A systematic study of PESQ's behavior in simulated VoIP environment (from reference signal characteristics perspective)

Peter Počta[1]*, Miroslava Mrvová*[1]*, Peter Kortiš*[1]*, Peter Palúch*[2]*, Martin Vaculík*[1]

[1]*Dept. of Telecommunications and Multimedia, FEE, University of Žilina, Univerzitná 1,*
SK-01026, Žilina, Slovakia, {pocta, kortis, vaculik}@fel.uniza.sk

[2]*Dept. of InfoCom Networks, FMSI, University of Žilina, Univerzitná 1, SK-01026, Žilina Slovakia*

{paluch}@fri.uniza.sk

Abstract

In this work, we experimentally study how behaviour of the *PESQ*-estimate varies with reference signal characteristics. In particular we investigate the impact of different lengths of reference signal and active speech ratios on speech quality estimation in simulated *VoIP* environment. These two reference signal characteristics are defined very broadly by *ITU-T* Recommendation *P.862.3*. That is reason to investigate an impact of those characteristics on speech quality estimation more in depth. We assess the variability of *PESQ* estimations with respect to the reference signal characteristics and network conditions and finally offer some proposals for the purpose of more accurate and reliable speech quality assessment from those reference signal characteristics point of view in *IP* networks.

Keywords

Perceptual Evaluation of Speech Quality (*PESQ*), speech quality, intrusive measurement, Voice over Internet Protocol (*VoIP*), reference signal characteristics, length of signal, active speech ratio.

1 Introduction

Voice over Internet Protocol (*VoIP*), the transmission of packetized voice over *IP* networks, has gained much attention in recent years. It is expected to carry more and more voice traffic for its cost-effective service. However, present-day Internet, which was originally designed for data communications, provides *best-effort* service only, posing several technical challenges for real time *VoIP* applications. Speech quality is impaired by packet loss, delay and jitter. Assessment of perceived speech quality in the *IP* networks becomes an imperative task to manufacturers and as well service providers.

Speech quality is judged by human listeners and hence it is inherently subjective. The Mean Opinion Score (*MOS*) test, defined by *ITU-T* Recommendation *P.800* [1], is widely accepted as a norm for speech quality assessment. However, such subjective test is expensive and time-consuming. It is impractical for the frequent testing such as routine network monitoring.

Objective test methods have been developed in recent years. They can be classified into two categories: signal-based methods and parameter-based methods. Signal based methods use two signals as the input to the measurements, namely, a reference signal and a degraded signal, which is the output of the system under test. They identify the audible distortions based on the perceptual domain representation of two signals incorporating human auditory models. These methods include Perceptual Speech Quality Measure (*PSQM*), Measuring Normalizing System (*MNB*), Perceptual Analysis Measurement System (*PAMS*), and Perceptual Evaluation of Speech Quality (*PESQ*). Among them, *PSQM* and *PESQ* [2] were standardized by the *ITU-T* recommendations such as *P.861* and *P.862* respectively. Parameter-based methods predict the speech quality through a computation model instead of using a real measurement. A typical model is the *E-model* as defined by *ITU-T* Recommendation *G.107*. The *E-model* includes a set of parameters characterizing end-to-end voice transmission as its input, and the output can be transformed into a *MOS* scale for prediction.

The algorithm *PSQM* is based on comparison of the power spectrum of the corresponding sections of reference and degraded signals. The results of this algorithm more correlate with the results of listening tests, in comparison with *E-model*. At present, this algorithm is no longer used due to a coarse time-alignment. Instead of it, the algorithm *PESQ* is rather used. *PESQ* combines merits of *PAMS* and *PSQM99* (an updated version *PSQM*), and adds new methods for transfer function equalization and averaging distortions over time. The algorithm *PESQ* facilitates with very fine time-alignment and one single interruption is also taken into account in the calculation of *MoS*. It can be used in wider range of network conditions, and gives higher correlation with subjective tests and the other objective algorithms [2, 3]. Unlike the conversational model, *PESQ* is a listening-only model; the degraded sample is time-aligned with the reference sample during pre-processing. The *PESQ MoS* values do not reflect the effects of delay on speech quality. The disadvantages include impossibility to use it for codec's with data rate lower than 4 kbps and higher calculation load what is caused by recursions in the algorithm.

