Statistical Analysis of a Subjective QoE Assessment for VVoIP Applications

Maria-Dolores Cano, Fernando Cerdan, and Sergio Almagro

A successful deployment of multimedia applications over wireless environments entails improving the quality of service (QoS), not only from a technical point of view, but also considering the quality of experience (QoE) from the final user's perception. Although objective QoE measure models avoid the difficulties of subjective surveys, subjective QoE assessments are essential to understand the way users evaluate the QoS. In this work, we study the effect of a wide range of parameters on the QoE of VVoIP applications in a real wireless scenario. Through a complete statistical analysis of users' ratings, we identify the following facts. Although the use of VVoIP in wireless networks does not yet represent an advantage for users, there are great expectations for all applications under study, and with greater popularity comes higher expectations. It is easier for respondents to identify good behavior than poor behavior. Whereas the respondents' frequency of Internet use does not impact on the scores, respondents' gender does. Finally, the most determining parameters of quality from a user's perspective were instability, video quality, voice distortion, usefulness, and graphical interface.

Keywords: QoE, QoS, VoIP, VVoIP, wireless networks.

Manuscript received Nov. 6, 2009; revised Aug. 3, 2010; accepted Sept. 16, 2010.

This work was supported by the project grant CON-PARTE-1 (TEC2007-67966-01/TCM, Spain), being also developed in the framework of "Programa de Ayudas a Grupos de Excelencia de la Región de Murcia, de la Fundación Séneca, Agencia de Ciencia y Tecnología de la RM (Plan Regional de Ciencia y Tecnología 2007/2010)."

Maria-Dolores Cano (phone: +34 968325953, email: mdolores.cano@upct.es), Fernando Cerdan (email: fernando.cerdan@upct.es), and Sergio Almagro (email: sergio.almagro@upct.es) are with the Department of Information and Communication Technologies, Technical University of Cartagena, Cartagena, Spain.

doi:10.4218/etrij.10.0109.0645

I. Introduction

So far, most research about quality of service (QoS) has focused on measuring the quality level offered by a service in terms of objective parameters. The ability to control these parameters allows guaranteeing the service level agreements (SLA) between providers and users. However, the final user, as an integral part of the service chain, perceives quality in a subjective way, not necessarily expressed in usual technical terms. The International Telecommunications Union, Telecommunication Standardization Sector (ITU-T) defines quality of experience (QoE) as the overall acceptability of an application or service, as perceived subjectively by the end-user [1]. To set the customer requirements for multimedia services according to the customers themselves, [2] recommends that performance should be expressed in a way that takes into account all aspects of the service from the customer's point of view, focuses on user-perceivable effects rather than their causes within the network, and is independent of the specific network architecture or technology.

QoE includes the complete end-to-end system effects, and overall acceptability may be influenced by users' expectations and context. Observe also that service acceptability does not always match with service quality. A service with a high quality level from an objective point of view (for example, delay, jitter, or noise) might be considered too expensive or have a graphical interface considered to be unattractive to users, for instance to customer segments highly influenced by marketing or brands. Consequently, it is essential to know a user's experience about the quality of the service, to improve it as much as possible within the contracted SLA, and hence, to increase the number of potential customers.

The ongoing work being developed by standardization organizations such as the ITU-T, the American Standards National Institute, the IEEE Standards Association, or the European Telecommunications Standards Institute reveals the importance of QoE. Whereas several models for voice quality prediction have been developed for wired environments and narrow band speech [3], [4], work is in progress for models for wideband speech and multimedia as a challenge for next generation networks [5], [6], and wireless environments have not even been addressed yet from a QoE perspective. For instance, in the work regarding the E-model [4], the author only mentions that an advantage factor could be taken into account as an additional factor in the model so that "the fact that customers may accept some decrease in quality for access advantage, e.g., mobility or connection into hard-to-reach regions" can be considered. However, the values of the advantage factor are provisional since they have not been confirmed by subjective investigations to date [4].

Several works that have studied the performance of multimedia applications from a subjective or objective QoE perspective are [7]-[14]. In [7], the authors study the effects of vertical handovers on speech quality in next generation networks. Within a mobile IPv4 testbed, experiments were carried out in listening-only mode, and participants rated the overall quality of the voice over internet protocol (VoIP samples). They found that packet loss rate and audio bandwidth are the most important network characteristics impacting on a user's QoE. Additionally, the authors conclude that current parametric quality prediction models like the E-model need some improvement in order to incorporate the special properties of wireless communications. Similarly, Perala and Varela [8] studied the performance of Skype, in terms of listening quality, when used in a converging networks context. They found that VoIP quality in a wireless context was lower and much more variable than its circuit-switched counterpart. In [9], the authors presented a complete comparative study of QoS and QoE in video and VoIP (VVoIP) applications. From the results, authors found that there was a significant divergence: whereas the objective parameters that were analyzed showed a good performance, users did not rate their QoE with the same level. This perception shows the need of improving QoE models. In [10], a multimedia quality integration function for QoE planning of videophone services was proposed. They observed that video display size has no effect on users' QoE scores, but video quality was the most important factor. Malfait and others [11] present the ITU-T Rec. P.563, a method for nonintrusive assessment of speech. However, they state that an improved wideband model will be necessary due to the wider adoption of VoIP. Other weaknesses of the classical measurement methods are related to the different properties of VVoIP as indicated by [12], and new methods more appropriate for VVoIP applications are appearing to objectively estimate QoE from QoS parameters [12]-[14]. However, they still lack a high correlation with subjective real scores.

In this work, we present a statistical analysis of QoE performance of five popular VVoIP applications. From a survey poll carried out with more than 80 respondents, we studied the effect of different parameters on users' QoE. Tests were conducted in the Wireless Local Area Network (WLAN) of one of the campus libraries at the Technical University of Cartagena (Spain). We decided to evaluate the VVoIP applications over a wireless environment as a first step to understand more complex scenarios, which usually involve several technologies (wired and wireless). Through the statistical analysis of users' MOS scores, we will discuss the performance of each application, identifying the importance of each parameter on the final QoE. Regarding previous works from related literature, our contribution lies on the inclusion of technical and non-technical parameters, and the identification of parameters that could affect users' willingness to use these types of services in a wireless environment; all of it is based on a subjective quality assessment carried out in conversational mode under real network circumstances.

