Speech Quality Evaluation and Benchmarking in Cellular Mobile Networks

Oliver Nipp[1], Marc Kuhn[1], Armin Wittneben[1], and Thomas Schweinhuber[2]
[1]Communication Technology Laboratory, ETH Zurich, 8092 Zurich, Switzerland
[2]T-Mobile International AG & Co. KG, Landgrabenweg 151, 53227 Bonn, Germany
{onipp, kuhn, wittneben}@nari.ee.ethz.ch, thomas.schweinhuber@t-mobile.cz

Abstract— **Speech quality is still and will be a major performance indicator for end-to-end QoS in (cellular) mobile telecommunication networks. Therefore, it is crucial to measure and to evaluate the behavior of speech quality in live networks either via commonly applied drive-tests or other kind of field trials. The evaluation of the measurement data is a key element in order to obtain representative and consistent results. We provide an overview of current techniques to evaluate and benchmark speech quality services, incorporating the current attempt from ETSI to evaluate speech quality per call. We show the deviations that occur when using different approaches to evaluate speech quality measurement data and conclude with a recommendation on how to efficiently design evaluations for comparative QoS benchmarks.**

I. Introduction

Independent of the technology used, it is necessary to evaluate the performance and end-to-end QoS level of a (wireless) network for maintaining, optimizing (for new as well as for expanding networks), troubleshooting and benchmarking purposes. During the last decades wireless systems have matured, but are still susceptible to failures, especially when different technologies like e.g. GSM and UMTS are concurrently involved (e.g. vertical handovers) and are therefore sources of quality degradation. The most important basic service is still and will be plain speech transmission, and with more and more of today's communication moving from fixed to mobile systems, the higher the quality demands will be. Users will expect a mobile speech quality which is close to the fixed line quality, independent of location or technology used.

The automatic assessment of speech quality in wired and wireless networks has been subject of extensive research, especially in the last decade. Industry standards like PAMS [1], PSQM/PSQM+ [2] and its successor PESQ [3], [4] compare an unprocessed original signal with the degraded version after sending it through, e.g., the mobile channel. PESQ is able to predict speech quality with good correlation to subjectively perceived quality in terms of MOS (mean opinion score) in a wide range of conditions, which includes coding distortions, errors, noise, filtering, delay, and variable delay [5]. PESQ is the industry standard for end-to-end speech quality measurements. On the other hand objective, non-intrusive, parameter based approaches have also shown to correlate very well with MOS values [6].

Even though, the mapping of perceived speech quality to the MOS score is well understood today, there exist quite a few methods to evaluate the measurement data. Key Performance Indicators (KPI) are defined by ETSI to provide a standardized way to evaluate and benchmark the performance of mobile cellular networks [7]. Network availability and accessibility is covered as well as application based indicators like the cut-off call ratio, setup time or data throughput measurements. Telephony speech quality is also defined as a KPI, whereas one should distinguish between an evaluation per sample (here: a short (5-12 s), predetermined sequence of a speech signal) or per call. PESQ should be used to measure the speech quality.

In live field trials as in drive testing, signal strength and/or signal quality is subject to large variations. This is, among others, due to the change in location and thus large and small scale fading. Further quality determining factors are the capacity and load of the current cell. Hence, there is a need to assess speech quality in short time intervals. PESQ has been validated for signals with a time duration of 8 to 12 seconds, but in drive testing a usual call duration is e.g. 120 s. Since the PESQ method is non-linear, the speech quality of an entire call cannot be calculated by simple averaging of the per sample speech quality evaluation. Hence, a method is presented in [8] to map ”per sample“ quality to ”per call“ quality assessments. We will provide evaluation results when applying this method on live measurement data and will compare these results to a ”per sample“ evaluation and benchmarking, respectively.

Once the method of speech quality assessment is determined, the next step will be to compare different results (e.g. ranking of competitors) based on the speech quality evaluation. One way to analyze PESQ measurement results on samples is to take the empirical CDF/PDF into account and use metrics like the expectation value, 10%, and 90% quantiles which will incorporate the overall information of the measured distribution. On the other hand a threshold is commonly determined that indicates the minimum PESQ value for which speech quality is acceptable for a user. Naturally, this depends on the used handset, voice codec, type of call (mobile terminated/mobile originated) and others. The ratio of good compared to all measured samples can be used to compare telecommunication network operators. Depending on the method of evaluation a comparative ranking can be given and will be derived and explained.