The characteristics of reference signals for objective speech quality measurements provided by *PESQ* are defined in Section 7 of the *ITU-T* Recommendation *P.862.3* [4]. Two reference signal characteristics are defined very broadly by this recommendation from our point of view that are length of reference signal and active speech ratio. The above mentioned recommendation recommends to use the reference signal in duration in the range from 8 seconds to 30 seconds for the purpose of *PESQ* measurement. The speech activity in the reference speech, which can be measured according to *ITU-T* Recommendation *P.56* [5], should be between 40% and 80% of their length. We suppose that those two characteristics can have an impact on the final *PESQ*'s estimation. That is a reason for exhaustive investigation of the impact of different lengths of reference signal and active speech ratios on speech quality estimation provided by *PESQ*.

Some works have been carried out on study of *PESQ*'s accuracy and behavior. Particularly, [6, 7, 8, 9] examined the *PESQ*'s accuracy in some cases. In [6], comparison between subjective test and *PESQ* score have been realized and mapping function known as *PESQ-LQ* has been proposed and verified. This function can significantly reduce the raw *RMS* error when compared to many subjective tests without using per-experiment mapping. In [7], the verification of *PESQ* performance in case of single frame losses has been conducted by means of formal listening only tests. The tests have proved that *PESQ* predicts the impact of single frame losses precisely. In [8], an investigation how subjects perceive bursty losses and how current objective measurement methods, such as *PSQM*, *MNB*, Enhanced Modified Bark Spectral Distance (*EMBSD*) and *PESQ,* correlate with subjective test results under burst loss conditions. Preliminary results show that *PESQ* displays an obvious sensitivity to bursty conditions compared to human subjects (it is more sensitive than subjects when loss burstiness is high and less sensitive when it is low). In [9], the effects of speech coder, packet loss concealment strategy, *IP* payload size, packet loss rate and burstiness of packet losses on *PESQ* accuracy were assessed. The results indicate that *PESQ* is a useful tool in helping to identify potential performance problems but is not accurate enough to specify speech quality requirements in Service Level Agreements (*SLA*s). In [10], a study of *PESQ*'s behavior from networking perspective (packet loss process) has been presented. It seems that *PESQ* maintains reasonable correlation with subjective scores even when the network conditions are bad. Also, the deviations seem to be systematic from subjective scores, which suggest that a simple compensation factor might be found (for instance, derived from network conditions) and used to improve the results. The using long reference signals for the speech quality assessment by *PESQ* have been investigated in [11]. Experimental results show that it is possible to use a longer reference signal for this purpose and propose extending the maximum length of reference signal to 30 seconds. The results of this work have been applied in *ITU-T* Recommendation *P.862.3*. In [12], we presented the investigation of an impact of duration of speech samples on speech quality from jitter rate and packet loss view point in *IP* networks. The *ITU-T G.729* and *ITU-T G.723.1* encoding schemes were used for the purpose of the simulations. The assessment of speech quality was realized by means of *PESQ*. The results show that the difference in duration of speech sequences has the impact on speech quality assessed by *PESQ*. That was our inspiration for conduct of more complex experiments in simulated *VoIP* environment in order to verify our preliminary simulations results.

Here we focus on an impact of different lengths of reference signal and active speech ratios on speech quality estimation provided by *PESQ* in simulated *VoIP* environment. The reference signals in length of 10, 20 and 30 seconds and reference signals with active speech ratios of 42, 62 and 82% are investigated in this study. We assess the variability of *PESQ*'s estimations with respect to the reference signal characteristics (length of signal, active speech ratio) and network conditions and finally offer some proposals for the purpose of more accurate and reliable speech quality assessment in *IP* networks from above mentioned reference signal characteristics point of view.

The rest of the paper is organized as follows: Section 2 introduces experimental scenario and experiments carried out in this study. In Section 3, the experimental results are presented and discussed. Section 4 concludes the paper and suggests some future studies.

2 Description of experiments

2.1 Experimental setup

One-way *VoIP* session was established between two hosts (*VoIP* Sender and *VoIP* Receiver), via the isolated *IP* network, in *IEEE 802.3i* 10Base-T Ethernet (Figure 1).