The rest of the paper is organized as follows. Section II presents the applications that users evaluate. Section III describes the procedure and the scenario to carry out the assessment. Results are discussed in section IV. The paper ends with a conclusion in section V.

II. VVoIP Applications

In this work, we focus on five popular VVoIP applications: Skype, Gizmo5, Damaka, ooVoo, and vBuzzer (only VoIP).

Skype [15] stands as one of the most popular VoIP applications, and it accounts for more than 4.4% of total VoIP traffic [16]. It offers phone calls, video calls, file transfer, and chat services. The main difference between most VoIP services and Skype is that the latter operates on a proprietary peer to peer (P2P) model. Gizmo5 [17] has clearly been Skype's competitor during the last years. It supports phone and video calls, instant messaging, two-way text messaging, group chat, voice mail, and file transfer. Gizmo5 is defined as a standards-based consumer calling service because it uses the Session Initiation Protocol (SIP), which notably increases its interoperability. It can also employ some proprietary codecs, such as iSAC [18].

As with previous applications, ooVoo [19] offers a variety of services ranging from video calls for up to six people to video chats, including a video message service. Two-caller (video or phone) calls work in a P2P fashion, whereas ooVoo servers are

employed for video calls with more than two participants. Differing from Gizmo5 and Damaka, ooVoo is not SIP-compliant, but uses proprietary protocols. Damaka [20] is a SIP-based IP telephony application with patent pending softswitching capabilities in a P2P environment. Some of the features provided are secure phone and video calls, instant messaging, file transfer, desktop and whiteboard sharing, and voice mail for both PC and mobile devices. As in Gizmo5, Damaka's interoperability is achieved thanks to the open standards compliance (it uses SIP). According to its developers, Damaka supports the same codecs as Skype does. Finally, vBuzzer [21] is a VoIP application for fixed or mobile devices that also offers an instant messenger service. It is SIP compliant and gives support to narrow band codecs, but it does not provide video calls.

In most cases, these applications support both standard codecs (such as iLBC [22]) and proprietary codecs (such as iSAC). Regarding video, Skype and ooVoo employ TrueMotion VP7 codec, a proprietary solution of [23], which provides a variable bitrate flow with a minimum bandwidth of 20 kbps. Gizmo5 uses Xvid [24], a MPEG4 video codec for PC. Damaka's video preferences are not publicly available. A more detailed description of the first four applications can be found in [9].

It is clear that the choice of a codec sets the limit of the best achievable quality if all network and environmental issues are disregarded. Speech codecs determine bit-rate and sampling rate, and can include features that influence delay and packet loss performance. In this study, we do not modify any application codecs' configuration. Also, if a codec is set by default or the application allows a dynamic codec selection, we have not modified the application configuration to force the use of a particular codec. Thus, the applications are evaluated as they would normally behave for common users.

III. QoE Evaluation Procedure

In this section, we describe the process for the QoE evaluation. First, we depict the participants who performed the video calls and answered the questionnaires. Then, we describe the real scenario where tests were conducted, and the methodology to obtain QoE scores.

1. Respondents

In order to empirically investigate the QoE of the five VVoIP applications, a questionnaire was administered to a sample of 85 engineering students from the Technical University of Cartagena (Spain) in September 2008. Of the respondents, 67% were male and 33% were female. Ages ranged from 19 to 28 years old, with an average of 22.4 years old and a standard deviation of 1.9 years. 62.4% of the respondents did not know any of these applications, 27.1% have heard about them, and only 10.5% had used Skype, Gizmo5, or other VoIP program at least once (but none of them had used ooVoo, Damaka, or vBuzzer before). In contrast, 64.7% of the respondents used the Internet daily, 24.7% frequently but not every day, and only 10.6% occasionally.

We followed recommendations [25] regarding the minimum number of respondents required for multimedia applications polling, as well as the need of not having previous experience in the area.

2. Testing Conditions

We followed recommendations regarding testing conditions for subjective QoE evaluations [25]-[27]. For more details about testing, refer to [9], where the same authors present the testing conditions of a QoS-QoE comparative study.

To be in compliance with the aforementioned testing conditions, the survey was conducted in the main library of the Technical University of Cartagena, using its 802.11g WLAN. A respondent was asked to make a call to another respondent located in a different room, but both within the same WLAN coverage. Coverage was excellent at endpoints and the environment was always quiet. We used 1.3 megapixel cameras with support for light variations and automatic white balance. Respondents were sat down in front of the camera at a distance of 30 cm to 40 cm. Calls were established with laptops whose screen resolution was 1,280×800 with 32-bit colors at a refresh rate of 60 Hz. Graphic cards were Mobile Intel(R) 945GM Express Chipset Family. Nevertheless, the video call image was shown in a small window (320×240) during the calls. Microphones were placed at 5 cm from the speaker's mouth. Cable headphones were chosen due to their better performance in terms of environmental noise, although noise levels were minimal.

Tests were conducted from Monday to Friday at the same time (from 9 am to 3 pm). In this time slot, the use of the WLAN is homogeneous as verified by the logs of the IT service of our university (see Table 1). In this way, we guaranteed that there was cross traffic but no network congestion.

3. Methodology

In this work, we use the absolute category rating (ACR) and the degradation category rating (DCR) methods recommended by ITU-T [25]-[27] in order to test the quality of the video and voice calls. The ACR is used to classify positive quality parameters (for example, video quality and usefulness) with

Table 1. Average downstream/upstream traffic at the access point of the WLAN used for the tests during the academic year 2008/2009.

