Transmitted Speech Quality versus Perceptual Annoyance and Service Acceptability Thresholds

Jan Holub [1], *Peter Počta* [2]

[1] Czech Technical University in Prague, Czech Republic
[2] University of Žilina, Slovak Republic
jan.holub@fel.cvut.cz, pocta@fel.uniza.sk

Abstract

In this contribution, two types of subjective listening test results are reported. The test goal was to link the P.800 MOS quality scale (1...5) to thresholds of service quality annoyance and service quality acceptability. Other tested aspect was trend analysis of these thresholds as identical experiment was performed 2 years ago.

Index Terms: Speech Quality, Mean Opinion Score, service acceptability

1. Introduction

Most of subjective listening tests are currently performed using ITU-T P.800 [1] methodology using Mean Opinion Score (MOS) scale. Such tests are used to subjectively assess speech samples impaired during transmission in the telecommunication channel. Most common impairments include coding and decoding distortions and artifacts, packet loss and PLC artifacts, background noise, temporal clipping, amplitude clipping etc. The final score 1(bad) to 5(excellent) is well understood and accepted by professionals, however, the score itself does not express anything about network user acceptability of the assessed speech sample. The network planners and operators need to know which quality level is yet acceptable and which will lead to user complains or even operator change.

Partial answer to this question is given by E-model [2,3], however, the mapping function between test methods attempting to replace subjective testing (like P.862, P.563 etc.) and E-model output R-factor is more than 10 years old and also the satisfaction levels of E-model are reported sometimes as too demanding as E-model is considered primarily as planning model.

Therefore the goal of our experiment was to understand relation between MOS scale and acceptability and noticeability levels of average network user.

2. Methods

To find the relation between MOS scale and acceptability and noticeability thresholds, we run two different subjective tests (as described further) on the same speech sample database.

The speech sample database contained 128 speech samples in Czech language coded by various contemporary coders (G.711, GSM 06.06 Full Rate, AMR) and affected by different types and levels of background noise (no, -10dB Hoth). The sample distortions were designed to reflect realistic distortion types but also to cover MOS scale approximately uniformly. The listening tests results were performed in accordance with ITU-T P.800 in listening chamber with reverberation time less than 190ms and background noise well below 20 dB SPL (A). The listening level was set to 79 dB SPL(A), keeping all the samples level equalized to -26 dBoV. In total, 24 listeners participated in each test, both men and women were represented equally. The age structure of listener group was 19-28 years. In the first run, they scored samples in a common way using Opinion Score approach. This experiment part resulted in traditional MOS-LQSn scores. After a short break the test was repeated in re-randomized order but the subjects scored each sample by letters A, B or C with the following meanings (translated from Czech language):

A – "This sample is OK for me"

B – "This sample is somehow distorted but it is not worth complaining"

C – "This is an unacceptable quality for me; I would consider to complain or even to change the operator"

The subjects were native Czech listeners, and all the samples and instructions were presented in Czech language. The obtained votes have been statistically processed and thresholds between A, B and C sample groups have been mapped to MOS scale.

Both tests as described above have been performed twice, in 2008 and 2010 to compare the results and to find possible adaptation trend.

3. Results

The MOS scores obtained during the listening tests have been grouped to intervals of 0.2 MOS width. The granularity level 0.2 has been identified empirically with respect to number of scores available. For each sample group, the relative occurrence of each letter (A, B or C) has been calculated (see Fig. 1 and 2 for 2008 results and Fig.3 and 4 for 2010 results). Then, polynomial regression of the 4th order has been used to find crossovers between the categories A, B and C. The crossover points have been calculated, see the Table 1.

