the analyzed observation has logarithmic nature. 10 bins were chosen mainly based on visual considerations – more bins would have enabled better means for balancing while making it harder to follow the differences in call duration distribution.

3.2.2. Linear regression model

To bring out the statistical significance and amount of variation explained this thesis relies on the linear regression modeling as following:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Equation 1: The theoretical model of linear regression

where y is response, ε is random error, k is the number of regressor dimensions, xs are the regressor variables, and βs are regression coefficients representing the expected change in response when all other regressor variables are held constant. (Montgomery et al. 2012: 67)

The R-squared, also called coefficient of determination, is used to quantify the fraction of variability explained by the model. The calculation formula:

$$R^2 = 1 - \frac{var(\varepsilon)}{var(Y)}$$

Equation 2: R-squared calculation formula

where *var(ε)* is the variance of the residuals and *var(Y)* is the variance in *Y*. (Garner 2015: 148)

To keep the figures comparable through the analysis the R-squared figures are derived only from raw data. Typically, the expected R-squared values are much higher than brought out in this thesis. This can be achieved on processed data, but it would interfere with the purpose to keep the impact of considered variables comparable.

There are 2 widely recognized problems with R-squared: 1) adding a predictor increases R-squared, 2) too many predictors can be modeling random noise and cause overfitting – that is reducing the predicting power.

In this thesis, the adjusted R-squared is used. It is a modified version of R-squared that increases only if the added predictor increases the model predicting power.

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - \#Xs - 1}$$

Equation 3: Adjusted R-squared calculation formula

where *n* is the number of points in data sample and *#Xs* is the number of variables. (Kottemann 2017: 185–186)

To describe the regression models F-statistic is also brought out. It is also describing the variance explained as following:

$$F = \frac{MSR}{MSE}$$

Equation 4: F-statistic calculation formula

where *MSR* is mean squares due to regression (or explained variability) and *MSE* is mean square error (or unexplained variability) (Sahay 2016).

The author acknowledges that for predicting purposes the generalized linear model is more appropriate. However, the purpose of this thesis is not to propose a predicting model, but to prove the relationship and estimate the impact of literature based selected variables. To overcome the issue of logarithmic nature of the relation, the simple linear regression model predictor is used as logarithm of the call duration.

3.2.3. Confidence intervals

For the studied relation 95% confidence intervals are plotted for MOS where the relation to call duration is shown. Having large dataset and bandwidths enables bringing out higher confidence intervals so that the neighboring bins would not have overlapping confidence intervals. The 95% is chosen as the most typical used.

The confidence intervals are calculated as following:

$$\left(\bar{X} - q\times\frac{\sigma}{\sqrt{n}}, \bar{X} + q\times\frac{\sigma}{\sqrt{n}}\right)$$

Equation 5: Confidence interval calculation formula

where $\bar{X}$ is the mean, q is the *(1-α/2)* quantile, n is sample size, and σ is standard deviation. In case of the 95% confidence interval the α is 0.05.

3.3. Description of variables in analysis

As in section 4 different regression formulas have been presented, the following Table 2 documents content and abbreviations of variables used in these regression formulas.

Variable	Description
DFF	Categorical variable that describes one of 3 acoustic device form factors used. The numerical value in dataset is recoded to be descriptive string: “Handset”, “Headphones” or “Speakerphone”. Note that “Headphones” also includes headsets (analog headphones together with microphone).
QoS_degradation	Continuous variable that quantifies the network degradation to subjectively perceived quality of audio calls. This variable is based on a model relating network degradations to user ratings based on a machine learned algorithm.
Rating_L	Discrete variable on 5-point scale that describes the subjective quality rating given by local user, in other words Android device user.
Rating_R	Discrete variable on 5-point scale that describes the subjective quality rating given by remote user.
Duration Log(Duration)	Continuous variables that describe call duration. Log(Duration) refers to logarithm of the call duration.
Device	Categorical variable describing the Android device used.
Country	Categorical variable describing the country of Android device user. (The country is available also for remote user, but only domestic calls are used for country related analysis meaning that this variable is the same for both call parties.)

Table 2: Variable names used in section 4

4. RELATION BETWEEN MEAN OPINION SCORE (MOS) AND CALL DURATION

4.1. Call duration distributions

The distributions are brought out for all calls, locally rated calls, and remotely rated calls to understand the data distribution and verify that the rated calls have similar distribution. That is to check potential biases related to how the ratings are collected.

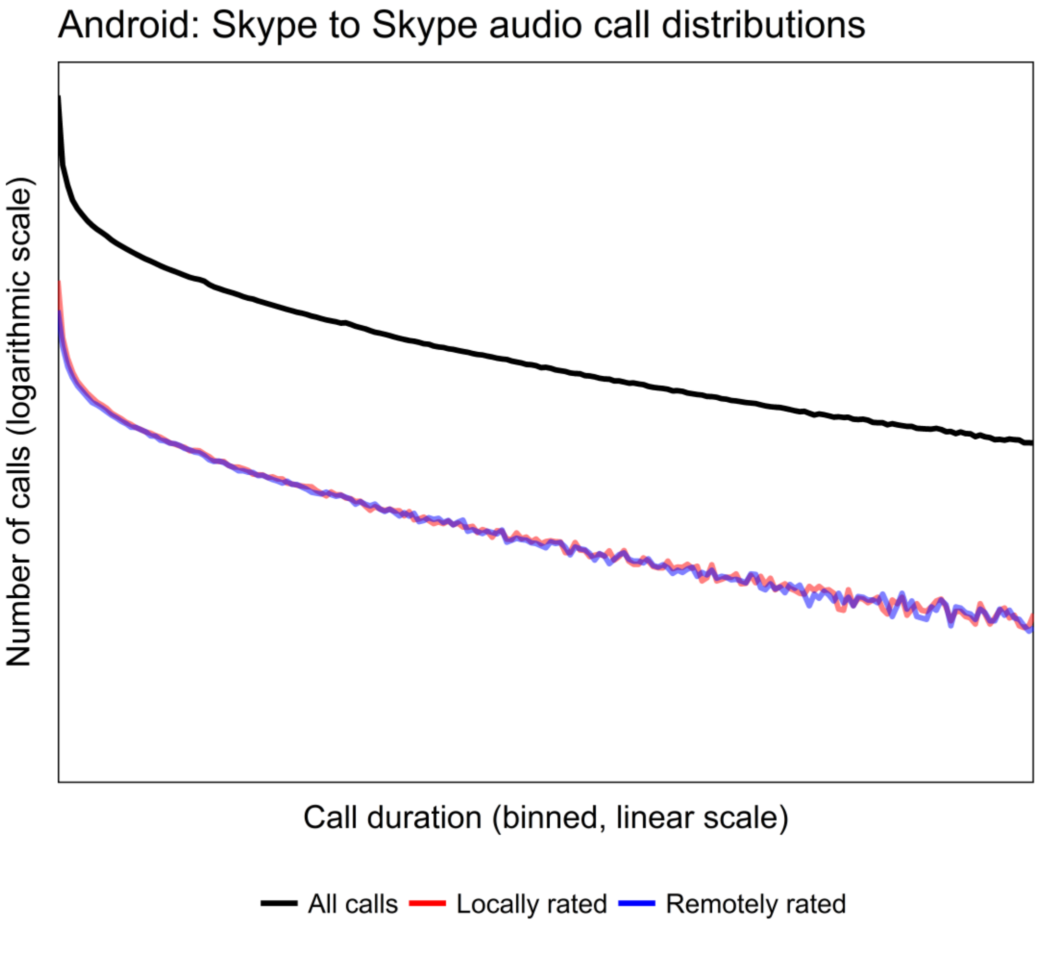

Figure 5: Skype to Skype audio call duration distribution on Android platform together with the distributions of locally and remotely rated calls

Although we can see that there are slightly more locally rated calls for shorter calls, it allows to conclude that the distributions are similar on both platforms and the rated calls are representative.

