ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 95 (2009) 950 – 957

DOI 10.3813/AAA.918225

Impact of Different Active-Speech-Ratios on PESQ's Predictions in Simulated VoIP Environment

Peter Počta[1)], Jan Holub[2)], Miroslava Mrvová[1)]

[1)] Dept. of Telecommunications and Multimedia, FEE, University of Žilina, Univerzitná 1, 01026, Žilina, Slovakia. pocta@fel.uniza.sk

[2)] Dept. of Measurement K13138, FEE, CTU Prague, Technicka 2, 16627, Prague 6, Czech Republic. holubjan@fel.cvut.cz

Summary

In this work, we experimentally study how behaviour of the *PESQ* predictions varies with reference signal characteristic. In particular we investigate the impact of different Active-Speech-Ratios on speech quality prediction in simulated *VoIP* environment from objective and subjective testing point of view. This reference signal characteristic is defined very broadly by *ITU-T* Recommendation *P.862.3*. That is the reason to investigate an impact of this characteristic on speech quality prediction more in-depth. We assess the variability of *PESQ*'s predictions with respect to Active-Speech-Ratio and network conditions, as well as their accuracy, by comparing the predictions with subjective assessments.

PACS no. 43.71.Gv, 43.72.Kb

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 as well as to 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. Subjective testing is expensive and time-consuming. That is the reason that subjective testing 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. Intrusive 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*) [2], Measuring Normalizing System (*MNB*) [3, 4], Perceptual Analysis Measurement System (*PAMS*) [5], and Perceptual Evaluation of Speech Quality (*PESQ*) [6, 7]. Among them, *PSQM* [8] and *PESQ* [9] 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. *E-Model* is a typical model, 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 (R-value) can be transformed into the *MOS*-Listening Quality Estimated narrowband (*MOS-LQEn*) values.

The *PSQM* algorithm 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 the 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

Received 18 November 2008,
accepted 20 May 2009.

the other objective algorithms [6, 7, 9]. Unlike the conversational model, *PESQ* is a listening-only model; the degraded sample is time-aligned with the reference sample during pre-processing. The *PESQMOS* 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* [10]. Two reference signal characteristics are defined very broadly by this Recommendation from our point of view, namely the length of reference signal and Active-Speech-Ratio. The above-mentioned recommendation recommends to use the reference signals in duration in the range from 8 seconds to 30 seconds for the purpose of *PESQ* measurement. The speech activity in the reference signals, which can be measured according to *ITU-T* Recommendation *P.56* [11], should be between 40% and 80% of their length. We suppose that those two characteristics can have an impact on final *PESQ*'s predictions. The detailed investigation of both characteristics has been proposed in [12] from *PESQ*'s prediction point of view. Some very important issues raises from [12] especially in the case of Active-Speech-Ratio experiment. That is the reason for exhaustive investigation of the impact of different Active-Speech-Ratios on speech quality assessment provided by *PESQ* and subjective tests.

Some works have been carried out on study of *PESQ*'s accuracy and behavior. Particularly, [13, 14, 15, 16] have examined the *PESQ*'s accuracy in some cases. In [13], the 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 [14], 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 [15], 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 has been reported. Preliminary results have shown 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 [16], the effects of speech coder, packet loss concealment strategy, IP payload size, packet loss rate and burstiness of packet losses on *PESQ* accuracy have been assessed. The results have indicated 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 (*SLAs*). In [17], 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.

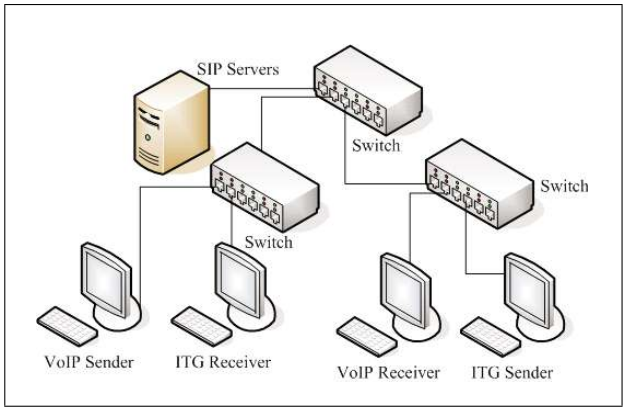

Figure 1. Experimental scenario.