	Downstream (bytes·10^3/s)		Upstream (bytes·10^3/s)	
	Mean	Standard deviation	Mean	Standard deviation
Mon	633.5	292.2	28.3	11.2
Tue	719.5	330.7	32.3	13.0
Wed	607.6	290.9	27.4	12.6
Thu	541.7	327.6	31.5	13.2
Fri	480.5	178.2	25.8	7.9

Table 2. ACR and DCR 5-point scales.

Score	ACR	DCR
5	Excellent	Very annoying
4	Good	Annoying
3	Fair	Slightly annoying
2	Poor	Perceptible but not annoying
1	Bad	Imperceptible

a 5-point scale that ranges from bad to excellent, as shown in Table 2. On the other hand, the DCR distinguishes among negative quality parameters using an annoyance 5-point scale that varies from imperceptible to very annoying (see Table 2).

Two subjects made five video calls each, one for each VVoIP application under study. During each call, subjects talked for 120 seconds. Afterwards, the respondents answered the questionnaire about the VVoIP applications used. We collected a total amount of 425 questionnaires from 85 subjects.

IV. Results

In this section, we present the outcomes of the subjective experiments by means of a statistical analysis. We first define the statistical measures, and then discuss the results.

1. Statistical Measures

We denote by $\bar{x}_k$ the average MOS value of a tested parameter for the k-th application under evaluation in a set S of size K. Then, we have

$$\bar{x}_k = \left(1/N\right) \sum_{i=1}^{N} x_{ik}, \tag{1}$$

where x_{ik} is the opinion score given to the tested parameter by the i-th respondent for the k-th application, and N is the total number of respondents. The standard deviation of x for the k-th application is defined as the square root of the variance as

$$s_k^2 = \left(1/N-1\right) \sum_{i=1}^{N} \left(x_{ik} - \bar{x}_k\right)^2. \tag{2}$$

The confidence interval (CI) associated with the MOS score of each evaluated parameter and application is given by

$$\left[\bar{x}_k - \delta_k,\ \bar{x}_k + \delta_k\right], \tag{3}$$

and the deviation term can be assessed using the standard deviation and the total number of respondents N. Thus, for a 95% CI and a normal distribution function, the deviation term is expressed as

$$\delta_k = 1.96 \left(s_k/\sqrt{N}\right). \tag{4}$$

The skewness, which measures the degree of asymmetry of data around the mean value of a distribution of samples, and the kurtosis, which measures how outlier-prone a distribution is, can be obtained using

$$\beta = m_3/m_2^{3/2}, \tag{5}$$

$$\gamma = m_4/m_2^2. \tag{6}$$

The central moment m is defined as

$$m_j = \left(1/N\right) \sum_{i=1}^{N} \left(x_i - \bar{x}\right)^j. \tag{7}$$

Analysis of variance (ANOVA) tests the hypothesis that the means among two or more groups are equal, under the assumption that the sample populations are normally distributed. For instance, for a two-way ANOVA, there are three possible null hypotheses: there is no difference in the means of factor f_1, there is no difference in the means of factor f_2, and there is no interaction between factors f_1 and f_2. The output test statistic F is calculated as

$$F = \left(\frac{\sum_{i=1}^{m} n_i \left(\bar{x}_i - \bar{x}\right)^2}{m-1}\right) \Bigg/ \left(\frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2}{N-m}\right), \tag{8}$$

where m is the number of specific combinations of factor levels whose effect is to be compared with other specific combinations. The number of observations, for the specific combination of factor levels i, is denoted by n_i, N is the total number of observations (the sum of the n_i), (m–1) are the degrees of freedom for treatment, and (N–m) are the degrees of freedom for error. The critical F value is given by the F distribution, the degrees of freedom, and the significance level (α-level).

The correlation coefficient r indicates the correlation between two parameters and is given by

Table 3. Users' opinion scores.

	Skype			Gizmo5			ooVoo			Damaka			vBuzzer		
	CI_{low}	MOS	CI_{up}	CI_{low}	MOS	CI_{up}	CI_{low}	MOS	CI_{up}	CI_{low}	MOS	CI_{up}	CI_{low}	MOS	CI_{up}
User expectations	4.04	4.25	4.45	3.96	4.13	4.30	3.70	3.91	4.12	3.05	3.33	3.60	3.29	3.51	3.72
Graphical interface	3.79	3.92	4.05	3.96	4.13	4.30	4.32	4.42	4.53	1.86	2.06	2.27	2.96	3.07	3.18
Ease of use	4.29	4.41	4.54	3.80	3.96	4.12	2.46	2.75	3.03	1.48	1.83	2.18	2.51	2.77	3.03
Video quality	3.06	3.34	3.62	1.85	2.15	2.44	3.76	4.01	4.26	1.00	1.26	1.52	-	-	-
A/V synchronization	4.27	4.45	4.62	3.87	4.15	4.44	4.30	4.48	4.67	3.46	3.75	4.05	-	-	-
Connection time	1.91	2.20	2.48	1.30	1.61	1.91	1.43	1.74	2.04	2.77	3.07	3.38	1.00	1.26	1.52
Instability	1.00	1.16	1.32	1.82	2.50	3.17	1.09	1.28	1.47	1.75	2.17	2.58	1.08	1.28	1.38
Delay	1.00	1.21	1.43	1.27	1.64	2.00	1.14	1.39	1.64	1.01	1.24	1.48	1.35	1.71	2.06
Echo	1.00	1.28	1.55	1.07	1.36	1.65	1.00	1.28	1.56	1.08	1.35	1.62	1.49	1.85	2.21
Noise	1.17	1.41	1.64	1.58	1.84	2.11	1.01	1.18	1.35	2.48	2.76	3.05	2.37	2.57	2.77
Voice distortion	1.15	1.36	1.56	1.04	1.33	1.62	1.00	1.23	1.46	1.55	1.91	2.27	1.24	1.50	1.77
Usefulness	3.75	3.89	4.04	2.83	3.12	3.42	3.99	4.12	4.26	1.95	2.22	4.26	2.74	2.88	3.02
Overall score	3.57	3.79	4.01	2.64	2.93	3.21	4.04	4.25	4.46	1.85	2.06	2.28	2.20	2.43	2.66

$$r = \left(1/N-1\right) \sum_{i=1}^{N} \left(\frac{x_{ik}-\bar{x}_k}{s_{xk}}\right)\left(\frac{y_{ik}-\bar{y}_k}{s_{yk}}\right), \quad (9)$$

where s_{xk} represents the standard deviation of parameter x for application k, and s_{yk} represents the standard deviation of parameter y for application k. The value of r is always in the range between –1 and 1. A positive value indicates a positive association between the two parameters. A negative value indicates a negative association.