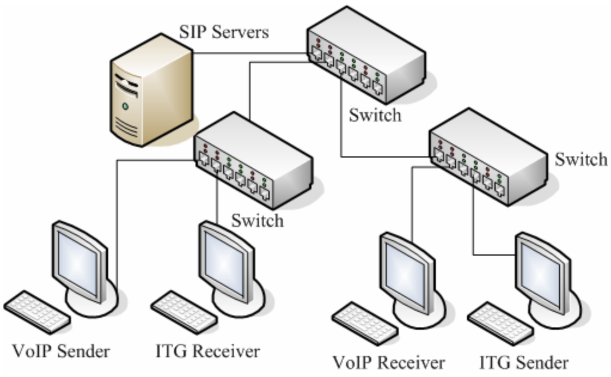

Fig.1 Measurement setup

Two stations (*ITG* Sender and *ITG* Receiver) equipped with the accomplished *D-ITG* traffic generator [13] were used to generate and receive background traffic. *ITG* Sender generated the User Datagram Protocol (*UDP*) and Transmission Control Protocol (*TCP*) packets of length 1024 bytes. Background traffic is described in the chapter 2.3. Voice traffic was generated using *VoIP* clients. Session Initiation Protocol (*SIP*) was used for established *VoIP* connection. For the measurement we chose the *ITU-T G.729A* encoding scheme [14]. In the measurement, two frames were encapsulated into a packet in turn; it corresponding to a packet size of 20 milliseconds. Adaptive jitter buffer, packet loss concealment, and Voice Activity Detection (*VAD*) / Discontinuous Transmission (*DTX*) are implemented in the *VoIP* clients used.

The measurements have been performed for six different testing conditions. The reference signals described in chapter 2.2 are utilized for transmission through the given *VoIP* connection. Finally, speech quality was measured by *PESQ* and then converted to *MOS-LQO* (*MOS*-listening quality objective) by the equation:

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945*x+4.6607}} \qquad (1)$$

where *x* and *y* represent the raw *PESQ* score and the mapped *MOS-LQO*, respectively. This equation is defined by *ITU-T* Recommendation *P.862.1* [15].

2.2 Description of the reference signals

The reference signals selection should follow the criteria given by *ITU-T* Recommendation *P.830* [16] and *ITU-T* Recommendation *P.800* [1]. The reference signals should include talkbursts separated by silence periods, and are normally of 1-3 seconds long. Also it should be active for 40-80% of their duration. The reference signals are composed from speech records. In our experiments, these speech records have been taken from a Slovak speech database. In each set, two female and two male speech utterances were used. The reference signals were stored in 16-bit, 8000 Hz linear *PCM*. Table 1 shows the active speech and background noise levels for each of used reference signals. The stationary background noise with no significant peaks in frequency spectrum is present from recording process.

Table 1 Active speech and background noise levels

Reference signal	**Active speech level [dB]**	**Background noise [dB]**
Male1	-22.09	-48.33
Male2	-30.05	-49.63
Female1	-20.98	-48.74
Female2	-23.47	-48.3

We conducted two types of experiments related to reference signal characteristics. Those experiments can be classified into:

- Reference-Signal-Length experiment,
- Active-Speech-Ratio experiment.

In the next subsections, the reference signals used for the purpose of above mentioned experiments are described in more details.

2.2.1 Description of a reference signal for the Reference-Signal-Length experiment

Reference speech signals of lengths of 10, 20 and 30 seconds were used for the purpose of this experiment. The average value of active speech ratio for each of signal lengths is about 63% of their lengths. The active speech ratio measurement process has to follow the criteria given by *ITU-T* Recommendation *P.56*. Those ratios have been measured by means of *ITU-T* Recommendation *G.191*'s software tool [17], known as *sv56*.

2.2.2 Description of a reference signal for the Active-Speech-Ratio experiment

In this case, the reference signals in length of 30 seconds with active speech ratios of 42, 62 and 82% were applied. The decision about using reference signals in length of 30 seconds comes from the first experiment results (presented in Section 3.1). This length provides more accurate results in comparison with other investigated lengths therefore enables more precise investigation of an impact of different active speech ratios on speech quality estimation, assessed by *PESQ*. The active speech ratios for each of used reference signals are presented in Table 2.

Table 2 Active speech ratios

Reference signal	**Active speech ratio of 42%**	**Active speech ratio of 62%**	**Active speech ratio of 82%**
Male1	42.731	57.106	81.142
Male2	41.749	60.808	81.609
Female1	41.525	64.401	83.071
Female2	42.746	65.743	82.199
Average value	**42.188**	**62.014**	**82.005**

2.3 Background traffic

Background traffic has been generated by *D-ITG* traffic generator. The primary goal of background traffic is two-fold. Firstly, it simulates the standard traffic that appears in current *IP* networks, which includes data transfer via Hypertext Transfer Protocol (*HTTP*) and File Transfer Protocol (*FTP*), multimedia streams for real-time applications. Secondly, it affects *VoIP* transmission by changing of *VoIP* connection network performance parameters such as delay, jitter and packet loss. The simulated background traffic includes the following three types of communication:

- "Data transfer service", which includes *FTP* and other non specified services, is represented as information stream with constant bit rate based on *TCP*.
- "Multimedia streaming service" represents real-time multimedia applications and therefore is based on information stream with a constant bit rate. The *UDP* is used in this case.
- "Web service" that is simulated as a sequence of separated data bursts with Poisson distribution of packet rate. The active period of the burst is 400 ms and the bursts appear periodically every two seconds. *TCP* was used for the purpose of this service.