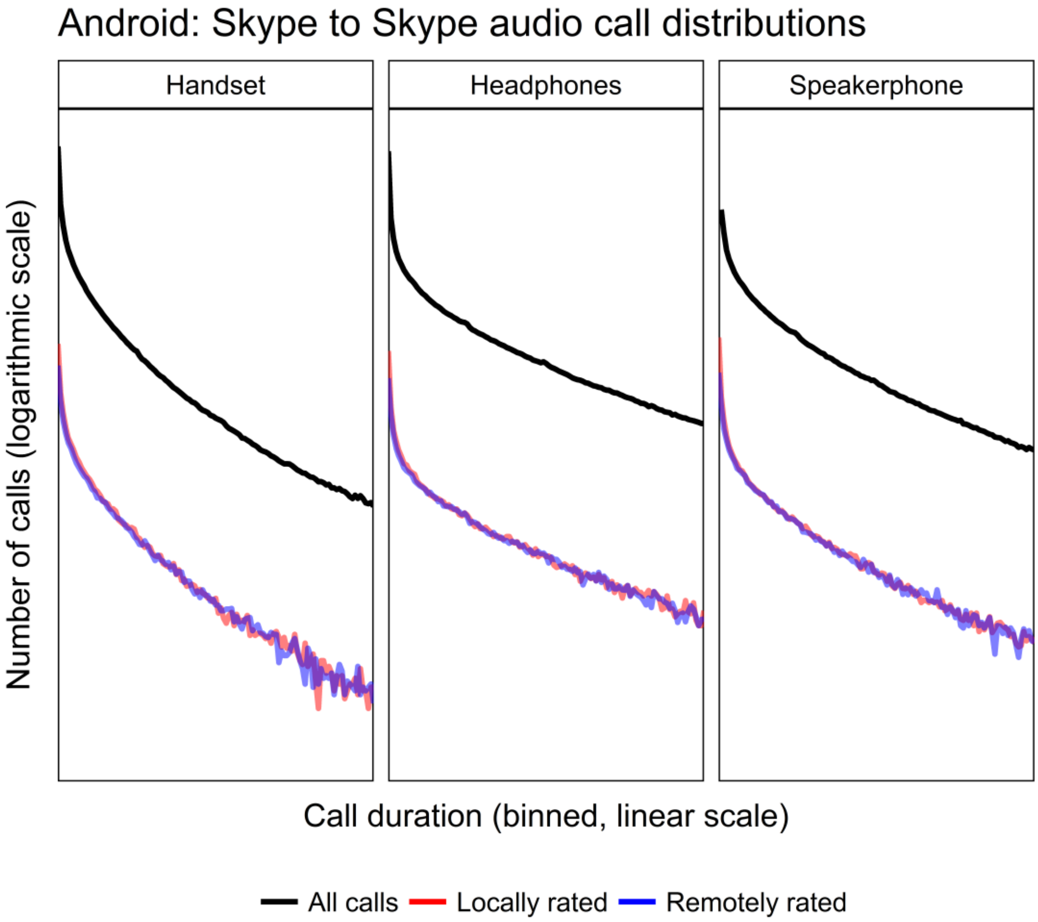

Figure 6: Skype to Skype audio call duration distribution on Android platform sliced by device form factor together with the distributions of locally and remotely rated calls

The figure above shows all calls and rated calls distributions per device form factor. We also see that the calls in handset mode are typically shorter and longer in headphones mode. This can be justified with the convenience of usage as the handset mode occupies one hand and speakerphone mode limits the distance between the device and user due to playback loudness from built-in loudspeaker(s).

Device form factor	Locally rated calls	Remotely rated calls
Handset	1.20	0.90
Headphones	1.23	1.08
Speakerphone	1.20	0.90

Table 3: Amounts of ratings available (%)

The Table 3 brings out the ratios of ratings available. Call durations are available for device form factors analyzed. The reason of higher amount of rated calls for headphones is not explained with the dataset available. It might be related to use case and/or platform on the other end of the call.

De Pessemier et al (2015: 5879) brought out that their dataset had 23.8% of calls related. The main explanation for the large difference here is that Skype is asking the feedback less frequently. That can be justified by user experience related matters that are not related to call quality.

4.2. Relation between mean opinion score (MOS) and call duration on the whole dataset

The VikingTalk study did not find a linear relation between the rating and duration of audio calls over the full range of ratings, but found that the calls rated as two lowest values were shorter than the calls rated higher on the 5-point scale; the analysis method relied on decision trees (De Pessemier et al. 2015: 5889–5891). The shortcoming of the method used was that the rating value seemed to be mapped to average call duration. Skype dataset also has the same problems of having higher ratios of extreme values and call durations not increasing continuously when averaged for discrete rating values.

However, in this thesis the author is binning the call duration and plotting the MOS score to prove that the relationship exists and is statistically significant. Due to the call distribution observed above binning on linear scale would provide us large sample sizes on lower call durations and leave small ones for the rest. This results in unreasonably large steps on MOS scale per linear bin for the shorter calls and large confidence intervals for the longer ones. The solution proposed below is using binning on a logarithmic scale.

In addition, the plotted 95% confidence intervals allow visual verification of the statistical significance. In a practical application the borderline cases can be checked with the t-test.

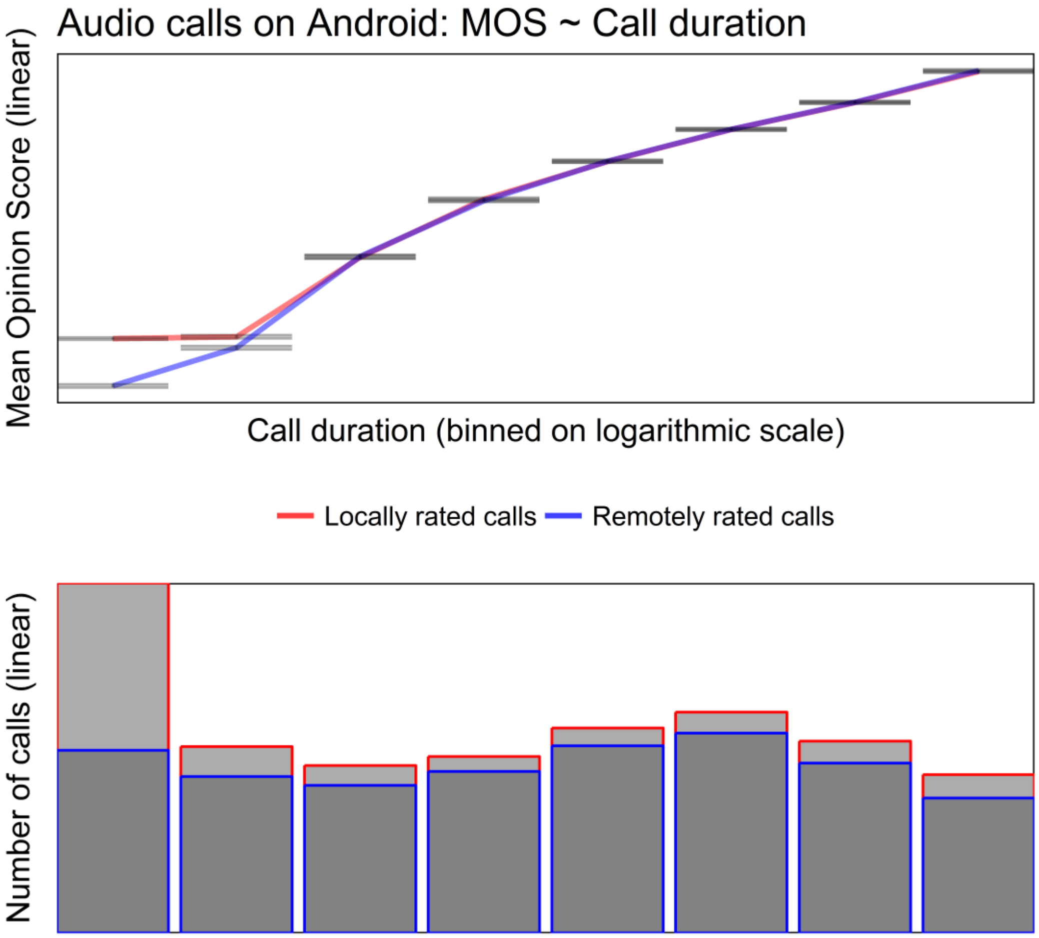

Figure 7: Skype to Skype audio call MOS and duration relation on the whole dataset

On the Figure 7 we can see the continuous relationship between call duration and perceived call quality. We can also see that the density of call count distributions per bin is matching well starting from the 2nd bin. As the dataset contains all calls the drop in MOS for very short calls might be related to technical issues.

Formula	Adj. R^2	p-value	F
Duration ~ Rating_L	1.74%	<0.001	5.27E+04
Duration ~ Rating_R	1.84%	<0.001	4.51E+04
Log(Duration) ~ Rating_L	5.47%	<0.001	1.72E+05
Log(Duration) ~ Rating_R	6.79%	<0.001	1.76E+05

Table 4: Linearly modeled relations between call duration and local/remote ratings

The Table 4 shows that the relation between call ratings and duration is much stronger when the duration is observed on a logarithmic scale. When looked together with the

Figure 7 then the higher R-squared value for remote ratings is mainly because due the difference with the first 2 bins where the local MOS is behaving differently than remote.