Last, the multiple linear regression model is given by

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_px_p + e, \quad (10)$$

where z is the dependent variable, x_i are the different independent variables (from 1 to p), a_0 is the intercept, and e represents an error term. The goodness of the model is characterized by the coefficient of multiple determination R^2, the significance F, the p-statistic, and the confidence intervals. We will use the regression to identify what parameters have the greatest effect on the final score of the VVoIP applications.

2. Mean Users' Opinion Scores

Through the ACR and the DCR methods (Table 2), users evaluated a wide range of quality parameters. By means of the ACR, respondents rated their expectations, usefulness, graphical interface, ease of use, video quality, and audio and video (A/V) synchronization. By means of the DCR, users rated their perception of connection time (the time to initiate the application), instability (dropped calls), echo, voice distortion, noise, and delay. From users' mean opinion scores and the corresponding confidence intervals, we can infer the following results (see Table 3). Note that the MOS obtained in this work agree with the MOS obtained in the QoE study carried out in [9], which reasserts the findings explained below.

Users have great expectations for all applications, and the more popular the application is, the higher the expectations for it are. Observe that the respondents are regular Internet users, and most of them use it daily. However, although VVoIP applications have already been in the market for several years, only a minority of respondents had used them at least once (10.5%). So, most respondents could not be classified as innovators from a customer behavior point of view [28], but rather as early or late majority users. Meaning that they are not quick to "purchase" the service and are less willing to take risks with it, or they enter the market when the newness has already declined. Consequently, even though they have great expectations, they will not easily accept degradation in transmission quality.

According to the respondents, the best graphical interface corresponds to ooVoo, followed by Skype, Gizmo5, vBuzzer, and Damaka. At first, graphical interface should be correlated with ease of use because we are working with applications whose users need not have a technical background; thus, it would be expected that the ranking of best interfaces would

Table 4. Correlation coefficients.

	Overall score	User expectations	Graphical interface	Ease of use	Video quality	A/V synchronization	Connection time	Instability	Delay	Echo	Noise	Voice distortion	Usefulness
Overall score	1	0.30	0.66	0.24	0.69	0.32	–0.23	–0.48	–0.25	–0.15	–0.59	–0.37	0.77
User expectations		1	0.29	0.18	0.22	0.47	0.30	0.22	–0.26	–0.35	–0.15	–0.32	0.22
Graphical interface			1	0.25	0.70	0.23	–0.26	–0.26	–0.13	–0.08	–0.55	–0.17	0.62
Ease of use				1	0.18	–0.02	–0.05	–0.09	0.07	0.04	–0.14	–0.15	0.24
Video quality					1	0.27	–0.29	–0.29	–0.22	–0.08	–0.52	–0.26	0.66
A/V synchronization						1	–0.33	–0.41	–0.63	–0.48	–0.29	–0.35	0.27
Connection time							1	0.20	0.23	0.10	0.11	0.22	–0.23
Instability								1	0.47	0.18	0.42	0.13	–0.56
Delay									1	0.54	0.39	0.36	–0.17
Echo										1	0.23	0.60	0.04
Noise											1	0.34	–0.57
Voice distortion												1	–0.21
Usefulness													1

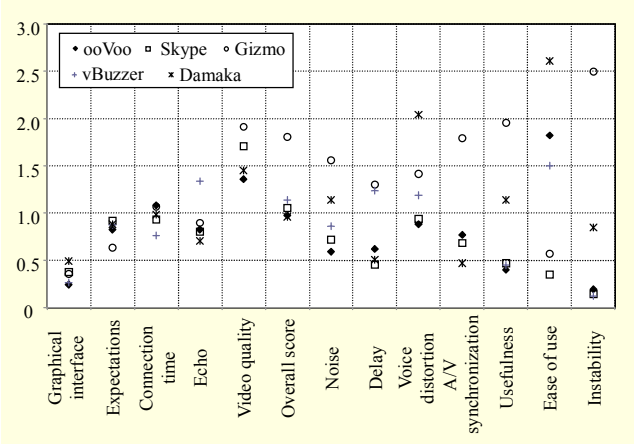

Fig. 1. Variance.

match the ranking of ease of use. This is true for all applications except ooVoo. Even though ooVoo achieved the highest score regarding graphical interface, it was poorly rated in terms of ease of use. Graphical interface and ease of use are parameters not usually included in the conventional QoE objective evaluation models. Nevertheless, they have a clear influence in the final overall score of each application, as we will discuss later.

Video quality perceived by the user and A/V synchronization are well known key parameters for multimedia QoE assessment. In our tests, we observed that although A/V synchronization was generally rated with good levels, the quality of the video set clear differences among applications. Indeed, ooVoo and Skype stand out over the other applications. Note that vBuzzer has not been included since video calls are not supported by this program.

All participants saw connection time as imperceptible with two exceptions: Skype's connection time was considered perceptible but not annoying and Damaka's connection time was perceptible and annoying. Instability (dropped calls) was significant only with Gizmo5 and Damaka. Delay and echo were mostly evaluated as imperceptible. In terms of noise, slightly annoying levels were perceived in Damaka and vBuzzer. Despite the wireless scenario, where it is not infrequent to have significant packet losses, Table 3 shows that the level of perceived voice distortion was low (imperceptible) for the five VVoIP applications.

Once all applications were tested, respondents rated the usefulness and gave an overall score to each application. Usefulness is related to functionality, and the general user's opinion about usefulness is only a little bit above the fair level. Therefore, in our opinion, results show that the use of VVoIP applications in a wireless environment do not yet represent an advantage for users compared with mobile telephony, that is, the global system for mobile telecommunications. This is due to the still current necessity of quality improvement, and at the same time, the demand for additional functionality compared to current telephony systems.