Table 3 Performance evaluation of testing conditions

Testing condition	**Data transfer Service [Mb/s]**	**Streaming service [Mb/s]**	**Web service [Mb/s]**	**Average offered traffic load [%]**
0	0	0	0	0
1	2	2.5	0.5	50
2	2.25	2.82	0.56	56.3
3	2.5	3.14	0.61	62.5
4	2.75	3.45	0.68	68.8
5	3	3.76	0.74	75

As mentioned in Section 2.1, the measurements have been performed for six different testing conditions. The selected bit rates of the three above mentioned types of communication, and average offered traffic load of background traffic, normalized to network capacity, are described in Table 3.

The simulation of the multimedia streaming service was carried out from the point of view of the impact of this service traffic on speech quality provided by *PESQ*. Note *D-ITG* traffic generator doesn't allow the simulation of the multimedia streaming service using Real-time Transport Protocol (*RTP*) but the *RTP* based streaming service has the same impact on speech quality as the streaming using *UDP*.

3 Experimental results

The measurements were independently performed 40 times under same testing condition for both the investigated reference signal characteristics (i.e. Reference-Signal-Length, Active-Speech-Ratio). The *MOS-LQO* scores were averaged and standard deviation values are described in Tables 4 and 5.

Table 4 Standard deviations of MOS-LQO for the Reference-Signal-Length experiment

Testing condition	**10 seconds**	**20 seconds**	**30 seconds**
0	0.2706	0.2134	0.0911
1	0.2430	0.2322	0.2252
2	0.2913	0.2734	0.2513
3	0.2962	0.2437	0.2350
4	0.2437	0.2066	0.1743
5	0.2961	0.2386	0.2314

Table 5 Standard deviations of MOS-LQO for the Active-Speech-Ratio experiment

Testing condition	**42%**	**62%**	**82%**
0	0.2501	0.3505	0.1462
1	0.2963	0.3352	0.2244
2	0.2116	0.2648	0.2775
3	0.2505	0.2184	0.2068
4	0.2642	0.2135	0.1925
5	0.1626	0.2443	0.2318

The next subsections describe experimental results for both examined reference signal characteristics in more details.

3.1 Results for the Reference-Signal-Length experiment

Figure 3 shows the results for all investigated Reference-Signal-Lengths. The graph represents the dependence of *MOS-LQO* change on the testing conditions. The testing conditions represent a few types of network conditions. Each network condition is described by traffic load. The increasing traffic load causes jitter and also packet loss increase. In general, speech quality drops with increasing packet loss and jitter. Figure 2 shows the traffic load for given testing conditions and investigated signal lengths. The traffic load was measured by means of the *Wireshark* network analyzer [18]. The transmission rates for given testing conditions are described in Table 3. The impact of background traffic on the packet loss in *VoIP* connection is shown in Figure 5.

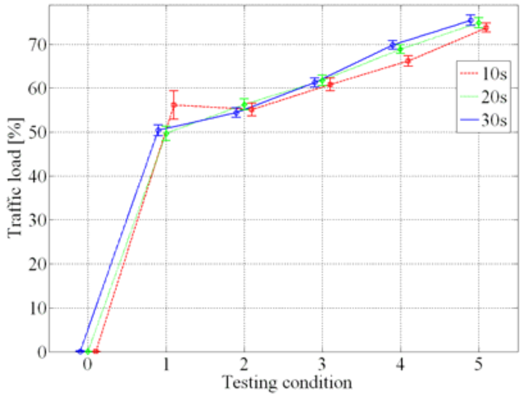

Fig.2 Traffic load for given testing conditions. The vertical bars show 95 % CI (derived from 40 measurements) for each testing condition. The testing condition numbers correspond to Table 3.

The 1035 voice packets were approximately transmitted during one 20 seconds long *VoIP* connection (20 seconds length of reference signal). Total packet loss ranged from 0.007 to 6.45% for these measurements. The total packet loss consists of two components. The first component is lost packets and the second component is dropped packets.