II. Speech Quality Evaluation

In this section we will compare two fundamental approaches to evaluate measurements of speech quality for a set of calls.

A. Speech Quality per Sample

The reason why speech quality is usually measured for short periods of time (5-10 s) is, among others, the highly varying (physical layer) transmission parameters such as signal strength (GSM: RxLev, UMTS: CPICH RSCP) or signal quality (GSM: RxQual, UMTS: CPICH $\mathrm{E_c/N_0}$)[1]. On the other hand, the per sample approach does not fully reflect the end-to-end user perception of speech quality because a call (time duration: $\mathrm{T_c}$) is usually much longer than some seconds. Therefore, and in order to model a real conversation, one call is separated into active and silent parts of length $\mathrm{T_s}$ as shown in Fig.1. The direction of speech quality determination is alternating, thus during the active part of the uplink the downlink is silent and vice versa.

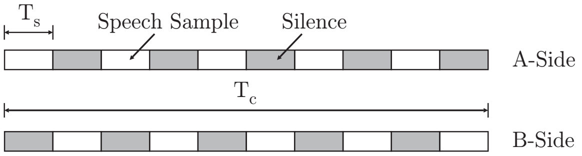


Fig. 1. Call divided into parts, $\mathrm{T_s}$: duration of one sample, $\mathrm{T_c}$: duration of one call

The number of samples that is included in one measurement call is therefore determined by the number of samples in up- and downlink direction.

There are mainly two possibilities to report and to evaluate a set of speech quality measurements. On the one hand, the PDF or CDF of the MOS scores achieved by different operators can be compared as in Fig. 2. Usual metrics for a comparison are the expectation value or the 10% quantile. For the latter one, lower values are desired, since better speech quality is achieved in a larger fraction of all tested samples. The main drawback with this type of comparison is the lack of verifiable statistical integrity, assuming that the PDF/CDF cannot be fully described with some kind of parameterized distribution. In some cases (e.g. similar measurement results) it is difficult to find reliable statements when comparing two PDFs/CDFs.

An alternative is provided by defining a threshold λ that is indicating the minimum PESQ value which still corresponds to an acceptable speech quality. This approach is statistically modeled by a Bernoulli experiment. Either the outcome z of a random variable Z is 0 (speech quality less than λ) or 1 (speech quality is greater or equal to λ). On the one hand, information is lost due to this approach because just a distinct value is taken out of the PDF/CDF. On the other hand, statistically verifiable statements can be given under certain conditions. Assuming uncorrelated samples, the following ratio can be interpreted as the probability to get at least acceptable speech quality

$$p = \frac{\text{No. of Samples with PESQ} \geq \lambda}{\text{No. of Samples}} = \frac{k}{N}. \qquad (1)$$

The probability p itself is equal to the ML-estimator $\hat{p}$ of a binomial distribution:

$$P(Z = k|N, p) = \binom{N}{k} p^k (1-p)^{N-k}. \qquad (2)$$

Here, Z is a random variable counting the number of samples k that show a PESQ value above λ, and N is the total number of measured and evaluated samples within all conducted calls. By means of this model we can calculate a measure of integrity, i.e., confidence intervals that show the goodness of our estimation and therefore the feasibility to compare different network providers with each other. In the following we will use a significance level of 0.05 to construct confidence intervals. This means that our estimated value is covered by the interval in 95% of all attempts; this interval is constructed by a specific method (confidence interval calculation is again an estimation).

Several ways to construct confidence intervals have been proposed in literature [9], [10]. Agresti and Coull constructed a confidence interval which has the advantage that its boundaries do not fall below zero or exceed one, independent of the numbers of successes or the sample size. The upper ($\hat{p}_{\text{UB}}$) and lower ($\hat{p}_{\text{LB}}$) bound are given by

$$\hat{p}_{\text{UB,LB}} = \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2N} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N} + \frac{z^2_{\alpha/2}}{4N^2}}}{1 + z^2_{\alpha/2}/N}, \qquad (3)$$

where $z_{\alpha/2}$ is the inverse standard normal distribution at half the significance level $\alpha/2$. In this section we implicitly assumed that successive samples are statistically independent. Because of the fact that a usual phone call lasts longer than the length of one speech sample, ETSI proposes a method to evaluate speech quality on an entire call in [8].

