Computer Standards & Interfaces 36 (2014) 626–630

Contents lists available at ScienceDirect

Computer Standards & Interfaces

journal homepage: www.elsevier.com/locate/csi

Evaluation of objective speech transmission quality measurements in packet-based networks

Oldřich Slavata *, Jan Holub

Dept. of Measurement, Faculty of Electrical Engineering, Czech Technical University in Prague, Technicka 2, CZ-166 27 Praha 6, Czech Republic

ARTICLE INFO

Article history:
Received 7 October 2011
Received in revised form 9 September 2013
Accepted 10 September 2013
Available online 17 September 2013

Keywords:
POLQA
speech quality
MOS
packet loss
jitter

ABSTRACT

This paper presents an analysis of the relation between IP channel characteristics and final voice transmission quality. The NISTNet emulator is used for adjusting the IP channel network. The transmission quality criterion is an MOS parameter investigated using the ITU-T P.862 PESQ, future P.863 POLQA and P.563 3SQM algorithms. Jitter and packet loss influence are investigated for the PCM codec and the Speex codec.

1. Introduction

At the beginning of the 21st century, increasing transmission capacity of the network and improved digital processing methods for video and acoustic signals enabled the Internet to be used for real time voice and video communication. VoIP (voice-over Internet Protocol) allows the transmission of voice in digital form in UDP/TCP/IP packets. Using the IP network to transfer a telephone call poses particular difficulties. Network parameters such as delay variations (jitter), packet loss and bandwidth affect the quality and clarity of the transferred audio signal. Other parameters do not affect the transmitted speech waveform directly but contribute to a decrease in the conversational quality score (e.g. delay).

To assess the quality of voice transmission we used the MOS (mean opinion score) scale (Table 1). The term MOS is defined in Recommendation ITU-T P.800 [15].

Several methods can be used to obtain MOS values. The most accurate method is a subjective test, where the MOS value is obtained directly from users. However, conducting subjective tests is time-consuming and expensive. It is therefore replaced by objective methods based on computer algorithms.

Intrusive methods provide results nearest to those provided by subjective tests. They are based on a comparison of the original and transferred sample. These algorithms use psychoacoustic models of human perception, seeking to offer a mathematical description of the human perception of sound, and to find variables which have a direct impact on the perceived quality of a voice signal. Intrusive methods include PAMS (Perceptual Analysis Measurement System), developed by British Telecommunications, PSQM (Perceptual Speech Quality Measurement), described in Recommendation ITU-T P.861, PESQ (Perceptual Speech Quality Evaluation of), according to ITU-T P.862 (P.862.1) and newly ITU-T P.863 — POLQA (Perceptual Objective Listening Quality Analysis) [5].

Non-intrusive methods are another type of quality measurement. These methods do not use the reference signal, and the final MOS is calculated using the parameters of the transferred sample only. A disadvantage of these methods is their lower accuracy and reliability. An example of a non-intrusive method is 3SQM, which is defined in recommendation ITU-T P.563.

2. Methods used to obtain MOS values

2.1. ITU-T P.862 — PESQ

PESQ is intrusive method of measuring speech transmission quality. It works on the principle of comparing the original and transferred sample.

Before the comparison, the amplitude equalization and time alignment of both samples must be done. Amplitude compensation only adjusts the volume to the level needed for further processing. It does not correct any errors caused by too high or low volume when recording the sample. For the final result of the PESQ algorithm is very important to have matched the corresponding sections of the signal. Therefore, it is important to align any delays of the degraded signal against the original. This part of the algorithm operates on the basis of correlation between the original and degraded signal. The algorithm first calculates the

* Corresponding author.
E-mail address: slavao1@fel.cvut.cz (O. Slavata).

0920-5489/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csi.2013.09.003

Table 1
MOS scale.

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

delay of the entire sample. Then the sample is divided to its sub-parts and the correlation is calculate for each part separately. Consequently, the sample is divided into shorter periods again and the various delays are recalculated until a segment is too short or the correlation is not better than in the previous step.

The most important part of the PESQ algorithm is psycho-acoustic transformation. Parameters of the original and degraded signal are evaluated using a mathematical model of the human auditory system.