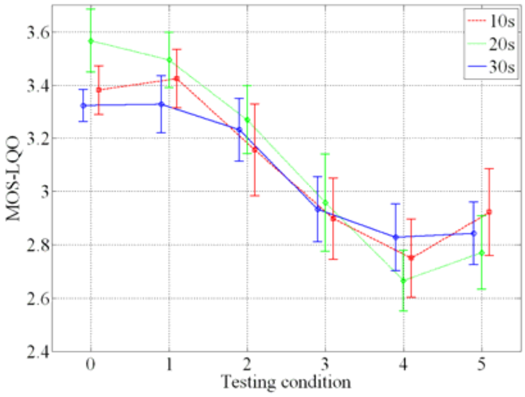

Fig.3 MOS-LQO as a function of background traffic for different reference signals lengths. Other detailed descriptions of Figure 2 apply appropriately.

In Figure 3, some differences among Reference-signals-Lengths can be seen from *MOS-LQO* perspective. They are caused by time-varying traffic load in the small scale (Figure 2), which results in variations of packet delay, so-called jitter. Jitter affects the probability that packets cannot be incorporated into speech reconstruction process because they have not been received in time. Adaptive buffers can alleviate this problem by buffering packets and delaying their playout time to compensate for varying network delay, while however, the absolute end-to-end delay must be limited to allow a fluent conversation. Figure 4 shows jitter values for all investigated testing conditions. The average value of jitter ranged from 2.55 to 14.71 milliseconds in the presented case.

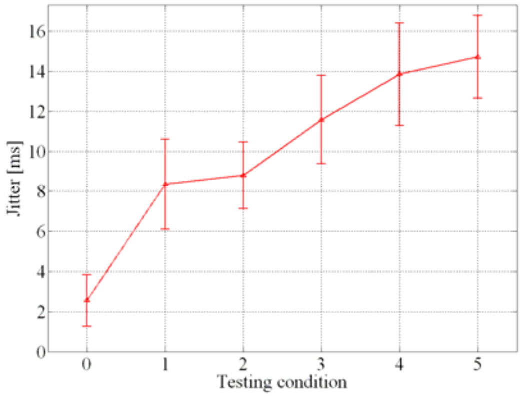

Fig.4 Impact of background traffic on jitter in VoIP connection for 20 seconds length of reference signal. Other detailed descriptions of Figure 2 apply appropriately.

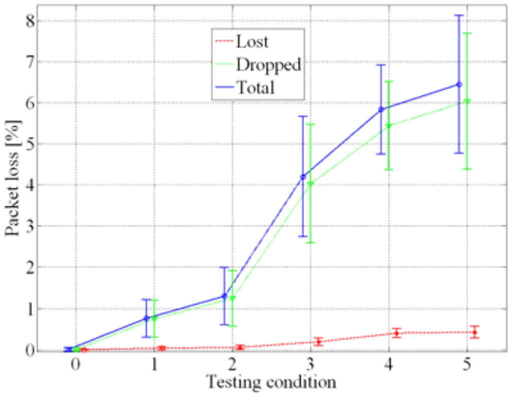

Fig.5 Impact of background traffic on packet loss in VoIP connection for 20 seconds length of reference signal. Other detailed descriptions of Figure 2 apply appropriately.

It was mentioned above that the packets delivered after playout deadline are not incorporated to speech reconstruction process. Those packets are called dropped packets and are depicted in Figure 5 (dashed green line). We can see from Figure 5 that dropped packets cover a big amount of total packet loss (solid blue line).

Table 6 Standard deviations of dropped packets for Reference–Signal-Length experiment

Testing condition	**10 seconds [%]**	**20 seconds [%]**	**30 seconds [%]**
0	0.3486	0.1676	0
1	1.9378	1.7262	1.1143
2	2.8558	2.4707	2.0914
3	3.3380	2.7540	2.4697
4	3.8097	2.4044	1.9511
5	4.0198	3.1270	2.7713

The standard deviations of dropped packets are described in Table 6. Values in the table show the standard deviations decline for longer lengths of the reference signal at all investigated testing conditions. The greater Reference-Signal-Lengths results in stable values of the network performance parameters (given by lower standard deviations); such as jitter, packet loss, etc.; and then naturally provide *MOS-LQO*'s more precisely. This fact is described in Table 4 and enables to realize more accurate objective evaluation of speech quality in *IP* networks by means of stable values of network performance parameters in the case of using a longer reference signal.

3.2 Results for Active-speech-ratio experiment

Figure 7 shows the results for all investigated Active-Speech-Ratios. The relationships between *MOS-LQO*'s and testing conditions for different Active-Speech-Ratios are depicted in this graph. Figure 6 shows the traffic load for given testing conditions. The impact of background traffic on the jitter and packet loss in *VoIP* connection is shown in Figures 8 and 9, respectively. Some additional information's about testing conditions are described above, in Section 3.1.