B. Speech Quality per Call

Mainly due to the dynamic channel conditions the evaluation of speech quality per call introduces some effects, especially in case of drive-testing, that are not covered by a ”per sample“ evaluation. The temporal structure, the inter sample correlation and the distribution of bad samples on a call (concentrated on some calls or distributed on many calls) are disregarded in this case. ETSI proposes a method to evaluate the speech quality of an entire call ($\mathrm{T_c}$ = 100 to 120 s) based on single speech quality samples

[1] RxQual: Reception Quality (based on bit error rates), RxLev: Reception Level (based on field strength of serving cell), CPICH: Common Pilot Channel, E_c: Chip energy, RSCP: Received Signal Code Power

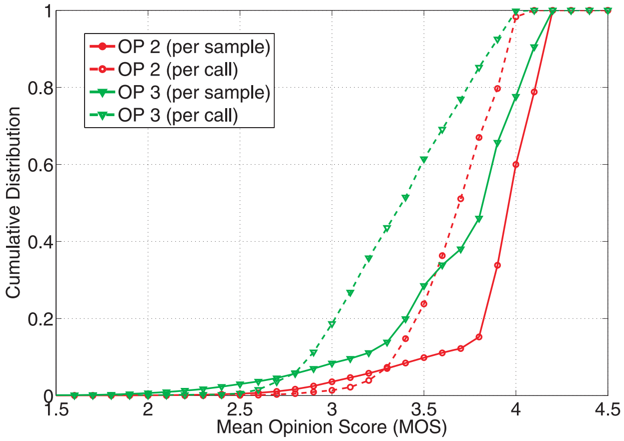


Fig. 2. CDF of Speech Quality (live network measurement data evaluated per sample and per call for equal threshold λ for both cases)

(T_s = 10 to 12 s). The up- and downlink speech sample transmission is alternating (uplink is in silent mode during downlink testing), which is taken into account by utilizing five speech samples for each direction. Furthermore, the following main effects are considered:

- ”Recency effect“: models the impact of the position of bad samples within a call
- Impact of a very bad speech quality sample within a call

It could be shown that bad speech quality samples have a stronger impact on the overall call quality rating when they are located close to the end of a call. This ”recency“ effect is taken into account and the overall MOS score for an aggregation of five speech quality samples is calculated by [8]

$$MOS_{\text{RE}} = \frac{\sum_{t=1}^{n} a_t MOS_t}{\sum_{t=1}^{n} a_t}, \tag{4}$$

where n is the number of speech samples (here: 5) and the coefficients a_t increasingly weight the successive MOS results towards the end of the call ($a_1 = 0.4, a_2 = 0.5, a_3 = 0.6, a_4 = 0.8, a_5 = 1$) [8]. The second effect (impact of the sample with the worst speech quality evaluation) is included in the model as well which results in the final MOS score for one call (either for uplink or downlink) [8]:

$$MOS_{\text{SpQ-C}} = MOS_{\text{RE}} - \frac{2}{5}(\overline{MOS} - \min_{t\in[1,5]}(MOS_t)), \tag{5}$$

with the average MOS score $\overline{MOS}$. This method shows a correlation of 85% with empirical data [8]. Again, a ratio can be defined that now counts the number of calls that are above a certain threshold λ rather than the number of samples that fulfill this condition. As has been done in the preceding section, the integrity of the ”per call“ evaluation can be calculated by means of confidence intervals. The method remains the same and therefore Eq. 3 can be used in the same way as with the ”per sample“ evaluations. Apparently, the number of events is heavily reduced. Two successive calls rather than two successive samples are considered to be statistically independent and therefore the confidence interval lengths will increase by a factor of approx. $\sqrt{T_c/(2T_s)}$ leading to a lower sensitivity of a statistical ranking between two or more operators.

Similar to the ”per sample“ method, a speech quality level can be given by either a certain metric of a PDF or CDF (see Fig. 2 ”per call“ method) or by means of a ratio like in Eq. 1 and in Fig. 3. As shown in Fig. 2 the evaluation ”per call“ is worse than for its ”per sample“ counterpart. In this case ”bad“ samples are likely distributed over many calls such that the call evaluation rates a call as ”bad“ rather than ”good“. If a large fraction of ”bad“ samples were concentrated on only a few calls, the ”per call“ evaluation would have been better than in the ”per sample“ case.

