A Large-Scale, Long-Term, User-Centric Evaluation of a Commercial Voice-over-IP Application

Toon De Pessemier, Luc Martens, Wout Joseph
iMinds - Ghent University, Dept. of Information Technology
G. Crommenlaan 8 box 201, B-9050 Ghent, Belgium
Tel: +32 9 33 14908, Fax:+32 9 33 14899
Email: {toon.depessemier, luc.martens, wout.joseph}@intec.ugent.be

***Abstract*—To make cheap voice calls, Voice-over-IP (VoIP) services are often used as an alternative to the traditional telephone service providers. Also on the mobile platform, these VoIP services are becoming increasingly popular due to the increased capabilities and connectivity of mobile devices. Understanding the user's usage behavior and quality assessment of the VoIP service is crucial to optimize the Quality of Experience (QoE) and making the service to succeed.**

Whereas multimedia services are often evaluated in a controlled laboratorium setting, with a selected group of test subjects, and during a short evaluation period, this study analyzes the service usage and quality assessments in a real environment, with more than thousand users, and over a period of 120 days. The influence of various parameters (such as audio codec, handovers, platform, and manufacturer of the device) on the subjective quality of the service is validated by analyzing the quality assessments of users, provided after each voice call. The time of the day showed to have a significant influence on the number of calls, the duration, and the subjective quality assessment.

Time-dependent patterns in the users' usage behavior are identified, thereby providing useful information to predict the system load. A regression analysis of the quality assessments over time shows that the perceived quality gradually decreases as users have utilized the service more, and get more familiar with it. In contrast, the mean duration of the calls increases as users get more familiar with it. This research is important in view of associating technical and contextual parameters with the QoE during service usage.

Keywords*—*Subjective evaluation techniques, Audio technology, VoIP, Mobile.

I. Introduction

Voice-over-IP (VoIP) services, such as Skype[1], Google Hangouts[2], and ooVoo[3], enable users to make free or cheap voice calls using the Internet and are becoming increasingly popular. Many VoIP services extend their applicability by offering a mobile application, giving users the opportunity to make VoIP calls with their tablet or smartphone. These VoIP calls, using the data connection of the mobile device, are an alternative for the traditional communication standard, GSM (Global System for Mobile Communications). A frequent criticism of VoIP applications is the poor quality. Although there has been much improvement due to better audio codecs and mobile data networks, people are still very demanding regarding the quality of VoIP calls because they are used for years to the high quality of landline telephones.

In view of a high quality and reliable communication over a best-effort network, end-to-end Quality of Service (QoS) management is becoming a challenge [1] for VoIP. Nevertheless, these purely technical parameters do not always reflect the perceived quality of a service correctly, since they only take into account network related aspects and neglect device characteristics, context parameters, and user aspects.

The Quality of Experience (QoE), or how a service is really perceived by the user, gives a more veracious understanding of the true quality [2], [3]. Assessing the QoE by asking users their opinion regarding the subjectively-perceived quality, could be costly and time consuming. Nevertheless, this kind of subjective quality measurement with actual test users considers how users perceive and experience a multimedia communication service as a whole [4], and it relates to the user-perceived experience directly rather than to the implied impact of QoS. As a result, the QoE as evaluated by the user, is considered as a more important metric than QoS [5]. 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' [6]. For new multimedia and communication services, identifying, understanding, and quantifying the most determining aspects of the QoE is of vital importance.

In many existing studies, the quality of VoIP solutions is analyzed using a private network [7], 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 controlled environment covering a limited area, such as a university campus [8], thereby limiting the freedom of the test subjects with the risk of obtaining results that are not generally applicable.

In addition, the number of test subjects participating in a (mobile) QoE experiment is often limited to a few dozen due to budget and time constraints [3]. In contrast, in this paper we investigate the usage behavior and quality assessments of more than thousand users of a VoIP service, making voice calls in their daily environment without any location, time, or usage constraint. More specifically, this analysis is based on data logging of real customers of a commercial VoIP application, developed and managed by a Belgian mobile network operator.

[1]http://www.skype.com
[2]http://www.google.com/hangouts
[3]http://www.oovoo.com

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.

