

Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Analysis of the Quality of Experience of a Commercial Voice-over-IP Service

Toon De Pessemier · Isabelle Stevens ·
Lieven De Marez · Luc Martens · Wout
Joseph

Received: date / Accepted: date

Abstract Voice-over-IP (VoIP) services, enabling users to make cheap phone calls using the Internet, are becoming increasingly popular, not only on desktop computers but also on mobile devices such as smartphones that are connected through mobile networks. Users' perception of the level of quality plays a key role in making a VoIP service to succeed or to fail.

This paper demonstrates the influence of technical parameters (such as the audio codec, type of data network, and handovers during the call), device characteristics (such as the platform, manufacturer, model, and operating system), and application aspects (such as the software version and configuration) on the subjective quality of a commercial VoIP service. The relative influence of all these parameters is determined and a decision tree combines these results in order to assess the subjective quality. Combining a large number of objective parameters in a decision tree to determine the user's subjective evaluation of the quality of a VoIP call is a novel and complex procedure. The subjective quality, in turn, has an influence on the duration of the call, and as a result an influence on the usage behavior of the service.

T. De Pessemier - L. Martens - W. Joseph
iMinds - Ghent University, Dept. of Information Technology, WiCa
G. Crommenlaan 8 box 201, 9050 Ghent, Belgium
Tel.: +32-09-33-14908
Fax: +32-09-33-14899
E-mail: toon.depessemier@ugent.be
E-mail: luc1.martens@ugent.be
E-mail: wout.joseph@ugent.be

I. Stevens - L. De Marez
iMinds - Ghent University, Dept. of Communication Sciences, MICT
Korte Meer 7-9-11, 9000 Ghent, Belgium
Tel.: +32-09-264 97 67
E-mail: lieven.demarez@ugent.be
E-mail: isabelle.stevens@ugent.be

The users' assessment of the service quality is not evaluated by merely taking a snapshot of the perceived quality at one moment in time but rather by analyzing the perceived quality over a longer period of time during service usage, which has not been done up to now. Analyzing the VoIP service using a regression analysis over a period of 120 days showed that the perceived quality decreases slightly when the user utilizes the service more often and gets more familiar with it.

Keywords VoIP · user experience · QoE · mobile

1 Introduction

Nowadays, a variety of Voice-over-IP (VoIP) services offers users the possibility to make free or cheap phone calls using the Internet. VoIP services such as Skype¹, ooVoo², and Google Hangouts³ are becoming increasingly popular due to the cost reduction benefit and its flexibility (VoIP is not tied to a specific address) [22]. Most of these services also provide a mobile application, enabling the users to make VoIP calls with their tablet or smartphone. This provides users an alternative for the traditional GSM (Global System for Mobile Communications) standard, i.e., the set of protocols for second generation (2G) digital cellular networks used by mobile phones for telephony. In recent years, cheap data plans of mobile operators and an increased coverage of cellular data networks further stimulated the growing popularity of VoIP applications on mobile devices.

Poor quality and un-reliability are two dark spots on the reputation of VoIP multimedia applications. Over the years, there has been much improvement due to better networks and audio codecs. But still, people are very finicky about voice quality in VoIP because they are used for years to the impeccable quality of landline phones [2].

As a result, end-to-end Quality of Service (QoS) management is becoming a challenge [15] in order to provide high quality and reliable communication over a best-effort network. Unfortunately, the QoS parameters do not always provide a good measure of the quality of the service as perceived by the user, since they only take into account network related aspects and neglect application and device characteristics, as well as non-technical and user aspects. In contrast, subjective quality measurements with actual test subjects can be performed to assess how a service is really perceived by the user [6, 7].

Although a subjective evaluation of the Quality of Experience (QoE) could be costly and time consuming, it gives more truthful results than an objective evaluation, which is merely based on QoS network performance parameters. QoE considers how users perceive and experience a multimedia communication service as a whole [19]. Since QoE relates to the user-perceived experience

¹ <http://www.skype.com>

² <http://www.oovoo.com>

³ <http://www.google.com/hangouts>

directly rather than to the implied impact of QoS, it is considered as a more important metric than QoS [20]. Both from a theoretical and empirical perspective, this concept has been broadened over the last years. As a result, different definitions of QoE exist, but all have similar notion, referring to user satisfaction [21]. By the ITU-T, QoE is defined as “the overall acceptability of an application or service, as perceived by the end-user”, which might be influenced by ‘user expectations’ and ‘context’ [1]. Identifying, understanding, and quantifying the most determining aspects making or breaking the QoE of individual (or communities of) users and translating these rich insights into service and application optimization recommendations, is considered to be essential.

The objective of this paper is to identify the technical parameters as well as device or application characteristics that influence the QoE during VoIP calls. For the first time, the evolution of the QoE during service usage was investigated by analyzing the users’ assessment of the quality of the VoIP service over a longer period of time.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work regarding VoIP services and the assessment of users’ QoE with VoIP. Section 3 presents VikingTalk, the commercial VoIP service that is used to study users’ QoE and the parameters that influence this QoE during VoIP calls. The data set obtained by logging the usage of this VoIP service, and the individual parameters that have been collected are discussed in Section 4. Section 5 elaborates on the influence of individual technical parameters, device characteristics, and application aspects on the subjective quality of the VoIP service. In Section 6, a decision tree combines all these parameters to perform a joint analysis in order to assess the subjective quality. The influence of the subjective quality on the duration of the call is discussed in Section 7. In Section 8, the users’ assessment of the quality of the VoIP service is considered as a measure that can evolve. This Section discusses the evolution of the quality rating over time and over the user’s experience or familiarity with the service. Finally, a brief conclusion on the research results is offered in Section 9.

2 Related Work

In many cases, the capabilities of VoIP technologies are analyzed using a private network [17], which enables the modification of the IP infrastructure and may have other characteristics regarding traffic or topology than the public Internet. Other experiments are performed in a predetermined environment covering a limited area, such as a university campus [4], thereby limiting the freedom of the test subjects with the risk of obtaining results that are not generally applicable.

These experiments showed that there is no precise match on both QoS and QoE assessments, because of QoE parameters that cannot be inferred from QoS analysis, because of the different effect of some QoS parameters which

prevail, or even because of the popularity of the VoIP application [4]. Still, the predetermined context and controlled setting of these experiments limit the applicability of the results. Moreover, the number of test subjects participating in a (mobile) QoE experiment is often limited to a few dozen due to time and budget constraints [7].

In contrast, this paper overcomes the limitations of previous studies, notably: 1) a limited number of users participating in the study; 2) tests performed in a controlled environment, thus results not directly applicable to everyday services in real world conditions. This work investigates the QoE during VoIP calls based on data of *a large number of test subjects* (more than thousand) making voice calls *in their daily environment* without any location, time or usage constraints.

These test subjects are real customers of a commercial VoIP application, namely VikingTalk⁴, developed and managed by a Belgian mobile network operator, namely Mobile Vikings⁵. This eliminates any possible bias that is associated with the recruiting of test subjects who are asked to use a service merely for the sake of evaluation purposes.

Measurements in public mobile data networks showed that the capacity offered by these networks is sufficient to make mobile VoIP calls possible [9]. In this context, the influence of QoS parameters, i.e., network related aspects, such as throughput, packet delay, or packet loss, on the QoE during VoIP calls has been extensively studied [5,9]. These network related aspects have shown to be determining factors for the QoE during VoIP, especially for slower data networks such as UMTS (Universal Mobile Telecommunications System) or EDGE (Enhanced Data Rates for GSM Evolution) networks.