4.3. Factors related to MOS and call duration

Based on the literature and author’s experience 4 factors (device form factor, QoS, device, and country) are analyzed to study the impact to factors of interest, namely: call duration and rating.

4.3.1. Quality of Service

The dataset contains QoS impact estimation to the locally rated audio. This is previously modeled using machine learning on call ratings and considering the network delays, packet loss, jitter, and the specifics of Skype processing handling the QoS problems.

Formula	Adj. R^2	p-value	F
Rating_L ~ QoS_degradation	4.21%	<0.001	1.06E+05
Rating_R~ QoS_degradation	4.95%	<0.001	1.13E+05
Log(Duration) ~ QoS_degradation	4.47%	<0.001	1.99E+05
Duration ~ QoS_degradation	2.31%	<0.001	1.01E+05

Table 5: Linearly modeled relations between QoS degradation and factors of interest

Linearly modeled QoS degradation variable is showing smaller R-squared to the factors of interest than modeled between them in Table 4.

4.3.2. Device form factor (DFF)

Based on Figure 6 we can already expect the form factor (handset, headphones, and speakerphone) impact to call duration.

Formula	Adj. R^2	p-value	F
Rating_L ~ DFF	0.05%	<0.001	721.5
Rating_R~ DFF	0.09%	<0.001	1.03E+03
Log(Duration) ~ DFF	1.92%	<0.001	4.94E+04
Duration ~ DFF	2.36%	<0.001	6.08E+04

Table 6: Linearly modeled relations between device form factor and factors of interest

From Table 6 we can see that the device form factor has statistically significant impact to ratings, but it explains very little of the variance. However, it is more strongly related to the linear call duration explaining more variance than regressor brought out in Table 4 and Table 5.

4.3.3. Device

The Android devices are mainly tablets and mobile handsets, ranging from low-end to high-end. Distortion, loudness, echo, and delay from objective quality parameters shown on Figure 2 are largely depending on the device. Device impact to MOS was also found significant on VikingVoip analysis (De Pessemier et al. 2015: 5886).

Linear modeling is computationally costly on categorical variable as each value is appearing as a separate variable. To handle this, the dataset is filtered keeping top 12 devices aggregated by the device user friendly name.

Formula	Adj. R^2	p-value (for 12 devices)	F
Rating_L ~ Device	0.10%	<0.001: 11/12; >0.05 1/12	67.16
Rating_R~ Device	0.02%	<0.001: 3/12; <0.01 3/12 <0.05: 1/12; >0.05: 5/12	13.07
Log(Duration) ~ Device	0.39%	<0.001: 10/12; <0.01 1/12; >0.05: 1/12	447.1
Duration ~ Device	0.24%	<0.001: 12/12	272.9

Table 7: Linearly modeled relations between top 12 devices aggregated by friendly name and factors of interest

From Table 7 we can see that most devices have statistically significant impact to factors of interest and are confirming the findings by De Pessemier et al. (2015: 5886).

4.3.4. Country

Section 2.3.2 covered the concerns related to cultural impact being a factor of MOS modeling, but standardized objective models not covering this. To study the cultural impact to the relation, the dataset is sliced by countries and filtered keeping only domestic calls.

As in section 4.3.3 the dataset is filtered keeping top 12 countries due to computational complexity and to ensure sufficient number of calls per country.

Formula	Adj. R^2	p-value (for 12 countries)	F
Rating_L ~ Country	0.41%	<0.001: 12/12	286.3

$\mathbf{Rating\_R} \sim \mathbf{Country}$	0.57%	<0.001: 11/12; <0.01 1/12	346.6
$\mathbf{Log(Duration)} \sim \mathbf{Country}$	2.06%	<0.001: 12/12	2550
$\mathbf{Duration} \sim \mathbf{Country}$	1.47%	<0.001: 11/12; <0.5: 1/12	1803

Table 8: Linearly modeled relations between top 12 countries and factors of interest

From Table 8 we can see that the country is a significant factor. This agrees with section 2.3.2.

4.4. The impact of related factors to the relation between call duration and MOS

To study the impact of factors analyzed in section 4.3, the basis model of the relation between logarithmic call duration and perceived quality is amended by each factor separately.

4.4.1. Quality of Service impact

To analyze the QoS impact to the relation studied the basis model is amended with QoS impact (and interaction) as a regressor.

Formula	Adj. R^2	F
$\mathbf{Log(Duration)} \sim \mathbf{QoS\_degradation} + \mathbf{Rating\_L} + \mathbf{QoS\_degradation} \times \mathbf{Rating\_L}$	6.77%	5.84E+04
$\mathbf{Log(Duration)} \sim \mathbf{QoS\_degradation} + \mathbf{Rating\_R} + \mathbf{QoS\_degradation} \times \mathbf{Rating\_R}$	7.16%	5.59E+04

Table 9: Linearly modeled relation of interest amended by QoS degradation and interactions

Comparing Table 4 and Table 9 we see that adding the QoS degradation as regressor the variance explained improved relatively by 23.8% and 5.4% when modeling call duration relation to local and remote ratings respectively.

However, the absolute difference in variance explained is rather small (1.3% and 0.37%) considering that Table 4 showed that the variance explained was >4% for all the studied relation components. This allows us to conclude that the network degradations are impacting the logarithmic call duration and perceived quality very similarly.

To visualize the QoS impact, the LMOS is plotted for all calls and calls in ideal network minimizing the impact. The criteria for ideal network conditions is selected to be modelled <0.1 audio MOS drop.

