Monitoring the User Perceived Quality of SILK-Based Voice Calls

Daniel Schlosser, Michael Jarschel, Valentin Burger, Rastin Pries
University of Würzburg, Institute of Computer Science,
Chair of Communication Networks, Würzburg, Germany
Email: {schlosser,michael.jarschel,burger,pries}@informatik.uni-wuerzburg.de

***Abstract*—Quality measurements are required to support VoIP traffic in the Internet. The widely used average Mean Opinion Score is, however, not sufficient for this task. In this paper, we provide a detailed analysis of the Skype SILK codec and compare it with the iLBC and GSM codec. The results show that the SILK codec is superior to the other codecs in scenarios with random and bulk packet loss. This increased tolerance of packet loss enables the option of QoE Monitoring under reasonable network conditions. From analysis, we derive an estimation, which can be used to monitor the MOS value of the users in the worst case. Furthermore, we show how sampling can effectively decrease the required measurement effort.**

***Index Terms*—QoE, Monitoring, SILK, Skype, Sampling**

I. Introduction

In the last few years, VoIP has come into its own. It is now used at universities, small and medium enterprises, and even in large companies on a professional level. With the deployment in these environments, it is necessary to guarantee the same or even better quality than the end user is accustomed to from classical telephone services. However, transmitting speech in packet switched networks has far more influence factors compared to a circuit switched telecommunication network. Hence, in the last years many studies have analyzed the influence of degraded Quality of Service (QoS) in packet switched networks on the user perceived quality. To enable this research, tools and models like PESQ [1] have been developed, which reliably calculate the Mean Opinion Score (MOS) the user would experience, considering the input signal and the output of the transmission. Further analyses with these tools have shown that these and other technical influence factors are correlated. For example, the influence of jitter and packet loss depend on the used codec.

As a result, there are many publications, e.g. [2]–[4], which look at different codecs and analyze the influence of certain Quality of Service (QoS) parameters on the Quality of Experience (QoE) in terms of the MOS. The models that are derived from these investigations consider mostly the average MOS. However, different realizations of the same QoS may lead to completely different MOS values. This is reasonable, because losing a single part of a transmission is perceived completely different to a constant noise, although the loss percentage is equal in both cases. Hence, results on the average MOS are not sufficient, if we have to guarantee a certain QoE for e.g. 95% of the users, as it is required in a professional environment. In this case, more detailed models are necessary.

In this paper, we address the monitoring problem described above. We focus on the SILK codec, the standard codec used by Skype, published in March 2010. In order to monitor the users' QoE, i.e., to make sure that under given network conditions a predefined percentage of all users perceive a good voice transmission quality, we establish a mapping between the QoS and a distribution describing a worst case assessment of the perceived MOS values. We present a detailed analysis of how the QoS influences the QoE perceived by the user. Therefore, we consider the impact of different loss patterns and show how equivalent loss patterns affect different transmissions durations. Combining these results, we demonstrate how the QoE perceived by the user can be safely monitored. Finally, we consider sampling techniques, which allow us to decrease the measurement effort drastically.

The paper is structured as follows. In Section II, we discuss the related work and our evaluation approach is described in Section III. We present the results for the SILK codec in Section IV and derive a model to assess the quality perceived by the user in Section V. Sampling is considered in Section VI. Finally, we draw conclusions in Section VII and give a brief outlook on future work.

II. Related Work

Many publications consider the user perceived quality of VoIP transmissions. Hence, in the following we give an overview of the most interesting areas citing some exemplary publications.

In [5] Deri published his work on an open source software to monitor VoIP traffic. The software, which has been further developed over the last years, is able to detect VoIP flows and export the flow characteristics using the netflow/IPFIX protocol. It does not consider the QoE of the monitored flows.

