Jan Holub – Pavel Soucek *

SUBJECTIVE TESTING AND OBJECTIVE MODELLING OF INFLUENCE OF DIFFERENT SOCIAL CLASSES TO VOICE CALL QUALITY PERCEPTION

The paper presents the analysis of different perception of quality of transmitted voice by different social classes in conference call. Two different social classes (namely people with great income, who are older than 35 and people with low income, who are younger than 25) have been examined by means of ITU-T P.800 conversational tests. Significant differences in perception between those two classes have been identified.

Keywords: *speech quality, subjective testing, Mean Opinion Score.*

1 Introduction

Modern telecommunication technology introduces several impairments into the voice transmission channel. Nowadays it is not enough to measure only physical parameters of the channel but it must be assessed how it affects resulting voice transmission quality.

To assess the quality of voice transmission the scale MOS (mean opinion score) is used (Table. 1). The term MOS is defined in Recommendation ITU-T P.800 [1].

MOS scale Tab. 1

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

There are several methods to obtain MOS values. The first and most accurate one is subjective testing, where the MOS value is obtained directly from users. Subjective tests are standardized by ITU in order to ensure repeatability of experimental results in [1].

Subjective tests are divided into the conversational and listening test. Conversational tests are more demanding on time and organization than listening. Therefore they are used mainly for testing the parameters of the transmission channel, which cannot be tested via a simple listening test, such as delay.

The second way to obtain MOS values is objective testing, which can be divided into two categories – intrusive and non-intrusive.

Intrusive methods usually deliver results nearest to subjective tests. They are based on a comparison of the original and transferred sample. These tests are based on algorithms, which use psychoacoustic models of human perception. They attempt to mathematically describe the human perception of sound and find variables which have a direct impact on the perceived quality of voice signal. Intrusive methods include several standardized algorithms, such as PESQ (Perceptual Evaluation of Speech Quality – ITU-T P.862 [2]) and POLQA (Perceptual Objective Listening Quality Assessment ITU-T P.863 [3]).

Another type of objective measurements is non-intrusive method. These methods do not use the reference signal and final MOS is calculated only from the parameters of the transferred sample. The disadvantage of these methods is a lower accuracy and reliability than in the case of intrusive methods. Example of non-intrusive method is 3SQM (Single Side Speech Quality Measurement – ITU-T P.563 [4]).

2 Work Performed

There are many impairment types that can affect conversation experience in telecommunications. One of the most significant is delay, because it directly influences user's experience. As we said

* **Jan Holub, Pavel Soucek**
Dept. of Measurement K13138, FEE, CTU Prague, Czech Republic, E-mail: holubjan@fel.cvut.cz

earlier effects of delay can be assessed by means of conversational tests. We focused on teleconference calls, which are used for business calls, or online lessons. Usually they are used when participants are far apart. Because there are recommendations only for conversation between two users, we tried to extend methods from [1] for conversation with three participants.

The relationship between delay and resulting MOS score is known (i.e. [5]) and is reflected by many algorithms. The way how delay affects users depends on many factors. We believe that among others it is socio-economic background of user.

Tests were performed in Czech language.

2.1 Test-bed

Experiment was conducted in four separate rooms (Fig. 1) to avoid direct contact between the respondents. One of the rooms fully meets the requirements of the standard (reverberation 182 ms, −60 dB). The second room is particularly suited for the listening tests, meets the requirements of reverberation time <500ms, the other parameters have not been measured in the room. The third room is undefined in terms of acoustic parameters, we can assume that even this room meets the reverberation time <500ms. In the fourth room the technical background of the experiment is situated, the network simulator and posts of experiment supervisor.