Furthermore, in traditional user experiments, the QoE is mostly evaluated by taking a snapshot of the subjective experience of the user at one moment in time during the complete use process of the service. The dynamic nature of the user's usage behavior and QoE is often ignored in research experiments, due to the lack of any time-related data in the analysis. Nevertheless, the user's usage behavior and QoE with a service continuously change over time, and are influenced by his or her prior expectations about the service [9]. Before people start using a particular product or service, they tend to already have some kind of preconception influencing their expectations [10].

After the adoption of the product or service, when the user is utilizing it more or less regularly, the actual use process evolves. As the user gains more experience with the usage, the user's familiarity with the product or service increases, and this has an impact on how it is being used [10]. As a result, the QoE should not be evaluated at a single point in time, but rather over a continuous period during the use process. This paper is the first to monitor trends in the user's usage behavior and analyze the QoE with a VoIP service over a longer period of time (four months).

II. The VoIP Service

The objective of this paper is to investigate a commercial VoIP service that is available on the mobile platform from a user point of view. The focus is on the user's quality assessment of the voice call and the user's usage behavior with the service. The investigated VoIP service is 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).

The added value of the investigated VoIP service compared to Skype is its transfer option that switches the voice call from the available data network (IP-based cellular network or WiFi) to the network of the user's mobile operator (GSM network) during the same call. In case of insufficient throughput or loss of coverage on the data network, the mobile operator automatically takes over the VoIP call.

Users are informed about the transfer of the voice call from the data network to the GSM network by a short beep sound during the voice call. Voice calls that are transferred to the GSM network of the mobile operator because of technical issues with the data network, are not transferred again to the data network (not even if the data connection is sufficiently recovered). So only one handover between data network and GSM network is allowed in order to limit the possible disruptions introduced by switching the network of the voice call.

All data about the VoIP service usage are internally stored and continuously monitored for customer services and billing purposes. These data can be used to identify service usage patterns and analyze the parameters that influence the QoE of the service. The data set used for this analysis contains the details of all voice calls made by the customers of the VoIP service over a period of 120 days. During this period, 127,826 voice calls are made by 1050 users of the VoIP service. Users of the VoIP service have the opportunity to evaluate the quality after each voice call using a 5-point scale rating mechanism, thereby providing a subjective evaluation of their QoE with the service. These users, not aware of this study, were not biased in any way during service usage or quality evaluation.

The tracking of users resulted in a data set consisting of objective technical parameters (such as the used audio codec, and the presence of handovers from data to GSM), contextual parameters (such as the platform used to make the call, the manufacturer of the user's device, and timestamps indicating the start and end time of the call), and a subjective quality assessment of the user (a 5-point scale rating of the quality of the call). The resulting data set allows on the one hand to investigate the influence of technical and contextual parameters on the user's usage behavior and QoE, and on the other hand, its time frame makes it possible to analyze trends over time.

III. Influential Parameters

In order to quantify and optimize the QoE of a service, identifying the parameters that most influence the quality assessment is essential. Table I lists the mean value of the quality assessments, as expressed by the users of the VoIP service, for different technical parameters and device characteristics. For each case, the table indicates the value of the parameter (1st column), the number of samples with this value (2nd column), and the mean quality assessment obtained for the samples with this value (3rd column). For all parameters in Table I, significant differences are found for the various options according to a statistical T-test on the level of 0.01.

For transmitting the sound of the voice call over the network, two different audio codecs are used during the evaluation period: GSM/8000 (8kHz mono) and GSM/32000 (32kHz mono). As expected, a more advanced audio codec results in a higher assessment for the quality. The difference in mean quality assessment for the two versions of the audio codec is 0.28.

In case of loss of coverage or insufficient throughput on the available data network (IP-based network), the service can fall back on the network of the user's mobile operator (GSM network). For most voice calls, the data network is stable, and such a data-to-GSM handover is not necessary. Since a data-to-GSM handover can introduce distortions or interruptions, calls in which such an handover occurs are characterized by a lower mean quality assessment than calls without handover. The difference in mean quality assessment between these two cases is 0.22 on the rating scale.