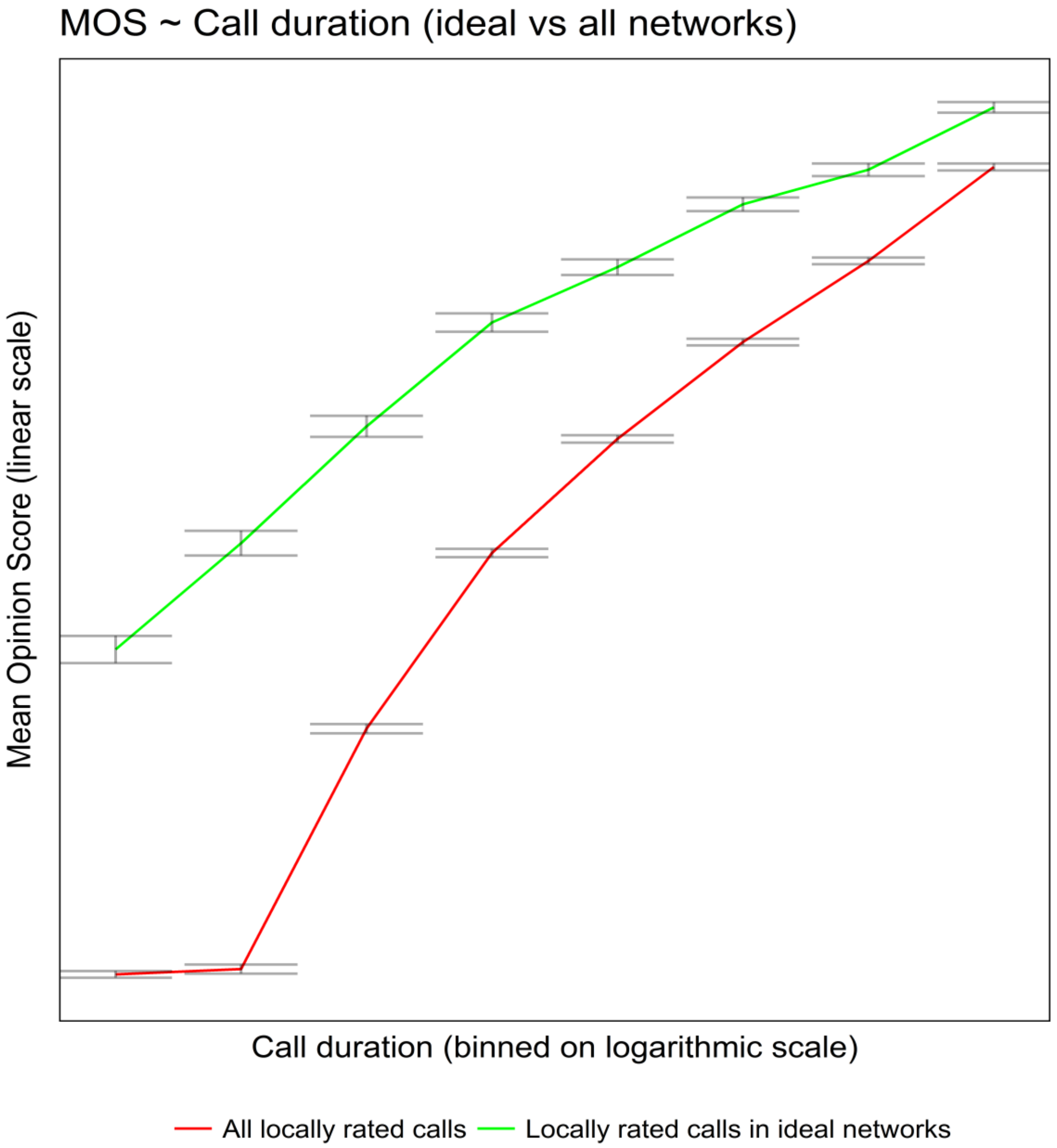

Figure 8: Locally rated calls in ideal and all networks

From Figure 8 we can see that the QoS problems mainly impact shorter calls and the relation remains continuous. We can observe severe MOS decrease due to QoS degradations that align with the respective literature brought out in section 2.3 Modeling quality of experience. Also, the figure shows that degradations impact shorter calls more hinting that this is likely a causal factor. The reason for higher relative improvement in ratings given on Android devices (Rating_L) results from making the first 2 bins align more linearly with others.

4.4.2. Device form factor impact

By adding the DFF as a regressor to the initial linear model we can see following:

Formula	Adj. R^2	F
$\mathbf{Log(Duration)} \sim \mathbf{DFF} + \mathbf{Rating\_L} + \mathbf{DFF} \times \mathbf{Rating\_L}$	7.01%	4.49E+04
$\mathbf{Log(Duration)} \sim \mathbf{DFF} + \mathbf{Rating\_R} + \mathbf{DFF} \times \mathbf{Rating\_R}$	10.54%	5.68E+04

Table 10: Linearly modeled relation of interest amended by DFF and interactions

Table 10 shows the large gap between the variation explained by the 2 models. It can be caused by the same issue observed earlier. Relative improvements in explaining the variation compared to Table 4 are 28.2% and 55.2% respectively roughly matching with the expectations based on section 4.3.2, however the increase of the gap needs further study.

To visualize the device form factor impact to the relation between MOS and call duration, the dataset is sliced accordingly. Then plot the MOS scores with confidence intervals with call duration on logarithmic scale and interpret observations.

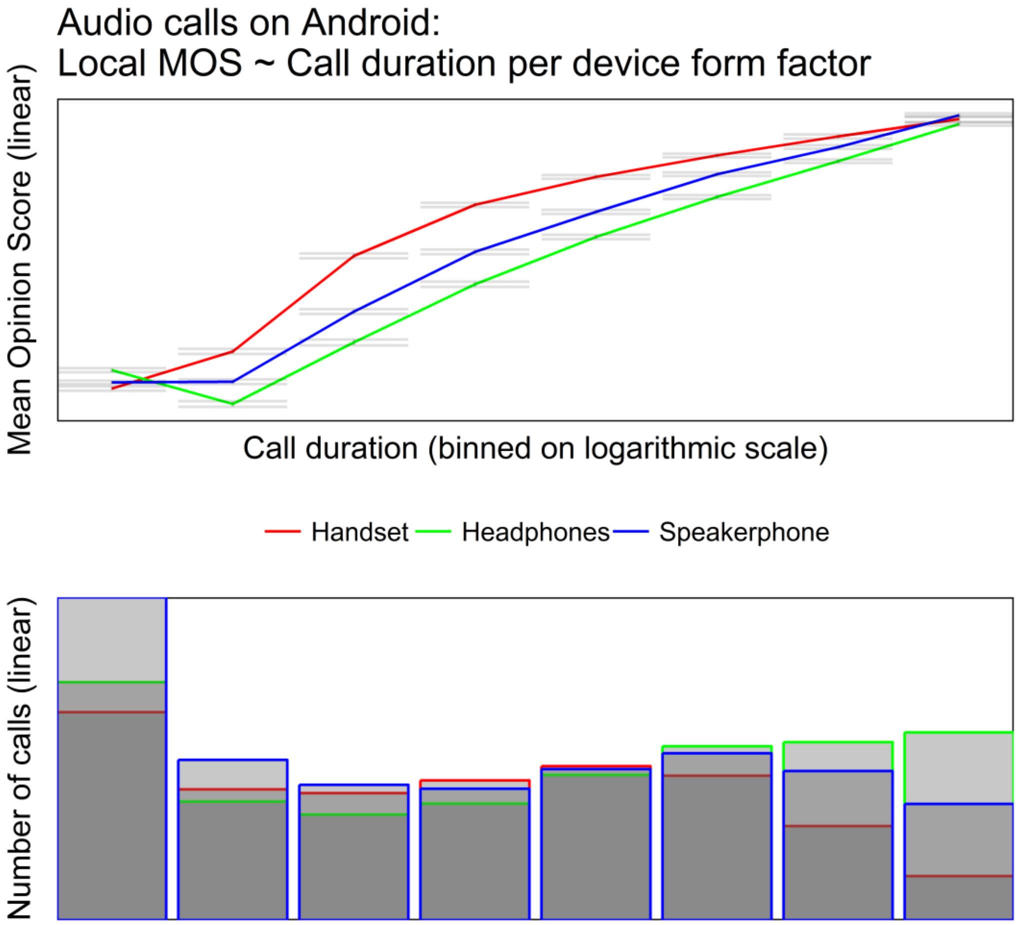

Figure 9: Skype to Skype audio call local MOS and duration relation sliced by DFF

On the Figure 9 the local MOS is overlaid for the device form factors in interest. The relation is statistically significant on a wide range of bins.

Figure 2 visualized the factors affecting VoIP call quality. For the same device acoustic interface related changes are in loudness, distortions, echo, conversational MOS, and possibly in delay.

It is expected that the handset mode receives higher scores compared to others as mobile phones are optimized for calling mainly in the handset mode. However, shorter calls and higher ratings in speakerphone mode compared to headphones might be hard to explain directly as the device user is expected to hear the other side more clearly (due to in-built loudspeaker limitations).

The required receive loudness rating values by ETSI standard are allowed to be several dB quieter for a mobile handset in speakerphone mode compared to handset and headphones modes in listening position. The frequency response requirement mask in speakerphone mode starts at 800Hz for speakerphone mode whereas the handset and headset mode ones start at 200Hz. It is also worth mentioning that there are no distortion requirements for the speakerphone mode in ETSI specification. It is defined only for handset and headset modes. (European Telecommunications Standards Institute 2015: 12–39)

Similarly in Skype for Business certification requirements the playback loudness in speakerphone mode is allowed to be quieter at lower distortion and echo requirements (Microsoft Corporation 2016: 38–58).