- The sample is divided into Sections 16 ms long with a 50% overlap.
- For each segment 256-point FFT is calculated.
- Series of FFT results are divided into 17 frequency bands called “bark bands”.
- For each of the seventeen bands the energy contained therein is summed.
- The energy is converted back to the volume level.
- Results are further threshold and weighted according to the sensitivity of the human ear to different frequencies.

The result of the transformation is a vector with 17 values for each 16 ms period. These vectors are then sorted into a matrix according to the time sequence in the signal. Matrixes of original and degraded signals are compared. Positive and negative differences are summed separately because the human ear is more sensitive to the added disturbance than the missing signal. The weighted sums of the differences are then subtracted from the maximum value of five and the resulting value is MOS for a given sample.

2.2. ITU-T P.863 — POLQA

POLQA is the successor of PESQ. Principle of the algorithm is similar to PESQ but it removes some of its disadvantages. Time alignment algorithm of POLQA can recognize new features of modern codecs such as “time warping” which PESQ evaluates as errors.

Similarly to PESQ POLQA supports measurements in the common telephony band (300–3400 Hz), but in addition it has a second operational mode for assessing HD-Voice in wideband and super-wideband speech signals (50–14000 Hz).

POLQA also examines the original signal and its possible errors (too much timbre, noise or reverberation) are taken into account in the final evaluation. This approximates the results of subjective tests where users compare the transmitted signal, with their subjective vision of the ideal.

2.3. ITU-T P.563 — 3SQM

3SQM is non-intrusive method for measuring listening quality of the voice signal. The algorithm consists of three separate parts which have different methods of calculating the MOS.

Part 1 In the sample are calculated parameters typical for computer signal processing such as: signal-to-noise ratio (SNR), the length of suspension and damping, time cropping … Range of values of these parameters are then used to estimate the value of MOS.

Part 2 A complex “cleaning” function is applied at the degraded sample. Missing parts are recalculated; the sample is filtered and further regulated. This purified sample, together with the original, is used as input signal for the simplified PESQ (without time alignment) and its output is an estimate of MOS.

Part 3 The main part of this block is a precision LPC model of the human vocal tract. This ‘synthesizer’ attempts to pronounce the degraded sample. The result is compared with the original sample. Everything different in the original sample is considered as unnatural to the human vocal tract and considered as damage caused during sample transfer. The sum of this added disturbance is used to calculate the MOS estimation.

The most distant of these three estimates of MOS is dropped and the arithmetic mean of the remaining two is the resulting estimate of MOS for the entire algorithm.

3. Experiment description

3.1. Test-bed

The test-bed (Fig. 1) consisted of three computers, an Opera audio analyser, and interconnecting cables. A concatenated speech file in WAV format (8kSa/S, 16bit), 16.75 s in length, was used. The file contained 4 short sentences spoken by 4 different speakers (two men, two women), and adequately covered the entire human speech spectra. Due to this fact, the concatenated file was used as an effective replacement for testing using multiple speech samples.

The signal was transferred from an audio output “line 1 out” of the Opera analyzer to an audio input (microphone) of PC 1. PC 1 and PC 2 were connected by a UTP network cable (subnet 192.168.0. X), as were PC 2 and PC 3 (subnet 192.168.1. X). PC 2 was therefore fitted with a two-port network interface card. The test signal was transferred from PC 1 to PC 3 using a VoIP call in the Linphone program, using PCM (G.711) and Speex codecs. From PC 3, the audio output (headphone) signal was led back into the audio input “line 2 in. of the Opera analyzer. The NISTNet emulator [16] was running on PC2, which (according to the specific settings) introduced transmission errors between PC 1 and PC 3. The results depend on the accuracy and repeatability of the network simulation. We proved by several experiments [8,9] that NISTNet suits these requirements satisfactorily. It was also used in other experiments [12]. The measured samples were adjusted in Adobe Audition 3.0 (converting stereo → mono) and then tested using the POLQA (ITU-T P.863) PESQ (ITU-T P.862) and 3SQM (ITU-T P.563) algorithms [7]. The PESQ algorithm output was recalculated to the value of MOS-LQO (Listening Quality Objective) according to a mathematical prescription defined in ITU-T P.862.1. According to the official wording of P.862,

