VoIP QUALITY ASSESSMENT USING THE STUDENT *t*-TEST

Z. GÁSPÁR[1] Gh. TOACŞE[1]

Abstract: *This paper presents a statistical method for determining the quality of service (QoS) of a voice over internet protocol (VoIP) carrier. The presented method introduces a new approach: the relation between speech transmission quality and call length average, for a given VoIP carrier and portfolio of customers. Based on real world data collection, the call length distribution is computed and then the mathematical expressions for the Student t-test are derived and used. By comparing the results of the statistical test with the customer service reports we show the viability of the statistical method for determining the voice quality degradation.*

Key words: *VoIP, voice quality, call length, Student t-test.*

1. Introduction

The public switched telephone network (PSTN) gives a high level of reliability and conversational quality to its users [1]. Unlike PSTN, due to the best effort nature of the Internet, a Voice over Internet Protocol (VoIP) carrier needs to handle fluctuations in the quality of calls for a certain destination, resulting in a lower total availability [3].

To meet its client's expectations, the VoIP telephony provider, that has several carriers for the same destination, needs to choose a carrier that provides adequate quality and route the calls to it. In order to achieve this, the telephony provider needs to monitor the quality of each carrier for that destination. Call quality evaluation is a field that received a broad attention from the scientific community. In chapter 2 we make a brief presentation of traditional and state of the art methods for call quality evaluation: subjective and objective ones; it is followed in chapter 3 by the presentation of the mathematical fundaments for statistical testing and the Student *t*-test.

The data used to evaluate our method is real life data collection of a VoIP long distance telephony provider, recorded from 2007-05 to 2007-09. It totals a number of 460 000 answered calls and during the same period the customer service department received 387 complaints that are the starting point for our quality assessment. Chapter 4 presents the results of the implementation expressed in the percentages of recognized situations. In the final chapter we present the conclusions of our study.

2. Voice Quality Measurement

Voice quality can be measured using subjective and objective methods. These methods were standardized by ITU-T and are presented in the following subchapters.

2.1. Subjective Voice Quality Measurement

Historically, the first developed method for quality evaluation was the ITU-T P.800,

[1] Dept. of Electronics and Computers, *Transilvania* University of Braşov.

recommendation that introduced the Absolute Category Rating (ACR) test method. The ACR gives scores from 5 (excellent) to 1 (bad). The measurement can be regarded as comparison of a test signal and a reference "in the mind" of the listener. This is because the listener is very familiar with the natural sound of human voice; one can compare the test signal with his/her representation of voice. In this way, based on statistical processing of individual results, a mean opinion score (MOS) can be calculated.

2.2. Objective Voice Quality

Intrusive measurements for voice quality like PSQM, PESQ [4] and PEAQ [5] for audio quality, per definition use two input signals for the quality evaluation. This is the undistorted original signal and the degraded signal. The MOS score is estimated based on a weighted sum of differences between the two signals.

In non-intrusive measurements, like the one defined in ITU-T recommendation P. 563 [6], only the degraded voice stream is used for the MOS estimation.

2.3. Statistical Methods

The statistical methods for determining the voice quality can only provide information regarding the average quality of a channel based on a large number of calls already sent. Holub et al. in [2] comes to the conclusion that there is a correlation between the call duration and the call quality in the low to medium range (Mean Opinion Scores between 1 and 3). Accordingly, Figure 1 shows the differences between the distributions of call duration for two carriers, clearly showing a shifting to the short duration call values for the poor quality carrier (carrier 1), 82% compared to 15.2% of the calls are shorter than 90 seconds, giving carrier 2 a higher average

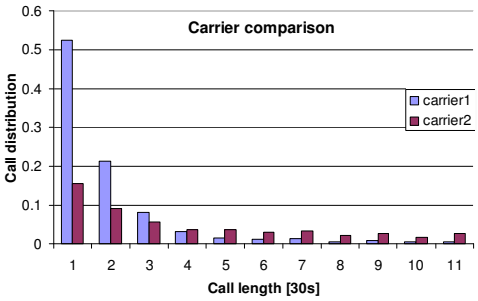

Fig. 1. *Distribution of call lengths for poor (carrier 1) and good (carrier 2) quality carriers. A number of 21694 calls were used for determining the distributions*

call duration. For this statistic, a number of 21694 real life calls were used, acquired during 2 years (2007-2008). The customers making the calls are prepaid long distance callers. The distributions were computed by calculating the relative frequency of each 30 seconds call duration interval.

3. Statistical Tests

Let us consider a one dimensional repartition whose probability density function $f(x;\Theta_1,\Theta_2,\ldots,\Theta_k)$ depends on k parameters $\Theta_1,\Theta_2,\ldots,\Theta_k$ that can be interpreted as the coordinates of a point $\Theta = (\Theta_1,\Theta_2,\ldots,\Theta_k)$ in the k dimensional Euclidian space R^k. Let A be a subset of R^k. If we presume that the Θ is included in A, we are in the presence of the null hypothesis (H_0), otherwise in the presence of the alternate hypothesis (H_1):

$$H_0 : \Theta \in A, \tag{1}$$

$$H_1 : \Theta \notin A, \tag{2}$$

$$P[\Theta \notin A \mid H_0 = true] = \alpha. \tag{3}$$

The confidence interval of a test is given by (3). If we have to test the average value of a random phenomenon ($\bar{x}$) we can use the Student t-test defined by (4):

$$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{\nu - 1}}}, \tag{4}$$

where: μ_0 is the call length average; υ is the number of samples; s is the estimated standard deviation.