The new generation of cellular data networks, such as HSPA (High Speed Packet Access) and LTE (Long Term Evolution) networks, and the traditional WiFi networks offer a higher bandwidth and reduce the risk of network impairments, such as packet loss and insufficient throughput, considerably. Moreover, forward error correction mechanisms using redundant data added to the voice stream can help to recover from these network impairments thereby sustaining audio quality [10]. In this context, the optimal level of redundancy for different network and codec settings has been explored in order to optimize user satisfaction with VoIP services.

As a result, QoS parameters do not always match with the user's QoE [4] during VoIP calls. The reduced risk of network impairments increases the relative influence of device characteristics, such as the platform, manufacturer, model, and operating system, as well as application aspects, such as the software version and configuration, on the QoE. However, the influence of these device characteristics and application aspects on the QoE during VoIP calls has never been investigated according to our knowledge.

Moreover, traditional user experiments evaluate the QoE by taking a snapshot of the subjective experience of the user at one moment in time during

⁴ <http://www.vikingtalk.com>

⁵ <http://www.mobilevikings.com>

the complete use process of the service. Nevertheless, the user's experience can change over time, and is influenced by his or her prior expectations about the service [8]. Before people start using a particular product or service, they tend to already have some kind of preconception influencing their expectations [12]. After a user has adopted a product, and is using it more or less regularly, the actual use process evolves. As a user has more experience with the service, familiarity of the user increases, and this has an impact on how the service is being used. Karapanos et al. describe different phases a user goes through when using a product or service going from "the initial experiences with a service" over "giving the service a meaningful place in life" to "integrating the service in the user's lifestyle" [12]. Therefore, it is important that the QoE is not evaluated on a single point in time but rather over a continuous period during the use process. The resulting data can be understood as a time series of one feature, of which the clue for evaluation is the detection of trends in several successive time points [16]. This paper is the first to monitor trends in and analyze the evolution of the user's QoE with a VoIP service over a longer period of time (four months).

3 The VikingTalk VoIP Service

Nowadays, a large variety of VoIP services is available. In this article, the user's experience with one of these services, namely VikingTalk, is investigated. VikingTalk offers users a multi-network VoIP telephony service, similar to the well-known Skype, enabling to call or receive calls from other VoIP users or people connected through the traditional fixed or mobile PSTN (Public Switched Telephone Network).

In contrast to Skype, the VikingTalk application has a transfer option that allows its customers to switch from VoIP (using the available cellular data connection or WiFi) to the customer's primary mobile operator (using GSM to connect) during the same call. If users have enabled this handover process in the configuration settings, the mobile operator automatically takes over the VoIP call in case of a poor voice quality or loss of coverage on the data network. By a short beep sound during the voice call, users are informed about the transfer of the voice call from VoIP over the available data network to non-VoIP over the GSM network of the mobile operator. If a voice call initiated using VoIP was transferred to the GSM network of the mobile operator because of technical issues, it is not transferred again to VoIP (not even if the data connection is sufficiently recovered) in order to limit the possible disruptions introduced by the switching during the voice call.

VoIP calls to another VoIP user are free for users of the VikingTalk application. VoIP calls to the fixed or mobile PSTN are charged based on the duration of the call. For voice calls that use the GSM network of the mobile operator (because of technical reasons), or voice calls initiated using VoIP that are transferred to the GSM network of the mobile operator, the rates offered by the mobile operator apply for the duration of the call over this GSM net-

work. Charges for data traffic are not included in the VikingTalk rates and are charged separately.

Besides the option to enable the automatic handover from VoIP to the network of the mobile operator, users have some additional configuration options in the application. Users can choose whether or not they want to use EDGE or 3G (3rd Generation) networks for their voice calls in case a WiFi connection is not available. For the VoIP service, a WiFi network is the first choice. If a WiFi network is not available, two alternative solutions exist for the voice call: either using the mobile data network (EDGE or 3G) that is available through the user's mobile data plan (data credit) or using the GSM network of the user's mobile operator thereby charging the user as for a traditional phone call (voice credit). If the user opts not to use EDGE or 3G, the voice call is transferred immediately to the GSM network of the mobile operator as soon as the user is out of range of the WiFi network. If the user opts to use EDGE or 3G, this data network is preferred above the GSM network of the mobile operator. Then, the available EDGE or 3G network is used by default in the absence of WiFi, and only in exceptional cases when the throughput of the mobile data network (EDGE or 3G) becomes insufficient, the call is still switched to the GSM network of the mobile operator. The use of EDGE or 3G can induce a poorer voice quality compared to the GSM network of the mobile operator, but on the other hand, the use of 3G or EDGE can reduce the cost charged by the mobile operator for voice calls.

In addition, users can configure whether or not they want to receive voice calls over VoIP. If this option is not enabled, all incoming voice calls are routed over the GSM network of the mobile operator and the caller is charged for this.

4 The VikingTalk VoIP Data Set

For billing purposes and customer services, data about the VikingTalk service usage are internally stored and continuously monitored. Analysis of these data can provide insights into the parameters that influence the service usage and users' QoE. The analysis of this paper was based on a data set containing the details of all voice calls made using VikingTalk over a period of nearly four months (120 days), from October 1, 2012 to January 28, 2013. This data set provides on the one hand a representative set of samples to investigate the influence of different parameters and on the other hand it allows to analyze trends in the users' QoE with the VoIP service over a longer period of time. The data set consists of objective, technical parameters regarding the call as well as subjective evaluations of the quality of the voice call.

The *objective*, technical parameters include an identification of the user who initiates the call and the user who receives the call, the audio codec, an indication of handovers during the call, the configuration settings of the application, the type and version of the operating system of the phone, the manufacturer and model of the phone, the version of the VikingTalk application, timestamps indicating the start and end of the call, and the duration of

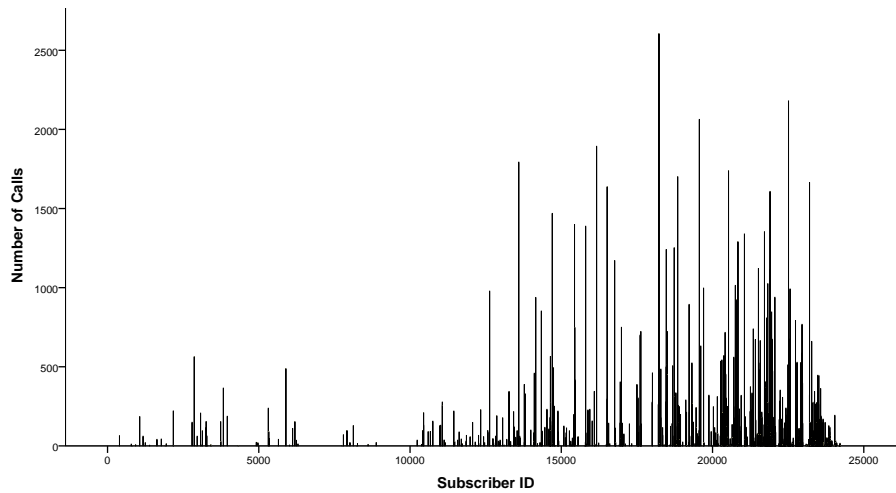


Fig. 1 Histogram of the number of calls made during the 120-day period per subscriber ID.

the voice call. All parameters that are important for this study are listed in Table 1, together with their possible values and the number of samples obtained for each value. (The duration of a call is handled separately in Table 3) After each voice call, the user has the opportunity to evaluate the quality of the VoIP service using a 5-point scale rating mechanism, thereby yielding a *subjective* evaluation of the user’s experience with the VoIP service.