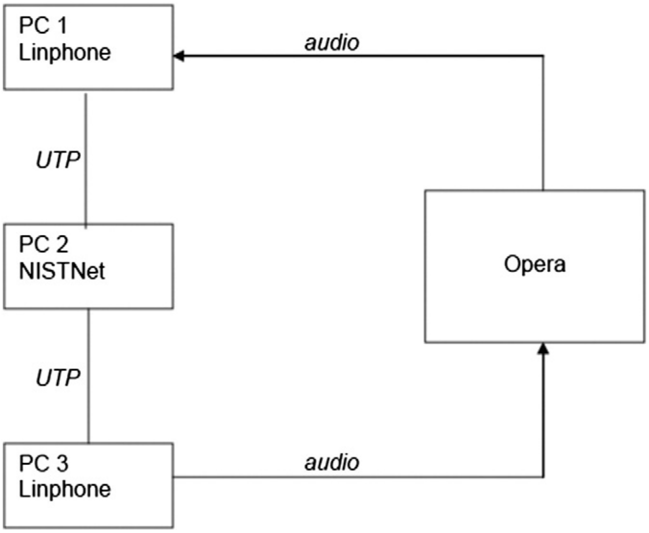


Fig. 1. Test-bed.

the effect of packet loss on CELP coded transmission can be tested in this way. It was not tested for PCM transmissions affected by packet loss, but the recommendation itself does not prevent any user making such tests [2,14].

3.2. Tested transfer parameters

3.2.1. Jitter

The limited speed of signal transmission in the network and signal processing components on the route, e.g. routers and converters, cause signal delay. The speed of the signal transmission is a particular problem when a call is made over a long distance or is led via satellite for part of the route. The delay alone does not affect the quality of the transferred signal. However, the delay is usually not constant. The sender generates packets at the same time intervals, but the network parameters may be changed during a call. Consequently, the transmission delay varies during the call. This phenomenon is called jitter, and may cause problems with the delivery of packets. It may change their order and thus impair the signal quality. It can be buffered to some extent to compensate on the recipient site. In this experiment, the jitter buffer of Linphone was set to default 60 ms. The NISTNet emulator allows the mean value of the delay to be set (parameter delsigma). The delay of each packet is randomly generated, with normal distribution around this value. In this experiment the following values (in ms) were adjusted: 0, 10, 20, 40, 60, 80, 100, 125.

3.2.2. Packet loss

The root cause of packet loss during transmission may be a route failure (drop-out of the satellite or microwave links), or saturation of the router buffer. Sometimes the packet is not used in the reconstruction of the signal, due to its excessive delay. Losses may be dependent (the probability of packet loss depends on whether the previous packet was lost) or independent. A suitably long speech sample with a high speech activity factor should be used in the case of independent losses, in order to assure uniform distribution of impairments in different measured samples. However, the sample length is limited by the requirements of the recommendation (20s as given by P.862.3). Independent losses were used in this experiment, and the following values were adjusted (in %) 0, 1, 2, 4, 6, 8, 10, 12.

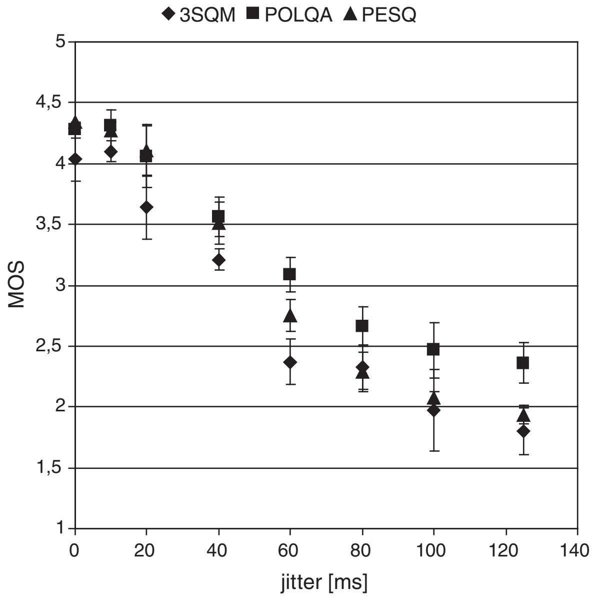


Fig. 2. MOS as a function of jitter (PCM codec).