Besides technical characteristics, also contextual characteristics, such as the platform and manufacturer of the user's device, can have an influence on the user's quality assessment, which is a subjective measure. The VoIP service is available on the iOS platform (for iPhones) as well as on Android (for a wide variety of phones). In terms of the number of voice calls, the majority of the logged voice calls (73.6%) is made on the iOS platform. In terms of quality assessment, the comparison between the Android and iOS platform reveals that the mean quality assessment on the Android platform is 0.32 lower than

Audio Codec	#Samples	Mean Quality Assessment
GSM/8000	3276	2.75
GSM/32000	4745	3.03
Data-to-GSM Handover	**#Samples**	**Mean Quality Assessment**
Yes	336	2.93
No	30048	3.15
Platform	**#Samples**	**Mean Quality Assessment**
Android	8021	2.91
iOS	22363	3.23
Manufacturer	**#Samples**	**Mean Quality Assessment**
LGE	31	2.65
SEMC	888	2.69
Samsung	5873	2.85
HTC	877	3.22
Apple	22363	3.23

TABLE I. THE DIFFERENCES IN MEAN QUALITY ASSESSMENT OF THE VOICE CALLS, OBTAINED FOR DIFFERENT TECHNICAL PARAMETERS AND DEVICE CHARACTERISTICS.

the mean quality assessment achieved on iOS. Noteworthy is that the absolute difference in mean quality assessment for different platforms is higher than the absolute difference in mean quality assessment for different audio codecs or the difference induced by a handover.

Also regarding the manufacturer of the device, significant differences in mean quality assessment are obtained. The subset of calls made using an Apple device has obtained the highest mean quality assessment. Approximately the same mean quality assessment was achieved by using phones of HTC. No significant difference was found between these two manufacturers in terms of the mean quality assessment for the VoIP service. On the other hand, the mean quality assessment achieved by using phones produced by Samsung is significantly lower according to statistical T-tests: a difference of 0.38 with phones produced by Apple, and a difference of 0.37 with phones of HTC. The lowest quality assessments are achieved by using phones produced by SEMC and LGE. The mean quality assessment achieved by phones of SEMC and LGE is lagging behind with a significant difference of respectively 0.53 and 0.57 with phones produced by HTC. Compared to Apple's phones, which achieved the highest mean quality assessment, the difference is 0.54 and 0.58 for respectively SEMC and LGE.

This difference in quality assessment can be explained by the types of phones produced by the various manufacturers. Whereas HTC and Apple are specialized in the production of mid-range to high-end phones, Samsung sells a large variety of phones: mid-range, high-end as well as a lot of low-end phones. The devices of SEMC and LG that were used in this experiment are mainly low-end phones. These cheap, low-end phones typically have a restricted number of features, limited hardware resources, and limited capabilities, which might deteriorate the user's QoE with resource demanding applications such as a VoIP service. So, the general QoE of users with their device can prevail and be reflected in the subjective quality assessment for an individual session of a mobile service. This emphasizes the subjective character of the QoE as reflected in the quality assessment that is specified after each voice call.

IV. PATTERNS IN USAGE BEHAVIOR

The users' usage behavior with the VoIP service is analyzed by monitoring the number of calls made by the users, the duration of the calls, and the quality assessment of the calls as provided by the users.

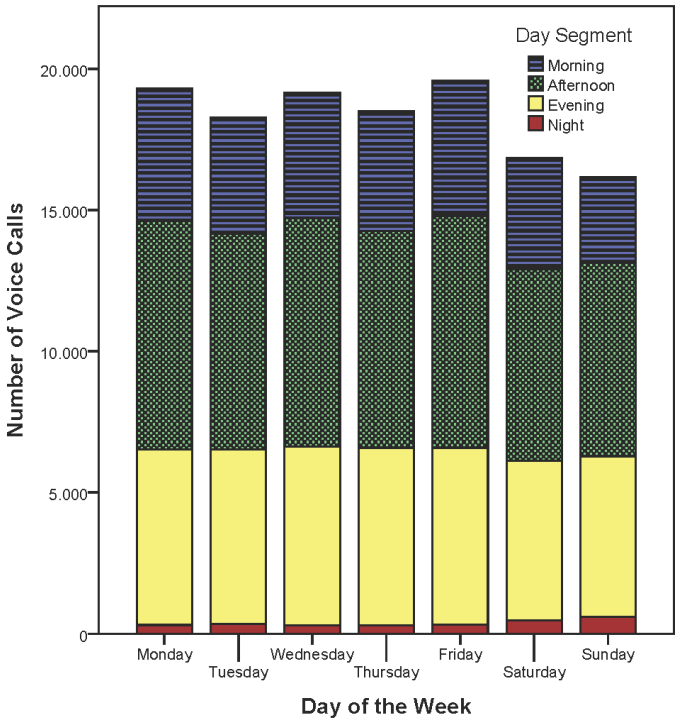

Fig. 1. Histogram of the number of calls, partitioned according to the segment of the day (morning, afternoon, evening or night) and the day of the week