The overall score correlates well with previously evaluated

parameters as is indicated by the correlation coefficients r in Table 4. Those applications that perform better in the subjective analysis receive higher overall scores. Additionally, we observe in Table 4 that the parameters are clearly correlated. For instance, video quality is correlated with graphical interface, and instability is correlated with A/V synchronization, delay, and echo.

From the top down, the global ranking based on the QoE assessment is ooVoo, Skype, Gizmo5, vBuzzer, and Damaka. Relating this result with users' expectations, we see that whereas ooVoo improves, Skype and Gizmo5 slightly decrease. Thus, we can state that expectations are lower for unknown applications than popular ones. Recall that 27.1% of participants knew about VVoIP applications, and 10.5% of them had used Skype or Gizmo5 at least once.

3. Agreement and Asymmetry in Users' Opinion Scores

To get a better understanding of the effect of these parameters, we have studied the variance, the skewness, and the kurtosis of each subjectively evaluated parameter for the five VoIP applications.

Through the variance, we can address the respondents' difficulty or ease in rating a specific parameter and the level of agreement among respondents. A lower variance may be interpreted as a higher agreement among the users, so it was easy for them to give a score. On the other hand, a higher variance represents a disagreement among the users, so it was difficult for them to evaluate the corresponding parameter. Figure 1 depicts the variance of the opinion scores for each parameter. The parameters with a more pronounced agreement among respondents (that is, variance below 1) were connection time, echo, delay, graphical interface, and overall score. This does not mean that the answers were similar; for instance, whereas graphical interface was rated as good or excellent for ooVoo, it got a poor score for Damaka (Table 3). The higher discrepancies among respondents' answers were in the evaluation of usefulness, ease of use, and instability.

From the analysis, we can state that obtaining an excellent subjective quality rating corresponds to a lower variance. However, this effect is not reproduced for poor ratings. This means that it is easier for respondents to identify a good behavior, but have difficulties (or more variable opinions) when a bad score is to be given. For the overall score, both excellent and bad ratings have a relatively low variance.

We also observe in Fig. 1 that the variance of all parameters is almost the same for the two best scoring applications (ooVoo and Skype), with the only exception being ease of use. Nevertheless, the variance of ease of use for Skype could be higher if that minority of respondents (10.5%) had not used it at least once. Consequently, we can affirm that respondents have the same level of agreement in the same parameters for the two best scored applications. On the other hand, respondents hold strong opposing views when evaluating Damaka or Gizmo5, seen in the high variances in Fig. 1. Hence, we confirm that good scores present a stronger agreement than lower scores.

Figure 2 plots the skewness of each rated parameter for all applications under study. In the context of subjective ratings, skewness shows the degree of asymmetry of the scores around the MOS value of each distribution of samples for a given parameter. A normal distribution has a perfect symmetry and a skewness of zero. Despite the initial asymmetry revealed by these results (see Fig. 2), we should take into account the following facts. First, the standard error for skewness can be assessed as twice the square root of $6/N$ regardless of the sign, where N is the number of respondents [29]. Consequently, in this study, we can accept as natural for a normal distribution a skewness whose value is below 0.53. Second, skewness decreases as the population increases [30]; hence, even though the number of respondents used in this work is reliable for a subjective quality test [25], it might not be large enough for a strong conclusion based only on skewness. Third, test scores are limited to the small range from 1 to 5, thus subjective scores have to approach to the maximum or minimum values in extreme quality cases. Finally, the presence of outliers in the test will be detected by inspection on skewness and kurtosis. In this sense, it should be noted though that, since we carry out our experiments in a real scenario, it is likely for a specific user to experience conditions different than the others. Thus, it does not mean that the respondent is an outlier, but that he/she really suffered a change in the application's operation compared to other participants.

Bearing this in mind, we examine the tendency of the scores given for each application. As we see in Fig. 2, most users rated connection time with scores that asymmetrically spread around the MOS towards lower values (positive skewness). Damaka is the exception because the skewness of its connection time is negligible.

Regarding the skewness of instability, all applications have a positive value. Although most scores were very low, indicating an imperceptible instability level, the fact of having a few dropped calls makes the mean appear higher, whereas the mass of the distribution is located on the left side of the mean. Voice distortion, delay, and echo show similar results with a small positive bias for all applications. A low negative skewness is observed for noise levels in Damaka and vBuzzer, so the scores are skewed right, which means they behaved worse than average value. In ooVoo, Skype, and Gizmo5, the perceived noise shows a positive skewness value in the range from 1 to 2.

Figure 2 also shows that users' expectations have a low

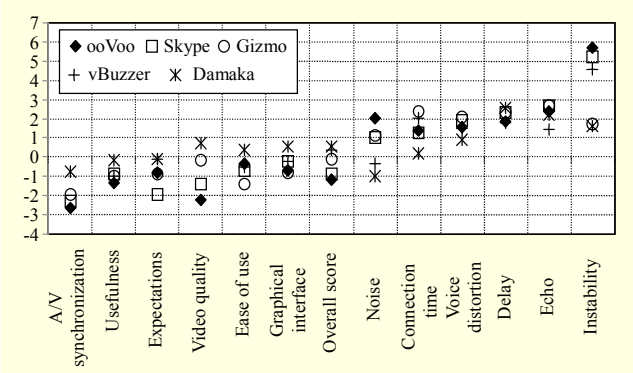

Fig. 2. Skewness.