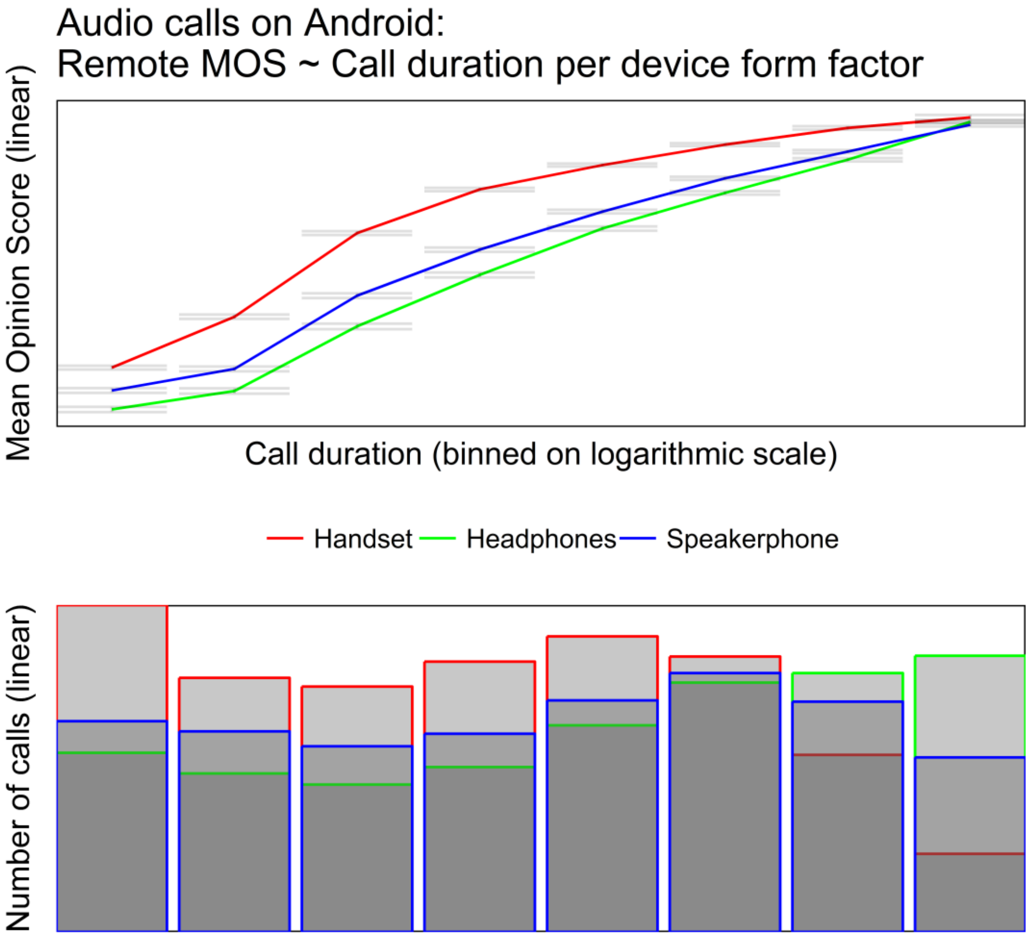

Figure 10: Skype to Skype audio call remote MOS and duration relation sliced by device form factor.

On the Figure 10 the remote MOS is similarly overlaid as on the Figure 9. The relation is very similar indicating that the conversation related aspects affect both call parties.

Considering this it is possible to answer the outstanding question why one device form factor is better than another. Analog headphones are likely headsets with the inline microphones often performing worse than the ones built into mobile phones. Although speakerphone mode is likely to pass echo (Kelloniemi et al. 2015: 8–11), the playback loudness limitations described above are also setting restrictions to the usage distance and background noise of the usage conditions. The difference between using local or remote rating as regressor is large indicating that the QoS degradation cannot be ignored because there is a similar phenomenon appearing with the first 2 call duration bins on Figure 9 (being the reason of different variances explained between raters).

Formula	**Adj. R^2**	**F**
$\mathbf{Log(Duration) \sim DFF + Rating\_L + DFF * Rating\_L + QoS\_degradation + QoS\_degradation \times Rating\_L}$	10.55%	4.06E+04
$\mathbf{Log(Duration) \sim DFF + Rating\_R + DFF * Rating\_R + QoS\_degradation + QoS\_degradation \times Rating\_R}$	11.91%	4.20E+04

Table 11: Linearly modeled relation of interest amended by DFF and QoS degradation

Including the network degradations as regressor reduced the relative the gap between the models.

4.4.3. Device impact

To study the device impact, the top 12 most frequently used devices are used as in section 4.3.3. Filtering is likely to cause a bias. To compensate the bias, the base models brought out in Table 4 are recalculated.

Formula	**Adj. R^2**	**F**
$\mathbf{Log(Duration) \sim Rating\_L}$	6.54%	4.99E+04
$\mathbf{Log(Duration) \sim Rating\_R}$	7.45%	4.94E+04
$\mathbf{Log(Duration) \sim Device + Rating\_L + Device \times Rating\_L}$	7.25%	2425
$\mathbf{Log(Duration) \sim Device + Rating\_R + Device \times Rating\_R}$	7.76%	2247

Table 12: Linearly modeled relation of interest and the models amended by country and interactions

Comparing the absolute improvements in Table 12 to variations explained in Table 7 we can see that although the variation devices are explaining is small, it improves the relation between call duration and ratings more comparing to explaining the variation of those variables separately. This indicates that aggregating the means of call durations and ratings by device will show a larger dispersion on respective axis.

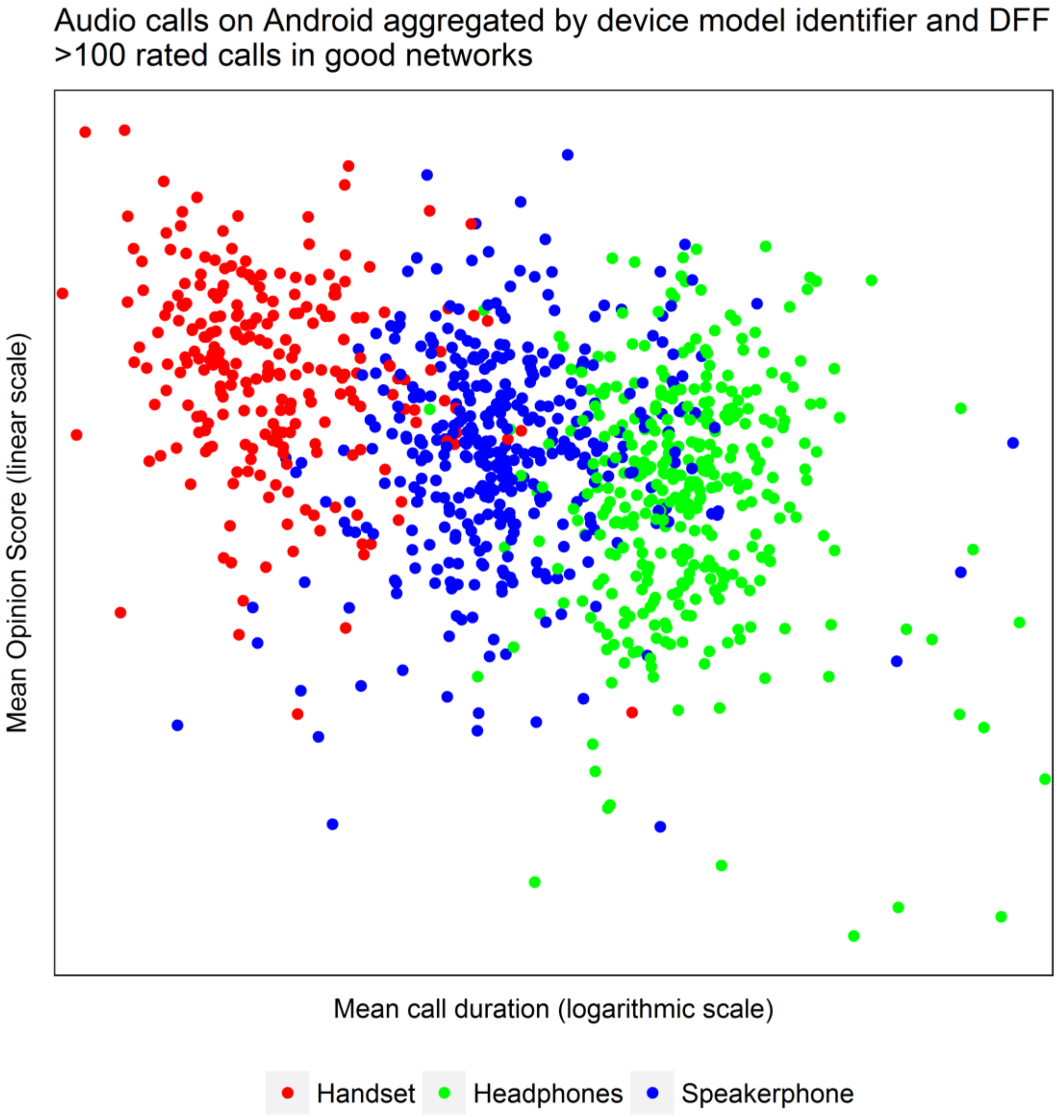

Figure 11: MOS relation to call duration, aggregated by device model identifier and form factor for devices with more than 100 locally rated calls in low QoS impact networks.

Figure 11 shows a large dispersion on both axis when aggregated by device. The dispersion on MOS axis confirms device impact to perceived quality as reported by De Pessemier (2015: 5884).

On Figure 11 we can also see the device form factor impact to call duration. It aligns with analysis conducted in sections 4.3.2 and 4.4.2. The same is also visible on call count histograms on Figure 9 and Figure 10. However, aggregating by device masks the relation as brought out in section 4.2 Relation between mean opinion score (MOS) and call duration on the whole dataset.

4.4.4. Country impact

To study the country impact, the top 12 most frequently appearing countries (keeping only domestic calls) are used as in section 4.3.4. To compensate the bias, the basis models brought out in Table 4 are recalculated.