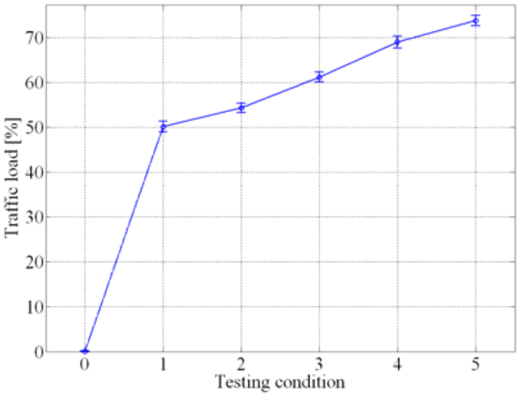

Fig.6 Traffic load for given testing conditions (30 seconds length of reference signal with Active-Speech-Ratio of 62%). Other detailed descriptions of Figure 2 apply appropriately.

Figure 7 depicts differences between investigated Active-Speech-Ratios in speech quality evaluation, provided by *PESQ*. It can be seen from Figure 7 that the difference in Active-Speech-Ratio has a significant impact on overall speech quality. This fact contributes our preliminary assumption that an increasing amount of speech (Active-Speech-Ratio) in reference signal has to result in increasing of reference signal sensitivity to packet loss change. That may be explained by increasing/decreasing of information (speech) loss probability at the same packet loss ratio in the case of higher/lower Active-Speech-Ratio. It is caused by a greater number of active speech periods in reference signals with higher Active-Speech-Ratio. The probability of information loss is greater if more periods are available. It means that it is possible to capture more impairments of speech quality in such case. By capturing majority of existing impairments, we are able to get a better insight about speech quality in investigated telecommunication network (especially in *VoIP* case) which turns to more reliable evaluation of investigated transmission line from this point of view. This effect is depicted in Figure 7. In more detail, it can be seen in this figure that *MOS-LQO* for higher Active –Speech-Ratio (82%) decreases faster in comparison with ratios 42% and 62%, but only for low values of traffic load (Testing conditions No.0-3). It was mentioned above that the reference signals with higher Active-Speech-Ratio contain more speech periods; it results in increasing of information loss probability and that is account for above mentioned *MOS-LQO* decreasing in the case of same packet loss ratio or the same testing condition.

However, it can be seen from Figure 7, that solid blue line (Active-Speech-Ratio of 82%) has a steeper slope than the other lines to the left of testing condition No.3. On the other hand, the slope of solid blue line is the same or slightly smaller than the other two lines to the right of testing condition No.3. It seems from these experimental results that an increment of reference signal sensitivity to packet loss change by higher Active-Speech-Ratios is only achieved for packet loss below 4% (Testing conditions No.0-3) (Figure 9). Probably, that is caused by decreasing of difference among captured number of speech quality impairments during active speech periods by higher packet loss ratios in the case of different Active-Speech-Ratios used. In other words, total number of captured impairments for different Active-Speech-Ratios will not be markedly changed by higher packet loss ratios. It causes that the change of information (speech) loss probability will not be achieved by Active-Speech-Ratios modification for higher packet loss ratios. It can result in similar slopes of *MOS-LQO* lines for all investigated Active-Speech-Ratios and for testing conditions above No.3 (Figure 7). Naturally, that is a point for a future investigation in this area because it requires a more precise elaboration.

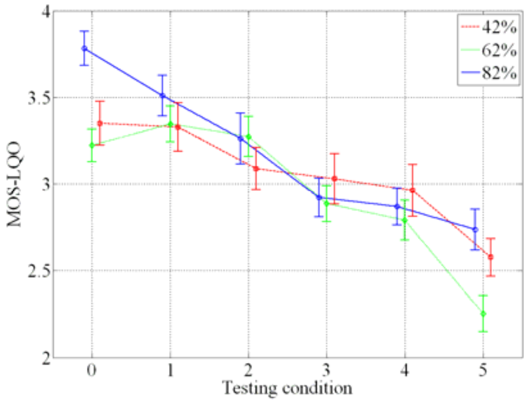

Fig.7 MOS-LQO as a function of background traffic for different active speech ratios. Other detailed descriptions of Figure 2 apply appropriately.

As can be seen from Figure 7, substantial difference among *MOS-LQO* values of investigated Active-Speech-Ratios has been obtained in the case of testing condition No.0 (0% traffic load, 0% packet loss), especially between 82% Active-Speech-Ratio and the other two cases. At this time, we have no theory that could explain this phenomenon. Naturally, that is a point for a future investigation because exhaustive measurements are needed to validate, study and interpret this phenomenon.