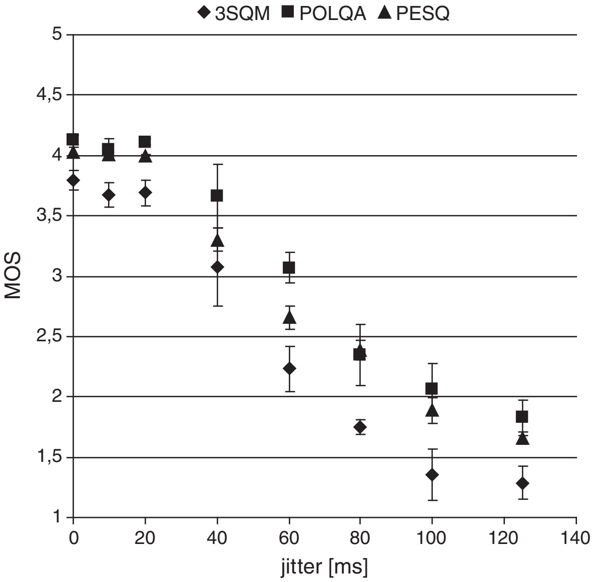


Fig. 3. MOS as a function of jitter (Speex codec).

4. Results

For each IP channel parameter setting, five samples were measured and processed. Using statistical processing of the results, a confidence interval of CI95 was calculated. It is displayed in the graph as error bars. The speech sample that was used is long enough even for the low packet loss values that were tested. Five repetitions are enough to achieve satisfactorily low result dispersion and uncertainties. The results for PESQ [2,3] and 3SQM [1] are in agreement with previous experiments [6,10,11,13].

4.1. Jitter

Changes in the delay in the transmission have a major impact on the quality of the transferred voice. The Linphone jitter buffer was set to its default value of 60 ms. When setting the parameter delsigma to 40 ms and higher, the jitter buffer on the receiver side can no longer fully

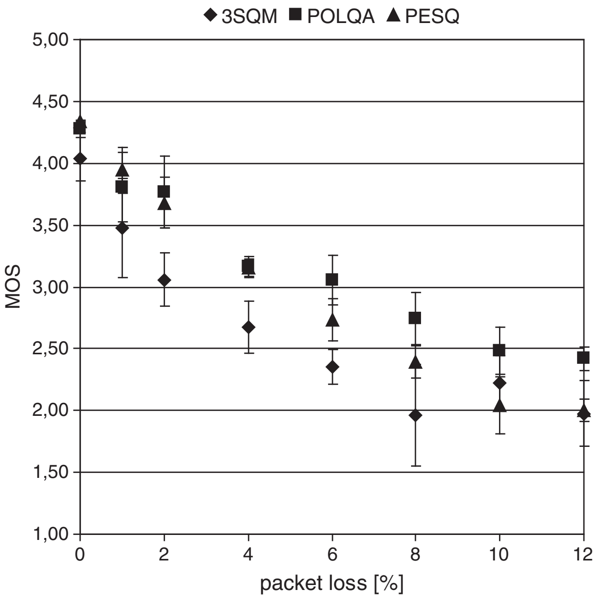


Fig. 4. MOS as a function of packet loss (PCM codec).

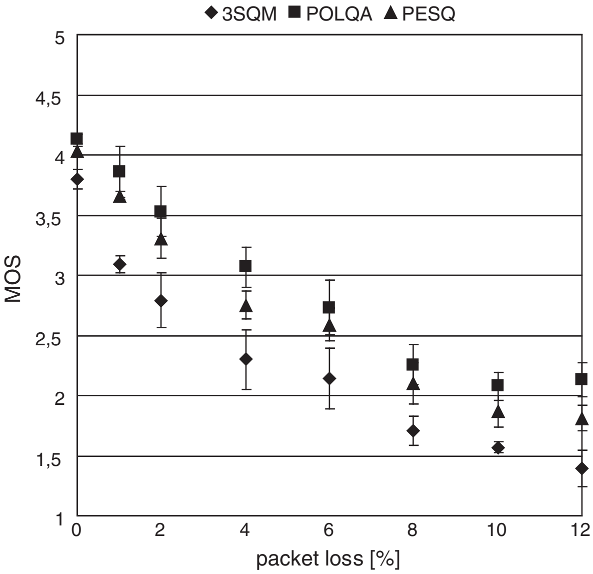


Fig. 5. MOS as a function of packet loss (Speex codec).