Other publications compare different VoIP applications considering different network characteristics. For example Chiang et al. [3] present a comparison between MSN and Skype considering the available bandwidth, packet loss, cross traffic, and NAT scenarios. One thing, which makes this publication special is that within this publication real user surveys have been conducted. In [2] Barbosa et al. compare Skype to

This work was funded by the Federal Ministry of Education and Research of the Federal Republic of Germany (Förderkennzeichen 01BK0917, GLab). The authors alone are responsible for the content of the paper.

Google Talk. Besides the effects of network QoS parameters on the QoE perceived by the user, this paper focuses on the strategies these applications use to cope with degraded network conditions, e.g., loss and low available bandwidth. However, both publications do not provide a model mapping the QoS of the network and the QoE perceived by the user. In order to assess the perceived quality, the authors use the PESQ [1] tool characterized by Rix [6]. Pennox [7] analyzes the accuracy of the PESQ tool and shows that the calculated MOS values are not necessarily precise. Considering these results, an additional gap between the targeted lower MOS values and the results of the PESQ tools can be introduced.

Chen et al. [8] provide such a model. It is based on the assumption that the call duration and the QoE are correlated. In the paper, they develop a model that can estimate the mean QoE a user perceives under a given network QoS. It is claimed that the latency of the network between the communication partners might influence the QoE of a user. However, this impairment does very seldom decrease the QoE in a way that the user would quit the call. They show that if there where disturbances in the network, which made the user hang up, in 46% this was caused by a low bit rate of the transmission. In 53% the hang up was caused by jitter, and only in 1% of all cases it was caused by the network latency. However, the model does not contain information about the QoE range different users may perceive.

Sat and Wah summarize their work in [9], [10]. Besides their work on coding schemes for VoIP transmissions and special features a peer-to-peer VoIP network should implement, they discuss in detail how latency affects the QoE of VoIP conferences with two or more partners. They conclude that there are currently no models which can correctly map these latency issues to the user perceived QoE.

The hypothesis of an exponential interdependency between QoS and QoE is postulated by Hoßfeld et al. [4]. The authors study the impairment of loss and jitter in the network and proof that there is an exponential mapping between random loss and QoE.

Our work is inspired by the last three publications, but extends their work in many ways. First of all, we focus on the wideband codec SILK, which is used since Skype version 4.0 and was revealed to the public as an IETF draft in March 2010. We consider random and bulk loss as well as different speech durations to provide a model which can be used to monitor the QoE. It considers not only the medium MOS but also the range in which different realizations of the same loss percentage are perceived by the user.

III. Evaluation Approach

The quality of a voice transmission is affected by many different parameters. If we follow the voice signal from speaker to listener, we find the following factors, which might influence the quality of the transmission, cf. Fig. 1. The first factor is the quality of the microphone and the analog-digital-converter. A cheap microphone, digital-analog-converter, or speaker will introduce noise to the system, which has nothing

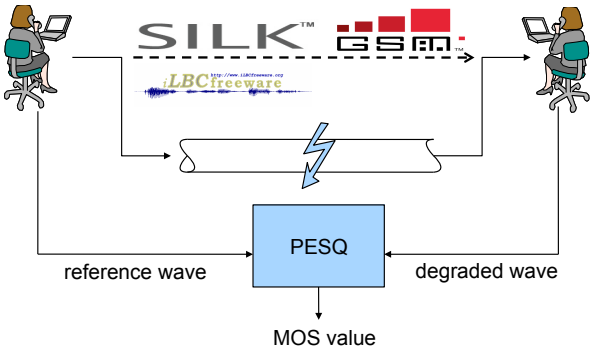

Fig. 1. VoIP quality evaluation set up

to do with the transmission of the voice stream itself. In our work, we focus on the influence of the network QoS on transmissions using a given codec. Therefore, we try to eliminate these impairments and use sentences of studio audio productions with CD-quality for our evaluation. We consider six different speech samples including English and German speakers. These samples include a high and a low voice from men and women.

The next influence factor is the codec, which is used to convert the digitized voice into a binary stream. We compare the quality of the well known GSM codec and the former Skype codec iLBC with the Skype SILK codec. Skype calls SILK an 'ultra wideband audio codec', c.f. [11], which means that this codec should achieve high QoE and should be less affected by network degradations.

