Impact of Different Active-Speech-Ratios on PESQ's Predictions in Case of Independent and Dependent Losses (in Presence of Receiver-Side Comfort-Noise)

Peter POČTA[1], *Jan HOLUB*[2], *Helena VLČKOVÁ*[1], *Zuzana POLKOVÁ*[1]

[1] Dept. of Telecommunications and Multimedia, FEE, University of Žilina, Univerzitná 1, SK-01026, Žilina, Slovakia
[2] Dept. of Measurement K13138, FEE, CTU Prague, Technická 2, CZ-16627, Prague 6, Czech Republic

pocta@fel.uniza.sk, holubjan@fel.cvut.cz

Abstract. *This paper deals with the investigation of PESQ's behavior under independent and dependent loss conditions from an Active-Speech-Ratio perspective in presence of receiver-side comfort-noise. 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-Ratios and loss conditions, as well as their accuracy, by comparing the predictions with subjective assessments. Our results show 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. Interestingly, we have found two additional effects in this investigated case. The use of higher Active-Speech-Ratios may lead to negative shifting effect in MOS domain and also PESQ's predictions accuracy declining. Predictions accuracy could be improved by higher packet losses.*

Keywords

Perceptual Evaluation of Speech Quality (PESQ), speech quality, intrusive measurement, Voice over Internet Protocol (VoIP), reference signal characteristic, Active-Speech-Ratio, comfort-noise.

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 mainly 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 and PESQ were standardized by the ITU-T Recommendations such as P.861 [8] and P.862 [9] respectively. In contrast to intrusive methods, the idea of the single-ended (non-intrusive) signal-based methods is to generate an artificial reference (i.e., an "ideal" undistorted signal) from degraded speech signal and to use this reference in a signal-comparison approach. Once a reference is available, a signal comparison similar to the one of PESQ can be performed. The result of this comparison can further be modified by a parametric degradation analysis and integrated into an assessment of overall quality. The most widely used algorithms include Auditory Non-Intrusive QUality Estimation (ANIQUE) [10] and standardized P.563 [11, 12]. 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) then 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. PESQ also facilitates with very fine time-alignment. It can be used in wider range of network conditions, and gives higher correlation with subjective tests and 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 [13]. 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's measurements. The speech activity in the reference signals, which can be measured according to ITU-T Recommendation P.56 [14], 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 [15] from PESQ's prediction perspective. Some very important issues raise from [15] 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 prediction provided by PESQ from dependent and independent losses perspective, also in presence of receiver-side comfort-noise.

Some works have been carried out on study of PESQ's behavior under single frame, uniform and dependent losses. In [16], 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 [17], 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 [18], a study of PESQ's behavior from networking perspective (dependent and uniform losses) 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. In addition, other works have been focused on impact of noisy circumstances on speech quality and also on noise reduction schemes performance. In [19], different objective quality measures for the performance prediction of noise reduction schemes have been compared to subjective data from listening tests. The results have shown that objective measures are able to predict subjective ratings in noise reduction schemes. In [20], the subjective and objective quality assessments for noise-reduced speech from the viewpoints of opinion rating and word intelligibility have been described. The results have confirmed that the PESQMOS correlates relatively well with the subjective MOS in the mentioned case.

Here we focus on an impact of different Active-Speech-Ratios on speech quality prediction provided by PESQ in case of independent and dependent losses in presence of receiver-side comfort-noise. 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-Ratios and loss 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 loss simulator (Fig. 1). In case of loss simulator, two currently most widely used models were deployed for the purpose of packet loss modeling, namely Bernoulli and Gilbert loss model. More details about loss models can be found below. For this experiment the ITU-T G.729AB encoding scheme [21] 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 implemented in the VoIP clients used. The jitter buffer does not play any role in case of this experiment because of small constant jitter inserted by the loss simulator during the measurement. The Comfort Noise Generator (CNG) usage was enabled in case of this experiment.

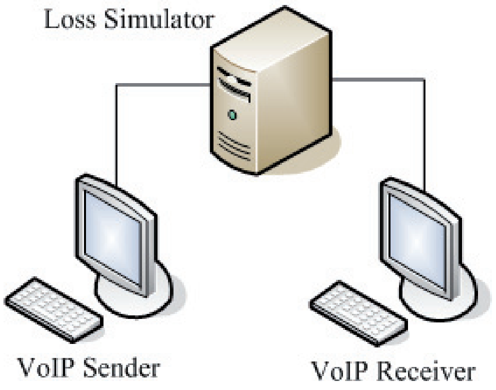

Fig. 1. Experimental scenario.

The reference signals described below were utilized for transmission through the given VoIP connection. Finally, speech quality was assessed by PESQ and then converted to MOS-Listening Quality Objective narrowband (MOS-LQOn) values by this equation:

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

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 [22]. In case of PESQ score calculation, we used some batch data processing techniques proposed in [23].