A. Number of Calls

The number of calls that users make using the VoIP service is analyzed by partitioning the calls according to the day of the week and the time of day when the call is made. Figure 1 shows a histogram of the total number of calls, made by all users of the VoIP service during the entire evaluation period. For each day of the week, the histogram shows the number of calls that were made and makes a subdivision according to the time of the day. The histogram partitions the day into four periods of six hours: morning, from 6:00 to 12:00, afternoon, from 12:00 to 18:00, evening, from 18:00 to 0:00, and night, from 0:00 to 6:00.

Most calls, $19,584$ in total, are made on Fridays, but the difference with other weekdays is small. For weekdays (Monday to Friday), the total number of calls is on average $18,963$. Given that the evaluation ran for 17 weeks, this comes down to an average of 1115 calls for a single day in the week. To compare, in total $16,840$ calls are made on Saturdays during the evaluation period; or on average 991 for a single Saturday. The total number of calls on Sundays is $16,173$, which amounts to 951 calls for an average Sunday. So, the average number of calls made on Saturday is 11.1% lower than on weekdays. On Sunday, 14.7% less calls are made compared to an average weekday. This significant difference in the number of calls per day demonstrates that users are utilizing the VoIP service more often on weekdays than on weekend days.

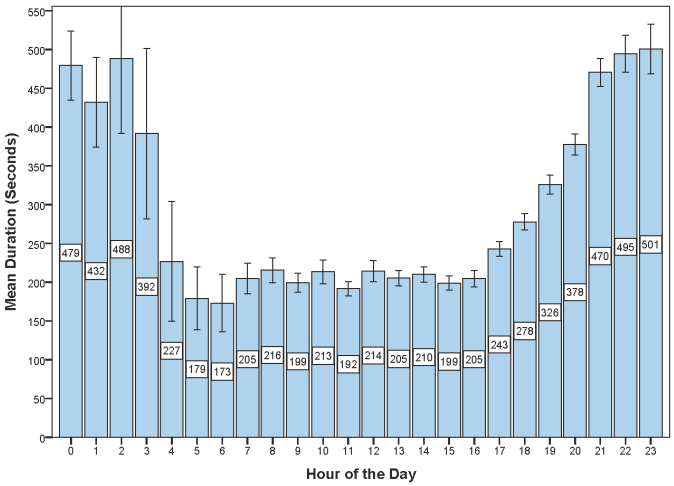

Fig. 2. The mean duration of the calls per hour of the day.

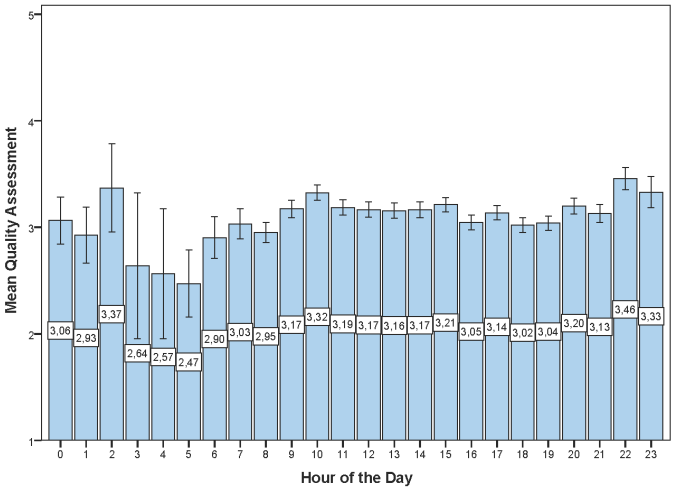

Fig. 3. The mean quality assessment of the calls as provided by the users per hour of the day.