Here we focus on an impact of different Active-Speech-Ratios on speech quality assessment provided by *PESQ* and subjective tests in simulated *VoIP* environment. The reference signals with Active-Speech-Ratios of 42, 62 and 82% are investigated in this study. We assess the variability of *PESQ*'s predictions with respect to Active-Speech-Ratio and network conditions, and also their accuracy, by comparing the predictions with subjective assessments.

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. Experiment description

2.1. Experimental scenario

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

Two stations (*ITG* Sender and *ITG* Receiver) equipped with the accomplished D-*ITG* traffic generator [18] were used to generate and receive background traffic. *ITG* Sender generated the User Datagram Protocol (*UDP*) and Transmission Control Protocol (*TCP*) packets of 1024 byte length. Background traffic is described in section 2.3 in more details. Voice traffic was generated using *VoIP* clients. Session Initiation Protocol (*SIP*) was used for established *VoIP* connection. For this experiment the *ITU-T* G.729AB encoding scheme [19] was chosen. In the measurements, two frames were encapsulated into a single packet; thus corresponding to a packet size of 20 milliseconds. Adaptive jitter buffer, G. 729AB's native packet loss concealment (*PLC*), and Voice Activity Detection (*VAD*)/Discontinuous Transmission (*DTX*) were im-

plemented in the *VoIP* clients used. The Comfort Noise Generator (*CNG*) was disabled in case of this experiment.

The measurements were performed for six different testing conditions. The reference signals described in section 2.2 were utilized for transmission through the given *VoIP* connection. Finally, speech quality was measured by *PESQ* and then converted to *MOS*-Listening Quality Objective narrowband (*MOS-LQOn*) by

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

where x and y represent the raw *PESQ* score and the mapped *MOS-LQOn*, respectively. The equation mentioned is defined by *ITU-T* Recommendation P.862.1 [20]. Afterwards, the accuracy of *PESQ* predictions was assessed by comparing the predictions with subjective assessments. Detailed information about subjective assessment can be found in section 2.4.

2.2. Reference signals

The reference signals selection should follow the criteria given by *ITU-T* Recommendations *P.830* [21] and *P.800* [1]. The reference signals should include talkbursts separated by silence periods, and are normally of 1–3 seconds long. They should also be active for 40–80% of their duration. The reference signals are composed of speech records. In our experiments, these speech records were 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 I 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 was present from recording process.

Reference signals in length of 30 seconds with Active-Speech-Ratios of 42, 62 and 82% were applied. All reference signals used were spoken by the same people (as defined in Table II), also for different Active-Speech-Ratios. The differences between reference signals used are only in case of number of talkspurts (sentences), resulting in different Active-Speech-Ratios. In case of higher Active-Speech-Ratios, the new sentences were added, as an extension. The decision about using reference signals in length of 30 seconds came from our previous published work [12]. The tests have proved that 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 prediction, assessed by *PESQ*. The long reference signals usage for the speech quality assessment by *PESQ* has been investigated in [22]. The experimental results have shown that for this purpose it is possible to use a longer reference signals and author has proposed extending the maximum length of reference signals to 30 seconds. The results of this work have been included in *ITU-T* Recommendation *P.862.3*.

The Active-Speech-Ratios and numbers of talkspurts (active speech periods) for each of used reference signals are presented in Table II. The Active-Speech-Ratio measurement process has to follow the criteria given by *ITU-T* Recommendation *P.56*. Those ratios were measured by means of *ITU-T* Recommendation G.191's software tool [23], known as sv56.

Table I. Active speech and background noise levels of the reference signals.

Reference signal	Active speech level [dB]	Background noise [dB]
Male1	−22.09	−48.33
Male2	−30.05	−49.63
Female 1	−20.98	−48.74
Female 2	−23.47	−48.30

Table II. Active-Speech-Ratios (ASR) and numbers of talkspurts of the reference signals.