2.2 Packet Loss Models

Packet loss is a major source of speech impairment in VoIP. Such a loss may be caused by discarding packets in the IP networks (network loss) or by dropping packets at the gateway/terminal due to late arrival (late loss).

Several models [24, 25] have been proposed for modeling network losses, the currently most widely used of them will be briefly discussed in the following sections.

2.2.1 Bernoulli Model

In the Bernoulli loss model, each packet loss is independent (memoryless), regardless of whether the previous packet was lost or not. In this case, there is only one parameter, the average packet loss rate, which is the number of lost packets divided by the total number of transmitted packets in a trace.

2.2.2 Gilbert Model

Most research in VoIP networks uses a Gilbert model to represent packet loss characteristics [24-26]. In 2-state Gilbert model as shown in Fig. 2, State 0 is for a packet received (no loss) and State 1 is for a packet dropped (loss). p is the probability that a packet will be dropped given that the previous packet was received. 1-q is the probability that a packet will be dropped given that the previous packet was dropped. 1-q is also referred to as the conditional loss probability (*clp*). The probability of being in State 1 is referred to as unconditional loss probability (*ulp*). The *ulp* provides a measure of the average packet loss rate and is given by [31]:

$$ulp = \frac{p}{p+q}. \tag{2}$$

The *clp* and *ulp* are used in the paper to characterize the loss behavior of the network.

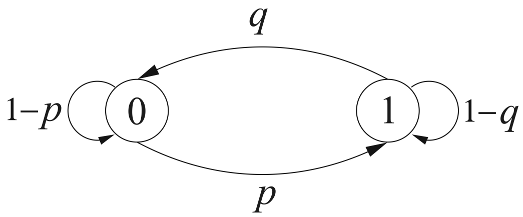

Fig. 2. Gilbert model.

Sixteen independent loss and dependent loss conditions were chosen to cover cases of interest. They consist of combinations of packet loss rates (from 0% to 15%) in case of independent losses and unconditional loss probabilities (*ulp*, 0%, 1.5%, 3%, 5%, 10% and 15%), conditional loss probabilities (*clp*, 15%, 30% and 50%) in case of dependent losses and 20 initial seeds to simulate different loss locations in both cases.

2.3 Signal Design

The reference signals selection should follow the criteria given by ITU-T Recommendations P.830 [27] and P.800 [1]. The reference signals should include talkspursts 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. Background noise was not present in case of reference signals. Degraded signals have contained special type of background noise, called comfort-noise. The comfort-noise mentioned could be defined as the stationary noise with no significant peaks in frequency spectrum and was present from CNG at receiver-side, only during silence periods. More details about CNG principle can be found in [28]. The resulting SNR was about 15 dB for degraded signals.

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 Tab. 1), 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 [15]. 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 also investigated in [29]. The experimental results have shown that for this purpose it is possible to use a longer reference signals and the 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.

Reference signal	Active-Speech-Ratio of 42%	Active- Speech-Ratio of 62%	Active- Speech-Ratio of 82%
Male1	40.183 (6)	60.333 (8)	82.516 (10)
Male2	43.249 (6)	63.861 (9)	80.612 (11)
Female1	43.677 (6)	62.065 (8)	84.180 (10)
Female2	41.780 (4)	62.054 (6)	82.121 (8)
Average value	**42.222 (5.5)**	**62.078 (7.75)**	**82.375 (9.75)**

Tab. 1. Active-Speech-Ratios and numbers of talkspurts of the reference signals.

The Active-Speech-Ratios and numbers of talkspurts (active speech periods) for each of the reference signals used are presented in Tab. 1. 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 [30], known as sv56.

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, 18 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.

Because of huge amount of objective measurement data, we had to make the decision which condition is the closest to real network conditions in order to limit the number of samples used in subjective tests. Finally, we decided on the basis of the available measurement results [24, 25, 31] that one of the dependent loss conditions is the best one for this purpose, namely *clp* = 30%. The subjective tests were done only for this condition.

All together, 108 speech samples were selected for subjective testing. Always 6 samples represented one network testing condition (the combination of *ulp*'s and *clp*) 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 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 80 times (20 different loss locations/patterns and 4 reference signals) under the same packet loss (independent losses), the same pair of *ulp* and *clp* (dependent losses) and the same Active-Speech-Ratio. The average MOS-LQOn score, 95% Confidence Interval (CI) and Mean Absolute Deviation (MAD) were calculated. The next subsections provide the detailed description of experimental results for the both examined types of losses.

3.1.1 Experimental Results for Independent Losses

Using a Bernoulli model gives us the possibility to analyze PESQ's behavior only from two perspectives, namely packet loss and Active-Speech-Ratio. Fig. 3 and 4 depict differences between investigated Active-Speech-Ratios in speech quality evaluation, provided by PESQ. It can be seen from the 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 an 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 rate in the case of using 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 the 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 and accurate evaluation of investigated transmission line from this perspective. The sensitivity effect mentioned is depicted in Fig. 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%.