B. Call Duration

The call duration is analyzed by partitioning the calls according to the hour of the day during which the call is made. Differences in call duration per day of the week showed to be limited and are therefore not shown in this paper. Figure 1 shows the mean call duration, together with the 95% confidence interval of this mean value, for each hour of the day. E.g., the bar corresponding to the value 0 on the X-axis indicates the mean duration of calls made between 0:00 and 0:59. A pattern in the call duration is visible in Figure 1, reflecting the typical calling behavior of people. During the early hours of the day (from 5:00 to 7:00), calls have the shortest duration (mean duration of 176 seconds). But also during the morning and the afternoon (from 7:00 to 17:00) the call duration is quite short (mean duration of 205 seconds). So, during daytime, people keep the conversations short. They may be busy with their daily activities and have limited time to chat. After office hours as people come home (from 17:00 to 0:00), the mean call duration increases. The mean duration during the evening is 384 seconds, with a peak of 500 seconds from 23:00 to 0:00. During the evening, people typically have more time to call friends or family and chat for a little longer. During the first hours of the night (from 0:00 to 3:00), the calls are still quite long (with a mean value of 467 seconds). Later on during the night (3:00 to 5:00), as more people are asleep, the mean call duration decreases. As expected, most voice calls are brief during these hours of the night.

C. Quality Assessment

As with the call duration, variations in the quality assessment are analyzed by partitioning the calls according to the hour of the day during which the call is made. Figure 3 shows the mean quality assessment as provided by the users, together with the 95% confidence interval of this mean value, for each hour of the day. The bar graph shows significant differences in mean quality assessment for different hours of the day. During daytime and the early hours of the evening (from 9:00 to 22:00), the variation is limited with a mean quality assessment between 3.02 and 3.32. Late in the evening, a small increase in quality assessment can be witnessed with a peak of 3.46 at 10:00.

The difference with the quality assessments obtained during the night is noteworthy. From 3:00 to 6:00, the mean quality assessments are 2.64, 2.57, 2.47 for the subsequent hours, significantly lower than during daytime. The limited number of calls during the night (Figure 1) results in a low bandwidth requirement on the cellular data network, which is often shared among the users. As a result, the technical conditions of the network are optimal during the night for a VoIP call. So, in a purely objective evaluation of the quality of the VoIP service, a higher assessment would be expected for calls during the night. In contrast, a lower subjective assessment is obtained for night calls, because of non-technical influences such as the context or the mood of the user. Various aspects might have a negative influence on the user's subjective assessment of the quality, e.g., tiredness or frustration due to be awake, an inappropriate timing of the call, etc. Because of the small number of calls during the night, the confidence intervals are larger from 3:00 to 6:00 than during daytime.

V. Evolution over Time

As users utilize the VoIP service for a longer period of time, they gain more experience with the usage, they become more familiar with it, and their expectations about the quality may change based on prior experiences. As a result, the user's usage behavior and QoE with a service should not be evaluated by a short-term experiment at a single point in time, but rather over a continuous period during the use process. In this paper, the usage of the VoIP service is tracked during a period of 120 days, allowing to analyze possible evolutions in the QoE (Section V-A) and usage behavior (Section V-B) of the users. Such a long-term evaluation with a large number of users, in a realistic environment without the constraints of a laboratory experiment, has never been performed up to now for a mobile VoIP service.

A. Trends in the Quality Assessment

To analyze the evolution of the usage behavior and QoE while users get more experienced with the VoIP service, the data of the voice calls are partitioned according to the

number of calls that the user has already made. The first partition contains the data of the first call of every user of the VoIP service. The data of the second call of every user are collected in the second partition. For this group of voice calls, users already have the experience of one previous VoIP call. Analogously, partition N contains the data of the N-th voice call of every user, or in other words, the data of calls made by users who have made $N-1$ voice calls in the past, which characterizes their experience with the service.

For each of the partitions, Figure 4 shows the mean quality assessment of the voice calls made by all users of the VoIP service. The number of calls made by the user is an indicator for the user's familiarity and experience with the service. Hence, Figure 4 illustrates the evolution of the mean quality assessment, as an indicator for the QoE, while users get more familiar and experienced with the VoIP service.

Because of the large user basis, the fluctuations of the mean quality assessment over the different partitions are limited. The standard deviation of the mean quality assessment is 0.14. But more important is the considerable decrease of the mean quality assessment, as the user has made more voice calls. This trend is visualized in Figure 4 by the line of best fit, which is the result of a linear regression analysis.

This linear regression analysis confirms the significance of the decrease in quality assessment as users get more familiar and experienced with the service. The mean difference in quality assessment between new users without any experience and users who have already made a hundred calls is 0.26, or more than a quarter of a point on a 5-point rating scale.