Reference	ASR 42%	ASR 62%	ASR 82%
Male1	42.731 (8)	57.106 (9)	81.142 (12)
Male2	41.749 (4)	60.808 (6)	81.609 (10)
Female 1	41.525 (3)	64.401 (5)	83.071 (6)
Female 2	42.746 (4)	65.743 (5)	82.199 (6)
Average	42.188 (4.75)	62.014 (6.25)	82.005 (8.5)

2.3. Background traffic

Background traffic was generated by D-*ITG* traffic generator. The primary task 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.

As mentioned in section 2.1, the measurements were 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 III.

The simulation of the multimedia streaming service was carried out from the aspect of the impact of this service

Table III. Performance evaluation of testing conditions. Tc: Testing conditions, DtS: Data transfer Service [Mb/s], Ss: Streaming service [Mb/s], Ws: Web service [Mb/s], Aotl: Average offered traffic load [%].

Tc	DtS	Ss	Ws	Aotl
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

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*.

2.4. Subjective assessment

The subjective listening tests were performed in accordance to *ITU-T* Recommendation *P.800* [1]. Always up to 8 listeners were seated in listening chamber with reverberation time less than 190 ms and background noise well below 20 dB SPL (A). All together, 21 listeners in the age of 19–30 years participated in the tests, the number of male and female listeners being balanced.

The samples were played out using high quality studio equipment in random order. Results in Opinion Score 1 to 5 were averaged to obtain *MOS*-Listening Quality Subjective narrowband (*MOS-LQSn*) values for each sample.

All together, 108 speech samples were selected for subjective testing. Always 6 samples represented one network testing condition and Active-Speech-Ratio. The 6 samples mentioned were composed of 3 male and 3 female samples. In each sample collection, the best, average and worst cases were chosen from speech quality and packet loss perspective. These were selected out of all recorded samples by expert listening.

3. Experimental results

In this section, we describe and explain experimental results for objective assessment and comparison with subjective scores in more details, respectively.

3.1. Experimental results for objective assessment

The measurements were independently performed 40 times under the same testing condition. The *MOS-LQOn* scores were averaged and standard deviations are described in Table IV.

Figure 3 shows the results for all the investigated Active-Speech-Ratios. The relationships between *MOS-LQOn*'s and testing conditions for different Active-Speech-Ratios are depicted in this graph. The testing conditions represent a few types of network conditions. Each

Table IV. Standard deviations of *MOS-LQOn*'s.

Tc	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

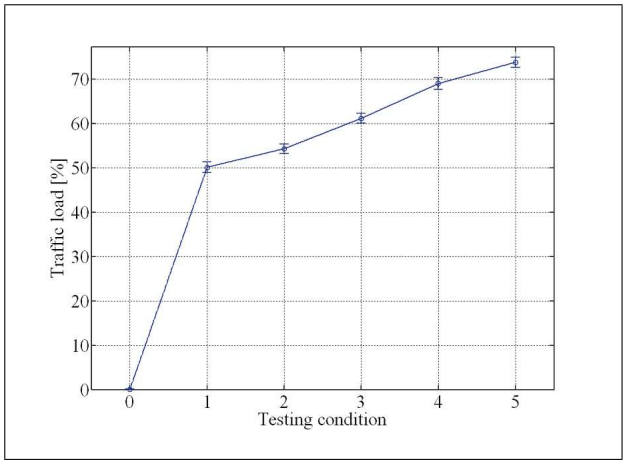

Figure 2. Traffic load for given testing conditions (30 seconds length of reference signal with Active-Speech-Ratio of 62%).The vertical bars show 95% CI (derived from 40 measurements) for each testing condition. The testing condition numbers correspond to Table III.

network condition is described by a 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. The traffic load was measured by means of the Wireshark network analyzer [24]. The transmission rates for given testing conditions are described in Table III. The impact of background traffic on the jitter and packet loss in *VoIP* connection is shown in Figures 4 and 5, respectively.