The data set contains 127, 826 samples each corresponding to one voice call made by during the 120-day period. The quality of 30, 384 of these voice calls, or 23.8% of the calls, is evaluated by the users using the 5-point scale rating mechanism. During the 120-day period, 1050 subscribers of the VoIP service were active, who each made on average 121.7 voice calls, or around 1 voice call per day.

However, not all 1050 subscribers of the service are equally active. Figure 1 shows a histogram of the usage of the VoIP service, indicating the number of voice calls made by each user during the time window of the data set. Most active subscribers use the service occasionally to make a voice call. In contrast, a small group of users used the service very intensively during the 120-day period. Three users made more than 2000 calls, and one of them, the most active user, made 2605 voice calls, or almost 22 calls a day. Although it is actually not allowed according to the terms of use of VikingTalk, some of these active subscribers might use the VikingTalk service as part of their business (e.g., call centers, phone shops).

Since evaluating the quality of the voice call is an optional feature that users can disable in the application, most records in the data set (76.2%) do not contain such a subjective evaluation. Figure 2 shows the distribution of the ratings for the voice calls that received an evaluation from the users. This histogram indicates that users tend to provide ‘extreme’ ratings, 1 or 5, rather than moderate ratings, 2, 3, or 4. This is a typical user behavior

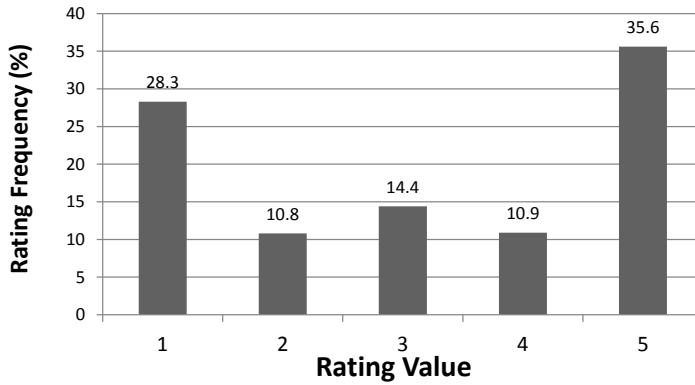


Fig. 2 The distribution of the subjective ratings of the user evaluating the voice call.

for 5-point scale mechanisms, and was already witnessed for ratings on the YouTube website [18].

5 Influencing Technical Parameters

In order to quantify the user's QoE and in the end improving the QoE, analyzing the influence of the different technical parameters is of vital importance. This section discusses in detail the influence of objective, technical parameters related to network handovers, the user's device, and the mobile VoIP application on the user's rating for the quality of the call, which indicates the subjective experience of the user with the VoIP service. Table 1 summarizes the effect of all the considered parameters on the subjective experience of the user.

5.1 Codec

An important technical parameter is the codec that is used to transfer audio over the network. For the VikingTalk service, two different codecs are used during the time period of the data set: GSM/8000 (8kHz mono) for approximately 40.8% and GSM/32000 (32kHz mono) for the remaining 59.2% of the calls for which this information is available. As expected, Table 1 shows that the quality of the call is rated higher in case GSM/32000 is used than in case of GSM/8000. The difference in mean rating achieved for both codecs is noteworthy (0.28, cfr. Table 1). Moreover, Table 2 lists the results of the T-test, showing that this difference in mean rating is statistically significant. This indicates an influence of the audio codec on the resulting quality of the voice call as perceived by the user. The T-test is a statistical hypothesis test which can be used to determine if two sets of data are significantly different from each other. The test statistic follows a Student's t-distribution if the null hypothesis is supported, which is in this case the equality of the mean values of the

two sets [13]. If the resulting p-value is lower than 0.05, the null hypothesis is rejected, and the two sets are considered as significantly different.

5.2 Data to GSM Handover

Another important technical parameter that can have a direct influence on the voice quality is the occurrence of handovers during the call. As explained in Section 3, the VikingTalk service falls back on the network of the user's mobile operator in case of a poor voice quality or loss of coverage. For most voice calls, network conditions are stable and these handovers are not necessary. Only in 1.1% of the voice calls in the data set, a data-to-GSM handover was registered. Because such a data-to-GSM handover can introduce distortions during the call, calls in which such an handover occur receive generally a lower rating from the user than calls without handovers. Table 1 indicates that the mean difference between a call with and a call without handover is 0.22 on the rating scale and Table 2 shows that this difference is significant. As a result, the occurrence of handovers during the voice call has a significant impact on the subjectively-perceived quality of the call.

5.3 Setting: Handovers to GSM

This handover process in case of a poor quality can be disabled by the users in the settings of the mobile application in order to reduce the cost charged by the mobile operator (Setting: Handovers To GSM). In case the handover process is enabled, users can notice a short distortion during the handover. In case it is disabled, users might experience a poor quality during the whole call due to a bad data network.

Analyzing the data set shows that the automatic handover process was enabled for only 10.0% of the calls, which also explains why during only 1.1% of the calls an handover occurs. Table 1 shows that if the handover process is enabled, the mean rating is higher (difference of 0.09) compared to the voice calls in which the handover process was disabled. This difference, which showed to be significant (Table 2), can be explained by two effects that strengthen each other. Firstly, users might expect a better quality if handovers are disabled, since handovers are usually associated with interruptions or distortions. These higher expectations might induce a lower subjective rating. Secondly, if a handover to the network of the mobile operator is not allowed by the user, all voice traffic has to be transmitted over the available cellular data (or WiFi) network, even if this connection provides a limited throughput. This increases the risk of a poor voice quality, or even an interruption of the call in case of a network disconnection. As a result, the configuration of the automatic handover process has a significant influence on the subjectively-perceived quality.

5.4 Setting: EDGE / 3G Calls

Through another application setting called “Setting: EDGE / 3G Calls”, users can specify whether or not they want to use EDGE or 3G networks for their voice calls in case a WiFi connection is not available. For approximately 23.8% of the voice calls, users enabled the use of EDGE or 3G. Table 1 shows that if this setting is enabled, the mean rating is 0.35 lower compared to the cases in which EDGE or 3G is not used. If EDGE or 3G is not used, the voice call is transferred to the GSM network of the mobile operator as soon as the user is out of range of the WiFi network. As a result, the higher mean rating for calls in which this setting is disabled can be explained by the superior quality of the GSM network of the mobile operator compared to the resulting quality of voice calls using EDGE or 3G networks. Table 2 shows that this difference in subjective quality is significant, indicating that enabling or disabling EDGE and 3G has a significant influence on the subjectively-perceived quality.