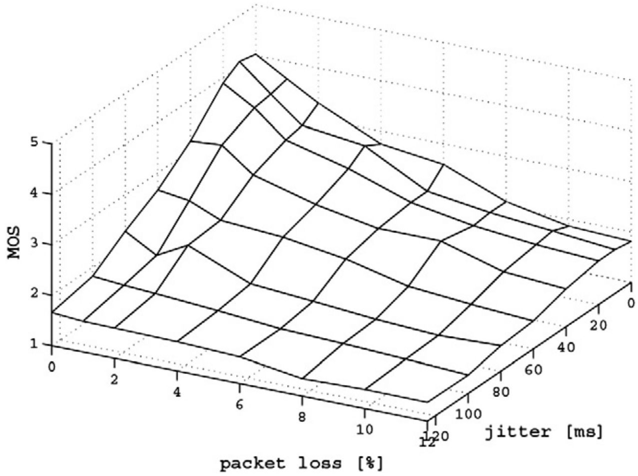


Fig. 7. MOS as a function of packet loss and jitter (PESQ algorithm, Speex codec).

compensate for errors caused by the disorderly packet delivery. When the jitter is higher than 100 ms, the transferred signal is almost unintelligible. The results for the PCM codec are depicted in Fig. 2, and for Speex codec in Fig. 3. In both cases the new POLQA algorithm predicts a higher MOS value than PESQ. Non-monotonicity of the graph in the range of 0–20 ms, particularly evident in the POLQA algorithm, is probably caused by an outlying result from one sample. It can be assumed that the measurement of significantly larger numbers of samples would cause extermination of the graphs. This may be a subject of further experiments. The graphs also show that the 20–40 ms samples have significantly greater variance than in the rest of the chart. This is a critical area for the jitter buffer on the receiver side, where one sample is buffered and the other, which is slightly different, is not buffered.

4.2. Packet loss

Packet loss also affected the quality of the transferred voice. The results can also be affected by Packet Loss Concealment (PLC), implemented by Linphone. Comparing our results with [4], it is obvious that our Linphone had no PLC implemented. Already for 1% packet loss the MOS value drops below 4, and for losses greater than 10% the transferred signal was unintelligible. The results for the PCM codec are shown in Fig. 4, and for the Speex codec in Fig. 5. Both graphs show an evident decrease already for 1% packet loss. Again, it can be seen that the POLQA algorithm predicts higher MOS values than PESQ. The POLQA result dispersion is probably due to the measurement procedure (number of samples). It is possible that POLQA is more sensitive to disturbances, and therefore provides larger variance of the results for the same samples than PESQ. Using more samples would probably cause extermination of the graphs.

4.3. Combination of packet loss and jitter

In real traffic, all kinds of defects occur simultaneously. The following charts show the dependence of MOS on a combination of jitter and packet loss. When the receiver is decoding the transmitted signal, an excessively delayed packet has the same effect as a lost packet, because neither can any longer be used for reconstructing the transmitted signal. As a result, exposure to both disorders simultaneously causes a faster decline in quality than the separate effects of only one of them. The results for the PESQ algorithm and the PCM codec are shown in Fig. 6, while the results for the PESQ algorithm and the Speex codec are depicted in Fig. 7. The Speex codec, designed specifically for VoIP, provides only a slightly lower quality of the transmitted voice signal than