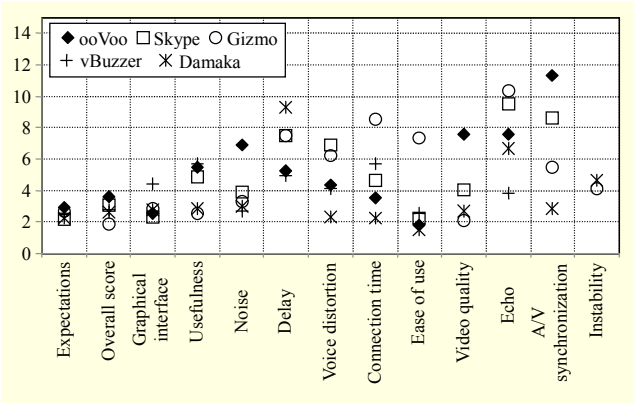

Fig. 3. Kurtosis.

negative skewness, which translates to the subjective scores being spread a little bit more towards higher values than the MOS. Respondents' answers tended to be confident about VVoIP applications. For graphical interface, there is hardly any asymmetry in the distribution of answers. On the other hand, whereas Damaka's distribution for the ease of use parameter is symmetric, results show an increasingly negative trend for ooVoo, vBuzzer, Skype, and Gizmo5; a bias towards being more and more easy to use, respectively.

Since there is a negative skewness of value 1 for all applications, users' perceptions of VVoIP applications' usefulness are spread a little towards higher values than the MOS. In addition, whereas Gizmo5 and Damaka hardly show skewness for video quality, ooVoo and Skype present a tendency towards values greater than the corresponding mean. A/V synchronization also shows a negative skewness. Finally, in the overall score, Gizmo5, vBuzzer, and Damaka have symmetric distributions, whereas Skype and ooVoo have their scores concentrated on the right part of the distribution.

If we observe the trend of the skewness, we can easily identify that all the parameters follow the same tendency. For most applications, there are parameters with positive skewness (instability, echo, delay, voice distortion, and connection time), and parameters with a negative skewness (overall score, ease of use, graphical interface, video quality, usefulness, expectations, and A/V synchronization). Note that parameters with positive skewness are those classified as poor (rated following the DCR scores), and parameters with a negative skewness are those classified as good (rated following the ACR scores). Because of the complementary nature of both rating scores, the small tendency towards low scores in DCR and high scores in ACR can be translated as an optimistic user's opinion regarding the QoE of these VVoIP applications.

Next, Fig. 3 depicts the kurtosis of each rated parameter. Kurtosis measures how outlier-prone a distribution is. For a normal distribution, the default kurtosis value is 3. The lower the kurtosis, the more prone a distribution is. As it occurs with the skewness, we can consider as normal kurtosis values two times below the standard error of the kurtosis, which is calculated as the square root of 24 divided by N [29]. Therefore, kurtosis below 4.06 (3 plus 1.06) will be negligible in this study. As shown in Fig. 3, the distributions of some subjective scores give kurtosis values much greater than that of a normal distribution. Similarly to skewness, this can be interpreted in two different ways. First, it can be a clear sign of outliers, meaning that a few of the respondents gave low (high) ratings whereas the majority of the respondents agreed on a higher (or lower) quality. Second, it can be the result of a different situation experienced by the user because of particular application misbehavior. The most extreme case is the instability parameter for ooVoo, Skype, and vBuzzer (kurtosis>15). Matching also with skewness results, there were respondents that rated the application's instability as annoying (at different levels), whereas most respondents rated it as imperceptible or perceptible but not annoying. Thus, either they suffered a dropped call, likely due to the wireless environment, or they are outliers.

4. Effect of Gender and Frequency of Internet Use in Users' Opinion Scores

At this point, we analyzed the variances between the groups, which would identify if gender or frequency of Internet use cause differences in the means of the overall score of each VVoIP application under study. The gender factor has two levels: male or female. In the frequency of Internet use factor, we take into account three levels: daily, frequent (but not daily), or occasional. Recall that if probability p (Table 5) is less than or equal to the significance level ($\alpha = 0.05$), then one or more means are significantly different (and the F statistic will be larger than the critical F). Otherwise, if p is larger than the α-level, then the means are not significantly different.

First, we identify the interaction among factors to determine

Table 5. ANOVA: (a) gender (2) × frequency of Internet use (3); (b) gender (2); and (c) frequency of Internet use (3).

	(a)		(b) $F_{0.05, 85}$=3.96		(c) $F_{0.05, 85}$=3.11	
	F	*p*	F	*p*	F	*p*
ooVoo	0.61	0.55	5.88	0.02	2.12	0.13
Skype	0.41	0.67	4.20	0.04	2.29	0.11
Gizmo5	1.06	0.35	9.41	0.00	3.20	0.05
Damaka	0.06	0.50	1.33	0.25	0.29	0.75
vBuzzer	0.21	0.82	0.06	0.83	3.44	0.04

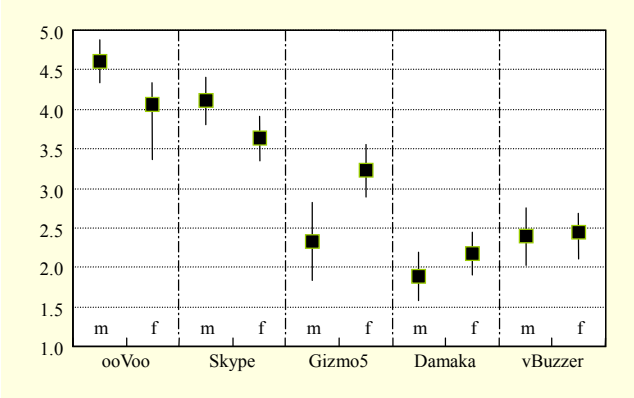

Fig. 4. Overall MOS means by gender (m: male and f: female).

if we can consider the effects of individual factors separately. Table 5(a) summarizes the results of the two-way ANOVA (gender (2) × frequency of Internet use (3)). We observe that the interaction is not significant for any of the VVoIP applications. Then, we examine the effect of the gender factor through a one-way ANOVA. The null hypothesis is that there is not a statistical difference among the overall scores given by female and male respondents. Results are included in Table 5(b). As observed, Skype, Gizmo5, and ooVoo, the best rated applications in the QoE assessment, have a test statistic F higher than the critical value, and a p-value lower than the α-level. Hence, we reject the null hypothesis. For these applications, there is a statistically significant difference among the population means. Figure 4 depicts the means and confidence intervals of the overall score of each application classified by the participants' gender. We can see that males and females do not share the same opinion of the highest scoring applications. In contrast, the results of the one-way ANOVA gender score for vBuzzer and Damaka, the worst rated applications in the subjective QoE assessment, indicate that there is not a statistically significant difference among the testing participants' means for the overall scores.