5.5 Setting: Incoming VoIP Calls

The last application setting allows users to configure whether or not they want to receive voice calls over VoIP. Users might want to initiate calls over VoIP in suitable conditions (e.g., if WiFi is available), but prefer to receive calls over the traditional GSM network of their mobile operator to optimize voice quality in less-suitable conditions (e.g., on the move). Table 1 shows a very small (0.07) difference in mean rating between cases in which incoming VoIP calls are enabled (approximately 91.2%) and cases in which VoIP is not used for incoming calls (approximately 8.8%). Furthermore, according to Table 2, this difference is not significant. So, the subjectively-perceived quality of a call is not influenced by the fact that users allow or not allow the application to receive incoming VoIP calls.

5.6 Mobile Platform

Also device characteristics, such as the platform, version of the operating system, manufacturer and model of the phone, as well as application aspects, such as the software version of the mobile application, can influence the subjectively-perceived quality of the VoIP service and as a result have an impact on user’s QoE. The majority of the voice calls (73.6%) is made on the iOS platform using an iPhone. The remaining 26.4% of the voice calls is made using an Android phone. The VikingTalk application is not (yet) available for other mobile platforms such as Blackberry or Windows Mobile. A comparison between the ratings achieved on the Android and iOS platform (Table 1) shows that the mean rating on the Android platform is 0.32 lower than the mean rating achieved on iOS. Table 2 reveals that this difference is significant, and as a result that (the user’s experience with) the mobile platform has an

influence on the user's evaluation of the quality of the VoIP service. This effect may in part be due to the fact that all iPhones (also the older ones) are well capable to run the VoIP application smoothly, whereas many low-end Android phone might induce a lower user experience due to limited resources in terms of memory and processing power.

5.7 Version of the Operating System

Further investigation reveals that even different versions of the mobile operating system can induce significant differences in the achieved ratings for the quality of the VoIP service. Table 1 compares the mean rating for five different versions of Android ranging from Eclair (2.0-2.1) to Jelly Bean (4.1.x) and shows that a higher version typically corresponds to a higher mean rating. One exception is the Froyo (2.2) version of Android, which does not follow this gradual increase of the mean rating.

An ANOVA (ANalysis Of VAriance) is used to determine whether or not the means of several groups are all equal, and therefore generalizes the T-test to more than two groups [13]. An ANOVA hypothesis test showed that a significant difference ($p = 0.00$) in quality rating exists between the different Android versions; all versions can be considered as mutually different in terms of the resulting quality rating. For iOS, eight different versions ranging from 4.3.1 to 6.0.2 are compared. Again, the results show a slight increase in the mean rating for subsequent version of the operating system, with exceptions for versions 4.3.3 and 6.0.1 that result in a higher mean rating than expected. An ANOVA test shows that a significant difference ($p = 0.00$) in quality rating exists between the different versions of iOS, except for the versions 4.3.5, 5.0.1, and 6.0, which are very close to each other in terms of mean rating. The increase in mean rating for subsequent versions of the operating systems may be due to, on the one hand improvements of the software in terms of fluency, user experience, and bug fixing. On the other hand, newer versions of the operating system often go together with newer devices having more resources and thereby proving a better experience to the users.

5.8 Manufacturer

Regarding the manufacturer, 73.6% of the voice calls in the data set is made using an iPhone running iOS produced by Apple. For the Android platform (covering the remaining 26.4% of the voice calls) different manufacturers have a market share in the production of mobile phones. 19.3% of the voice calls in the data set is made using a phone produced by Samsung; 2.9% of the calls is made using a phone of HTC; for 2.9% of the calls, a phone of SEMC was used, and the remaining 1.3% of the calls were made using phones of a variety of manufacturers. Table 1 shows that the mean rating of the service quality varies for phones of different manufacturers and an ANOVA hypothesis

test confirmed the significance of these differences with a p-value of 0.00. The highest mean rating is achieved by using iPhones of Apple. The set of calls made using an HTC device has approximately the same mean rating. No significant difference was found between these two manufacturers in terms of the mean rating for the VoIP service.

In contrast, the mean rating achieved by using phones produced by Samsung is significantly lower according to the T-tests: a difference of 0.38 with phones produced by Apple, and a difference of 0.37 with phones of HTC. The reason for this might be the large variety of phones produced by Samsung. Whereas HTC and Apple are specialized in the production of mid-range to high-end phones, Samsung sells mid-range, high-end as well as a lot of low-end phones. These cheap, low-end phones typically have limited hardware resources, a restricted number of features, and limited capabilities, which might deteriorate the user's experience with mobile services such as VikingTalk. The lowest mean rating for the VoIP service is achieved by using phones produced by SEMC. The mean rating achieved by phones of SEMC is lagging behind with a significant difference of respectively 0.54 and 0.53 with phones produced by Apple and HTC. Again, this can be explained by SEMC's low-end phones, which might have a negative influence on the user's subjective evaluation.

5.9 Phone Model

The influence of the device on the user's subjective evaluation for the quality of the VoIP service is confirmed by Figure 3, which shows the mean rating achieved for the most popular phone models. In this figure, the phone models are arbitrary classified according to their hardware specifications. The high-end phones, typically having a quad-core processor and a big screen, are represented by light green bars. The orange bars indicate the mid-range phones, typically having a dual-core processor and a big screen. The low-end phones, typically having a single-core processor and a small screen, are represented by dark red bars. Since no information about the exact iPhone model was available in the data set, all iPhones are classified in the same category, which is considered as mid-range. Figure 3 demonstrates that large and significant differences in mean rating between the various phone models exist (ANOVA hypothesis test with $p = 0.00$). In general, the ratings achieved by using the high-end and mid-range phones are higher than the ratings achieved by using a low-end phone. However, since many other factors influence the user's rating, exceptions exist: low-end phones that achieved a high mean rating, such as Acer's E310 model, and high-end or mid-range phones with a low mean rating, such as the Galaxy Nexus.

5.10 Version of the Mobile Application

The last technical parameter that was investigated is the version of VikingTalk, the mobile application that enables the users to make VoIP calls. Table 1 shows

Codec	Samples	Mean Rating
GSM/8000	3276	2.75
GSM/32000	4745	3.03
Data to GSM Handover	Samples	Mean Rating
Yes	336	2.93
No	30048	3.15
Setting: Handovers to GSM	Samples	Mean Rating
Enabled	3045	3.23
Disabled	27339	3.14
Setting: EDGE / 3G Calls	Samples	Mean Rating
Enabled	7242	2.88
Disabled	23142	3.23
Setting: Incoming VoIP Calls	Samples	Mean Rating
Enabled	27701	3.15
Disabled	2683	3.08
Mobile Platform	Samples	Mean Rating
Android	8021	2.91
iOS (iPhone)	22363	3.23
Android Version	Samples	Mean Rating
API Level 7: Eclair, 2.0-2.1	187	2.27
API Level 8: Froyo, 2.2	360	4.09
API Level 10: Gingerbread, 2.3.3-2.3.7	4526	2.74
API Level 15: Ice Cream Sandwich, 4.0.3-4.0.4	1920	2.92
API Level 16: Jelly Bean, 4.1.x	1022	3.38
iOS Version	Samples	Mean Rating
4.3.1	51	1.47
4.3.3	1061	4.62
4.3.5	55	2.67
5.0.1	1795	2.82
5.1.1	6636	3.36
6.0	4378	2.78
6.0.1	7952	3.26
6.0.2	396	3.63
Manufacturer	Samples	Mean Rating
Apple	22363	3.23
HTC	877	3.22
Samsung	5873	2.85
SEMC	888	2.69
VikingTalk App Version	Samples	Mean Rating
4.0.4.25	59	3.90
4.2.2.18	8016	2.91
4.2.2.22	22284	3.23