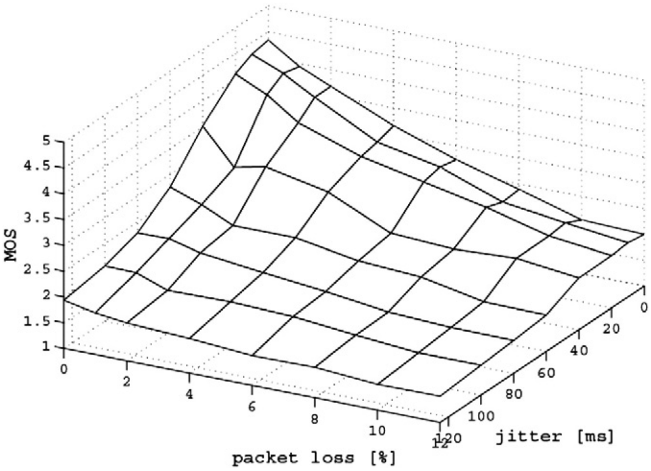


Fig. 6. MOS as a function of packet loss and jitter (PESQ algorithm, PCM codec).

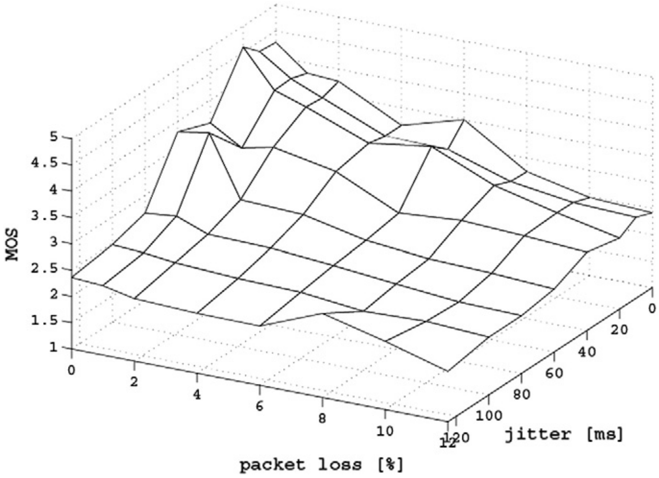


Fig. 8. MOS as a function of packet loss and jitter (POLQA algorithm, PCM codec).

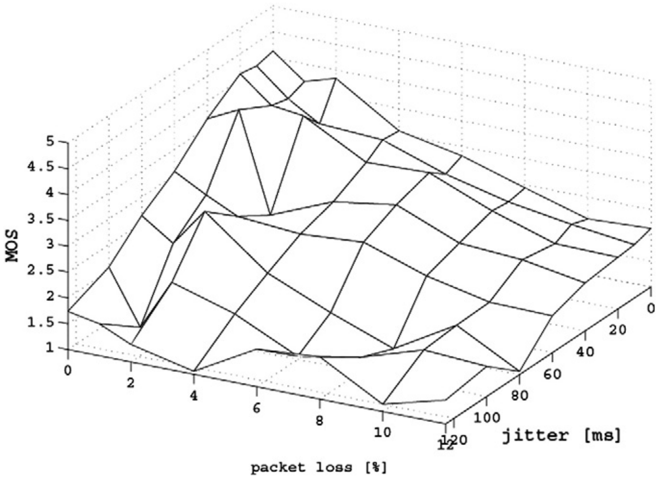


Fig. 9. MOS as a function of packet loss and jitter (POLQA algorithm, Speex codec).

the common and widely quoted PCM codec. However, Speex needs only half the bit rate (32 Kbps compared to 64Kbps PCM).

The results for the POLQA algorithm and the PCM codec are depicted in Fig. 8, and the results for the POLQA algorithm and the Speex codec are shown in Fig. 9. A slight difference can also be seen between the PCM and Speex codecs. The dispersion of the results of the POLQA algorithm is also clearly visible. Due to its greater sensitivity than PESQ, the POLQA algorithm probably requires a larger number of samples for statistical analysis.

5. Conclusion and future work

This paper has verified the impact of changes in delay and packet loss on the quality of voice transmission in IP networks, and compares the results of the Speex codec with PCM. It has also compared the results of the new POLQA testing standard with older algorithms.

The main objective was to identify the relation between the network parameters and the MOS values as delivered by different objective algorithms. A second objective was to compare the Speex codec with the PCM reference codec. This codec, designed specifically for VoIP, provides almost the same transmission quality at half the required transmission rate.

