Comparison of Subjective and Objective Speech Quality Testing Methods in the VoIP Networks

Z. Becvar, J. Zelenka, M. Brada and T. Valenta
Department of Telecommunication Engineering
Czech Technical University
Technicka 2, Prague, 16627, Czech Republic
Phone: (+420) 2-2435 5994 Fax: (+420) 2-3333 9810 E-mail: becvaz1@fel.cvut.cz

Keywords: Quality of speech, VoIP, MOS, PESQ, 3SQM

Abstract – The main requirement in the Voice over IP technology is a good quality of transmitted signal between calling subscribers. A signal quality on a receiving side can be measured through a number of methods. The main purpose of this paper is a comparison of the results obtained by subjective listening tests and objective measuring methods. There exist several types of objective testing methods. This paper deals with two of them: PESQ and 3SQM. English and Czech language speeches were applied into subjective and objective tests.

1. INTRODUCTION

Main problems in the VoIP (Voice over Internet Protocol) are the packet loss and the varying delay of packets (jitter) due to distinct path of each packet in network and different network load. On the receiving side of a transmission chain there is a need of continuous data flow in order to reconstruct a voice waveform correctly. If the data packet is not available at the proper time, it has the same effect as if the packet has been lost. This causes decreasing quality of recovered signal on the receiving side. The result of decreasing in quality is rising of customer's dissatisfaction and shortening of calls.

Measuring of signal quality can be performed by subjective or objective test. Subjective listening test is the most exact method for quality measurement. This is also time-consuming, expensive and labor intensive method. Because of difficult preparation and a need of great number of listeners is hard to realize these tests to frequently. In some cases it can not be used because it is an intrusive technique of measuring.

In opposite to these facts objective methods do not depend on complex preparation phase. Unfortunately objective methods do not obviously provide accurate results. Objective methods are much more suitable for repeating tests and they can also be used in hardware implementation as an In-service Non-intrusive Measurement Device (INMD). This kind of device can be used as a real-time speech quality indicator that is a very important statement for telecommunication operators.

This paper is focused on a comparison of results of subjective and two types of objective tests. The tested utterances were in two languages: Czech and English. Each of those languages was tested separately.

The rest of the paper is organized subsequently. Next chapter describes two objective testing methods. The third section is focused on the arrangement of the subjective test. Further section provides the results of the subjective test and compares them with results obtained by objective tests. Last section presents our conclusions.

2. OBJECTIVE QUALITY ASSESSMENT METHODS

ITU-T recommendations are describing, among others, two most used objective methods: PESQ and 3SQM.

2.1 PESQ

PESQ (Perceptual Evaluation of Speech Quality) is the objective method (described in ITU-T recommendation P.862 [1]) developed for end-to-end speech quality assessment in conversational voice communication. PESQ can be used for narrow-band and wide-band telephone networks and speech codecs. The principle of PESQ is based on the comparison of original, non-degraded signal X(t), with degraded signal Y(t). Y(t) is the result of passing signal X(t) through a communication system. PESQ generates a prediction of the quality which would be given to signal Y(t) in subjective listening test.

The signal is rated with MOS (Mean Opinion Score) [2], [3]. PESQ MOS value range is defined according to ITU-T P.862 between -0.5 and 4.5.

For comparison of PESQ and subjective results the subjective score must be acquired from listening tests that meet the recommendation ITU-T P.830 [4]. Correlation coefficient is calculated with Pearson's formula [1].

The average correlation was 0.935 in 22 ITU benchmark experiments. Absolute residual error for ITU benchmarks was less then 0.25 MOS for 72.3% and less then 0.5 MOS for 91.1% of the conditions.

The ITU-T P.862 recommendation describes all requirements on the tested speech signal. Signal must have a character as a real signal carried by communications network. The tested speech signal must include speech burst with length between 1-3 s. These burst must be separated with silence. Speech must be active between 40% and 80% of the signal length. These parameters are depending on the language. In most experiments are used two or three burst with total duration 8 s. There can be used sentences with 8 – 20 s of speech in the test. Frequency characteristics and level alignment must be in accordance with recommendation ITU-T P.830. PESQ can be used to assess the quality of systems carrying speech signals with background noise. The noise must be added before the signal is passing through the communication

system. It is necessary to beware of any distortions by quantization, amplitude clipping or resampling. Tested speech must be 16-bit linear PCM (Pulse Code Modulation) sampled with 8 kHz sample rate (this rate is preferred) and it is possible to use speech sampled with 16 kHz.