Figure 3 depicts differences between investigated Active-Speech-Ratios in speech quality evaluation, provided by *PESQ*. It can be seen from above-mentioned figure 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 in reference signal (expressed by the Active-Speech-Ratio characteristic) has to result in increase of reference signal sensitivity to packet loss change. That may be explained by increase/decrease 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 a case. By capturing majority of ex-

isting 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 3. In more detail, it can be seen in this figure that *MOS-LQOn* 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 increase of information loss probability and that is account for above-mentioned *MOS-LQOn* decreasing in the case of same packet loss ratio or the same testing condition.

However, it can be seen from Figure 3, 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. From these experimental results it seems 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 5). Probably, that is caused by decrease 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-LQOn* lines for all the investigated Active-Speech-Ratios and for testing conditions above No.3 (Figure 3). Naturally, that is a point for a future investigation in this area because it requires a more precise elaboration.

On the other hand, the fact that the speech quality is better for higher Active-Speech-Ratios (Figure 3) may simply be caused by a low amount of additive stationary background noise in the recordings used in the case of this experiment. It is widely known that the type and level of background noise that is present between active speech intervals might have a strong influence on the results. This background noise becomes more audible in the longer silent periods of speech (reference signal) with lower Active-Speech-Ratio, especially in case of 0% packet loss (Figure 3, Testing condition No.0). More information about the background noise levels for each of the reference signals used can be found in Table I. The low amount of background noise mentioned is normally present from recording process.

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 the presented results (Figures 4, 5).

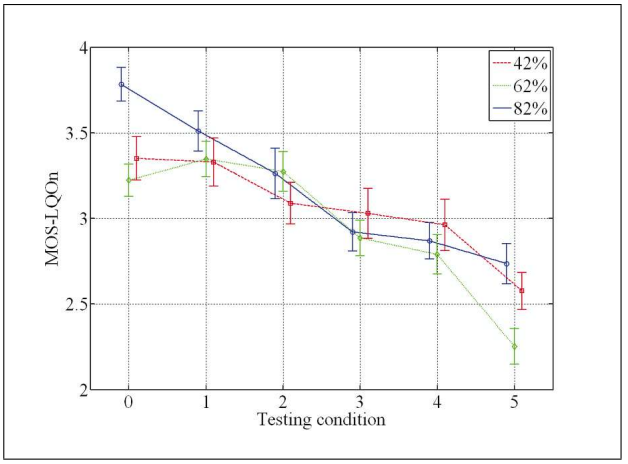

Figure 3. *MOS-LQOn* as a function of background traffic for different Active-Speech-Ratios. Other detailed descriptions of Figure 2 apply appropriately.

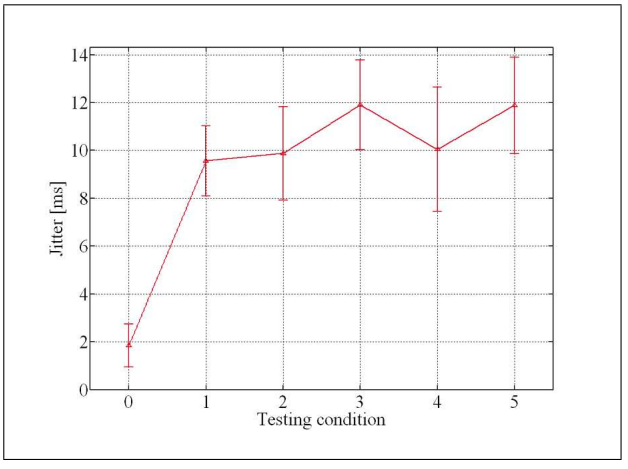

Figure 4. Impact of background traffic on jitter in *VoIP* connection for 30 seconds length of reference signal with Active-Speech-Ratio of 62%. Other detailed descriptions of Figure 2 apply appropriately.

Table V describes the standard deviations of dropped packets which have been captured for this experiment. From Tables IV and V it can be seen 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.