The third objective was to explore the MOS predictions of the new POLQA testing standard. This algorithm predicts a higher MOS value for most samples, but its results are highly scattered. It appears that more samples are required for proper function when testing packet loss or jitter.

There are several options for continuing this work. First, as mentioned above, in order to further decrease the final result dispersion, the POLQA algorithm may be tested on a larger sample database. Second, when a suitable subjective test methodology is standardized (the currently used ITU-T P.800 does not seem to be suitable for samples affected by low value packet loss, due to the short sample length that is required), the measured samples may be tested on a group of listeners using a subjective method. Then the results can be compared with objective algorithms.

References

[1] ITU-T P.563, Single-ended method for objective speech quality assessment in narrow-band telephony applications, May 2004.

[2] ITU-T 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, February 2001.

[3] ITU-T P.862., 1 Mapping function for transforming P.862 raw result scores to MOS-LQO, November 2003.

[4] ITU-T P.863 — trial executable code of POLQA algorithm as obtained from POLQA developers team, www.polqa.info.

[5] POLQA, Technical White Paper, Opticom, July 2010.

[6] E. Gündüzhan, K. Momtahan, A Linear Prediction Based Packet Loss Concealment. Algorithm for PCM Coded Speech, IEEE Trans. Speech Audio Process. 9 (8) (November 2001) 778–785.

[7] I. Vondrka, Implementation of the P.563 (3SQM) standard in PC's using Lab/Windows CVI, May 2005. (Diploma Thesis CTU FEE).

[8] O. Slavata, Měření kvality přenosu hlasu pro sítě typu VoIP, Bachelor project, FEE CTU, August 2007.

[9] O. Slavata, Neintruzivní měření kvality přenosu hlasu v telekomunikačních sítích, Diplomová práce, January 2010. (Diploma Thesis, CTU FEE).

[10] P. Počta, J. Holub, Predicting the Quality of Synthesized and Natural Speech Impaired by Packet Loss and Coding Using PESQ and P.563 Models, Acta Acust. United Acust. 97 (5) (September/October 2011) 852–868.

[11] P. Počta, J. Holub, Effect of speech activity parameter on PESQ's predictions in presence of independent and dependent losses, Comput. Stand. Interfaces 36 (1) (November 2013) 143–153.

[12] Sankaranarayanan, G., Hannaford, B.: Comparison of Performance of Virtual Coupling Schemes for Haptic Collaboration using Real and Emulated Internet Connections, University of Washington, Seattle. 2007.

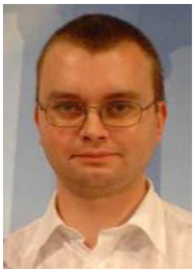
[13] Y. Stein, I. Druker, The Effect of Packet Loss on Voice Quality for TDM over Pseudowires, RAD Data Communications, October 2003.

[14] Ditech networks, Limitations of PESQ for Measuring Voice Quality in Mobile and VoIP Networks, December 2007.

[15] http://www.itu.ch.

[16] http://www-x.antd.nist.gov/nistnet/.

Ing. Oldřich Slavata was born in Prague, Czech Republic, in 1984. The Bc. (2007) Ing. (2010) in Measuring Technology in Czech Technical University in Prague, Faculty of Electrical Engineering. Ph.D. student on Department of Measurement since 2010. His research interests cover packet-based networks, digital signal processing, speech coding and processing, psychoacoustics and measurements in telecommunication networks. Ing. Slavata published four papers in international conferences.



Assoc. Prof. Ing. Jan Holub, Ph.D. was born in Prague, Czech Republic, in 1973. The Ing. (1996) Ph.D. (1999) and Assoc. Prof. (2004) in Measuring Technology in Czech Technical University in Prague, Faculty of Electrical Engineering. His research interests cover AD and DA converters, digital signal processing, speech coding and processing, psychoacoustics and measurements in telecommunication networks. Dr. Holub authored more than 90 conference papers and 18 journal papers. He is the Chairman of the Organizing Committee for the MESAQIN Conference since 2001, a member of program committee of WTS 2006–11, and the chair of IMEKO TC-1.