Table1 – Threshold Results

	MOS threshold between A and B	MOS threshold between B and C
2008	3,84	2,07
2010	3,45	2,16

10.21437/PQS.2010-16

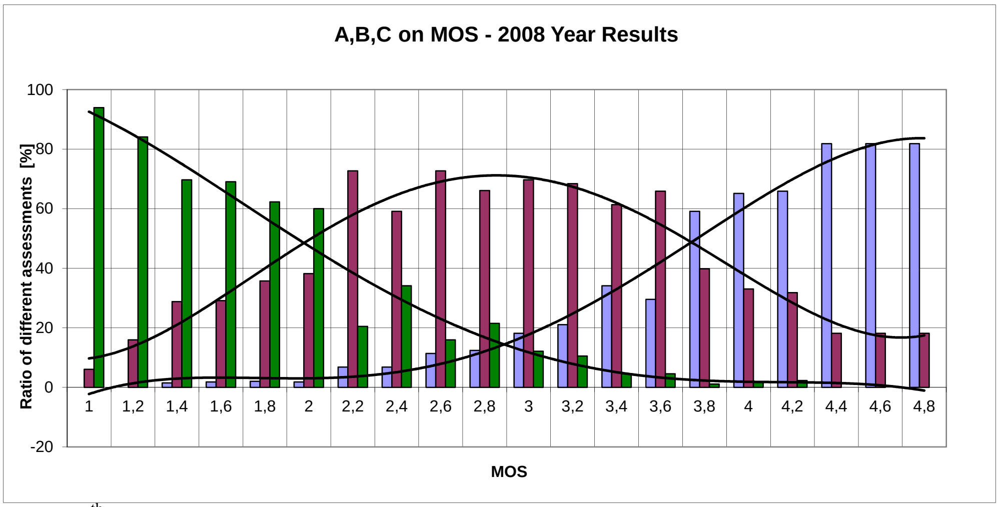

Fig. 1 – The 4th polynomial regression of data measured in 2008. Opinion curves A (right), B (middle) and C (left)