Formula	Adj. R^2	F
$\mathbf{Log(Duration)} \sim \mathbf{Rating\_L}$	4.35%	3.50E+04
$\mathbf{Log(Duration)} \sim \mathbf{Rating\_R}$	4.42%	3.07E+04
$\mathbf{Log(Duration)} \sim \mathbf{Country} + \mathbf{Rating\_L} + \mathbf{Country} \times \mathbf{Rating\_L}$	6.52%	4472
$\mathbf{Log(Duration)} \sim \mathbf{Country} + \mathbf{Rating\_R} + \mathbf{Country} \times \mathbf{Rating\_R}$	6.33%	3742

Table 13: Linearly modeled relation of interest and the models amended by country and interactions

We can see a large improvement in variation explained (relatively 49.9% and 43.2%) and there is no large gap appearing between the models. Appended models in Table 13 are matching with the expectations from Table 8. However, it should not be explained only by cultural impact because the network infrastructure and devices used are also different.

To understand the cultural impact the dataset is filtered keeping only the calls without network impact. Then filter out top 12 countries (note that these are not the same as previously analyzed ones) keeping only domestic calls.

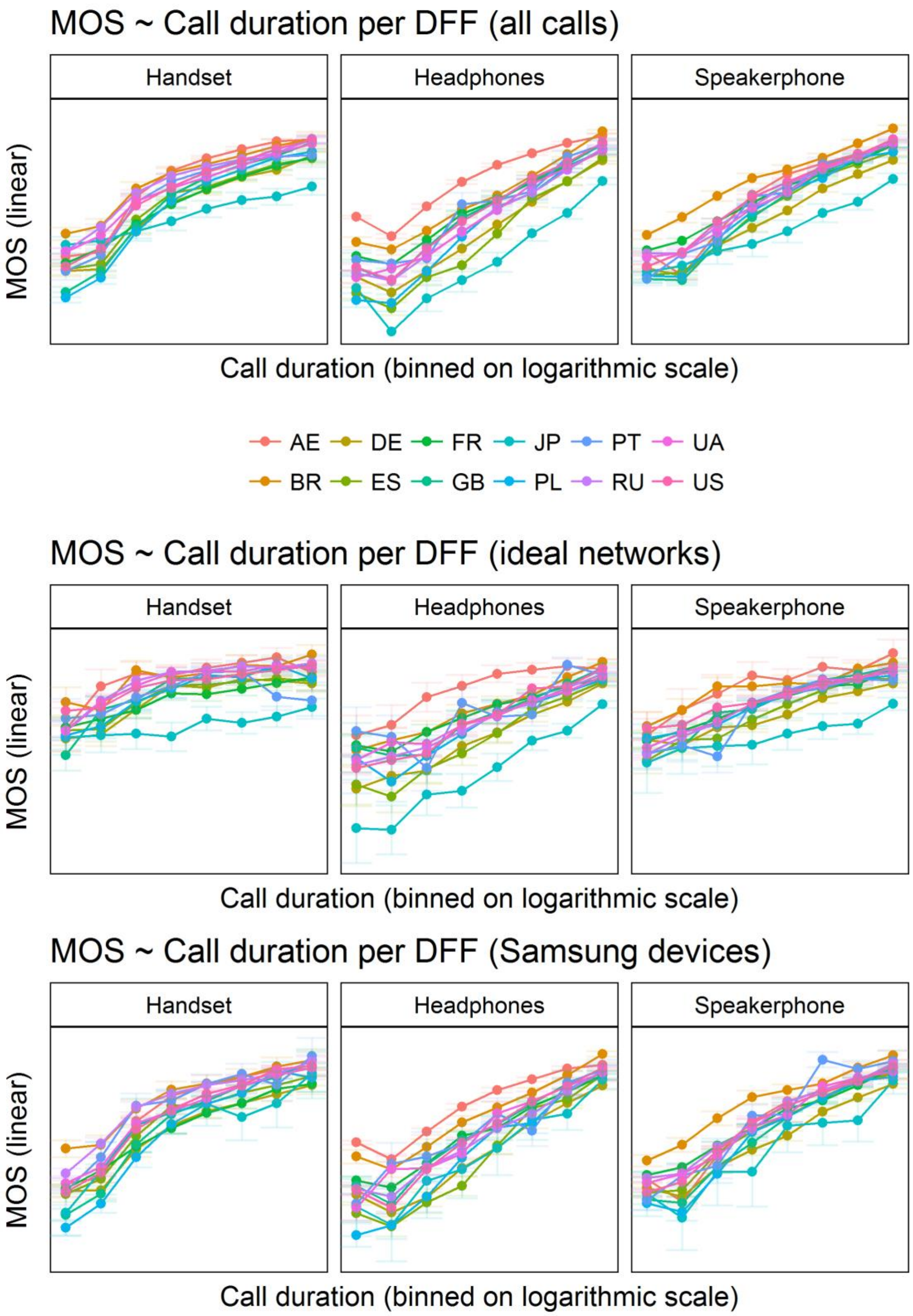

Figure 12: Local MOS and call duration relation in 3 views: 1) all calls, 2) calls in ideal networks, 3) all calls on Samsung devices

Due to data slicing the amounts of calls are narrowing down, which causes some discontinuity on the graphs, and the confidence intervals are rather large. Device impact is checked on third plot to have an idea if the device impact might be causing the difference. Samsung as a device vendor is chosen as it has the highest market share on Android devices. Slicing the dataset by specific device model was not providing statistically significant results due to low call counts in at least one country.

The upper 2 rows of graphs show that Japan is statistically different in large range from all other countries in selection. However, when non-Samsung devices are filtered out, then Japan is still on the lower side more than others, but not an outlier anymore. This is likely to indicate that the difference with Japan can be device related and the interactions between country and device needs further investigation. Brazil is showing high MOS scores on all views and based on the plots we could tentatively conclude that there are statistically relevant differences between some countries. However, there is not enough data available to bring out the binned relation for specific devices in ideal network conditions and we cannot draw definitive conclusions.

4.5. Analysis results summarized

The analysis in sections 4.2, 4.3, and 4.4 reached to several results. These are summarized below in Table 14 to provide an overview.

Analysis	Results
Relation between call duration and subjectively perceived call quality	The relation is continuous and statistically significant. Logarithmic call duration explains larger part of the perceived call quality than linear. Subjectively perceived call quality is the largest factor analyzed, explaining the variation in call duration, also indicating a causal relation.
Relation between subjectively perceived call quality and all factors separately	All factors showed statistically significant relation. By variation explained the ranking is the following: 1) Logarithmic call duration, 2) Network degradations, 3) Linear call duration, 4) Country. The variation explained by device and DFF is lower by magnitude.
Relation between logarithmic call duration and all factors separately	All factors showed statistically significant relation. By variation explained the ranking is the following: 1) Subjectively perceived call quality, 2) Network degradations, 3) Country, 4) DFF, 5) Device
Network degradations impact to the relation studied	Adding network degradation as a regressor when linearly modeling the relation improves the variation, but from the relation perspective this parameter is not as important as it might be projected based on relations to ratings and logarithmic call duration separately. That is likely due to impacting both the same way as the relation studied. Filtering to minimize the impact of network degradations reduced the number or short calls, made the aggregated relation more linear for the shorter calls (on logarithmic scale of binned call duration).
DFF impact to the relation	The relationship was different (due to differences in call duration distributions) for all 3 categories, but remained significant and continuous. Adding DFF as a regressor improves the variation explained by the relation analyzed up to 86% relatively.
Device impact to the relation	Device is impacting the relation, but the additional variation explained is minimal. Plotting out the aggregation by device showed large dispersion on both axis and masked the found relationship.
Country impact to the relation	Country affects the relationship and should be used in practical applications of the relation. The analysis remained inconclusive to tell if the cause is cultural. Compensating for the network degradations and devices used left too narrow data slices to draw conclusions.

Table 14: Analysis results summarized

4.6. Practical implications and study limitations

The relation between perceived call quality and call duration provides a practical solution to measure the immediate return of investments made into quality enhancement. If the user ratings are not collectable (as for traditional calling services) then using the longitudinal changes in call duration distribution is a workaround to measure changes in customer satisfaction. As brought out in section 2.1 the satisfaction leads to retention.

Even if the user feedback is available then there is more data about the call durations. This can result in a faster way to get actionable feedback. That is especially important if a change has unwanted effect resulting in reduced satisfaction. In case the user ratings are available the practical benefit of using call duration instead is to speed up the A/B testing. The choice depends on the following:

1. Ratio of rated calls.
2. Expected change on MOS scale as A/B test success criteria.
3. Standard deviation of the ratings to estimate the minimum number of user ratings needed.
4. Function to interpret the MOS criteria as call duration change success criteria. For better estimation of the expected change in duration it is useful to choose binning step size reasonably to magnify the area of interest.
5. Distribution of the call durations to derive standard deviation. That will help to determine the sample size needed and compare it with the sample size of user ratings.