PESQ also supports major part of known coding techniques such as G.711, G.726, G.727, G.728, GSM codecs, etc.

2.2 3SQM

3SQM evaluates the signal speech quality and predicts the MOS-LQO score [3] the same as PESQ that is considered to be a precedent version of 3SQM.

Unlike PESQ, only the degraded signal is considered in measurement and the original input signal is not needed any more. Then, it can be used in continual or long-term measurements. The final quality coefficient is determined from much more values than the PESQ does. Especially important to us are the items related to network transmission characteristics. Ignoring others the most important are the packet loss and packet loss concealment factor, transmission channel errors or effects of varying delay. According to this, 3SQM should much better express and interpret the main causes of quality degradation affecting VoIP calls.

The format of input signal is strictly defined in the recommendation. Sixteen bit linear PCM with sampling frequency at 8 kHz. If the frequency is higher, then the down sampling must be accomplished with a low pass filter. Minimum signal length is 3 seconds and maximum 20 seconds. Speech must be present in more then 25% of the time and it should not be longer than 75% of the time of recorded audio signal. The speech level must fit into limits from -36 to -16 dBo [5], otherwise the low SNR will affect the result.

The algorithm can be divided into three main parts: pre-processing, calculation of characteristic speech parameters and evaluation of speech quality model.

In pre-processing step, the voice activity detector tries to identify the noise and speech parts of signal. After that the normalization is performed. Also some other coefficients are discovered and remembered for next steps processing.

The calculation of speech parameters can be also divided into three independent areas: vocal tract analysis and unnaturalness of speech, analysis of strong additional noise and interruptions, mutes and time clipping.

There are several parameters, but in the following paragraphs, only those related to the VoIP problematic are mentioned and described.

Detection of unnaturalness is based on detection of repeated signal samples. Some very simple packet loss concealment methods tend to repeat the last received packet in case of drop out. It often only degrades the listening experience with no improvement in quality. The other non-human sounds, such as DTMF tones are also recognized even they occur in the middle of spoken word.

Noise is not the main problem of VoIP networks. The noise level generally depends on the speech digitalizing device and the used codec, but it doesn't vary in time. So the noise level is one factor having impact on the final result, but has nothing to do with packet loss problem.

Interrupts are the direct consequence of packet loss, of course when no other attempt to recover is used. The 3SQM is able to distinguish between natural end of words and sentences and unnatural speech clipping. The pitch period is detected and the signal is sliced into short frames. In each frame the maximum level of signal is determined and compared with neighboring frames. The interruption will rise as a rapid change in value of this level.

After all of the coefficients have been evaluated, they are multiplied by weighting vector and the resulting MOS-LQO is predicted. The 3SQM generally takes into part much more quality affecting events then PESQ. Many of them are useless in VoIP environment, but those which have been mentioned are very suitable for determining how the packet loss affects the speech signal quality.

3. COMPOSITION OF THE SUBJECTIVE TEST

3.1 Subjective test definition

The parameters for subjective testing are given by ITU-T recommendations P.800 [3] and P.830 [4]. This standard defines procedures of selecting speakers for source speeches, methods for recording and preparing of input samples, required sample quantities and formats, parameters of testing environment and methods for selection and guidance of test attendants. The phases of the experiment and their recommended content are not defined.

3.2 Speaker selection

The type of speaker is not defined by ITU, only minimal quantity of 2 male and 2 female speakers is required. Also no distracting stress in voice should be present. Three random male speakers and two female speakers were selected for the test and instructed to produce the speech adequate to recommendation demands.

3.3 Recording methods, signal processing

Recording procedure was realized in a quiet small sized room, equipped with fan-less computer recording system and audio mixing system and microphones. Used environment was a set of small radio recording studio, which fulfills the needs of recording methods specifications.

All acquired signals were recorded in maximal quality and then digitally processed to required form and separated segments for each part of experiment.

Signal qualities and their content description are staged in section for experiments with PESQ, 3SQM.