If the value Z is less than the value of the Student distribution, according to $\nu-1$ degrees of freedom and α-confidence interval, we need to accept the test hypothesis (H_0), otherwise the alternate hypothesis (H_1).

4. Implementation of the Statistical Test

The Student *t*-test requires the random variable to be normally distributed. Figure 1 shows us that the call length is an exponential type distribution, inappropriate for this test. To overcome this we group the calls in sets containing N successive calls, and for the calls in each set we compute the call length average (CLA).

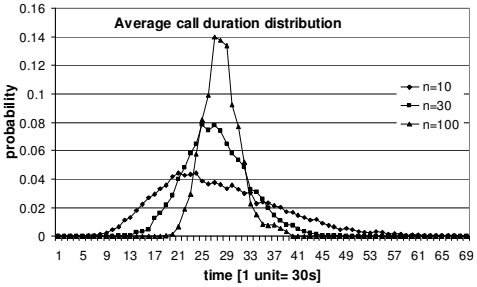

Fig. 2. *Average call duration distribution as a function of sample number. We can make the normal distribution assumption starting from a sample size of 30*

In Figure 2 it is shown how the distribution of the CLA converges to a normal distribution as the number of calls in the set is increased. After the variable change we take ν averages and compute the Student *t*-test to determine if there is a statistically significant difference from the expected average. If there is a statistically significant difference, with confidence level of α, when noticing lowering of the average we can presume that the carrier is having quality issues. In order to keep the call length average samples independent, a condition of the Student *t*-test, one call duration is used in only one average calculation, so the total number of calls is:

$$N_{calls} = \nu \cdot N . \tag{5}$$

We confirm the results of the test by analysing the customer service tickets for that period. If we had at least one customer complaint for that specific destination we could say that our presumption is confirmed and we have a confirmed event. Because not all customers call to report the problem immediately, we took an additional 48 hours period after the last analysed call to look for customer service tickets. Figure 3 shows the percentage of detected cases from all the reported cases between 2007/05 and 2007/09.

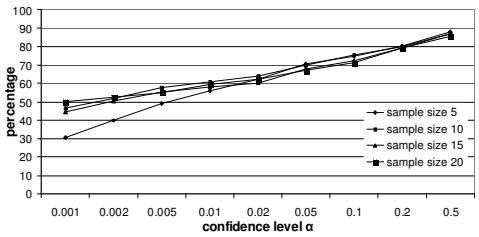

Fig. 3. *Confirmed detection percentage from all reported cases to customer service*

The other parameter of the test we need to take into account is the percentage of confirmed events from the total number of detected events shown in Figure 4.

5. Conclusions

This paper presents a novel statistical method for voice quality assessment. The simulation shows that a large percentage

(over 80%) of low quality situations can be detected using the presented statistical method.

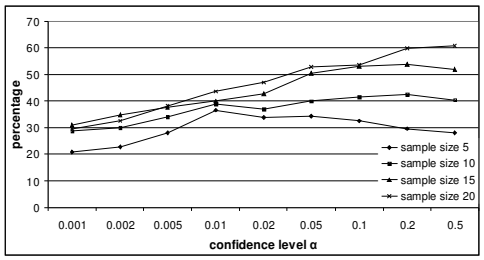

Fig. 4. *Percentage of confirmed events from the total number of detected events*

As shown in [7] the voice stream analysis is a computationally intensive operation requiring tens (in the case of intrusive objective measurements) and hundreds (in the case of non-intrusive objective measurement) of MFLOP for each second of analysed call. This voice stream analysis phase is followed by a summation of the partial results, giving the final MOS score for the call. The fundamental difference brought by the new method consists of completely eliminating the need of the voice stream analysis only requiring the statistical processing phase, independent from the call length. No longer requiring voice stream analysis, this approach introduces a computationally feasible way to implement a close to real time voice quality assessment method. The number of calls required for this test is given by (10) and only medium or high (tens, hundreds of calls per hour) traffic destinations can be tested close to real time using this method.

By using this new method, medium to large size VoIP providers have the possibility to assess the voice quality.

The existence of this new call quality assessment method opens the possibility to develop automated real time call routing algorithms that take into account the call quality factor.

Acknowledgements

The authors would like to thank Florin Miron and Silvana Santa for providing the data and for many insightful discussions that contributed to the ideas of this paper.

References

1. Carolyn, R.J., Yakov, K., et al.: *VoIP Reliability: A Service Provider's Perspective*. In: IEEE Communications Magazine **7** (2004), p. 48-54.
2. Holub, B., Beerends, J.G., et al.: *Dependence between Average Call Duration and Voice Transmission Quality*. In: Wireless Telecommunication Symposium, 2004, p. 75-81.
3. Jiang, W., Schulzrine, H.: *Assessment of VoIP Service Availability in the Current Internet*. In: Proceedings of the 4th International Workshop on Passive and Active Network Measurement, 2003, p. 150-159.
4. ITU-T Rec. P.861: *Objective Quality Measurement of Telephone-Band Speech Codecs*. In: International Telecommunication Union, Geneva, 1996, p. 1-17.
5. ITU-T Rec. P.862: *Perceptual Evaluation of Speech Quality*. In: International Telecommunication Union, Geneva, 2001, p. 1-30.
6. ITU-T Rec. P.563: *Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications*. In: International Telecommunication Union, Geneva, 2004, p. 1-25.
7. Opticom: *User Manual 3SQM™ OEM. Version 2.0.1*. Available at: www.opticom.de.