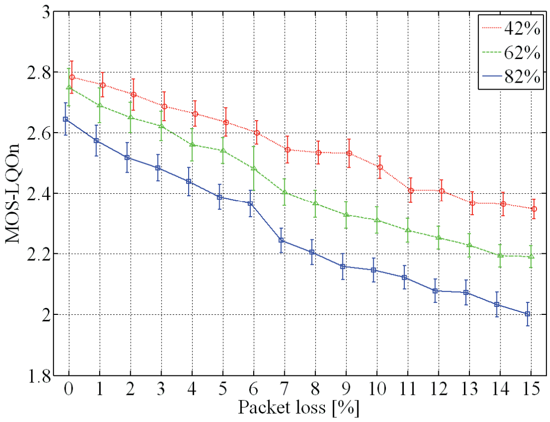

Fig. 3. MOS-LQOn as a function of packet loss for different Active-Speech-Ratios in case of independent losses. The vertical bars show 95 % CI (derived from 80 measurements) for each loss.

Interestingly, a consecutive accruing shifting in MOS-LQOn domain occurs between the investigated Active-Speech-Ratios in this experiment case (Fig. 3). The captured shiftings (42%-62%, 42%-82%) are between 0.033 and 0.374 MOS-LQOn's. Probably, it is caused by the interaction of the sensitivity effect mentioned and PESQ's time-alignment problem related to usage of degraded signals with comfort-noise, as mentioned above. If some kind of background noise (like comfort-noise, etc.) is present in a degraded sample, PESQ has a problem to clearly distinguish the bounds of the silence periods and active speech intervals. The problem mentioned occurs more frequently if the signals contain more speech periods. This leads to additional extension of noted shifting. It could be seen in Fig. 3, for instance 82% curve. In case of our experiment, the numbers of talkspurts for the investigated signals are referred to in Tab. 1. In addition, it can be seen from Fig. 5, that the above-mentioned interaction could also cause the PESQ's score variation increasing in case of higher Active-Speech-Ratios using. Fig. 5 shows MAD of MOS-LQOn's, which has been obtained for this experiment. We can see from this figure that the deviation of predictions is much smaller in case of lower Active-Speech-Ratios, especially when network conditions degrade (packet loss increase). As was early mentioned, the first fact is related to the above-described interaction and the second one could be explained by higher probability of losses obtained at active speech intervals (effective loss probability) at higher packet losses. More effective losses may lead to small variation in PESQ score. If the higher Active-Speech-Ratio is used, this effect could be more markedly achieved. This effect could be characterized as the sensitivity effect gain. On other hand, it could also alleviate the time-alignment impact and finally the interaction of both described facts cause that the similar MAD values have been obtained in case of higher packet losses for all investigated Active-Speech-Ratios.

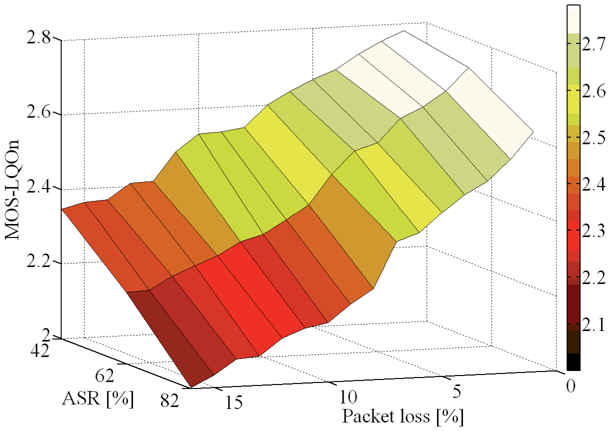

Fig. 4. MOS-LQOn versus packet loss and Active-Speech-Ratio for independent losses.

A two-way analysis of variance (ANOVA) was conducted on MOS-LQOn's using packet loss and Active-Speech-Ratio as fixed factors (Appendix A.1, Tab. A.1). We found clearly the highest F-ratio for the Active-Speech-Ratio ($F = 2123.11$, $p < 0.01$). Moreover, the packet loss factor showed a little bit weaker effect on quality than Active-Speech-Ratio, with $F = 385.68$, $p < 0.01$.