III. Benchmarking

In the last section methods were proposed to evaluate the level of speech quality in (wireless) telecommunication networks. For benchmarking purposes a comparison between the results of two or more independent network providers is necessary. In this section some possible methods for hypothesis testing are briefly reviewed and an alternative formulation for an exact Binomial Test is given.

A. Ranking based on PDF/CDF

As described before, a ranking based on PDFs/CDFs can just be obtained by comparing one or a combination of several metrics (e.g. the expectation value in combination with the 10% quantile). Essentially, the differences of such metrics between two or more operators have to relate to a deviation in user perception which has to be known before a reliable statement about this difference can be given. Alternatively, approaches are described which are based on statistical assumptions in the following section.

B. Ranking of Binomial Quantities

Other than in the last paragraph, we now consider proportions, i.e., ratios of ”good“ speech quality samples or calls (PESQ $> \lambda$). In the following, different hypothesis tests are described. The χ^2 and *Fisher's exact test* are widely used, if two or more groups of rates are to be compared. Additionally, an alternative test that is directly based on the binomial distribution is presented. The objective is to find out if there is enough evidence that the null-hypothesis H_0 (the performance of two competitors does not differ significantly and is not independent) can be rejected in favor of the alternative hypothesis H_a (the performance of two competitors differs significantly and is independent). The significance level throughout this section is 0.05.

1) χ^2 Test: This test assesses whether paired observations on two variables (No. of ”good“ speech samples/calls, No. of ”bad“ samples/calls), expressed in a contingency table, are independent of each other. In the 2×2 case, the χ^2 test is equivalent to the *fourfold test*, whose χ^2 test statistic results in (using Tab.1):

$$\chi^2 = \frac{(N_1+N_2)(k_2N_1-k_1N_2)^2}{N_1N_2(k_1+k_2)(N_1+N_2-k_1-k_2)}. \tag{6}$$

If the test statistic exceeds a value of 3.841 (χ^2 distribution with 1 degree of freedom at 0.95), H_0 will be rejected in favor of the alternative and the populations of Operator 1 and Operator 2 are defined as independent with significance level 0.05.

	"good" samples/calls	"bad" samples/calls
Operator 1	k_1	N_1-k_1
Operator 2	k_2	N_2-k_2

TABLE I

2×2 CONTINGENCY TABLE

Since the χ^2 test belongs to the class of asymptotic tests (test statistic is asymptotically χ^2 distributed) it needs a minimum number of observations in the contingency table to be reliable. The number of observations have to exceed 5. For smaller numbers exact tests have to be utilized, which are usually more complex.

2) Fisher's Test: Fisher's exact test is used for 2×2 contingency tables, in case the total number of observations is less than 20 or there is a cell-value less than 5. The probability of the exact arrangement as shown in Tab. I is [11]

$$p = \frac{\binom{N_1}{k_1}\binom{N_2}{k_2}}{\binom{N_1+N_2}{k_1+k_2}}. \tag{7}$$

Let us assume that the null hypothesis is true, which means that k_1 and k_2 are chosen out of the same population. The objective is now to find the probability of the arrangement in Tab. I and of more extreme cases (in terms of a greater inequality between k_1/N_1 and k_2/N_2). If this sum of probabilities is below the significance level, the null hypothesis is rejected. Due to the factorial involved in the test statistic, the evaluation gets more complex for a high number of observations and asymptotic tests will be more attractive.

3) Binomial Test: Another exact test can be stated which directly uses the binomial probability as in Eq. 2:

$$P(Z=k|N,a) = \binom{N}{k} a^k(1-a)^{N-k}, \tag{8}$$

with the random variable Z whose observations k are the number of "successes" or samples/calls that show a PESQ value greater than λ, respectively. The variable N is the overall number of samples/calls and a denotes the probability of a "success". The variable a can be bounded depending on the application. For speech quality evaluation the usual proportion of k/N is between 0.8 (a_{min}) and 1 (a_{max}). No further a-priori knowledge is provided, thus a uniform distribution between a_{min} and a_{max} is assumed as its PDF:

$$p(a) = \begin{cases} \frac{1}{\Delta a} & a_{\text{min}} \le a < a_{\text{max}} \\ 0 & \text{else} \end{cases}, \tag{9}$$

where $\Delta a = a_{\text{max}} - a_{\text{min}}$. The joint PDF of a and k is given by

$$\begin{aligned} p(a,k|N) &= P(k|N,a)\cdot p(a) \\ &= \binom{N}{k} a^k(1-a)^{N-k}\frac{1}{\Delta a} \\ &\forall \quad a_{\text{min}} \le a < a_{\text{max}}. \end{aligned} \tag{10}$$