3.4 Testing environment

The testing environment is the place where the listening process is in progress. It was used a same place for recording and listening according to recommendations. The test was done with high-quality headphones in this testing place and prevented from any interruption or distortion.

3.5 Listeners

Applicable set of listeners was acquired from local university and final number of test attendants was 41 persons. Most of them were PhD students aged from 24 to 30 years.

3.6 Test phases and elements

Final listening test was designed to endure between 20 to 30 minutes.

Input samples used in test were distorted - audio segment of specific length was removed and the percentage of removed segments varied in those samples. The parameters of these speech samples were selected to be in accordance with ITU-T recommendations P.862, P.562, P.800 and P.830.

Used segments durations were selected to fit typical packetization lengths used in VoIP communication and the lengths were 10 and 20 ms. Percent occurrences of removed segments were 2, 5, 7, 10, 20 and 30%. Distribution of those losses was selected from measurements of typical networks packet loss.

4. RESULTS OF TEST

In the following Table 1 are results of the subjective test and both objective tests. There are utterances in Czech and English language.

Language	Length of packet	Percentage of lost packets	Subjective	PESQ	3SQM
	(ms)	(%)	MOS	MOS	MOS
Czech	10	2	3.546	3.381	3.543
		5	3.132	2.988	2.807
		7	2.943	2.859	2.846
		10	2.664	2.535	2.493
		20	2.222	2.079	2.009
		30	1.727	1.676	1.619
	20	2	4.146	3.511	3.579
		5	3.634	2.962	3.1
		7	3.317	2.819	2.832
		10	3.122	2.44	2.842
		20	2.439	1.729	2.255
		30	2.146	1.3	2.16
English	10	2	4.195	3.822	3.468
		5	3.512	3.534	2.812
		7	3.439	3.349	2.806
		10	3.268	3.223	2.643
		20	2.732	2.666	2.334
		30	2.024	2.179	2.398
	20	2	4.39	3.723	3.71
		5	3.561	3.32	3.322
		7	2.976	3.11	2.92
		10	3.317	2.896	2.812
		20	2.512	2.058	2.691
		30	2	1.602	2,759

Table 1. Absolute MOS score of subjective and objective tests

For every language there are two variants of length of packet. In most cases it is used 10 ms and 20 ms length of packet in VoIP. It means that there is 10 ms or 20 ms of speech waveform that is compressed in every data packet transmitted in IP network. In the third column is a percentage of lost packets. There are represented low loss networks and networks with very high loss of packets.

In the Fig. 1 are three decreasing curves of speech quality represented by MOS score. PESQ testing method is about 0.5 MOS below subjective test in all packet loss quantities. 3SQM test has a similar character as PESQ but its results have greater deviations.

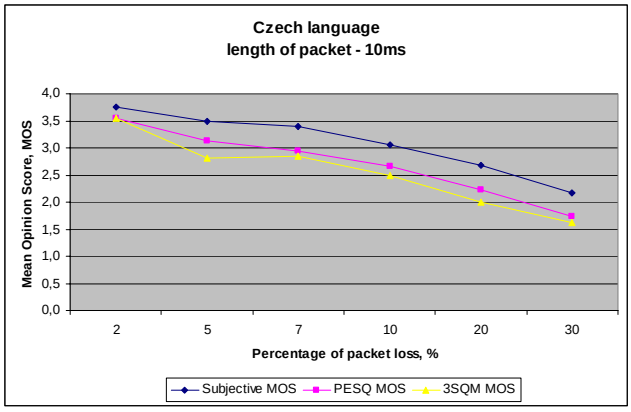

Fig. 1. MOS score of Czech language; length of packet is 10ms

For 20 ms Czech language there is a larger difference between PESQ and subjective test and this difference is about 0.7 MOS. 3SQM method gives better results for higher percentage of lost packets. This situation is shown in Fig. 2.