After the voice sample has been encoded, the stream is cut into frames, which are transmitted over the IP network. Different network QoS parameters, e.g. packet loss, jitter, delay, and bandwidth restrictions, affect the quality perceived by the user. The two main influence factors according to [8] are packet loss and jitter. Although these factors are completely different on the network level, their effects are the same. This is, because all modern VoIP systems integrate a play-out or jitter buffer. The behavior of this buffer is very simple. Whenever a packet arrives at the destination, the information in the packet is extracted and stored in this buffer. The replay of the voice stream is initially delayed for a short time. Hence, the buffer enables smooth replay of packet streams, which suffer from jitter, if all frames reach the buffer before they are accessed by the replay function. However, if the replay algorithm accesses a buffer place, for which the frame has not yet arrived, the algorithm considers this frame to be lost. Therefore, jitter and loss can be considered to have the same consequences from the application layer perspective. From the monitoring perspective, we can conclude that it is necessary to check whether a frame arrives in time at the destination to cover jitter and packet loss. Other influences of the network are bandwidth bottlenecks and latency. Bandwidth bottlenecks, which reduce the available bandwidth below the transmission rate of the sender, are detected and Skype reacts with adapting

the transmission and bit rate of the codec. The influence of latency is not yet researched in detail, especially for multi-user voice conferences, as these influences are not completely clear yet. At the moment, there are no tools to map these influences to MOS values. Therefore, we focus on loss and jitter and leave bandwidth restrictions and latency to future work.

In order to analyze the influence of the network on the QoE, we encode the original .wav file with the different codecs using their default parameters for sampling and encoding bit rate. The resulting bitstreams are saved and packet loss is applied by removing the information from selected frames. We decode the bitstream generated in the previous step and compare the resulting .wav files with the original ones using the PESQ tool [1]. This way, we have full control of what happens on the network layer. We can exactly determine, which frame is lost on the network and make sure that a considered transmission has exactly the number of lost frames, which we choose. Another advantage of this method is that we do not have to resynchronize the send and received .wav file, because we know exactly the beginning and the end of the transmission.

The PESQ tool, which we use to calculate the MOS value of each disturbed and undisturbed voice transmission, has a specified input voice duration between 5 and 30 seconds. In order to examine how an equivalent amount of loss is perceived over different time durations, we use voice excerpts of 5 seconds, 10 seconds, and 20 seconds length. Each sample contains a short silence period at the beginning and the end of the file. This is a requirement of the PESQ tool as explained in the PESQ application guide [12].

IV. Monitoring Silk QoE

Determining the user perceived QoE from measured QoS in the network requires a mapping between these two metrics. For voice traffic the tool PESQ provides such a mapping. It derives user QoE by comparing recorded voice tracks before and after a transmission via a network. The resulting QoE values are based on the MOS, which assigns numbers from one for unacceptable quality, over three for acceptable quality, to five for very good quality. We performed tests using PESQ on the influence of random and bursty packet loss on the voice codec SILK as well as its predecessors iLBC and GSM. It has to be noted that comparing two binary identical files with PESQ leads to a maximal score of 4.5.

Fig. 2 illustrates the test results for random packet loss. For each loss value, we considered 2000 different random loss pattern with exactly the same number of lost frames. In Fig. 2 the boxes give the inter quartile range of the MOS for each value of loss and the triangles show the maximum and minimum values. As a modern voice codec, which is optimized for higher bandwidths than iLBC and GSM, SILK achieves the highest PESQ-based MOS value of 4.5, when not exposed to any loss. iLBC and GSM score 0.9 and 1.3 points lower respectively. With increasing loss, the achieved MOS values decrease for all codecs. However, at 5.2% packet loss, 75% of the SILK connections still have an acceptable quality, i.e., the 25% quantile is above a MOS value of 3.0. Each of these connections are rated with a higher MOS score than 75% of the iLBC and all GSM connections. Although the range between the connection with the highest and the connection with the lowest MOS score increases rapidly, the inter quartile range stays relatively small with a range of about 0.35 MOS values. This shows that uniformly distributed loss affects many transmissions in a similar way. For Fig. 2 we only show to the results of the five second voice duration tests. The results for longer voice periods reveal slightly different quantile values, but mainly the same influences.