Table 1 The influence of technical parameters on the subjective evaluation of the quality of the voice call.

that different mean ratings of the quality are achieved for the different versions of the software and an ANOVA showed that these differences are significant ($p = 0.00$). As with the version of the operating system, we expected an increase in mean rating for subsequent versions of the application. However, the first version, 4.0.4.25, achieved the highest mean rating, which is significantly

Technical Parameter	t-statistic	Degrees of Freedom	p-value
Codec	8.19	7604.52	0.00
Data To GSM Handover	2.54	343.92	0.01
Setting: Handovers to GSM	-3.09	3821.82	0.00
Setting: 3G / EDGE Calls	16.33	13333.69	0.00
Setting: Incoming VoIP Calls	1.95	3408.84	0.05
Mobile Platform	15.31	15465.22	0.00

Table 2 T-tests indicating whether or not the influence of a technical parameter on the subjective evaluation of the quality of the call is significant.

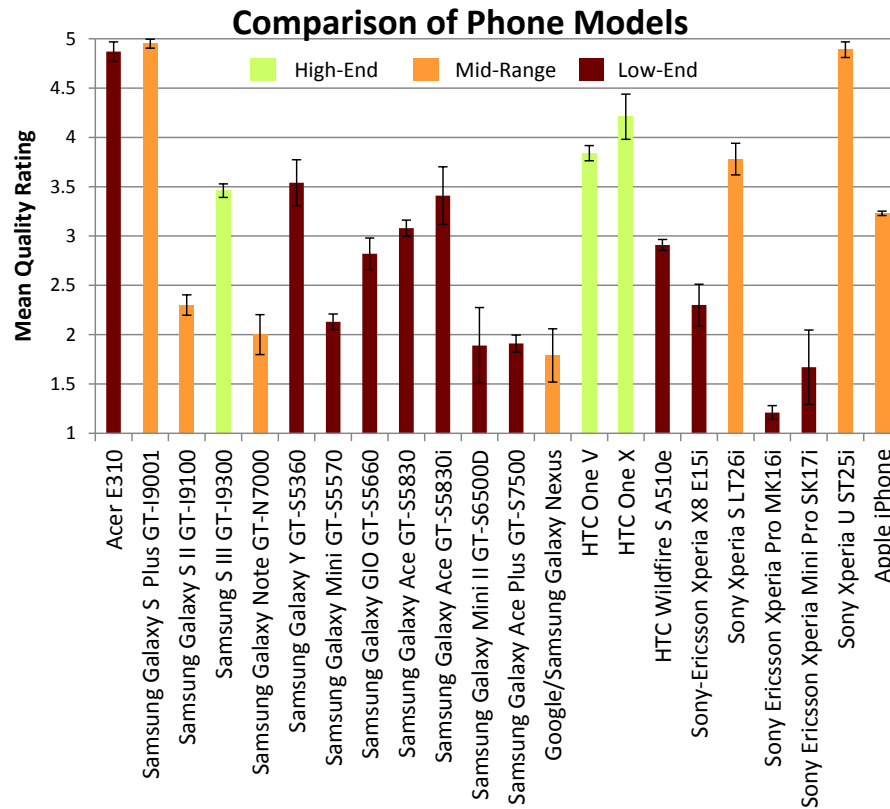


Fig. 3 Comparison of the mean ratings for the different phone models. The light green bars stand for high-end phones; the orange bars correspond to midrange phones; and the dark red bars stand for low-end phones.

higher than the mean rating of the two subsequent versions. This might be due to the fact that the users' expectations are the lowest for the first version of the application, thereby inducing higher subjective ratings. In contrast, when new versions of the application became available, users already had the experience of using the first version of the application, and they might have higher expectations for a new version. The latest version, 4.2.2.22, achieved

a mean rating that is significantly higher than the previous version, 4.2.2.18, probably because of improvements in the service.

6 Combining Technical Parameters in a Decision Tree

All the technical parameters that showed to have a significant influence on the user's subjective rating (cfr. Section 5) can be combined into a *decision tree*. A *decision tree* is a classification technique that uses a tree-like graph or model of decisions and their possible consequences [3]. This decision support tool is in many cases preferred over other non-parametric techniques because of the readability of their learned hypotheses and the efficiency of training and evaluation. The output is highly visual and easy to interpret. This makes it possible to reason on the data and investigate the direct influence of the adjustment of a specific parameter.

For this analysis, the SPSS⁶ predictive analytics software and the CHAID (CHi-squared Automatic Interaction Detection) technique was used to compose the decision tree. CHAID is a technique that is based upon adjusted significance testing (Bonferroni testing) for composing a decision tree [14]. By default, CHAID uses multiway splits to make decisions and classify the samples into different classes. In this analysis, the dependent variable is the quality rating, indicating the subjective experience of the user during the voice call; the significant technical parameters (cfr. Section 5) are considered as the independent variables. The large sample size of the data set allowed to use the CHAID technique to effectively compose a decision tree thereby partitioning the data samples into different cases.

Figure 4 to 7 show the decision tree consisting of the root, which is the starting point, the leaves, which are the end points and represent different classes, and branches, which represent conjunctions of technical parameters that lead to those classes. Figure 4 shows the root and the first branch of the decision tree. Figure 5 to 7 show the subtrees for the children of the root. For each class, the decision tree indicates the mean and standard deviation (Std. Dev.) of the ratings, the number of samples (n), the frequency of occurrence (%), and a prediction of the rating for future calls. Each split is characterized by the technical parameter used to make a decision, the adjusted p-value (Adj. p-value), the F-statistic (F), and the degrees of freedom (df).

Figure 4 shows that the root node of the tree has a mean rating of 3.15 with a standard deviation of 1.66. In this root node, all samples with a rating are considered and no distinction has been made based on the technical parameters. By making a distinction based on the technical parameters, more specific situations can be distinguished, e.g., in the leaves (Figure 5, 6, and 7). Node 14, visible in Figure 7, achieved the highest mean rating (4.81), which is an estimation of the quality of the service that is significantly higher than the estimation of the root. This node represents calls without handovers made

⁶ <http://www-01.ibm.com/software/analytics/spss>

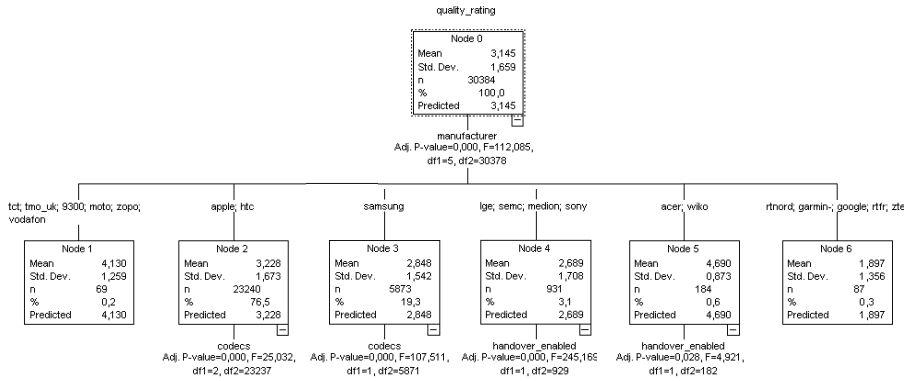


Fig. 4 The root and the first branch of the decision tree typifying the samples based on the objective, technical parameters.

by using an Acer or Wiko phone. The standard deviation of this class is 0.69, significantly lower than the standard deviation of the root.