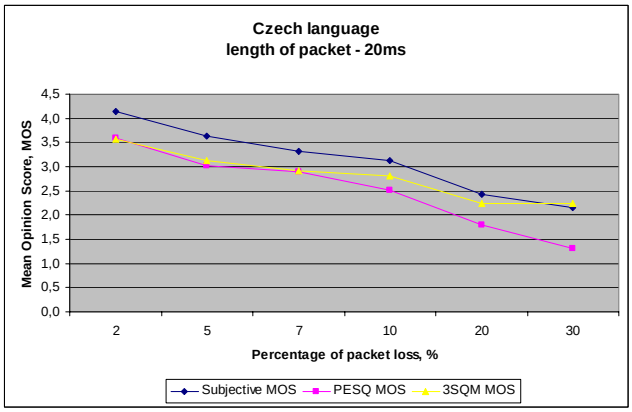

Fig. 2. MOS score of Czech language; length of packet is 20ms

If an English utterance of 10 ms packet length is tested, PESQ gives very good results. In opposite, 3SQM has very poor result with 0.7 MOS deviation for lower percentages of packet loss. This can be seen in Fig. 3.

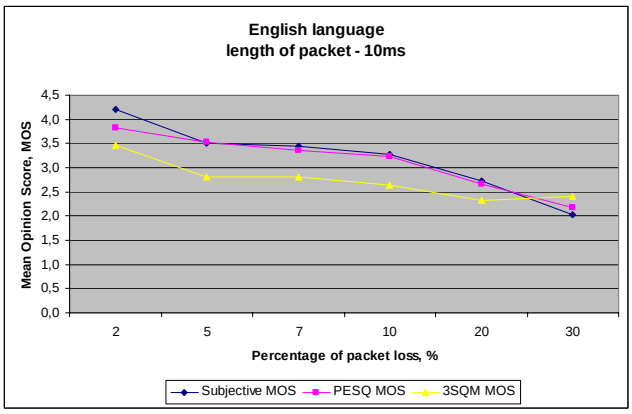

Fig. 3. MOS score of English language; length of packet is 10ms

In Fig. 4 there is an English utterance with 20 ms length of packet. Objective methods, PESQ and 3SQM have very similar results for lower packet loss percentage. In higher percentage, PESQ method has lower results according to subjective test.

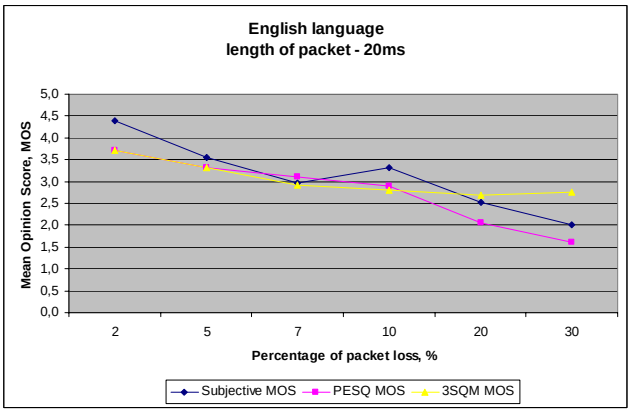

Fig. 4. MOS score of English language; length of packet is 20ms

When results of Czech and English utterances are compared, the objective tests of English utterances give better results than in case of Czech utterances. If the PESQ results of Czech speeches were recounted with suitable relation, they could give a better outcome.

5. CONCLUSION

Essentially, character of PESQ results corresponds to the shape of the subjective test much better than those ones of 3SQM test results; nevertheless it gives an undervalued quality of the speech. Larger undervaluation is in the case of Czech language. The results were quite similar in the English utterances but with shorter length of packets.

3SQM has got a better correspondence with subjective speech quality at higher values than at lower values of packet loss percentage; on the other hand it doesn't provide good results for any language.

The conclusion is that both of these methods are not as accurate as it is required for implementation in practical services applications. Subjective measurement is non-replaceable way to evaluate quality of speech. It offers the accuracy, which objective methods are not able to reach.

ACKNOWLEDGEMENT

This work has been supported by the grant "Research of perspective information and communication technologies" No. MSM6840770014 funded by Ministry of Education of the Czech Republic.

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] J.D. Gibson, "Speech Coding Methods, Standards and Applications", *IEEE Circuits and Systems Magazine*, p. 30-49, Fourth quarter 2005.

[3] ITU-T Recommendation P.800.1, *Mean Opinion Score (MOS) terminology*, 2003.

[4] ITU-T Recommendation P.830, *Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs*, 1996.

[5] ITU-T Recommendation P.563, *Single Ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications*, 2004.