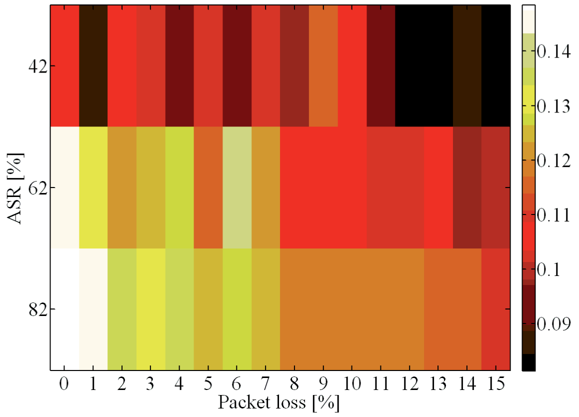

Fig. 5. MAD of MOS-LQOn's at each point of loss space and Active-Speech-Ratio in case of independent losses.

3.1.2 Experimental Results for Dependent Losses

Using a Gilbert model extends our possibilities to investigate PESQ's behavior to three perspectives, namely *ulp*, *clp* and naturally Active-Speech-Ratio. The experimental results for all investigated *clp*'s are depicted in Fig. 6, 7 and 8. We can observe how speech quality drops, as expected, with both *clp* and *ulp*. Also, it is clear that the different Active-Speech-Ratios could seriously influence the quality in case of dependent losses. Obviously, we obtained same effects as in the first case (independent losses). It means that using higher Active-Speech-Ratio leads to increase of reference signal sensitivity to packet loss change and negative shifting in MOS-LQOn domain, also in case of dependent losses. The captured shiftings are in the same range as stated above.

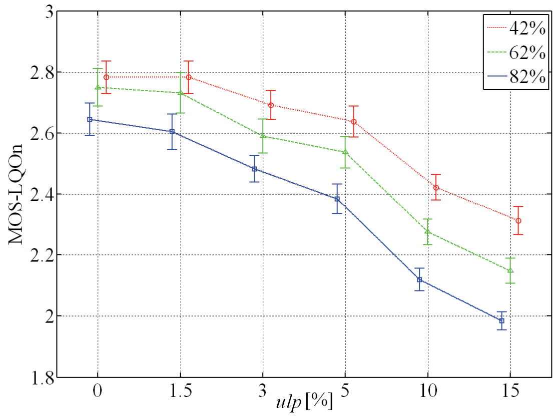

Fig. 6. MOS-LQOn as a function of unconditional loss probability for different Active-Speech-Ratios in case of dependent losses (*clp* = 30%). Other detailed descriptions of Fig. 3 apply appropriately.

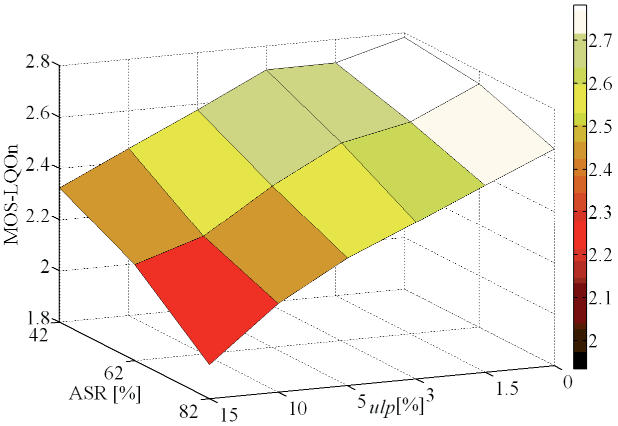

Fig. 7. MOS-LQOn versus unconditional loss probability and Active-Speech-Ratio for dependent losses (*clp* = 50%).

In Fig. 9, we can see the MAD of MOS-LQOn's for 30% *clp*. Unsurprisingly, PESQ's predictions deviation behavior is also similar as obtained in the previous case. Interestingly, the highest deviation has been obtained at 0% packet loss. At this time, we have no theory that could explain this phenomenon. Naturally, that is a point for a future investigation because exhaustive study is needed to validate, and interpret this phenomenon. On basis of those results, we can pronounce that higher Active-Speech-Ratio usage may lead to PESQ's predictions deviation increasing, especially in case of lower packet losses.

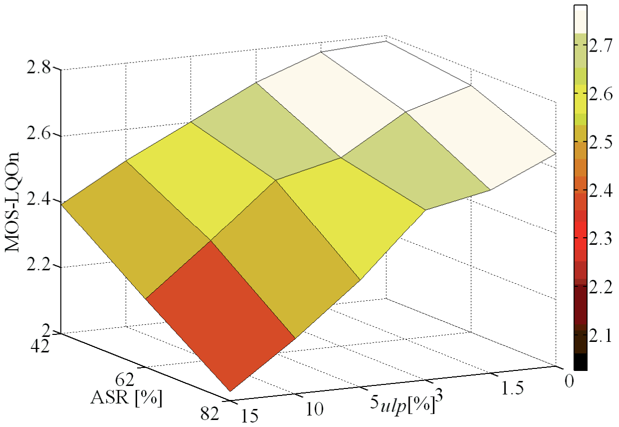

Fig. 8. MOS-LQOn versus unconditional loss probability and Active-Speech-Ratio for dependent losses (*clp* = 15%).