In terms of frequency of Internet use, the one-way ANOVA

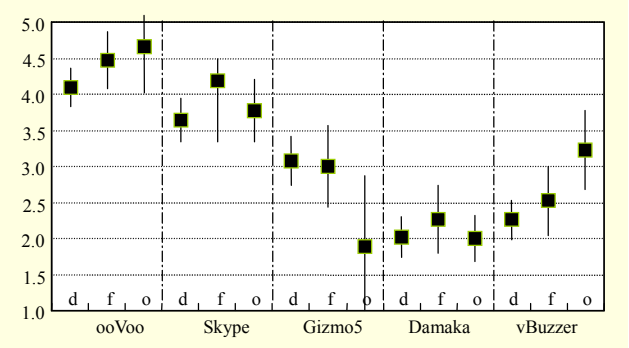

Fig. 5. Overall MOS means by frequency of Internet use (d: daily, f: frequently, and o: occasionally).

shows that there is not a statistical difference among the population means for the overall scores based on the respondents' frequency of Internet use (see Table 5(c) and Fig. 5). At first, this result may seem surprising, since we might expect that those users more familiar with Internet would be more demanding. However, despite being Internet users, they had not yet adopted these types of applications, and consequently, they shared a common point of view regarding them.

5. Identification of Key Parameters for the Overall Score

Through multiple-linear regression analysis, we identified what parameters had the strongest influence on the overall scores given by the participants.

Table 6 includes the regression coefficients. Examining the results, we see that graphical interface, usefulness, video quality, instability, and voice distortion are the five major elements that affect the subjective overall score for the VVoIP applications under study (excluding vBuzzer). The other parameters were discarded after showing a p-value above 0.1. Using only these five factors, the determination coefficient R^2 presents a value of 0.71, multiple correlation coefficient R is 0.844, adjusted R^2 is 0.71, and significance F is 0. Regression coefficients are positive for good parameters (graphical interface, usefulness, and video quality), and negative for poor parameters (instability and voice distortion), which is reasonable. Note that the validity of these statistical inferences from regression analysis rests on the following assumptions: linearity is assumed to hold, and errors in the regression model follow a normal distribution, are mutually independent, and have the same variance. These statements were confirmed through the scanning of the residuals.

We already said that our participants belong to the early or late majority users. Thus, according to the diffusion theory accepted for describing consumer behavior [28], they would not accept as much degradation in quality as innovators do.

Table 6. Regression coefficients.

	Coeff.	Stand. Err.	t statistic	p-value	CI_{low}	CI_{up}
Intercept	0.616	0.209	2.944	0.35%	0.204	1.027
Graphical interface	0.283	0.059	4.806	0.00%	0.167	0.399
Usefulness	0.494	0.056	8.898	0.00%	0.385	0.604
Video quality	0.152	0.038	3.946	0.00%	0.076	0.227
Instability	–0.147	0.050	–2.955	0.30%	–0.245	–0.049
Voice distortion	–0.219	0.036	–6.023	0.00%	–0.291	–0.148

Table 7. Comparing multiple linear regression coefficients between individual applications and general model.

	ooVoo	Skype	Gizmo5	Damaka	General
Usefulness	✓			✓	✓
Graphical interface			✓		✓
Ease of use	✓				
Video quality		✓	✓		✓
A/V synchronization		✓		✓	
Instability		✓	✓	✓	✓
Noise	✓	✓		✓	
Echo	✓				
Voice distortion		✓			✓
Delay	✓				

This is corroborated by the regression analysis since video quality, instability and voice distortion are incorporated as coefficients into our preliminary model.

In addition, we detect some small differences between global regression analysis using all observations and particular regression analysis of each individual application (see Table 7). Note that, as mentioned previously, ease of use and graphical interface are not usually taken into account in QoE models, but they have an important weight in the overall score as we found in this study. Moreover, usefulness is not included as a regression coefficient for the most popular applications of this study (Gizmo5 and Skype), but it is for the unknown ones.

V. Conclusion

Through statistical analysis of subjective QoE opinion scores, we searched for new aspects of QoE assessment, not included in previous models, which rise due to the inherent properties of broadband wireless communications, new services acceptability issues, and customer behavior. We conducted a survey in a real WLAN, with 85 respondents from whom we collected a total of 425 observations. For each observation, participants rated expectations, usefulness, graphical interface, ease of use, video quality, A/V synchronization, connection time, instability, echo, voice distortion, noise, delay, usefulness, and a final overall score. The five VVoIP applications selected for this study (based on popularity) were Skype, Gizmo5, ooVoo, Damaka, and vBuzzer. We found that users' QoE is mainly given by usefulness, graphical interface, video quality, instability (dropped calls), and voice distortion. We obtained a clear difference between the MOS of the overall quality scores depending of the respondents' gender, but the frequency of Internet use did not affect the results. Agreement among participants was higher for good ratings than for poor ones. From the results obtained in this work, we conclude that current QoE assessment models for multimedia applications in a wireless environment need to take into account not only technical measurements but also non-technical ones. Thus, the advantage of access factors such as those used in the E-model should be updated according to the application level of popularity or new functionalities (usefulness) among other features. Nevertheless, this work should be seen as a first subjective approach for QoE evaluation of VVoIP in wireless networks, and we will delve further into this area in future works.

References

[1] ITU-T P.10/G.100 "Vocabulary for Performance and Quality of Service. Amendment 2. New Definitions for Inclusion in Recommendation ITU-T P.10/G.100," July 2008.

[2] ITU-T Rec G.1010 "End-User Multimedia QoS Categories," Nov. 2001.