In the analysis above most of the listed parameters are not brought out as the call duration distribution is depending on several parameters. These parameters are not generic for our dataset. Each slicing of the data introduces a need to recalculate call distribution and the relation between call duration and ratings.

When collecting the data, it can be useful to look for periodic patterns to avoid related effects. This can affect the first results from an A/B test when a short period is observed. The periodic pattern in Skype call quality feedback is not explained in this study, but it is acknowledged and considered when preparing the dataset.

The relation between MOS and call duration is significant on raw data. The technical issues like QoS related ones are affecting mainly short calls and might be related to the differences in the distributions between local and remote ratings for shorter calls. It is likely to improve the strength of the relationship if technical issues could be considered by accompanying models. Like in the analysis above we used separately modelled QoS impact prediction to audio call MOS that was derived from the user ratings and QoS parameters using machine learning based modeling. Useful things to monitor in parallel are for example call dropping, gap durations, concealment rates.

However, the described relation does not cover the many aspects that cannot be derived from the call ratings and should be taken into account with other QoE metrics. Such could be calls dropped in the initiation phase or usability issues that complicate to place or accept a call.

The dataset contained millions of calls, but this was not sufficient to create adequate views comparing the relation between MOS and call duration slicing by country, device, QoS impact, and device form factor. This problem of small sample sizes could be solvable by grouping (for example focusing on languages instead of countries when trying to find cultural impact) or compensating the QoS impact instead of discarding problematic calls. Another possibility to increase the sample sizes per bin would be to increase bin widths, but those experiments were not producing presentable results to confirm nor reject the impact.

More data is needed for further analysis, but it can also be that device and cultural components are affecting the call duration and ratings simultaneously and the relation between these is not changing significantly. The device form factor is relevant as on Figure 11 we can see rather orthogonal impact – clearly impacting mainly the call duration. Although the differences can be >1 MOS in extreme cases it might be explainable by difference in call duration distribution between devices and looking only the mean value is not descriptive enough in this case.

5. CONCLUSION

In this thesis, the relation between user ratings and call duration was demonstrated by using a different method compared to an earlier study (De Pessemier et al. 2015) on VoIP. The method used was averaging the user ratings on binned call duration. Regardless of the related work published earlier, the relation was surprisingly strong also on noisy raw data from Skype calls just by linearly fitting the logarithmic call duration. The relation differs for local and remote user ratings, but the differences were marginal after call duration exceeded a certain threshold.

User convenience impact to call duration was covered through slicing the dataset by device acoustic interface used. The relation and call duration distributions were brought out for handset, headphones, and speakerphone device form factor showing that these have statistically significant differences.

The practical usage of call durations instead of user ratings is assessing the impact of controlled experiments. It can speed up the test depending on the ratio of ratings available. If a tested improvement impacts a very specific subset of user base like a specific device in a specific mode, then collecting enough user ratings can be too time consuming (as the experiment might interfere with other experiments or approved releases) and this relation becomes handy in proving that the treatment has an impact. If there are no immediate gains from using the relation in parallel with ratings collecting, then the experiment could stop when either criteria is met. It might be the only method to model perceived QoE impact for platforms where it is not possible to collect the user ratings.

Future work should focus on investigating if the factors analyzed are casual, accumulate more data or use different methods to investigate the device and cultural impacts, and use the data from other telecommunication service providers to make generic models about the relation demonstrated.

6. REFERENCES

1. **Chatterjee S.** (2010) “Modeling, debugging, and tuning QoE issues in live stream-based applications - A case study with VoIP”. ITNG2010 - 7th International Conference on Information Technology: New Generations, pp.1044–1050. DOI: 10.1109/ITNG.2010.44
2. **Chen C.-N., Chu C.-Y., Yeh S.-L., Chu H.-H., Huang P.** (2012) “Measuring the perceptual quality of skype sources”. ACM SIGCOMM 2012 Computer Communication Review, Vol. 42, Issue 4. pp. 521–526. DOI: 10.1145/2377677.2377779
3. **Chen K.-T., Huang C.-Y., Huang P., Lei C.L.** (2006) “Quantifying Skype user satisfaction”, ACM SIGCOMM Computer Communication Review, Vol. 36, Issue 4, pp.399-410. DOI: 10.1145/1151659.1159959
4. **Chen Y., Wu K., Zhang Q.** (2015) “From QoS to QoE: A tutorial on video quality assessment”. IEEE Communications Surveys and Tutorials, Vol. 17, No. 2, pp.1126–1165. DOI: 10.1109/COMST.2014.2363139
5. **Conway A. E.** (2004) “Output-based method of applying PESQ to measure the perceptual quality of framed speech signals”. 2004 IEEE Wireless Communications and Networking Conference (IEEE Cat. No.04TH8733), pp.2521–2526. DOI: 10.1109/WCNC.2004.1311485
6. **Cooil B., Keiningham T. L. Aksoy L., Hsu M.** (2007) “A Longitudinal Analysis of Customer Satisfaction and Share of Wallet: Investigating the Moderating Effect of Customer Characteristics”. Journal of Marketing, Vol. 71 (January 2007), pp.67–83. DOI: 10.1509/jmkg.71.1.67
7. **Daengsi T., Wuttidittachotti P.** (2013) “VoIP Quality of Experience: A proposed subjective MOS estimation model based-on Thai users”. 2013 Fifth International Conference on Ubiquitous and Future Networks (ICUFN). pp. 407–412. DOI: 10.1109/ICUFN.2013.6614851
8. European Telecommunications Standards Institute (2015) “TS 126 131 - V12.3.0 (2015-01)”. Available at: http://www.etsi.org/deliver/etsi_ts/126100_126199/126131/12.03.00_60/ts_126131v120300p.pdf [Accessed February 15, 2017].

9. **Falk T.H., Chan W.Y.** (2009) “Performance study of objective speech quality measurement for modern wireless-VoIP communications”. EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2009, Issue 1, Article No. 171, 11 p. DOI: 10.1155/2009/104382

10. **Garner H. (**2015) “Clojure for data science”. Birmingham: Packt Publishing Ltd., 608 p.

11. **Gerpott T. J., Rams W., Schindler A.** (2001) “Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market”. Telecommunications Policy, Vol. 25, pp.249-269. DOI: 10.1016/S0308-5961(00)00097-5

12. **Hines A., Skoglund J., Kokaram A. C., Harte N.** (2015) “ViSQOL: an objective speech quality model”. EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2015, Issue 1, Article No. 13, 18 p. DOI: 10.1186/s13636-015-0054-9

13. **Hogg D.W.** (2008) “Data analysis recipes: Choosing the binning for a histogram”. Available at: https://arxiv.org/pdf/0807.4820.pdf [Accessed March 24, 2017]. ArXiv ID: arXiv:0807.4820v1

14. International Telecommunication Union (2015a) “ITU-T Rec. G.107.1: Wideband E-model”. ITU-T Recommendations. Available at: http://www.itu.int/rec/recommendation.asp?lang=en&parent=T-REC-G.107.1-201506-I [Accessed February 2017].

15. International Telecommunication Union (2015b) “ITU-T Rec. G.107: The E-model: a computational model for use in transmission planning”. ITU-T Recommendations. Available at: http://www.itu.int/rec/recommendation.asp?lang=en&parent=T-REC-G.107.1-201506-I [Accessed February 2017].

16. International Telecommunication Union (2014) “ITU-T Rec P.863: Perceptual objective listening quality assessment”. ITU-T Recommendations. Available at: http://www.itu.int/rec/recommendation.asp?lang=en&parent=T-REC-P.863-201409-I [Accessed February 2017].

17. International Telecommunication Union (2004) “ITU-T Rec P.563: Single-ended method for objective speech quality assessment in narrow-band telephony

applications". ITU-T Recommendations. Available at: https://www.itu.int/rec/T-REC-P.563-200405-I/en [Accessed February 2017].