The objective experimental results show that the change of Active-Speech-Ratio 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

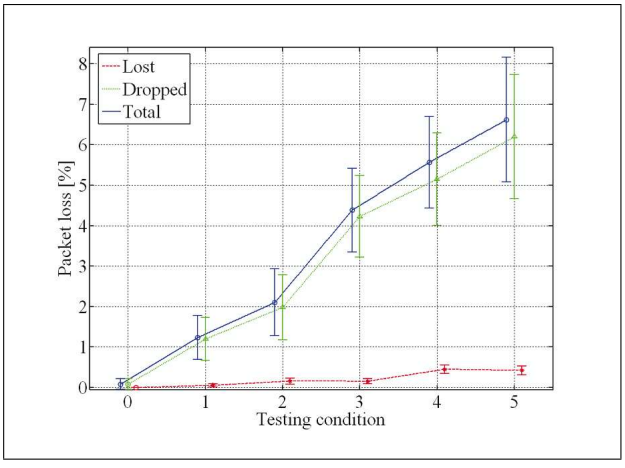

Figure 5. Impact of background traffic on packet loss in *VoIP* connection for 30 seconds length of reference signal with Active-Speech-Ratio of 62%. Other detailed descriptions of Figure 2 apply appropriately.

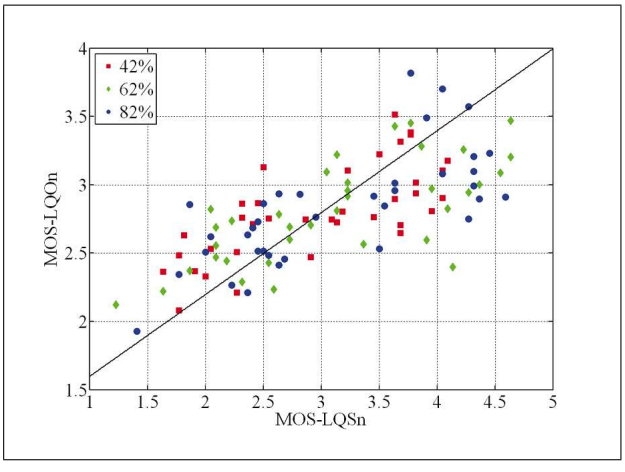

Figure 6. Subjective results (*MOS-LQSn*) versus *MOS-LQOn* output (not regressed).

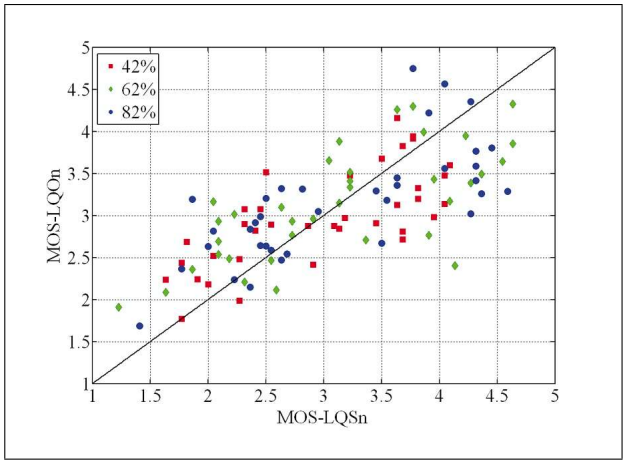

Figure 7. Subjective results (*MOS-LQSn*) versus *MOS-LQOn* output (1st order regression).

Table V. Standard deviations of dropped packets.

Tc	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

Table VI. Pearson correlation coefficient and Root Mean Square Error between *MOS-LQSn* and *MOS-LQOn* before and after 1st order regression.

ASR	42%	62%	82%
ρ before regression	0.7159	0.7046	0.7219
δ after regression	0.6146	0.7841	0.7517
ρ before regression	0.7159	0.7046	0.7219
δ after regression	0.5465	0.5630	0.6329

into Quality of Service in real *VoIP* scenarios to make comparison among network providers more feasible.

3.2. Comparison with subjective scores

The results obtained by means of subjective testing (*MOS-LQSn*) are compared with *MOS-LQOn* results in Figures 6 and 7. Obviously, the sensitivity to Active-Speech-Ratio modification of *PESQ* is much weaker than that of human subjects (see Figure 6). As attempts to use 2-nd or 3-rd order regression (as recommended in *ITU-T* Recommendation *P.862*) lead to non-monotonic results, the 1-st order regression was used instead. Figure 7 depicts the results after the 1-st order regression.