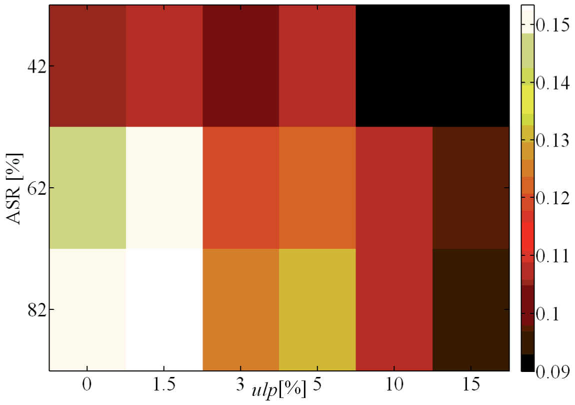

Fig. 9. MAD of MOS-LQOn's at each point of loss space and Active-Speech-Ratio in case of dependent losses (*clp* = 30%).

Three two-way ANOVA's were similarly carried out on MOS-LQOn's for all investigated *clp*'s, using *ulp* and Active-Speech-Ratio as fixed factors (Appendix A.2, Tab. A.2-4). We obtained similar results as in the case of independent losses.

3.2 Experimental Results for Subjective Assessment

As mentioned above, the subjective tests were realized for dependent losses (*clp* = 30%). The results obtained by means of subjective testing (MOS-LQSn) are compared with MOS-LQOn results in Fig. 10 and 11. Obviously, the sensitivity to Active-Speech-Ratio modification of PESQ is a bit weaker than that of human subjects (see Fig. 10). As attempts to use the 3-rd order regression (as recommended in ITU-T Recommendation P.862) lead to non-monotonic results, the 2-nd order regression was used instead. Fig. 11 depicts the results after the 2-nd order regression.

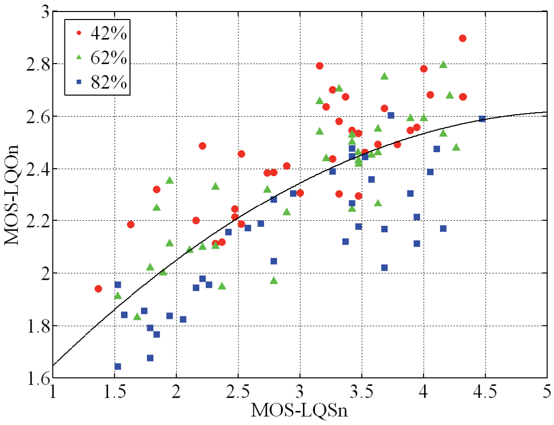

Fig. 10. Subjective results (MOS-LQSn) versus MOS-LQOn output (not regressed).

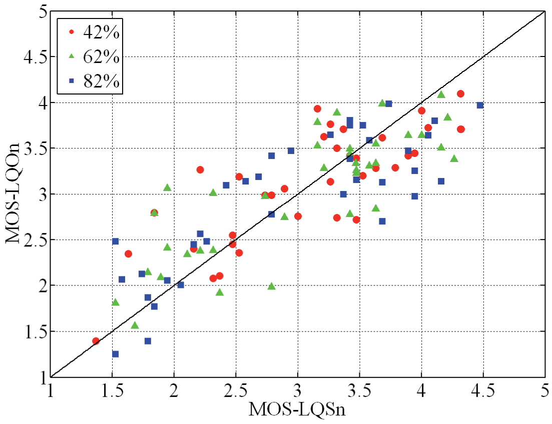

Fig.11. Subjective results (MOS-LQSn) versus MOS-LQOn output (2nd order regression).

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 Tab. 2. It is seen that higher Active-Speech-Ratio usage may lead to higher correlation with subjective test but on the other hand to PESQ's predictions accuracy declining (expressed by RMSE (after regression)). The prediction accuracy can be increased by higher packet losses, as can be seen in Fig. 11. The results obtained in case of smaller MOS values are mainly closer to diagonal line than higher MOS are. In general, high packet losses generate lower MOS scores.

Active-Speech-Ratio	42%	62%	82%
ρ before regression	0.7953	0.8159	0.8211
ρ after regression	0.8029	0.8174	0.8313
δ before regression	0.8907	0.9161	1.0573
δ after regression	0.4518	0.4699	0.4917

Tab. 2. Pearson correlation coefficient and Root Mean Square Error between MOS-LQSn and MOS-LQOn before and after 2nd order regression.