In contrast to uniformly distributed loss, which spreads the lost frames more or less equally over the complete transmission, we now consider bulk loss. In order to capture all effects of a burst with a predefined length n, we remove n consecutive frames at each possible position of the transaction. The resulting MOS values for the five second transmissions are shown in Fig. 3. While the main trend is the same as in the random loss case, we notice two main differences. First, for all considered loss values there are some transmissions, which seem to be undistorted, i.e., the maximum MOS value in all

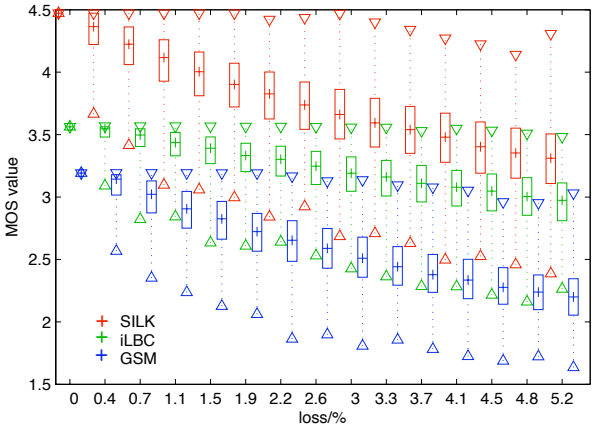

Fig. 2. QoE of GSM, iLBC, and Silk for increasing random loss

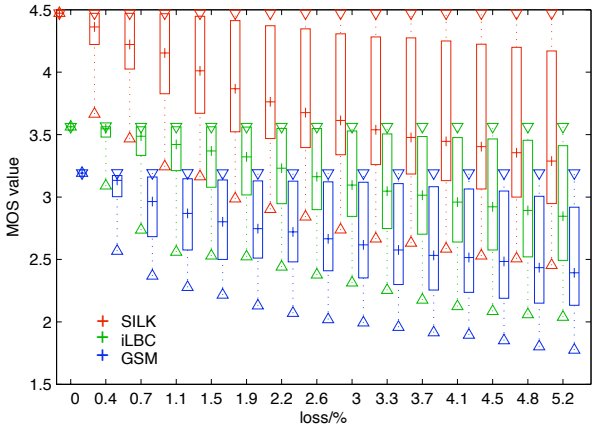

Fig. 3. QoE of GSM, iLBC, and Silk for increasing bulk loss

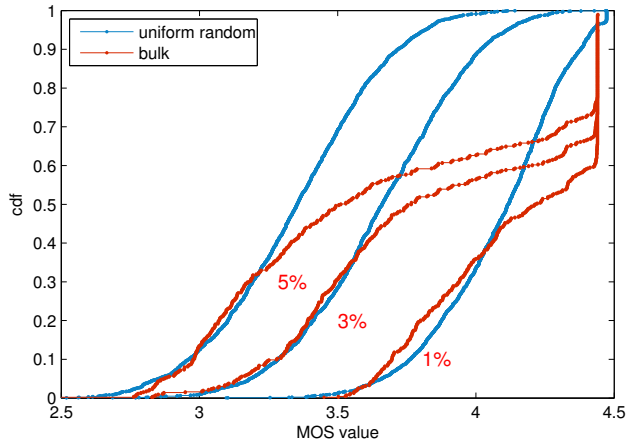

Fig. 4. QoE comparison between random and bulk loss for Silk Codec

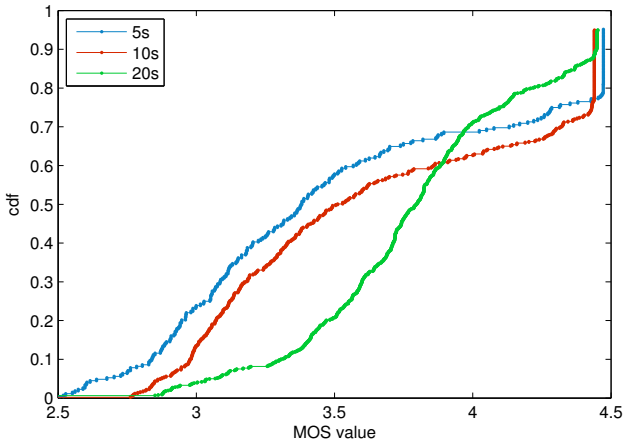

Fig. 5. Comparison of loss bursts to different voice transmission durations

cases is 4.5. This phenomenon can be explained by the silence periods at the beginning and the end of the transmissions, which are needed by the PESQ tool. If the lost frames fall within these periods, they do not disturb the transmission and the MOS value is unchanged. Secondly, we recognize that the inter quartile ranges increase in size. This means that the QoE differs more, if we consider bulk loss.

It is not surprising that SILK provides better quality in both scenarios. The interesting point is that SILK can provide MOS values above 3 for most of the customers, even if 2% of all packets get lost in the network. This feature makes it interesting for QoE monitoring. The other codecs provide only fair quality in an undisturbed network. This means that every network degradation leads to unacceptable quality, which makes them uninteresting for QoE monitoring.

In order to create a worst case assessment, which covers random and bulk loss, we need to compare these influences in more detail. Fig. 4 shows a comparison of the Cumulative Distribution Functions (CDFs) derived from the SILK PESQ measurements for one, three, and five percent of random and bursty loss. The plots of the burst loss show that burst loss leads to better user experience in most cases. For all other cases, the gap between the plots is quite small. Only in cases with low packet loss, i.e., the plots for 1% packet loss, it is clearly visible. However, in cases with low packet loss, the user perceived quality is relatively high. In the critical areas between a MOS of 3.5 and 2.5, a model describing the influence of random loss is sufficiently precise. Hence, we will use the influence of random loss for our monitoring solution, as it is easier to model and provides sufficient precision.