The decision tree also provides insight into the relative influence of the different technical parameters. E.g., the first branch makes a decision based on the device manufacturer (Figure 4). The mean ratings of the difference nodes indicate that this technical parameter has a significant influence on the resulting quality rating: a mean rating of 1.90 for devices of Garmin, Google, ZTE, and others, whereas devices of Acer and Wiko obtained a mean rating of 4.69. In contrast, the impact of the version of the VikingTalk app is smaller. E.g., in the branch corresponding with devices of Apple or HTC and the codec unspecified (Figure 5), the difference in mean rating between version 4.2.2.22 and 4.0.4.25 is limited (0.67).

This decision tree offers the service provider a tool to estimate the QoE based on the user’s situational context (phone, software version, codec, handovers, etc.). Moreover, it provides insights into the relative influence of technical parameters on the resulting QoE and the effect of changes in the values of these parameters on the QoE. To suit future situational contexts, the decision tree can iteratively be refined based on the knowledge of new examples [11].

7 Impact on the Call Duration

The user’s QoE during the voice call may have an influence on the user’s usage pattern and interaction behavior with the service. The assumption is that if the quality of the voice call is bad, users will have a worse experience and stop using the service earlier compared to a situation with a perfect quality. So, in case of a low quality as expressed by the user’s subjective rating, the duration of the voice call will be shorter.

To investigate this influence, the voice calls are partitioned based on the user’s quality rating and for each class, Table 3 lists the mean call duration.

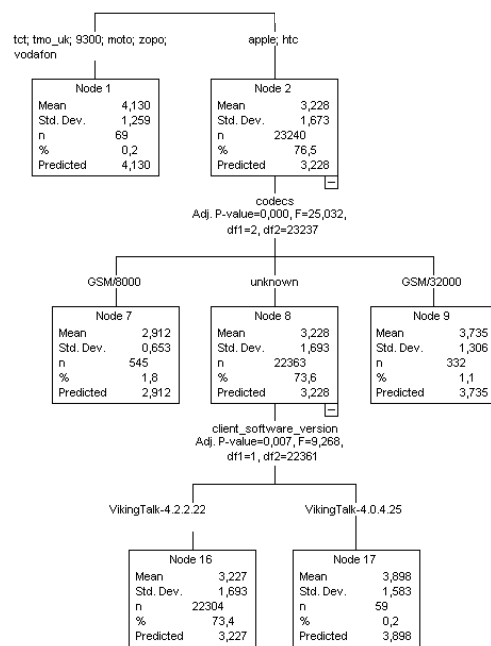


Fig. 5 The part of the decision tree showing the subtree for the first two children of the root.

Although there is no linear relationship between the rating and the duration of the call over the full rating scale, the influence of the quality on the duration of the voice call is revealed by the mean values. The voice calls with low quality ratings (1 or 2) have the shortest mean duration (less than 4 minutes). Calls with a higher rating (3, 4, or 5) have a longer duration (more than 4 minutes), but for these classes no relationship between the quality and the duration is discovered. So as soon as the quality is fair, the duration of the voice call will not increase as the quality is further improved.

The T-test confirms that voice calls that received a low quality rating from the users (1 or 2) have a significantly shorter duration than voice calls that received a fair or high rating (3, 4, or 5). The following parameters were obtained: degrees of freedom = 29495.01, $t = 8.23$, $p = 0.00$. The difference in duration between the class of calls with rating = 1 and the calls with rating = 2 is not significant. Also for the classes receiving higher ratings from the users, differences in call duration are not significant. So, if users are not satisfied with the quality, voice calls will be shortened compared to the situation of a perfect quality. In contrast, if users are satisfied with the quality of the call, the duration of their call is not influenced by further quality improvements.

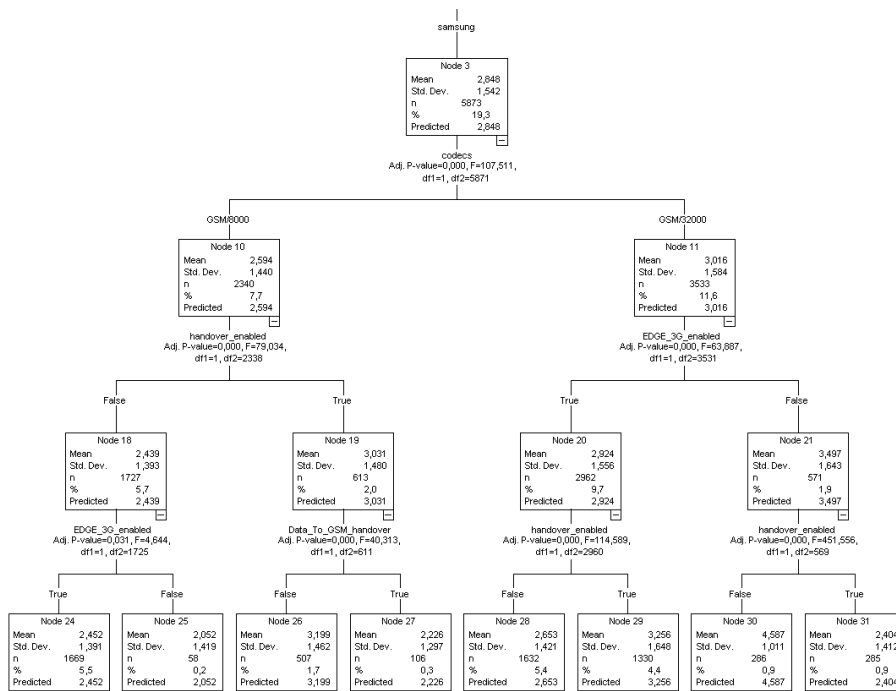


Fig. 6 The part of the decision tree showing the subtree for the third child of the root.

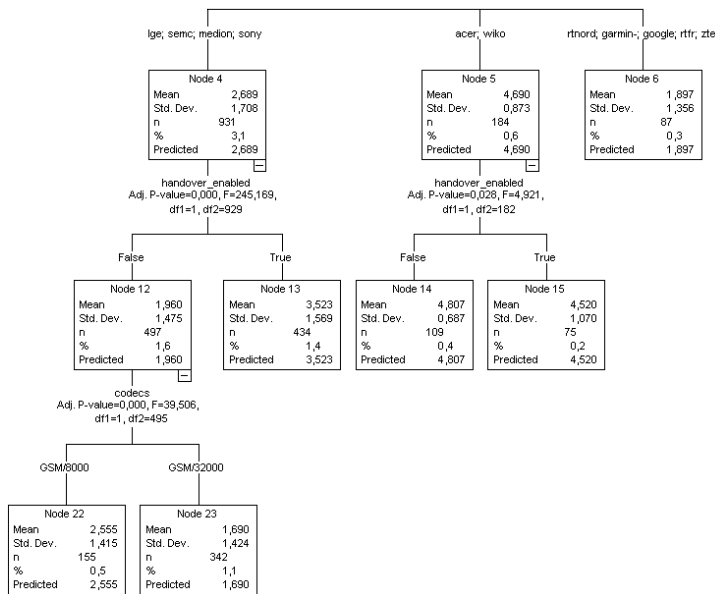


Fig. 7 The part of the decision tree showing the subtree for the last three children of the root.