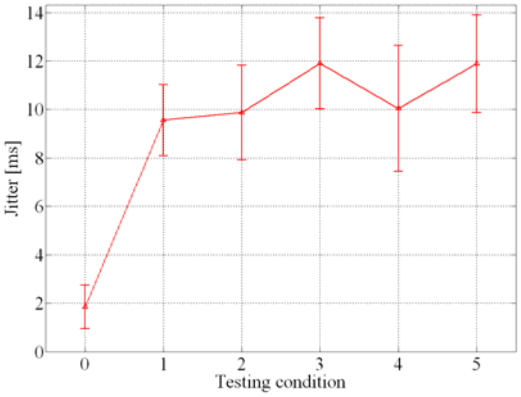

Fig.8 Impact of background traffic on jitter in VoIP connection for 30 seconds length of reference signal with Active-Speech-Ratio of 62%. Other detailed descriptions of Figure 2 apply appropriately.

The 1520 voice packets were approximately transmitted during one 30 seconds long *VoIP* connection (30 seconds length of reference signal). Total packet loss ranged from 0.08 to 6.62% and the average jitter values ranged from 1.84 to 11.88 milliseconds in the case of presented results (Figures 8, 9).

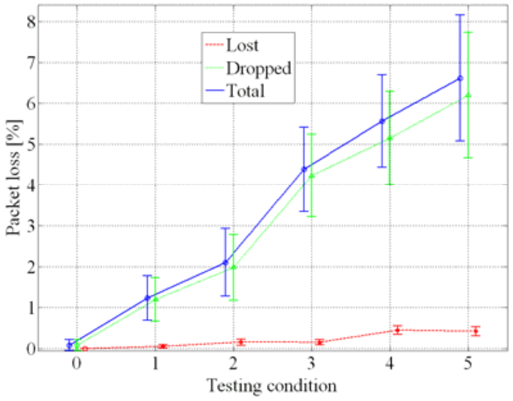

Fig.9 Impact of background traffic on packet loss in VoIP connection for 30 seconds length of reference signal with Active-Speech-Ratio of 62%. Other detailed descriptions of Figure 2 apply appropriately.

Table 7 describes the standard deviations of dropped packets which have been captured for this experiment. It can be seen from Tables 5 and 7 that it is not possible to improve speech quality assessment accuracy by means of an Active-Speech-Ratio modification but an increment in reference signal sensitivity to packet loss change can be achieved by this approach.

Table 7 Standard deviations of dropped packets for Active–Speech-Ratio experiment

Testing condition	42% [%]	62% [%]	82% [%]
0	0.7456	0.1634	0.4895
1	2.5078	1.7405	2.4401
2	1.7671	2.4492	2.1151
3	2.4142	2.3107	2.2398
4	2.8441	1.9391	2.1746
5	3.5247	2.8579	3.5247

Our experimental results show that the change of Active-Speech-Ratios has a significant impact on overall speech quality, especially in the case of lower values of packet loss, below 4%. This fact is our motivation for finding of the feasible average Active-Speech-Ratios for some languages or types of languages and conversational scenarios. Naturally, an issue of Active-Speech-Ratio setup with regards to different languages and conversational scenarios is also open for discussion. Average Active-Speech-Ratios adjustment might enable to provide an assessment of speech quality more reliably.

Nowadays, such improved assessment of speech quality is demanded to be involved into Quality of Service (*QoS*) in real *VoIP* scenarios to become comparison among network providers more feasible.

4 Conclusion and future work

This paper investigated an impact of different lengths and Active-Speech-Ratios of an input reference signals in *PESQ* based speech quality estimation in simulated *VoIP* environment. The results presented in the paper confirm our preliminary assumption that both the investigated characteristics of the reference signals (Reference-Signal-Length and Active-Speech-Ratio) may have an impact on the final *PESQ*'s estimation.

In case of the Reference-Signal-Length characteristic, lower standard deviations of dropped packets and of *MOS-LQO*'s scores have been achieved for longer durations of the reference signal at all investigated testing conditions. This fact allows more accurate objective evaluation of speech quality in *VoIP* scenarios by means of stable values of network performance parameters (packet loss, etc).

An investigation of an impact of the Active-Speech-Ratio on *MOS-LQO* approve our second hypothesis that an increase in amount of speech in the reference signal (expressed by the Active-Speech-Ratio characteristic) may result in an increase of the reference signal sensitivity to packet loss change. In this experiment, this effect has been observed only for packet loss below 4%. This outcome inspires us to find out the feasible average Active-Speech-Ratios for some languages or types of languages and conversational scenarios. We suppose that adjustment of average Active-Speech-Ratios may help to provide for more reliable assessment of speech quality.