In the performance comparison of the different codecs and for deriving a suitable model we used the results, which we generated using five seconds of voice transmission. The cause for this is presented in Fig. 5. It provides a comparison of the MOS value CDFs for 5, 10, and 20 seconds of voice transmission given five percent bursty loss. It has to be noted that due to different voice transmission durations, different numbers of frames are consecutively removed at each point of the encoded file. Due to the higher frame number of a longer voice transmission, the CDFs of longer files contain more values, as we consider every possible starting point for the loss period.

For the lower 70% of all results, the plot describing the results of the 5 second voice duration shows the lowest MOS values. Only for the upper 30% of the resulting MOS values, the results from the 20 seconds file predict lower QoE values. This is because the silence periods at the beginning and the end of the files have to be considered. In a file containing 20 seconds voice, the relative part of the silence is smaller than in a 5 second file. However, the MOS values in this area are all above 3.9 and therefore not critical for our monitoring solution. Thus, we are able to use the model for the influence on a five second voice file for our monitoring strategy.

V. Modeling the QoE of a SILK Transmission

In the previous section, we concluded that we can use the MOS distribution for a certain level of random loss as a worst case assessment of the user perceived quality for a QoS to QoE mapping. Hence, we now take a closer look on how to analytically describe these distributions.

Fig. 6 depicts the CDFs of our measured MOS distribution for increasing random loss in bright lines. The dark thin curves show the results, which we can achieve with a normal distribution, for which we adapted the mean value and the standard deviation to fit the corresponding measurement results. The graphs mostly overlap, as the mean squared relative error between the measured results and the approximation is below $1.3 \cdot 10^{-3}$ for all fits.