Quality Rating	Samples	Mean Call Duration (seconds)
1	8613	238
2	3276	224
3	4382	320
4	3309	265
5	10804	284

Table 3 The mean duration of the voice calls partitioned based on the user’s quality rating.

8 Evolution over Time

As users become more familiar with the service, their perceptions of the service quality may change over time. Therefore, evaluating the QoE over a longer period of time is essential to quantify the user’s experience with a service. This has never been investigated up to now for a widely-used, mobile, multimedia service, such as the VikingTalk VoIP service.

Figure 8 shows the mean quality rating that was achieved during each day since the start of the experiment. The graph also illustrates the result of a linear regression analysis, fitting the measured mean ratings to a linear function. The independent variable of this regression analysis is the day since the start of the experiment; the dependent is the obtained mean rating. However, the slope of this linear function showed not to be significantly different from zero ($p = 0.33$). So, the mean quality rating measured during each day does not increase or decrease significantly over time.

The absence of a trend in the quality assessment can be explained by the fact that users of the service have a different usage pattern, as explained in Section 4. On average, users make 1 voice call a day, but many users utilize the service only sporadically, and some users utilize the service very intensively (up to 22 calls a day). As a result, different users get familiar with the service and gain experience with the usage of the application at a different rate. Users who utilize the service very intensively are familiar with it after a few days. In contrast, users who utilize the service sporadically need more time to get familiar with it, and their experience with the service will evolve slower.

Therefore, the evolution of the quality ratings can be investigated as a function of the user’s usage and familiarity with the service. Figure 9 shows the mean quality rating over the subsequent voice calls made by the user. The *call sequence number* can be seen as an indication of the user’s *familiarity* with the service. The graph clearly shows a decrease of the mean quality rating over the first 120 voice calls made by the users. This is confirmed by a linear regression analysis, resulting in a linear model with a significant slope for the *call sequence number* ($p = 0.00$, $R^2 = 0.29$). Equation (1) shows that the mean quality rating is 3.22 for the first call of the users and that this mean quality rating slightly decreases as users have utilized the service more often. After making 120 voice calls, the mean quality rating is 0.312 lower compared to the first call.

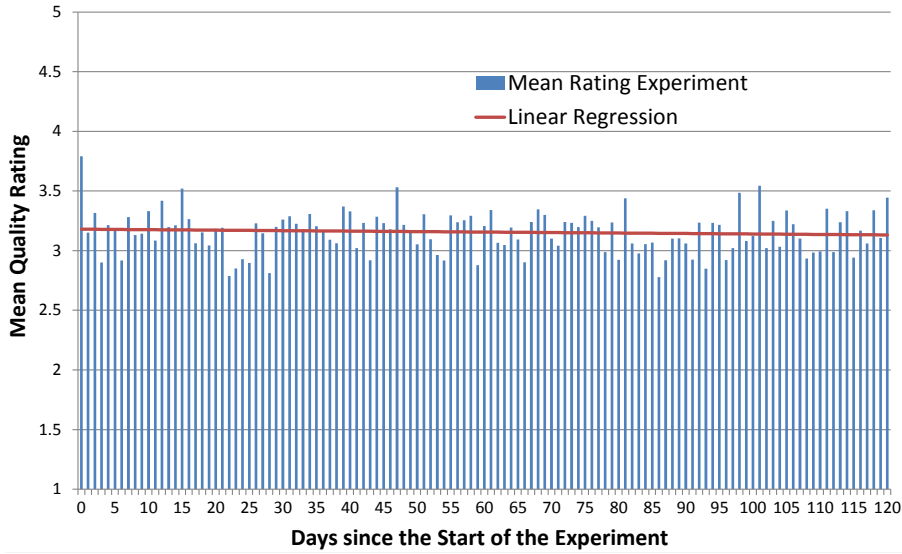


Fig. 8 The evolution of the quality rating over time.

$$\text{MeanQualityRating} = 3.22 - 0.0026 * \text{CallSequenceNumber} \quad (1)$$

This decrease in mean quality rating can be explained by the evolution in the user's familiarity and experience with the service. When users start utilizing a service, they have certain expectations, which influence their judgment of the service. When users utilize the service more often, they become more experienced and familiar with the service, thereby adjusting their expectations. Familiarity with the service might induce higher expectations, based on the users' previous experiences with the service. Higher expectations may in turn lead to a lower quality assessment. In addition, after using a service several times, users might pay more attention to the quality, thereby noticing more artifacts in the audio, and as a result perceiving the quality of the service differently. Important to note is that this trend does not continue during extended use of the service. For intensive users, we notice that the subjective quality stagnates after 120 calls; users are fully familiar with the service and the mean rating remains constant.

9 Conclusions

In this paper, the QoE of a commercial VoIP service is investigated by analyzing the subjective quality ratings and usage patterns of more than thousand actual users of the service in their daily environment without any restrictions. More specifically, this paper focuses on the influence of application and device characteristics on the perceived quality as assessed by the users' ratings.

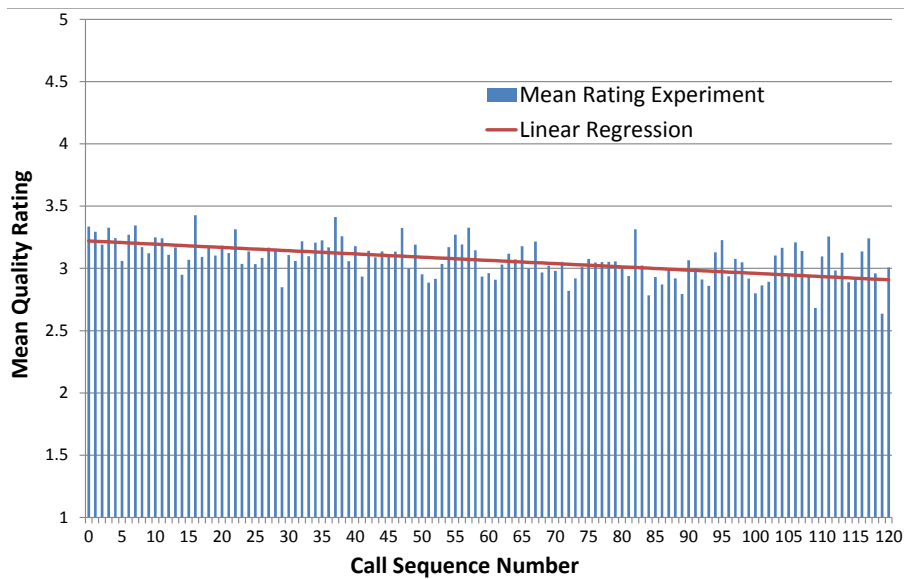


Fig. 9 The evolution of the quality rating over the subsequent voice calls made by the user.