[3] ANIQUE+, "Auditory Non-Intrusive Quality Estimation Plus (ANIQUE+) Perceptual Model for Non-Intrusive Estimation of Narrowband Speech Quality," ATIS-0100005.2006, ANSI, 2006.

[4] ITU-T Rec. G.107, "The E-Model: A Computational Model for Use in Transmission Planning," Apr. 2009.

[5] IEEE P1903™/D1 "Draft White Paper for Next Generation Service Overlay Network," May, 2008.

[6] "Challenge to All: Roadmap for the Usage of Collaboration Tools," Keith Dickerson, ETSI Board, Nov. 26, 2009.

[7] S. Möller et al., "Speech Quality while Roaming in Next Generation Networks," *Proc. ICC*, 2009, pp. 1-5.

[8] P. Perala and M. Varela, "Some experiences with VoIP over Converging Networks," *Proc. MESAQIN*, 2007, pp. 1-7.

[9] M.-D. Cano and F. Cerdan, "Subjective QoE Analysis of VoIP Applications in a Wireless Campus Environment," Telecommunication Systems, in press, 2010. Available: http://www. springerlink.com/content/7420510843151356/

[10] T. Hayashi et al., "Multimedia Quality Integration Function for Videophone Services," *Proc. GLOBECOM*, 2007, pp. 2735-2739.
[11] L. Malfait, J. Berger, and M. Kastner, "P.563–The ITU-T Standard for Single-Ended Speech Quality Assessment," *IEEE Trans. Audio, Speech, and Language Process*., vol. 14, no. 6, 2006, pp. 1924-1934.
[12] P. Calyam et al.,"A "GAP-Model"-Based Framework for Online VVoIP QoE Measurement," *J. Commun. Networks*, vol. 9, no. 4, 2007, pp. 446-456.
[13] S. Tao, J. Apostolopoulus, and R. Guerin, "Real Time Monitoring of Video Quality in IP Networks,"*Proc. ACM NOSSDAV*, 2005.
[14] G. Rubino and M. Varela, "A New Approach for the Prediction of End-to-End Performance of Multimedia Streams," *Proc. QEST*, 2004.
[15] Skype. Available: http://www.skype.com
[16] TeleGeography. International Carriers' Traffic Grows Despite Skype Popularity, TeleGeography Report and Database, 2006. Available: http://www.telegeography.com
[17] Gizmo5. Available: http://www.gizmo5.com. Last visited Oct. 28, 2009.
[18] Global IP Solutions. Available: http://www.gipscorps.com. Last visited Oct. 28, 2009.
[19] ooVoo. Available: http://www.oovoo.com. Last visited Oct. 28, 2009.
[20] Damaka. Available: http://www.damaka.com. Last visited Oct. 28, 2009.
[21] vBuzzer. Available: http://www.vbuzzer.com. Last visited Oct. 28, 2009.
[22] S. Andersen et al., "Internet Low Bit Rate Codec (iLBC)," IETF RFC 3951, 2004.
[23] On2 web site. Available: http://www.on2.com. Last visited Oct. 28, 2009.
[24] F. Fitzek and M. Reisslein, "MPEG–4 and H.263 Video Traces for Network Performance Evaluation," *IEEE Network*, vol. 15, no. 6, 2001, pp. 40-54.
[25] ITU-T Rec P.910 "Subjective Video Quality Assessment Methods for Multimedia Applications," Sept. 1999.
[26] ITU-T Rec. 920, "Interactive Test Methods for Audiovisual Communications," Mar. 2000.
[27] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," Aug. 1996.
[28] W.L. Wilkie, *Consumer Behaviour*, John Wiley & Sons Inc., New York, 1994.
[29] B.G. Tabachnick and L.S. Fidell, *Using Multivariate Statistics*, 3rd ed., New York: Harper Collins, 1996.
[30] NIST/SEMATECH, "e-Handbook of Statistical Methods," 2009. Available: http://www.itl.nist.gov/div898/handbook/

Maria-Dolores Cano received the telecommunications engineering degree in 2000 from the Universidad Politécnica de Valencia (UPV), Spain, and obtained the PhD from the Universidad Politécnica de Cartagena (UPCT), Spain, in 2004. She joined the UPCT in 2000, where she is currently an associate professor at the Department of Information Technologies and Communications. She has published more than 25 papers in international journals and conferences in the areas of quality of service (QoS), traffic control, and security. Her current research interests include QoS and security provisioning in telecommunications networks. Dr. Cano was awarded a Fulbright grant as a postdoctoral researcher in 2006 at Columbia University, USA, and since 2007 has collaborated with several universities in South America. She was awarded the Best Paper Award at the *IEEE International Symposium on Computers and Communications ISCC'05*. She is member of the Editorial Board of the *IET Journal on Wireless Sensor Systems* as well as of the Technical Program Committee of several IEEE international conferences. She also collaborates as a reviewer for international journals included in the JCR.

Fernando Cerdan obtained the Telecommunications Engineering degree in 1994 from the Polytechnic University of Catalunya (UPC), Spain, and received the PhD from the same University in 2000. In 1996, he started working towards a PhD in telecommunication supported by a national grant. He is currently a full professor at the Universidad Politécnica de Cartagena (UPCT), Spain. From 2002 to 2005, he was vice-dean at the Telecommunications School at UPCT in charge of international and company relationships. Since 2006, he has been the Head of the Department of Information Technologies and Communications. Most of his published papers are in the field of traffic control, service integration, and QoS in wired and wireless data networks, although his current interests also includes learning technologies and application development.

Sergio Almagro received the BS in telematics engineering in 2002 from the Universidad Politécnica de Cartagena (UPCT), Spain. Afterwards, he served as a researcher in the Department of Information Technologies and Communications of the same university, from where he obtained the telecommunications engineering degree in 2004. Since 2008, he has been working towards his PhD supported by a regional grant. His main research interests are in the area of modeling, analysis, and the optimization of protocols and architectures for broadband wireless networks. Most of his published papers are focused on the field of traffic control and wireless data networks simulation.