The two-way ANOVA was also conducted on MOS-LQSn's using *ulp* and Active-Speech-Ratio as fixed factors (Appendix B, Tab. B.1). We got clearly the highest F-ratio for the Active-Speech-Ratio ($F = 306.8$, $p < 0.01$). Moreover, the *ulp* factor showed a little bit weaker effect on quality than the Active-Speech-Ratio, with $F = 256.04$, $p < 0.01$. In comparison with the same test for objective data, a bit higher F-ratios were obtained for the *ulp* as well as for the Active-Speech-Ratio (Appendix A.2, Tab. A.3). On the other hand, the same F-ratio was observed in case of an interaction of both parameters. Based on similar F-ratios, we can say that the results of this ANOVA test confirmed the sensitivity and negative shifting effects defined in Section 3.1 but a bit weaker than obtained in objective assessment.

On the basis of this comparison, we can pronounce that the subjective tests confirm the objective experimental results (presented in Section 3.1) but the behavior of PESQ should be modified to better model the impact of the investigated reference signal characteristic on speech quality in the noted case.

The experimental results show that the change of Active-Speech-Ratio has a significant impact on overall speech quality. This fact is our motivation for finding of the feasible average Active-Speech-Ratios for some conversational scenarios. Naturally, an issue of Active-Speech-Ratio setup with regards to different 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 in real VoIP scenarios to make comparison among network providers more feasible.

4. Conclusions and Future Works

This paper has investigated an impact of different Active-Speech-Ratios of an input reference signals in PESQ based speech quality prediction in case of dependent and independent losses and in presence of receiver-side comfort-noise. The main goal of this study is to gain a better understanding of behavior of the PESQ's predictions under different Active-Speech-Ratios in noted case as well as to assess their accuracy by comparing the predictions with subjective assessments.

The results presented in the paper 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. Interestingly, we have found two additional effects in this investigated case. The use of higher Active-Speech-Ratios may lead to negative shifting effect in MOS domain and also PESQ's predictions accuracy declining. Predictions accuracy could be improved by higher packet losses.

A future work will focus towards the following issues. At first, we will attempt to find out an appropriate average Active-Speech-Ratios for some 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. Secondly, we plan to exhaustively study the highest MAD at 0% packet loss and find out the reason for that.

Acknowledgements

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 the Czech Ministry of Education: MSM 6840770014 "Research in the Area of the Prospective Information and Navigation Technologies".

In addition, 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 but not least the anonymous reviewers for their helpful comments on an earlier version of this paper.

References

[1] *ITU-T Rec. P.800: Methods for Subjective Determination of Transmission Quality*. International Telecommunication Union, Geneva (Switzerland), 1996.

[2] BEERENDS, J. G., STEMERDINK, J. A. A perceptual speech quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc*., 1994, vol. 42, p. 115-123, ISSN 1549-4950.

[3] VORAN, S. Objective estimation of perceived speech quality - Part I: Development of the measuring normalizing block technique. *IEEE Trans. on Speech and Audio Processing*, 1999, vol. 7, p. 371-382, ISSN 1063-6676.

[4] VORAN, S. Objective estimation of perceived speech quality - Part II: Evaluation of the measuring normalizing block technique. *IEEE Trans. on Speech and Audio Processing*, 1999, vol. 7, p. 383-390, ISSN 1063-6676.

[5] RIX, A. W., HOLLIER, M. P. The perceptual analysis measurement system for robust end-to-end speech quality assessment. In *Proceedings of IEEE ICASSP 2000*. Istanbul (Turkey), 2000, vol. 3, p. 1515-1518.

[6] RIX, A. W., HOLLIER, M.P., HEKSTRA, A.P., BEERENDS, J.G. Perceptual evaluation of speech quality (PESQ) - The new ITU standard for objective measurement of perceived speech quality, Part I – Time-delay compensation. *J. Audio Eng. Soc*., 2002, vol. 50, p. 755-764, ISSN 1549-4950.

[7] BEERENDS, J.G., HEKSTRA, A.P., RIX, A. W., HOLLIER, M.P. Perceptual evaluation of speech quality (PESQ) - The new ITU standard for objective measurement of perceived speech quality, Part II – Psychoacoustic model. *J. Audio Eng. Soc*., 2002, vol. 50, p. 765-778, ISSN 1549-4950.

[8] *ITU-T Rec. P.861: Objective Quality Measurement of Telephone-Band (300-3400 Hz) Speech Codecs*. International Telecommunication Union, Geneva (Switzerland), 1998.

[9] *ITU-T Rec. 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*. International Telecommunication Union, Geneva (Switzerland), 2001.

[10] KIM, D.-S. ANIQUE: An auditory model for single-ended speech quality estimation. *IEEE Transaction on Speech and Audio Processing*, 2005, vol. 13, no.5, p. 821- 831, ISSN 1063-6676.

[11] MALFAIT, L., BERGER, J., KASTNER, M. P.563 – The ITU-T standard for single-ended speech quality assessment. *IEEE Transaction on Audio, Speech and Language Processing*, 2006, vol. 14, no. 6, p. 1924-1934, ISSN 1558-7916.