A more advanced audio codec and a recent version of the mobile operating system have a significantly positive effect on the perceived quality. The occurrence of handovers from data to GSM network during the call and the use of EDGE/3G instead of the GSM network of the mobile operator have a significantly negative effect. Still, the analysis of the setting “enable/disable handovers to GSM” showed that enabling this handover process is the best solution to avoid bad quality over a low-throughput network connection. Also the type of phone and mobile platform have an influence on the subjective quality of the VoIP call. Low-end phones induce a lower user experience due to limited hardware resources, which is reflected by a lower subjectively-perceived quality for Android phones of certain brands. The fact that users use the application to receive incoming VoIP calls has no significant influence.

This paper provides an innovative analysis of the relative influence of a large number of objective parameters on the subjective quality by means of a decision tree. This analysis proves that besides the traditional QoS and network related parameters, such as throughput, packet delay, or packet loss, also less obvious parameters, such as application and device characteristics, have a significant influence on the QoE during VoIP calls.

The perceived quality of the VoIP service showed to have an influence on the duration of the call. The mean duration of voice calls that received a low quality rating from the users (1 or 2) is significantly shorter than the mean duration of voice calls that received a fair or high rating (3, 4, or 5). This proves that users’ interaction behavior and usage pattern with a VoIP service is influenced by the subjectively-perceived quality.

To estimate and quantify the user's experience with a service, it is essential to evaluate the perceived quality not at one specific moment in time, but rather over a longer period of time during service usage. This study is the first to investigate the evolution of the perceived quality for a widely-used, mobile, multimedia service. For the studied VoIP service, the perceived quality showed to decrease slightly as the user has utilized the service more often and got more familiar with it. This result proves that the QoE can evolve during the entire use process of the service due to adjusted expectations, previous experiences with the service, and a change in user focus leading to a change in the impact of the quality. These results can be used as a guidance for future, user-centric evaluations of the QoE of mobile multimedia services.

References

1. Definition of Quality of Experience (QoE). Liaison statement, ITU-T, International Telecommunication Union (2007). Ref.: TD 109 rev 2 (PLEN/12)
2. About.com: What Affects Voice Quality in VoIP Calls (2013). Available at <http://voip.about.com/od/voipbasics/a/factorsquality.htm> Accessed on 5 July 2013.
3. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees, first edn. Chapman and Hall/CRC (1984)
4. Cano, M.D., Cerdan, F.: Subjective qoe analysis of voip applications in a wireless campus environment. *Telecommunication Systems* **49**(1), 5–15 (2012). DOI 10.1007/s11235-010-9348-5. URL <http://dx.doi.org/10.1007/s11235-010-9348-5>
5. Cardeal, S., Neves, F., Soares, S., Tavares, F., Assuncao, P.: Arqos : System to monitor qos/qoe in voip. In: EUROCON - International Conference on Computer as a Tool (EUROCON), 2011 IEEE, pp. 1–2 (2011). DOI 10.1109/EUROCON.2011.5929310
6. De Pessemier, T., De Moor, K., Joseph, W., De Marez, L., Martens, L.: Quantifying subjective quality evaluations for mobile video watching in a semi-living lab context. *Broadcasting, IEEE Transactions on* **58**(4), 580–589 (2012). DOI 10.1109/TBC.2012.2199590
7. De Pessemier, T., De Moor, K., Ketykó, I., Joseph, W., De Marez, L., Martens, L.: Investigating the influence of qos on personal evaluation behaviour in a mobile context. *Multimedia Tools and Applications* **57**(2), 335–358 (2012). DOI 10.1007/s11042-010-0712-y. URL <http://dx.doi.org/10.1007/s11042-010-0712-y>
8. Geerts, D., De Moor, K., Ketykó, I., Jacobs, A., Van den Bergh, J., Joseph, W., Martens, L., De Marez, L.: Linking an integrated framework with appropriate methods for measuring qoe. In: Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on, pp. 158–163 (2010). DOI 10.1109/QOMEX.2010.5516292
9. Hofffeld, T., Binzenhfer, A.: Analysis of skype voip traffic in umts: End-to-end qos and qoe measurements. *Computer Networks* **52**(3), 650 – 666 (2008). DOI <http://dx.doi.org/10.1016/j.comnet.2007.10.008>. URL <http://www.sciencedirect.com/science/article/pii/S138912860700299X>
10. Huang, T.Y., Huang, P., Chen, K.T., Wang, P.J.: Could skype be more satisfying? a qoe-centric study of the fec mechanism in an internet-scale voip system. *Network, IEEE* **24**(2), 42–48 (2010). DOI 10.1109/MNET.2010.5430143
11. Imam, I., Michalski, R.: Learning decision trees from decision rules: A method and initial results from a comparative study. *Journal of Intelligent Information Systems* **2**(3), 279–304 (1993). DOI 10.1007/BF00962072. URL <http://dx.doi.org/10.1007/BF00962072>
12. Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.B.: User experience over time: an initial framework. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09, pp. 729–738. ACM, New York, NY, USA (2009). DOI 10.1145/1518701.1518814. URL <http://doi.acm.org/10.1145/1518701.1518814>
13. Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W.: Applied Linear Statistical Models, fifth edn. McGraw-Hill (2005)

14. Magidson, J.: The chaid approach to segmentation modeling: chi-squared automatic interaction detection. In: Bagozzi, Richard P. (ed); *Advanced Methods of Marketing Research*, p. 118159. Blackwell, Oxford, GB (1994)
15. Manousos, M., Apostolacos, S., Grammatikakis, I., Mexis, D., Kagklis, D., Sykas, E.: Voice-quality monitoring and control for voip. *Internet Computing, IEEE* **9**(4), 35–42 (2005). DOI 10.1109/MIC.2005.92
16. Novkov, L., tepnkov, O.: Visualization of trends using radviz. *Journal of Intelligent Information Systems* **37**(3), 355–369 (2011). DOI 10.1007/s10844-011-0157-4. URL <http://dx.doi.org/10.1007/s10844-011-0157-4>
17. Palmieri, F.: Large scale voice over ip experiences on high performance intranets. In: S. Chaudhuri, S. Das, H. Paul, S. Tirthapura (eds.) *Distributed Computing and Networking, Lecture Notes in Computer Science*, vol. 4308, pp. 355–366. Springer Berlin Heidelberg (2006)
18. Rajaraman, S.: Five Stars Dominate Ratings . YouTube Official Blog (2009). Available at <http://youtube-global.blogspot.be/2009/09/five-stars-dominate-ratings.html> Accessed on 15 July 2013.
19. Reiter, U.: Overall perceived audiovisual quality - what people pay attention to. In: *IEEE 15th International Symposium on Consumer Electronics 2011 (ISCE)*, pp. 513–517 (2011). DOI 10.1109/ISCE.2011.5973883
20. Rowe, L.A., Jain, R.: Acm sigmm retreat report on future directions in multimedia research. *ACM Transactions on Multimedia Computing, Communications, and Applications* **1**(1), 3–13 (2005). DOI 10.1145/1047936.1047938. URL <http://doi.acm.org/10.1145/1047936.1047938>
21. Soldani, D., Li, M., Cuny, R.: *QoS and QoE Management in UMTS Cellular Systems*, pp. i–xxvii. John Wiley & Sons, Ltd (2006)
22. Support, B.A.: *The Growing Popularity of Hosted VoIP* (2013). Available at <http://www.advisorysupport.co.uk/the-growing-popularity-of-hosted-voip/> Accessed on 5 July 2013.