A future work will focus towards the following issues. First, we would like to verify our results by subjective tests and afterwards propose the final findings that could provide for objective speech quality evaluation more closely to reality in real *VoIP* scenarios. We are currently preparing the subjective tests in cooperation with *MESAQIN*'s laboratory in Prague (Czech Republic). Secondly, we plan to investigate an increase in reference signal sensitivity to a packet loss change by higher Active-Speech-Ratios for different languages and packet loss patterns. Thirdly, we want to investigate the impact of different Active-Speech-Ratios on speech quality in the case of 0% packet loss for different languages. Fourthly, we will attempt to find out an appropriate average Active-Speech-Ratios for some languages or type of languages and conversational scenarios. Apparently this point could be very interesting for other speech quality laboratories around the world. By this investigation, we might refine on the existing broadly recommended Active-Speech-Ratios (40% - 80%), defined by *ITU-T* Recommendation *P.862.3* and provide for more reliable speech quality assessment, provided by *PESQ*.

5 Acknowledgements

This work has been partially supported by the VEGA grant 1/0313/08 of Scientific Grant Agency of the Slovak Republic.

We would like to thank Joachim Pomy and Jan Holub for valuable comments and help in preparing of this paper, and last not least the reviewers for their invaluable comments and suggestions.

6 References

[1] ITU-T Rec. P.800 “Methods for subjective determination of transmission quality” International Telecommunications Union, Geneva, 1996.

[2] ITU-T Rec. P.862 “Perceptual Evaluation of Speech Quality “International Telecommunications Union, Geneva, 2001.

[3] 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 network and codecs, In *Proceedings of IEEE international conference on Acoustics, speech and signal processing, Salt Lake City, USA, 2001, pp.749-752.*

[4] ITU-T Rec. P.862.3 “Application guide for objective quality measurement based on

Recommendations P.862, P.862.1 and P.862.2 "International Telecommunications Union, Geneva, 2005.

[5] ITU-T Rec. P.56 "Objective measurement of active speech level "International Telecommunications Union, Geneva, 1993.

[6] Rix, A. W.: Comparison between subjective listening quality and P.862 PESQ score, In *Proceedings of conference MESAQIN 2003*, Prague (Czech Republic), 2003, ISBN 80-01-02822-4.

[7] Hoene, Ch., Dulamsuren-Lalla, E.: Predicting Performance of PESQ in Case of Single Frame Losses, In *Proceedings of conference MESAQIN 2004*, Prague (Czech Republic), 2004, ISBN 80-01-03017-2.

[8] Sun, L., Ifeachor, E.C..: Subjective and Objective Speech Quality Evaluation under Bursty Losses, In *Proceedings of conference MESAQIN 2002*, Prague (Czech Republic), 2002, ISBN 80-01-02515-2.

[9] Pennock, S.: Accuracy of the Perceptual Evaluation of Speech Quality (PESQ) algorithm, In *Proceedings of conference MESAQIN 2002*, Prague (Czech Republic), 2002, ISBN 80-01-02515-2.

[10]Varela, M., Marsh, I., Gronvall, B.: A systematic study of PESQ's behaviour, In *Proceedings of conference MESAQIN 2006*, Prague (Czech Republic), 2006, ISBN 80-01-03503-4.

[11]Takahashi, A.: Objective quality evaluation based on ITU-T Recommendation P.862 by using long reference speech (NTT), COM12-D008, Jan. 2005.

[12]Počta, P., Vaculík, M.: Impact of duration of speech sequences on speech quality, In *Journal of Telecommunications and Information Technology*, vol. 7, no. 4, pp.72-76, ISSN 1509-4553.

[13]Botta, A., Dainotti, A., Pescapè, A.: Multi-protocol and multi-platform traffic generation and measurement, In *Proceedings of conference INFOCOM 2007*, Anchorage (Alaska, USA), 2007.

[14]ITU-T Rec. G.729 "Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Exited Linear prediction (CS-ACELP)" International Telecommunications Union, Geneva, 1996.

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

[16]ITU-T Rec. P.830 "Subjective Performance Assessment of digital telephone-band and wideband digital codecs" International Telecommunications Union, Geneva, 1996.

[17]ITU-T Rec. G.191 "Software tools for speech and audio coding standardization" International Telecommunications Union, Geneva, 2005.

[18]Wireshark network analyzer, http://www.wireshark.org/