As can be seen from Figure 7, *PESQ* over-predicts speech quality when losses are low, and under-predicts it when the losses are high. In general, the low packet losses produce higher *MOS* scores and high losses generate lower *MOS* scores. More details about this fact can be found in Figure 3. Similar results (over- and under-prediction behaviours) however from bursty losses perspective have been obtained in [17].

The *PESQ*'s performance from Active-Speech-Ratio perspective is characterized by the Pearson correlation coefficient ρ and Root Mean Square Error (RMSE) δ. The statistics for ρ and δ are summarized in Table VI. It is seen that only δ has been changed by means of regression.

Figure 8 depicts differences between obtained *MOS-LQSn* and *MOS-LQOn* scores for Active-Speech-Ratio of 82%. In the case of other investigated Active-Speech-Ratios similar curves have been achieved. Because of restricted space of this paper and similar results, we present only results obtained for Active-Speech-Ratio of 82%. It is possible to seen from this picture that the shapes of the curves are the same but shifted in *MOS* scale (in negative or positive way). It is caused by *PESQ*'s over- and under-prediction behaviours, which were shown in Figure 7. It seems from those curves that there may be dependency between *PESQ*'s Active-Speech-Ratio and under- and over-

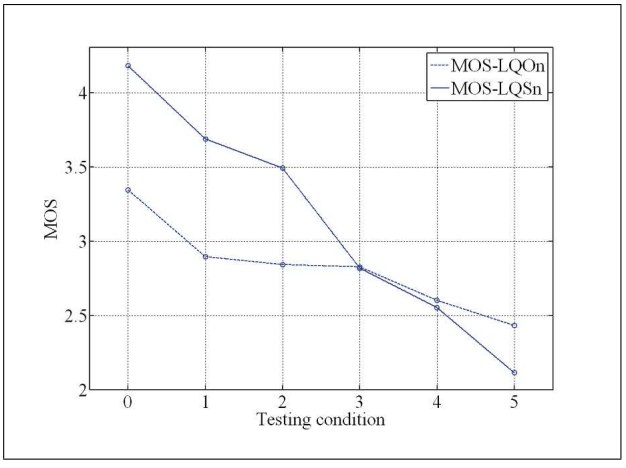

Figure 8. Comparison of the obtained *MOS-LQSn* and *MOS-LQOn* scores for Active-Speech-Ratio of 82%.

prediction behaviours. It means that if the problems related to under- and over-predictions will be resolved this can have positive impact on *PESQ*'s Active-Speech-Ratio sensitivity. Naturally, that is the point for future investigation because it requires a more precise elaboration.

On the basis of similar shapes of curves, we can pronounce that the subjective tests confirm our objective experimental results, presented in section 3.1. On the other hand, there is weak a correlation caused by *PESQ*'s under- and over-prediction behaviour and its possible dependency with Active-Speech-Ratio behaviour, as mentioned before.

4. Conclusions and future work

This paper has investigated an impact of different Active-Speech-Ratios of an input reference signals in *PESQ* based speech quality prediction in simulated *VoIP* environment. The main goals of this study are to gain a better understanding of behaviour of the *PESQ*'s predictions under different Active-Speech-Ratios and also to assess their accuracy by comparing the predictions with subjective assessments. The results presented in the paper have confirmed our preliminary assumption that the investigated characteristic of the reference signals (Active-Speech-Ratio) may have an impact on the final *PESQ*'s predictions and subjective scores.

The objective results have approved our 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%.

The subjective results have confirmed our objective results by curves comparison approach but on the other hand a weak correlation between subjective and objective scores has been achieved. From subjective results it is clear that *PESQ* under- and over-predicts speech quality in the case of low and high losses situations, respectively. The possible reason for the weak correlation is a potential dependency between *PESQ*'s under- and over-prediction and Active-Speech-Ratio behaviours.

A future work will focus towards the following issues. At first, we plan to exhaustively investigate an increase in reference signal sensitivity to a packet loss change by higher Active-Speech-Ratios for different languages and packet loss patterns in the case of dependent and independent losses. Secondly, 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*. Thirdly, *PESQ*'s over- and under-prediction and Active-Speech-Ratio behaviors dependency is also the important point for future investigation in this area.