[12] *ITU-T Rec. P.563: Single-Ended Method for Objective Speech quality Assessment in narrow-Band Telephony Applications.* International Telecommunication Union, Geneva (Switzerland), 2004.

[13] *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 (Switzerland), 2005.

[14] *ITU-T Rec. P.56: Objective Measurement of Active Speech Level.* International Telecommunication Union, Geneva (Switzerland), 1993.

[15] POČTA, P., MRVOVÁ, M., KORTIŠ, P., PALÚCH, P., VACULÍK, M. A systematic study of PESQ's behavior in simulated VoIP environment (from reference signals characteristics perspective). In *Proceedings of MESAQIN 2008 Conference*. Prague (Czech Republic), 2008, p. 13-21, 2008, ISBN 978-80-01-04193-2.

[16] HOENE, Ch., DULAMSUREN-LALLA, E. Predicting performance of PESQ in case of single frame losses. In *Proceedings of MESAQIN 2004 Conference*. Prague (Czech Republic), 2004, ISBN 80-01-03017-2.

[17] SUN, L., IFEACHOR, E. C. Subjective and objective speech quality evaluation under bursty losses. In *Proceedings of MESAQIN 2002 Conference.* Prague (Czech Republic), 2002, ISBN 80-01-02515-2.

[18] VARELA, M., MARSH, I., GRONVALL, B. A systematic study of PESQ's behavior. In *Proceedings of MESAQIN 2006 Conf.* Prague (Czech Republic), 2006, ISBN 80-01-03503-4.

[19] ROHDENBURG, T., HOHMANN, V., KOLLMEIER, B. Objective perceptual quality measures for the evaluation of noise reduction schemes. In *Proceedings of 9th International Workshop on Acoustic Echo and Noise Control (IWAENC 2005)*. Eindhoven (Netherlands), 2005, p. 169-172.

[20] KITAWAKI, N., YAMADA, T. Subjective and objective quality assessment for noise reduced speech. In *Proceedings of ETSI Workshop on Speech and Noise in Wideband Communication*. Sophia Antipolis (France), May 2007.

[21] *ITU-T Rec. G.729: Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Exited Linear Prediction (CS-ACELP).* Int. Telecommunication Union, Geneva (Switzerland), 2007.

[22] *ITU-T Rec. P.862.1: Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO*. International Telecommunication Union, Geneva (Switzerland), 2003.

[23] CHOCHLÍK, M., GRONDŽÁK, K., BABOŠ, Š. Windows operating system core programming. University of Žilina, Žilina (Slovakia), 2009, ISBN 978-80-8070-970-9 (in Slovak).

[24] YAJNIK, M., MOON, S., KUROSE, J., TOWSLEY, D. Measurement and modelling of the temporal dependence in packet loss. In *Proceedings of IEEE INFOCOM 99 Conference.* New York (USA), 1999, vol. 1, p. 345–352.

[25] JIANG W., SCHULZRINNE, H. *QoS Measurement of Internet Real-Time Multimedia Services.* Technical Report, CUCS-015-99, Columbia University (USA), Dec. 1999.

[26] SANNECK H., LE, N. T. L. Speech property-based FEC for Internet telephony applications. In *Proceedings of the SPIE/ACM SIGMM Multimedia Computing and Networking Conference.* San Jose (USA), 2000, p. 38–51.

[27] *ITU-T Rec. P.830: Subjective Performance Assessment of Digital Telephone-Band and Wideband Digital Codecs.* International Telecommunication Union, Geneva (Switzerland), 1996.

[28] RAAKE, A. *Speech Quality of VoIP: Assessment and Prediction*. Chichester (UK): John Wiley&Sons, 2006. Chapter 3, pp. 51-110, ISBN 0-470-03060-7.

[29] TAKAHASHI, A. *Objective Quality Evaluation Based on ITU-T Recommendation P.862 by Using Long Reference Speech (NTT), COM12-D008.* International Telecommunication Union, Geneva (Switzerland), January 2005.

[30] *ITU-T Rec. G.191: Software Tools for Speech and Audio Coding Standardization.* International Telecommunication Union, Geneva (Switzerland), 2005.

[31] JIANG W., SCHULZRINNE, H. Modelling of packet loss and delay and their effect on real-time multimedia service quality. In *Proceedings of 10th International Workshop Network and Operations System Support for Digital Audio and Video (NOSSDAV 2000)*. Chapel Hill (USA), 2000.

About Authors...