In order to use these estimations for monitoring, we need a model which is able to map the measured loss to the corresponding mean values and standard deviations. We consider linear, quadratic, exponential, and radical functions for fitting. Table I presents the Mean Squared Errors (MSEs) between the considered approximations and mean values. We see that the exponential fit outperforms the other.

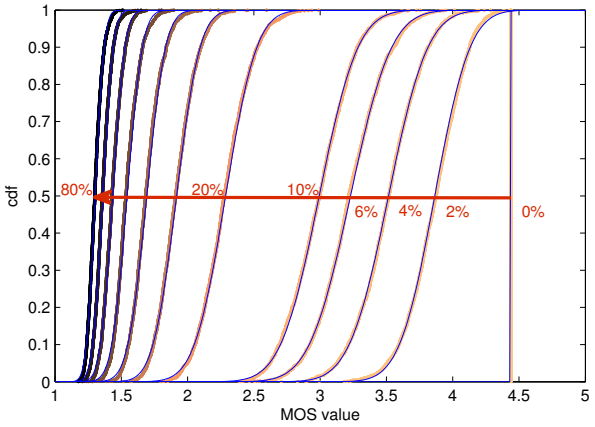

Fig. 6. Estimating the QoE of random loss with normal distributions

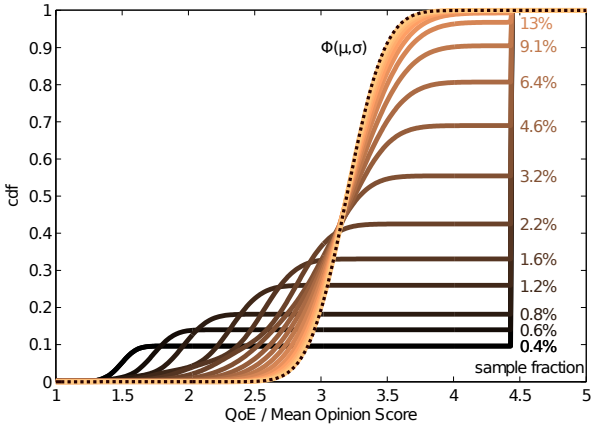

Fig. 7. Comparison of QoE estimations for an increasing number of samples

TABLE I
MSE OF DIFFERENT FUNCTIONS FITTED TO MEASURED MEAN VALUE

function	MSE	normalized MSE
exponential	0.0016	0.0038
radical	0.0126	0.0296
quadratic	0.0159	0.0376
linear	0.0853	0.2011

For modeling the standard deviation, we applied the same models. Again, the fit with an exponential function matches well. However, a quadratic fit is even better for modeling the standard deviation. We could not find any good fit with a radical function. Table II denotes the mean squared error for the best approximations of the standard deviation. We conclude from the low MSE values that we can model the worst case assessment of the QoE with normal distributions using the exponential function fit presented in Equation 1 for the mean value and an exponential or quadratic function fit given in Equation 2 and 3 for the standard deviation.

$$2.415\exp(-0.05332x) + 1.328 \tag{1}$$

$$0.268\exp(-0.01957x) - 0.009182 \tag{2}$$

$$2.652 * 10^{-05}x^2 - 0.004668x + 0.2563 \tag{3}$$

VI. SAMPLING

In the previous sections, we proposed a model to assess the QoE of a SILK call. The monitoring effort for this is quite high as we have to monitor each packet stream and look for packets that will not arrive at the destination in time. To lower the needed effort, sampling can be used. We focus on classical n out of N sampling, which means that we only analyse n random packets out of a group of N transmitted packets.