Acknowledgement

This work has been partially supported by the Slovak VEGA grant agency, Project No. 1/0313/08, "The investigation of the methods of detection of the critical conditions in telecommunication networks from the speech quality point of view", the Slovak Research and Development Agency under the contract No.APVV-0369-07 and by the Czech ministry of Education: MSM 6840770014 "Research in the Area of the Prospective Information and Navigation Technologies".

We would like to thank Joachim Pomy (ETSI/STQ, *ITU-T*/SG12) for valuable comments and help in preparation of this paper, all subjective tests participants for rating our samples and last not at least the reviewers for their invaluable comments and suggestions.

References

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

[2] J. G. Beerends, J. A. Stemerdink: A perceptual speech quality measure based on a psychoacoustic sound representation. J. Audio Eng. Soc. **42** (1994) 115–123.

[3] S. Voran: Objective estimation of perceived speech quality. Part I: Development of the measuring normalizing block technique. IEEE Trans. on Speech and Audio Processing **7** (1999) 371–382.

[4] S. Voran: Objective estimation of perceived speech quality. Part II: Evaluation of the measuring normalizing block technique. IEEE Trans. on Speech and Audio Processing **7** (1999) 383–390.

[5] A. W. Rix, M. P. Hollier: The perceptual analysis measurement system for robust end-to-end speech quality assessment. Proceedings of IEEE ICASSP 2000, June 2000.

[6] A. W. Rix, M. P. Hollier, A. P. Hekstra, J. G. Beerends: PESQ, the new ITU standard for objective measurement of perceived speech quality. Part 1: Time alignment. J. Audio Eng. Soc. **50** (2002) 755–764.

[7] J. G. Beerends, A. P. Hekstra, A. W. Rix, M. P. Hollier: PESQ, the new ITU standard for objective measurement of

perceived speech quality. Part II: Perceptual model. J. Audio Eng. Soc. **50** (2002) 765–778.

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

[9] ITU-T Rec. P.862: Perceptual evaluation of speech quality. International Telecommunication Union, Geneva, 2001.

[10] 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 Telecommunication Union, Geneva, 2005.

[11] ITU-T Rec. P.56: Objective measurement of active speech level. International Telecommunication Union, Geneva, 1993.

[12] P. Počta, M. Mrvová, P. Kortiš, P. Palúch, M. Vaculík: A systematic study of PESQ's behavior in simulated VoIP environment (from reference signals characteristics perspective). Proceedings of conference MESAQIN 2008, Prague, Czech Republic, 2008.

[13] A. W. Rix: Comparison between subjective listening quality and P.862 PESQ score. Proceedings of conference MESAQIN 2003, Prague, Czech Republic, 2003.

[14] C. Hoene, E. Dulamsuren-Lalla: Predicting performance of PESQ in case of single frame losses. Proceedings of conference MESAQIN 2004, Prague, Czech Republic, 2004.

[15] L. Sun, E. C. Ifeachor: Subjective and objective speech quality evaluation under bursty losses. Proceedings of conference MESAQIN 2002, Prague, Czech Republic, 2002.

[16] S. Pennock: Accuracy of the perceptual evaluation of speech quality (PESQ) algorithm. Proceedings of conference MESAQIN 2002, Prague, Czech Republic, 2002.

[17] M. Varela, I. Marsh, B. Gronvall: A systematic study of PESQ's behaviour. Proceedings of conference MESAQIN 2006, Prague, Czech Republic, 2006.

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

[19] ITU-T Rec. G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-exited linear prediction (CS-ACELP). International Telecommunication Union, Geneva, 1996.

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

[21] ITU-T Rec. P.830: Subjective performance assessment of digital telephone-band and wideband digital codecs. International Telecommunication Union, Geneva, 1996.

[22] A. Takahashi: Objective quality evaluation based on ITU-T recommendation P.862 by using long reference speech (NTT), COM12-D008. International Telecommunication Union, Geneva, Jan. 2005.

[23] ITU-T Rec. G.191: Software tools for speech and audio coding standardization. International Telecommunication Union, Geneva, 2005.

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