18. International Telecommunication Union (2001) "ITU-T Rec P.862. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs". ITU-T Recommendations. Available at: http://www.itu.int/rec/recommendation.asp?lang=en&parent=T-REC-P.862-200102-I [Accessed February 2017].
19. International Telecommunication Union (2000) "ITU-T Rec P.920: Interactive test methods for audiovisual communications". ITU-T Recommendations. Available at: http://www.itu.int/rec/T-REC-P.920-200005-I/en [Accessed February 2017].
20. International Telecommunication Union (1998) "ITU-T Rec P.911: Subjective audiovisual quality assessment methods for multimedia applications". ITU-T Recommendations. Available at: https://www.itu.int/rec/T-REC-P.911-199812-I/en [Accessed February 2017].
21. International Telecommunication Union (1996) "ITU-T Rec P.800: Methods for subjective determination of transmission quality". ITU-T Recommendations. Available at: https://www.itu.int/rec/T-REC-P.800-199608-I/en [Accessed February 2017].
22. **Jelassi S., Rubino G., Melvin H., Youssef H., Pujolle G.** (2012). "Quality of experience of VoIP service: A survey of assessment approaches and open issues". IEEE Communications Surveys and Tutorials, Vol. 14 No. 2, pp.491–513. DOI: 10.1109/SURV.2011.120811.00063
23. **Jiang C., Huang P.** (2011) "Research of monitoring VoIP voice QoS". ICICIS '11 Proceedings of the 2011 International Conference on Internet Computing and Information Services, pp. 499–502. DOI: 10.1109/ICICIS.2011.130
24. **Kelloniemi A., Esken E., Koivuniemi K., Vaalgamaa M.** (2015) "Echo attenuation issues in handheld mode and test specifications for smartphones". ETSI Workshop on Telecommunication Quality beyond 2015. Available at: https://docbox.etsi.org/workshop/2015/201510_STQWORKSHOP/S05_SPEECH_AUDIO_VIDEO_QUAL_3_INSTRUMENTAL_TESTING/ECHO_ATTEN

UAT_HANDHELD_SMARTPHON_KELLIONIEMI_SKYPE.pdf [Accessed February 20, 2017]

25. **Kohavi R., Longbotham R.** (2009) “Controlled experiments on the web: survey and practical guide”. Data Mining and Knowledge Discovery, Vol 18, Issue 1, pp.140–181. DOI: 10.1007/s10618-008-0114-1
26. **Kohavi R., Longbotham R.** (2010) “Unexpected Results in Online Controlled Experiments”. ACM SIGKDD Explorations Newsletter, Vol. 12, Issue 2, pp.31–35. DOI: 10.1145/1964897.1964905
27. **Kohavi R., Henne R.M.R., Sommerfield D.** (2007) “Practical guide to controlled experiments on the web: listen to your customers not to the HiPPO”. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’07, pp.959–967. DOI: 10.1145/1281192.1281295
28. **Kottemann J.E.** (2017) “Illuminating statistical analysis using scenarios and simulations”. Hoboken, New Jersey: John Wiley & Sons, 312 p.
29. **Lewis, R. C., Booms, B. H.** (1983) “The marketing aspects of service quality”. Emerging perspectives on services marketing, Vol. 65, No. 4, pp.99-107.
30. Microsoft Corporation (2016) “Skype for Business Audio Test Specification: For personal devices using custom software or hardware audio preprocessing”. Available at: http://download.microsoft.com/download/E/1/0/E108B62D-C15D-4C45-874F-42E785B10B99/SkypeforBusiness_Logo_3_0.zip [Accessed February 2017].
31. **Mittal V., Kamakura W.** (2001) “Satisfaction, repurchase intent, and repurchase behavior: investigating the moderating effect of customer characteristics”. Journal of Marketing Research, Vol. 38 (February 2001), pp.131–142. DOI: 10.1509/jmkr.38.1.131.18832
32. **Montgomery D.C., Peck E.A., Vining G.G. (**2012) “Introduction to Linear Regression Analysis, 5th Edition.” New Jersey: John Wiley & Sons, 672 p.
33. **Nolting M., von Seggern, J.E**. (2016) “Context-based A/B Test Validation”. Proceedings of the 25th International Conference Companion on World Wide Web, pp. 277–278. DOI: 10.1145/2872518.2889306

34. **Ogunfunmi T., Narasimha M.J.** (2012) “Speech over VoIP Networks: Advanced Signal Processing and System Implementation”. IEEE Circuits and Systems Magazine, Vol. 12, No. 2, pp.35–55. DOI: 10.1109/MCAS.2012.2193436
35. **Oh H., Kim K.** (2017) “Customer satisfaction, service quality, and customer value: years 2000-2015”. International Journal of Contemporary Hospitality Management, Vol. 29, Issue 1, pp.2–29. DOI:10.1108/IJCHM-10-2015-0594.
36. **Oliver R.L.** (1999) “Whence Consumer Loyalty?” Journal of Marketing, Vol. 63 (Special Issue 1999), pp.33–44. DOI: 10.2307/1252099
37. **Parasuraman, A., Zeithaml, V., Berry, L.** (1985) “Conceptual model of service quality and its implications for future research”. Journal of marketing, Vol. 49, No. 4, pp.41–50. DOI: 10.2307/1251430
38. **De Pessemier T., Stevens I., De Marez L., Martens L., Joseph W.** (2015) “Analysis of the quality of experience of a commercial voice-over-IP service”. Multimedia Tools and Applications, Vol. 74, pp.5873-5895. DOI: 10.1007/s11042-014-1895-4
39. **Qiao Z., Sun L., Ifeachor E.** (2008) “Case study of PESQ performance in live wireless mobile VoIP environment”. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, (IEEE Cannes, France, 2008), 6p. DOI: 10.1109/PIMRC.2008.4699880
40. **Sahay A.** (2016) “Applied regression and modeling: a computer integrated approach”. New York, NY: Business Expert Press. 195 p.
41. **Stankiewicz R., Jajszczyk A.** (2011) “A survey of QoE assurance in converged networks”. Computer Networks, Vol. 55, Issue 7, pp.1459–1473. DOI: 10.1016/j.comnet.2011.02.004
42. **Takahashi A., Yoshino H., Kitawaki N.** (2004) “Perceptual QoS assessment technologies for VoIP”. IEEE Communications Magazine, Vol. 42, No. 7, pp.28–34. DOI: 10.1109/MCOM.2004.1316526
43. **Taylor S., Baker T.** (1994) “An assessment of the relationship between service quality and customer satisfaction in the formation of consumers' purchase intentions”. Journal of Retailing, Vol. 70, Issue 2, pp.163-178. DOI: 10.1016/0022-4359(94)90013-2

44. **Teas R.K.** (1993) “Expectations, Performance , Evaluation, and Consumers’ Perceptions of Quality”. Journal of Marketing, Vol. 57, No. 4 (Oct. 1993), pp.18–34. DOI: 10.2307/1252216

45. **Tsolkas, D., Liotou E., Passas N., Merakos L.** (2017) “A survey on parametric QoE estimation for popular services”. Journal of Network and Computer Applications, Vol. 77, Issue C, pp.1–17. DOI: 10.1016/j.jnca.2016.10.016

46. **Vos K., Sørensen K.V., Jensen S.S., Spittka J.** (2010) “SILK”. IETF 77. Available at: https://www.ietf.org/proceedings/77/slides/codec-3.pdf [Accessed February 2017].

47. **Wuttidittachotti P., Daengsi T.** (2017) “VoIP-quality of experience modeling: E-model and simplified E-model enhancement using bias factor”. Multimedia Tools and Applications, Vol. 76, pp.8329-8354. DOI: 10.1007/s11042-016-3389-z

48. **Zhang, W., Chang Y., Liu Y., Xiao L., Tian Y.** (2015) “Study the voice QoE for speech codec in Chinese environment”. 2014 IEEE 79th Vehicular Technology Conference (VTC Spring), 5 p. DOI: 10.1109/VTCSpring.2014.7023112

7. APPENDIX

7.1. Appendix 1: Abbreviations

A/B (test)	Controlled experiment where user is exposed to control (A) or treatment (B).
ACR	Absolute Category Rating. Method defined by ITU-T to rate a single test condition.
ITU-T	International Telecommunication Union – Telecommunication Standardization Sector.
MOS	Mean Opinion Score. Arithmetic mean over all individual values.
PESQ	Perceptual Evaluation of Speech Quality. Method defined by ITU-T to objectively evaluate speech quality.
QoE	Quality of Experience. Measure of the overall level of customer satisfaction.
QoS	Quality of Service. Measure of the overall performance of a computer (or telephony) network.
VoIP	Voice over Internet Protocol. Enables calling over packet-switched network.

Table 15: Frequently used abbreviations.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Oliver Loper,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose Relationship between call duration and perceived call quality based on Skype audio call telemetry from Android devices, mille juhendajad on Oliver Lukason ja Kuldar Kõiv,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **22.05.2017**