Since the technical quality of the VoIP service was not altered, the significant decrease of the subjective quality assessment must be due to user aspects. When users utilize the service more often, they become more familiar and experienced with it, thereby adjusting their expectations. Familiarity and experience with the service might induce higher expectations, which are fed by previous VoIP sessions, which users regard as a reference point. Higher expectations may in turn lead to a lower quality assessment. Moreover, after using the VoIP service several times, users may become more demanding and pay more attention to the quality, thereby noticing more artifacts in the audio, and as a result have a lower perception of the quality. Further analysis showed that this decreasing trend stagnates during extended usage of the service. The mean quality assessment becomes constant after about 120 calls. Users are fully familiar with the service and additional voice calls have no significant influence on their quality assessment.

B. Trends in the Call Duration

An interesting characteristic of the user's usage behavior with the VoIP service is the mean call duration. The call duration is measured as the time between the moment when the person being called answers the call and the time when one of the two persons hangs up the phone. To analyze the evolution of the call duration while users get more experienced with the VoIP service, the data of the voice calls are partitioned according to the number of calls that the user has already made, just as in Section V-A.

For each of the partitions, Figure 5 shows the mean duration of the voice calls made by all users of the VoIP service.

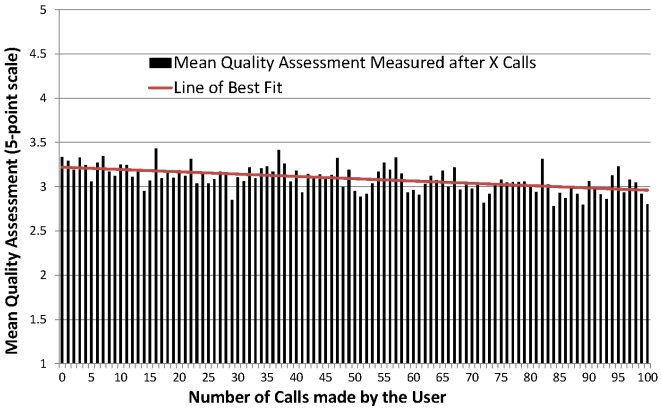

Fig. 4. The evolution of the mean quality assessment over time, together with the line of best fit.

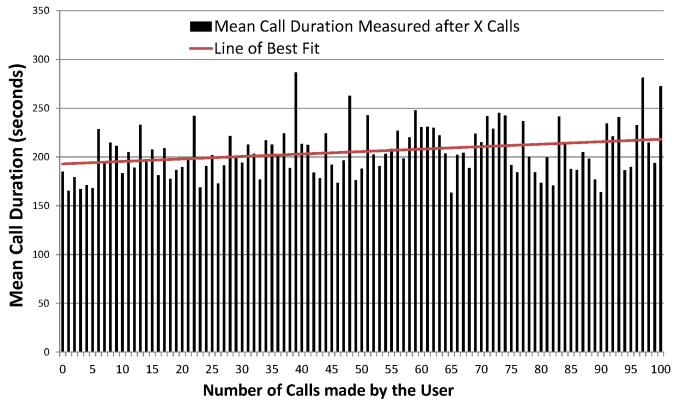

Fig. 5. The evolution of the mean call duration over time, together with the line of best fit.

The standard deviation of the call duration over the different partitions is limited to 25.35 but a trend in the duration can be witnessed. Through a linear regression analysis, the line of best fit is calculated, which visualizes the increasing call duration in Figure 5. The line of best fit, taking into account the variation of the duration of individual calls, shows that the mean duration of the first call of users is 193 seconds. For the users' hundredth call, the regression indicates a mean duration of 218 seconds, or 25 seconds longer than the duration of users' first call.

The call duration can be considered as an implicit feedback mechanism. A longer duration points to an extended use of the VoIP service, which might be an indication of a good user experience. As users have made more VoIP calls in the past and consequently are more familiar with the service, they will make longer calls. A possible explanation is that users start considering the VoIP service less as a nice-to-have gadget, but more as a valuable alternative for the traditional GSM network of their mobile operator to make longer calls.

So over time, users' quality assessment decreases (Section V-A) but the mean duration of their voice calls increases. Just as in Section V-A, this trend does not continue during a prolonged use of the service. As expected, the mean call duration does not persistently increase, but stabilizes when users are fully familiar and experienced with the service.