Peter POČTA was born in 1981, in Nové Zámky, Slovakia. He received his M.S. and Ph.D. degrees from University of Žilina, Faculty of Electrical Engineering, Slovakia in 2004 and 2007, respectively. During his Ph.D. study, he realized a few fellowships. Firstly, he spent three months as an Erasmus student in the Department of Electrical Engineering and Information Technology, Chair of Telecommunications at Dresdner University of Technology, Germany. He collaborated on testing principles over ADSL access lines. Secondly, he was with Alcatel-Lucent, R&D center, Network integration department, Stuttgart, Germany. He was entrusted with investigation of some impacts on speech quality in WiMAX system. He is currently an Assistant Professor at the Department of Telecommunications and Multimedia, University of Žilina, and ETSI STQ working group member. Areas of his interest include speech and video quality assessment, access networks, convergent networks, VoIP, VoWLAN and cross-layer optimization. He has published over 20 papers in famous journals and conferences in his areas of interest, for instance: Acta Acustica united with Acustica, Advances in Multimedia (Hindawi), AEÜ - International Journal of Electronics and Communications (Elsevier) and MESAQIN conference.

Jan HOLUB was born in Prague, Czech Republic, in 1973. He received his Ing. (1996), Ph.D. (1999) and Assoc. Prof. (2004) in Measuring Technology from the 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 chairman of the organizing committee for the MESAQIN conference since 2001, member of program committee of WTS 2006-9 and co-chairman of organizing committee of WTS 2009.

Helena VLČKOVÁ was born in Čadca, Slovakia, in 1987. She received her B.S. degree in Telecommunications from University of Žilina, Faculty of Electrical Engineering, Slovakia, in 2009. Currently, she is master student at the same faculty. Her areas of interest include speech and video quality assessment and networking.

Zuzana POLKOVÁ was born in Čadca, Slovakia, in 1987. She received her B.S. degree in Telecommunications from University of Žilina, Faculty of Electrical Engineering, Slovakia, in 2009. Currently, she is master student at the same faculty. Her areas of interest include speech quality assessment and networking.

Appendix

A. ANOVA for Objective Assessment

In the next subsections, the detailed results of the analysis of variance (ANOVA) conducted on MOS-LQOn for independent and dependent losses can be found.

A.1 Independent Losses

Tab. A.1 provides the results of ANOVA carried out on the independent losses test results (Dependent variable: MOS-LQOn) described in more detail in Section 3.1.1.

Effect	SS	df	MS	F	p
Packet loss (1)	106.949	15	7.1299	385.68	0.0000
Active-Speech-Ratio (2)	78. 497	2	39.2487	2123.11	0.0000
(1)*(2)	0.851	30	0.0284	1.53	0.0031
Error	70.101	3792	0.0185		
Total	256.398	3839			

Tab. A.1 Summary of ANOVA conducted on MOS-LQOn's in case of independent losses.

A.2 Dependent Losses

In Tab. A.2-4, the results of ANOVA for the dependent losses test results and all investigated *clp*'s (Dependent variable: MOS-LQOn) are shown. More details about this can be found in Section 3.1.2.

Effect	SS	df	MS	F	p
ulp (1)	44.941	5	8.9881	405.35	0.0000
Active-Speech-Ratio (2)	28.169	2	14.0843	635.18	0.0000
(1)*(2)	0.372	10	0.0372	1.68	0.0081
Error	31.531	1422	0.0222		
Total	105.012	1439			

Tab. A.2. Summary of ANOVA conducted on the MOS-LQOn's in case of dependent losses (*clp* = 50%).

Effect	SS	df	MS	F	p
ulp (1)	56.742	5	11.3484	552.01	0.0000
Active-Speech-Ratio (2)	26.67	2	13.3351	648.65	0.0000
(1)*(2)	0.337	10	0.0337	1.64	0.0089
Error	29.234	1422	0.0206		
Total	112.984	1439			

Tab. A.3. Summary of ANOVA conducted on the MOS-LQOn's in case of dependent losses (*clp* = 30%).

Effect	SS	df	MS	F	p
ulp (1)	43.797	5	8.7593	417.62	0.0000
Active-Speech-Ratio (2)	27.151	2	13.5756	647.24	0.0000
(1)*(2)	0.317	10	0.0317	1.51	0.0092
Error	29.826	1422	0.021		
Total	101.09	1439			

Tab. A.4. Summary of ANOVA conducted on the MOS-LQOn's in case of dependent losses (*clp* = 15%).

B. ANOVA for Subjective Assessment

Tab. B.1 provides the results of ANOVA carried out on the dependent losses test results (Dependent variable: MOS-LQSn) described in more detail in Section 3.2.

Effect	SS	df	MS	F	p
ulp (1)	975.77	5	195.155	256.04	0.0000
Active-Speech-Ratio (2)	10.36	2	5.182	306.8	0.0011
(1)*(2)	47.38	10	4.738	1.52	0.0000
Error	1550.31	2034	0.762		
Total	2583.83	2051			

Tab. B.1. Summary of ANOVA conducted on MOS-LQSn's in case of dependent losses (*clp* = 30%).