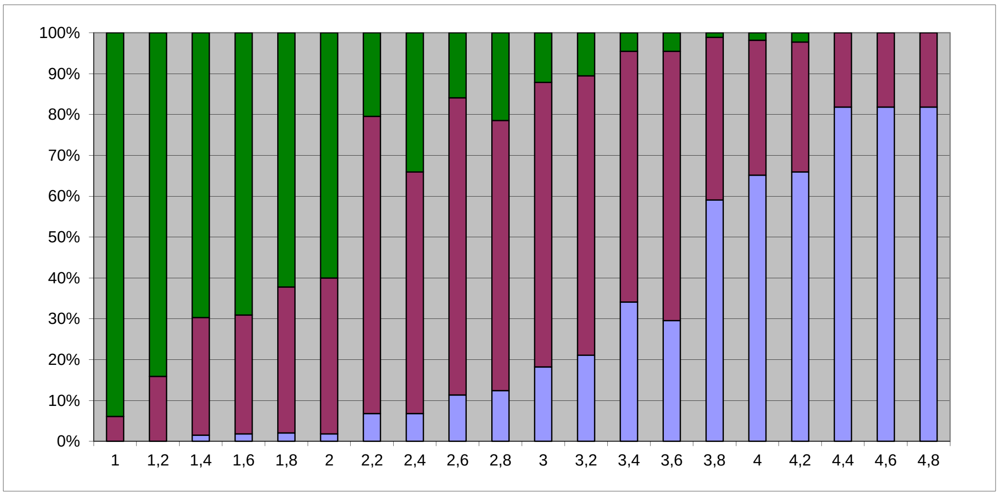

Fig. 2 – Relative occurrence of scores measured in 2008

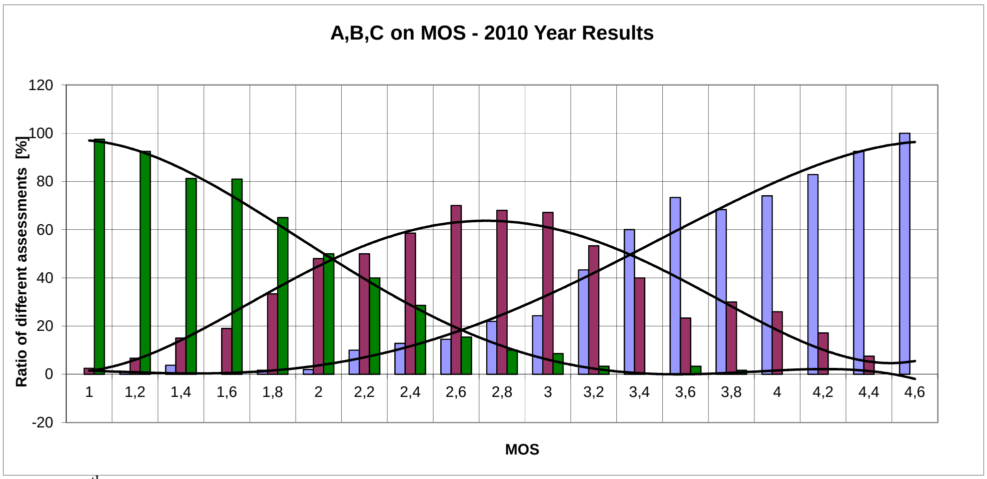

Fig. 3 – The 4th polynomial regression of data measured in 2010. Opinion curves A (right), B (middle) and C (left)

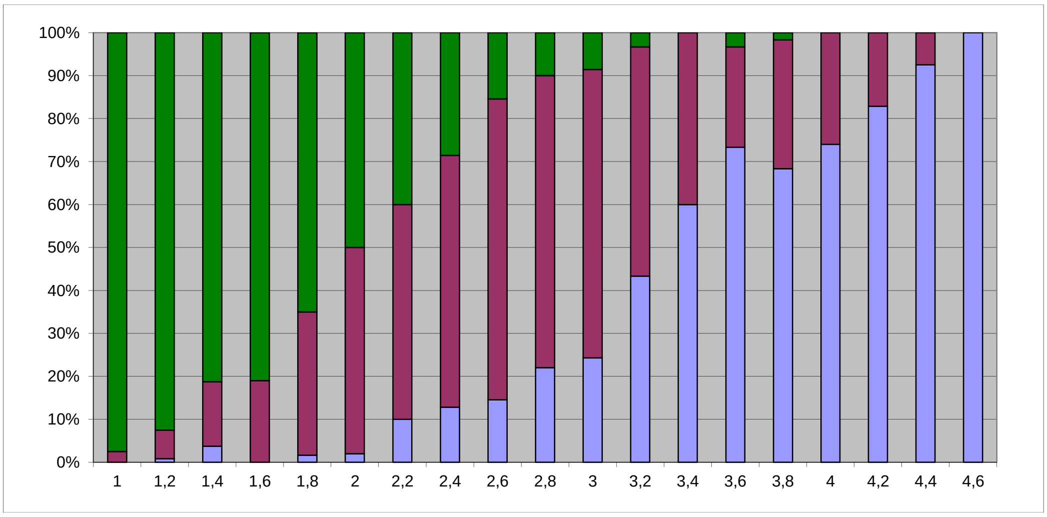

Fig. 4 – Relative occurrence of scores measured in 2010

4. Conclusions

The analysis of 2 years old test results (the previous tests were performed on different test subjects so the age structure of both tests is consistent) in comparison with contemporary results clearly shows threshold shift between groups A and B that means users are getting more tolerant to slight distortions and assess more samples as “OK” instead of “somehow distorted but not worth complaining”) while the lower threshold (between B and C groups) is slightly increased or remains the same.

Future research in this direction should focus to long-term trend analysis, running similar experiment each 2 years in the future. Another possibility would be to run several subjective tests for different age listener groups in parallel, focusing on differences in quality demands depending on human age.

Monitoring the real demands of telecommunication network users seems to be of vital importance of practical deployment of speech transmission quality monitoring systems where proper thresholds should be set to enable automatic indication of important quality decrease. Systems output using MOS-like output parameter (deploying e.g. P.861 PSQM, P.862+P.862 PESQ-LQ.1, P.561+P.562 CCI, P.563 3SQM, P.564 or proprietary PAMS, TOSQA etc.) can deploy thresholds similar to obtained by our research.

5. Acknowledgements

This work has been supported by the Czech ministry of Education: MSM 6840770014 ”Research in the Area of the Prospective Information and Navigation Technologies”, by the Slovak Research and Development Agency under the contract No.APVV-0369-07 and 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”. Author would like to thank to Ing. Oldřich Slavata and Ing. Pavel Souček for their invaluable help and assistance.

References

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

[2] ITU-T G.107 The E-Model, a computational model for use in transmission planning. Geneva 2003

[3] ITU-T G.108 Application of the E-model: A planning guide. Geneva 1999