TABLE II
MSE OF DIFFERENT FUNCTIONS FITTED TO MEASURED STANDARD DEVIATION

function	MSE	normalized MSE
quadratic	5.9841e-05	0.0165
exponential	7.1722e-05	0.0198
linear	2.1937e-04	0.0606

In theory, this sampling method can be understood by looking at the classical urn model. If we consider a loss probability of p, this means that we draw randomly independent n balls out of an urn with N balls. $M = p \cdot N$ balls are red, which means the sampled packet does not arrive in time or at all. $N-M$ are green, which models that the corresponding packet would arrive in time. Dependent on the number of samples n, different outcomes are possible. If we consider drawing $n = 2$ balls, we could draw 2 green, a green and a red, or two red balls. We would interpret the outcome of this results as 0% loss, 50%, and 100% loss respectively. Considering the proposed MOS assessment, this yields:

$$\begin{aligned} P(MOS = x) =& \sum_{i=0}^{n} P(V = i)P(MOS = x|V = i) \\ =& \delta(x - \mu_{n,0})h(0|N, M, n) \\ &+ \sum_{i=1}^{n} h(i|N, M, n)N(\mu_{n,i}, \sigma_{n,i})(x), \end{aligned} \tag{4}$$

where V ist the number of packets considered as lost, δ is the Dirac Impulse, h is the hypergeometric distribution, and N is the normal distribution assessment from Section V.

The resulting CDFs are shown in Fig. 7. Brighter colors are related to more sampled packets n and the black dashed graph denotes the target function of the model we want to approximate. If we consider only a small fraction of samples, i.e., $\frac{n}{N} < 1\%$, we can distinguish two areas. In the part right of the target function, representing 80% of all sampling outcomes, the result overestimates the MOS values. In the left part, the MOS values are clearly underestimated. This reflects the example presented before. Either the sampling suggests perfect quality or it overestimates the loss and underestimates

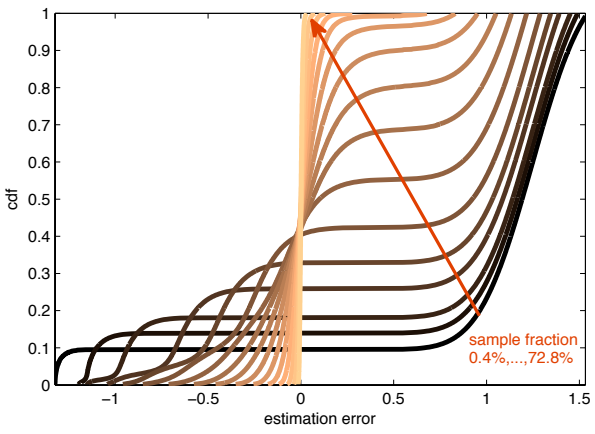

Fig. 8. CDF of the error caused by sampling

the QoE. For fractions $\frac{n}{N} > 10\%$ the accuracy increases.

To analyze the precision of the n out N sampling, we consider the difference of the quantiles of the original distribution and the results of Equation 4. Fig. 8 depicts these differences for $p = 5\%$ loss. We see that the absolute error is decreasing for an increasing number of samples. If we want to decrease the estimation error below a value of, e.g., 1%, we would need a sampling ratio $\frac{n}{N} > 25\%$. More details on the 1%, 5%, 25%, 50%, 75%, 95%, and 99% quantiles for our 5% loss example are given in Table III. It has to be noted that, due to the structure of Equation 4, results can only be calculated numerically.

TABLE III
QUANTILES OF THE ABSOLUTE ERROR DISTRIBUTION FOR DIFFERENT SAMPLING RATES

Fraction	1%	5%	25%	50%	75%	95%	99%
0.4%	-1.3234	-1.3111	1.0173	1.1918	1.3361	1.4852	1.5296
0.8%	-1.1156	-0.9734	0.9125	1.1399	1.2853	1.4122	1.4436
1.2%	-1.0016	-0.7424	-0.3993	1.0875	1.2409	1.3586	1.3851
2.2%	-0.7536	-0.6004	-0.2403	0.9238	1.1443	1.2608	1.2835
4.6%	-0.4836	-0.3624	-0.1323	0.0615	0.8984	1.1017	1.1284
9.1%	-0.2866	-0.2124	-0.0813	0.0225	0.1463	0.8676	0.9387
18.3%	-0.1436	-0.1072	-0.0436	0.0069	0.0623	0.1553	0.3065
36.4%	-0.0596	-0.0447	-0.0192	0.0019	0.0257	0.0643	0.0945
51.5%	-0.0326	-0.0245	-0.0106	0.0010	0.0143	0.0365	0.0545
72.8%	-0.0127	-0.0096	-0.0041	0.0006	0.0060	0.0153	0.0227

VII. Conclusion and Outlook

Using VoIP services on a professional level requires to monitor the quality perceived by the user. The average MOS is not sufficient for this task. More precise models are necessary, which also consider the complete spectrum of user perceived QoE caused by the same kind of network degradation.