VI. Conclusion

In this paper, a commercial VoIP service is investigated by analyzing the subjective quality assessments and usage patterns of more than thousand actual users of the service in their daily environment without any restrictions. More specifically, the focus of this paper is on the influence of technical characteristics (such as audio codec and handovers) and device characteristics (such as platform and manufacturer) on the perceived quality as assessed by the users. The results showed that differences in quality assessment due to device characteristics can be higher than difference due to technical characteristics.

Regarding the usage behavior, the results showed that the afternoon is the most popular time for using the VoIP service, followed by the evening. Furthermore, on a weekend day 12.9% less calls are made compared to week days. The time has a significant influence on the duration of the call. On average, calls have the shortest duration during the morning (around 3 minutes), but also calls made during daytime are quire short (around 3.5 minutes). In contrast, the mean call duration during the evening is 6.5 minutes and almost 8 minutes during the night. This confirms the expected influence of contextual parameters, such as time, on the users' usage behavior with online services such as VoIP.

Significant differences in the mean quality assessment are obtained during different hours of the day, without any technical cause. Subjective assessments for the quality showed to be lower during the night, just because of the timing of the call. The real reason why users provide a lower quality assessment during the night is an interesting topic for future research. The significant correlation between time and quality assessment demonstrates the influence of non-technical parameters and user aspects on the Quality of Experience.

In addition, this research emphasizes the importance of analyzing user behavior and subjective evaluation of an online service over a longer period of time. As users have utilized the service more, and get more familiar with it, expectations and perceptions may change. For the VoIP service, an analysis of the data over time showed a significant decrease in quality assessment, but a significant increase of the call duration.

References

[1] M. Manousos, S. Apostolacos, I. Grammatikakis, D. Mexis, D. Kagklis, and E. Sykas, "Voice-quality monitoring and control for voip," *Internet Computing, IEEE*, vol. 9, no. 4, pp. 35–42, 2005.

[2] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying subjective quality evaluations for mobile video watching in a semi-living lab context," *Broadcasting, IEEE Transactions on*, vol. 58, no. 4, pp. 580–589, 2012.

[3] T. De Pessemier, K. De Moor, I. Ketykó, W. Joseph, L. De Marez, and L. Martens, "Investigating the influence of qos on personal evaluation behaviour in a mobile context," *Multimedia Tools and Applications*, vol. 57, no. 2, pp. 335–358, 2012. [Online]. Available: http://dx.doi.org/10.1007/s11042-010-0712-y

[4] U. Reiter, "Overall perceived audiovisual quality - what people pay attention to," in *IEEE 15th International Symposium on Consumer Electronics 2011 (ISCE)*, June 2011, pp. 513–517.

[5] L. A. Rowe and R. Jain, "Acm sigmm retreat report on future directions in multimedia research," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 1, no. 1, pp. 3–13, 2005. [Online]. Available: http://doi.acm.org/10.1145/1047936.1047938

[6] International Telecommunication Union, "Definition of Quality of Experience (QoE)," ITU-T, International Telecommunication Union, Liaison Statement, 2007, ref.: TD 109 rev 2 (PLEN/12).

[7] F. Palmieri, "Large scale voice over ip experiences on high performance intranets," in *Distributed Computing and Networking*, ser. Lecture Notes in Computer Science, S. Chaudhuri, S. Das, H. Paul, and S. Tirthapura, Eds. Springer Berlin Heidelberg, 2006, vol. 4308, pp. 355–366.

[8] M.-D. Cano and F. Cerdan, "Subjective qoe analysis of voip applications in a wireless campus environment," *Telecommunication Systems*, vol. 49, no. 1, pp. 5–15, 2012. [Online]. Available: http://dx.doi.org/10.1007/s11235-010-9348-5

[9] D. Geerts, K. De Moor, I. Ketykó, A. Jacobs, J. Van den Bergh, W. Joseph, L. Martens, and L. De Marez, "Linking an integrated framework with appropriate methods for measuring qoe," in *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, 2010, pp. 158–163.

[10] E. Karapanos, J. Zimmerman, J. Forlizzi, and J.-B. Martens, "User experience over time: an initial framework," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 729–738. [Online]. Available: http://doi.acm.org/10.1145/1518701.1518814