The conditional PDF of a given k follows readily

$$\begin{aligned} p(a|k,N) &= \frac{p(a,k|N)}{P(k|N)} \\ &= \frac{1}{P(k|N)\Delta a}\binom{N}{k} a^k(1-a)^{N-k} \\ &\forall \quad a_{\text{min}} \le a < a_{\text{max}}. \end{aligned} \tag{11}$$

With the conditional PDF we can now define the probability $P(a_1 > a_2)$ that Operator 1 (sample size: N_1, number of "good" samples/calls: k_1, success probability a_1) performs better than Operator 2 (N_2, k_2, a_2). Assuming that a_1 and a_2 are independent leads to the probability that $a_1 > a_2$:

$$\begin{aligned} &P(a_1 > a_2|N_1,N_2,k_1,k_2) = \\ &\int_{a_{\text{min}}}^{a_{\text{max}}}\int_{a_{\text{min}}}^{a_1} p(a_1,a_2|N_1,N_2,k_1,k_2)\mathrm{d}a_2\mathrm{d}a_1 = \\ &\int_{a_{\text{min}}}^{a_{\text{max}}}\int_{a_{\text{min}}}^{a_1} p(a_1|N_1,k_1)p(a_2|N_2,k_2)\mathrm{d}a_2\mathrm{d}a_1. \end{aligned} \tag{12}$$

For probabilities $P(a_1 > a_2) > 0.95$, Operator 1 is defined to have a greater "success" probability than Operator 2.

IV. RESULTS

Following the approaches given in the preceding sections, we will provide results of live network quality measurements that have been obtained in the second half of 2006.

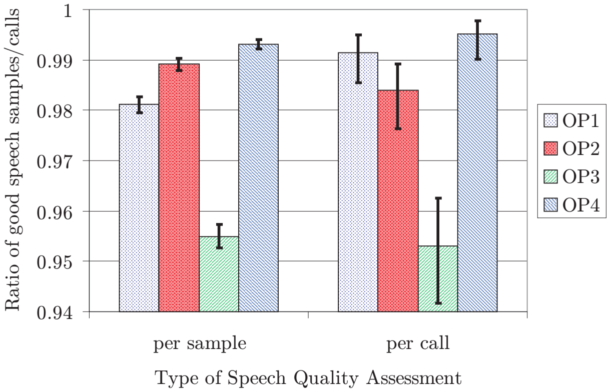
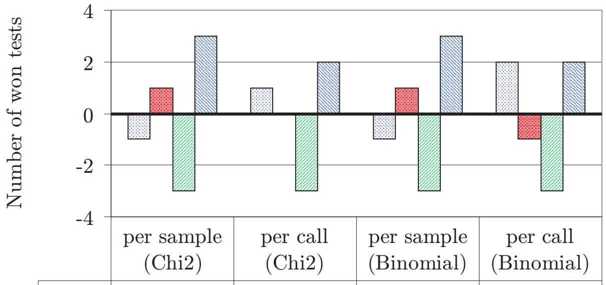


Fig. 3. Ratio based evaluation of speech quality per sample and per call including confidence intervals ($\alpha = 0.05$)

In Fig. 3 the speech quality evaluation of the per sample and per call approach is compared. Since it has been assumed that all samples (in this example approx. 30'000) are statistically independent, the confidence interval lengths are very small, thus leading to a high sensitivity in terms of ranking. On the



	per sample (Chi2)	per call (Chi2)	per sample (Binomial)	per call (Binomial)
░ OP1	-1	1	-1	2
🟥 OP2	1	0	1	-1
🟩 OP3	-3	-3	-3	-3
▨ OP4	3	2	3	2

Fig. 4. Pairwise ranking with χ^2 and binomial test

other hand, the per call evaluation just assumes that two calls in succession are independent rather than two samples within a call. The lower number of events (in this example approx. 1'600 calls) leads to much larger confidence intervals and hence to a lower ranking sensitivity.