In this paper, we provide a detailed analysis of the Skype SILK codec and compared it to its successor iLBC and GSM using an analysis based on the PESQ tool.We analyzed bulk and random loss by applying error patterns directly to the encoded VoIP frames. From the results, we showed that SILK provides a better QoE in all considered cases. Furthermore, we found out that for the SILK codec the impact of random loss is more severe and that distributions modeling these results can be used as a worst case assessment. We studied different speech durations. We revealed that equivalent loss percentages lead to a stronger degradation of the perceived quality, if they are applied to shorter speech transmission. Thus, we use models for short speech transmissions, in order not to overestimate the QoE.

We modeled these worst case assessments using a normal distribution and presented formulas to derive all necessary parameters from the loss measured in the network. Finally, we demonstrated that sampling can be used to decrease the monitoring effort. We provided numbers for the precision of different sampling rates, so that the sampling rate can be chosen to fit the needed precision of the monitoring system.

In future work, we want to implement the results of this paper in a monitoring system for VoIP virtual networks. Therefore, we will extend the model to consider bit rate adaptations, parameters, and aggregation, which enables us to monitor links with many VoIP flow as a whole. Furthermore, we want to design a flexible monitoring architecture, which adapts the measurement effort and precision at different points of the network. The goal is to build a network management control, which monitors and proactively adapts the network to guarantee high voice quality for all users.

Acknowledgment

The authors would like to thank Prof. Tran-Gia for the fruitful discussion and support on this work.

References

[1] ITU-T Recommendation, "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*, 2001.

[2] R. Barbosa, C. Kamienski, D. Mariz, A. Callado, S. Fernandes, and D. Sadok, "Performance Evaluation of P2P VoIP application," in *ACM NOSSDAV*, 2007.

[3] W. Chiang, W. Xiao, and C. Chou, "A Performance Study of VoIP Applications: MSN vs. Skype," in *MULTICOMM*, 2006.

[4] T. Hoßfeld, D. Hock, P. Tran-Gia, K. Tutschku, and M. Fiedler, "Testing the IQX Hypothesis for Exponential Interdependency between QoS and QoE of Voice Codecs iLBC and G.711," in *18th ITC Specialist Seminar on Quality of Experience*, 2008.

[5] L. Deri, "Open Source VoIP Traffic Monitoring," *SANE*, 2006.

[6] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) - a New Method for Speech Quality Assessment of Telephone Networks and Codecs," in *IEEE International Conference on Acoustic Speech and Signal Processing*, 2001.

[7] S. Pennock, "Accuracy of the perceptual evaluation of speech quality (pesq) algorithm," in *MESAQIN*, 2002.

[8] K. Chen, C. Huang, P. Huang, and C. Lei, "Quantifying skype user satisfaction," in *Applications, technologies, architectures, and protocols for computer communications*, 2006.

[9] B. Wah and B. Sat, "The Design of VoIP Systems With High Perceptual Conversational Quality," in *Ubiquitous Multimedia Computing*, 2009.

[10] B. Sat and B. Wah, "Analyzing Voice Quality in Popular VoIP Applications," in *IEEE MultiMedia*, 2009.

[11] Skype Limited, "Silk: Super wideband audio codec," last accessed May, 27th, 2010. [Online]. Available: http://developer.skype.com/silk

[12] ITU-T Recommendation, "P. 862.3," *Application Guide for Objective Quality Measurement Based on Recommendations P.862, P.862.1 and P.862.2*, 2007.