The results of a pairwise ranking between four operators is shown in Fig. 4 for different types of evaluation. For each test two operators are compared. In case of a significant difference in performance (ratio of "good" samples/calls to the overall number of samples/calls) the winning operator gets one point, whereas the other operator loses one point. If no significant difference is found, no points will be given at all. Therefore, a maximum of +3 (all tests won) and a minimum of -3 points (all tests lost) can be achieved in our example. Obviously, the ranking gives a slightly different picture in the "per call" and "per sample" case. The ranking of Operator 1 and Operator 2 based on the "per sample" approach deviates from the results of the ranking based on the "per call" approach. Additionally, it is shown that the binomial test detects a significant difference between Operator 1 and Operator 2 (per call evaluation), whereas the χ^2 test does not detect such a difference.

k_2	χ^2	Fisher	Binomial
1933	0	0	0
1932	0	0	1
...	0	0	1
1928	1	0	1
1927	1	1	1

TABLE II

TEST OF INDEPENDENCE WITH $N_1 = 2000$, $N_2 = 2000$, $k_1 = 1950$

In the following Tab. II all hypothesis tests which are based on ratios are applied on a set of parameters. A one indicates the rejection of the null hypothesis and thus the independence of two sample sets. Apparently, the binomial test indicates independence for $k_2 = 1932$ and a difference between k_1 and k_2 of 18. The χ^2 test follows at a difference of 22 and Fisher's exact test at a difference of 23. The observation that the binomial test indicates independence for the smallest difference of the tested ratios applies in general.

V. Conclusion

In this paper we showed different methods how speech quality measurements can be evaluated and benchmarked. Since two separate calls in a drive-test are more likely to be uncorrelated rather than samples, an evaluation of speech quality "per call" should be preferred. This approach is much better in line with the statistical model of an accumulation of Bernoulli experiments. The assumption of a Binomial distribution for the number of "good" speech samples or calls led to the definition of confidence intervals. To be consistent with this model we suggest to use the exact Binomial test for pairwise benchmarking or ranking, respectively. Even though, a significant difference in performance of two operators can be shown statistically, this does not necessarily result in a change of user quality perception, but gives a first indication. The application of several benchmarking methods have been demonstrated specifically for speech quality. Furthermore, these methods can also be applied on any kind of proportion, e.g., call drop, setup success or service availability rates.

Acknowledgment

The authors would like to thank T-Mobile International for their support and the provision of live network drive-test data.

References

[1] ITU-T Rec. P.800, "Methods for objective and subjective assessment of quality", International Telecommunication Union, Geneva, Aug. 1998.

[2] ITU-T Rec. P.861, "Objective quality measurement of telephone-band (300-3400 Hz) speech codecs", International Telecommunication Union, Geneva, Feb. 1998.

[3] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", International Telecommunication Union, Geneva, Feb. 2001.

[4] ITU-T Rec. P.862.1, "Mapping function for transforming P.862 raw result scores to MOS-LQO", International Telecommunication Union, Geneva, Nov. 2003.

[5] Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P., "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Proceedings, vol.2, no.pp.749-752, 2001.

[6] Werner, M.; Kamps, K.; Tuisel, U.; Beerends, J.G.; Vary, P.; "Parameter-based speech quality measures for GSM", *IEEE Proceedings on Personal, Indoor and Mobile Radio Communications*,vol.3, no.pp. 2611- 2615 vol.3, Sept. 2003.

[7] ETSI Technical Specification TS 102 250-2, "Technical Specification Speech Processing, Transmission and Quality Aspects (STQ); QoS aspects for popular services in GSM and 3G networks; Part 2: Definition of Quality of Service parameters and their computation", V1.4.1, 2006.

[8] ETSI Technical Report TR 102 506, "Speech Processing, Transmission and Quality Aspects (STQ); Estimating Speech Quality per Call", V1.1.1, Oct. 2006.

[9] Agresti, A. and Coull, B. A., "Approximate is better than "exact" for interval estimation of binomial proportions", *The American Statistician*, vol.52, no.2, pp.119-126, 1998.

[10] Brown, L.D.; Cai, T.T.; DasGupta, A., "Interval Estimation for a Binomial Proportion", *Statistical Science*, vol. 16, no. 2, pp. 101-133, 2001.

[11] Fisher, R.A., "On the interpretation of χ^2 from contingency tables, and the calculation of P", *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87-94, 1922.