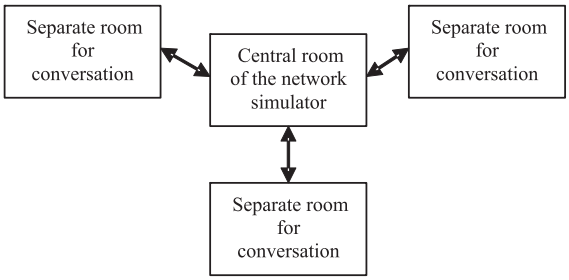

Fig. 1 Test-bed

2.2 Network Simulator

Respondent's posts include telephone chassis with standard handset. The signal from a handset microphone is pre-processed and routed into the central part of the simulator (Fig. 2). In the opposite direction is carried signal to the loudspeaker. The signal from the microphone is pre-processed in microphone amplifier (SHARK). The central part of the simulator consists of two digital signal processors. The first processor (DCX A) made a filtration with a Butterworth high-pass filter 48th level, 303Hz and low-pass filter Bessel 24th level, 3031Hz. Furthermore, the DCX A sets the first part of the variable delay in the range of 1-582 ms.

The processors are connected so that each of the three inputs of DCX B is the sum of the two different analogue outputs of DCX A. In the DCX the second part of the delay B is implemented and output signals are carried into the handsets of the respondents.

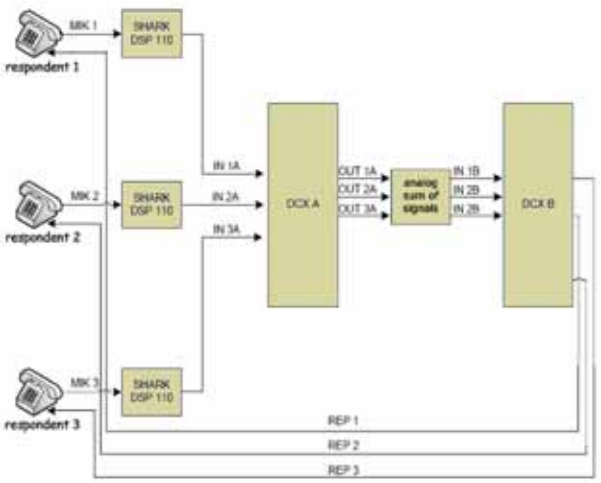

Fig. 2 Block diagram of network simulator

2.3 Delay

Delay is defined as time needed for voice signal to travel from talker to listener. Delay of telephone call in an IP network has several different causes. On the side of speaker it is particularly encoding, packetisation and controller interface. On the side of listener it is buffer, depacketisation and decoding. Causes of delays in the IP network itself are in particular: limited speed of signal transmission in the network and signal processing time of involved components such as routers and converters. The speed of the signal transmission is a particular problem when the call is made for long distance or part of the route led via satellite. In this experiment the following values of delay were adjusted: 62, 337, 612, 887 and 1176 ms.

2.4 Selection of Participants

Certain criteria must be met in selection of participants. They are described in detail in [1]. Among others participants should not be experts in area of telecommunication and they should have no hearing impairments.

As it was proven in previous experiments described in (i.e. [6]), participants are not able to distinguish between individual values of delay sometimes. Therefore all participants were instructed prior to testing that they should focus on delay.

Because of nature of our experiment we need participants with different socio-economic background. We decided to divide participants into two groups named Managers and Students.

Managers are people with higher education, with prestigious job position, above average income and their age is higher than 35. They are used to certain standards and they are willing to pay for quality. Also they expect to get quality they paid for.

In contrast, Students have lower income than managers and are used to get by with cheaper services. It was not necessary for participants from this group to be actually studying at the time, when tests took place.

In our experiment participated 55 people – 43 in group of Students, 9 in group of Managers. Other 3 participants were part of pre-test session, which was used for selection of conversation scenarios and proper way instruction session. The numbers of participants in both groups clearly show that Managers are much more difficult to acquire for participation in subjective tests.

2.5 Conversation Scenarios

There are several ways to conduct conversational tests. We used loosely defined scenarios based on real everyday life situations, which were from 2 to 3 minutes long. We tried to find scenarios which will be interactive enough, symmetric if possible. In the end we picked 5 following scenarios:

- Selecting gift – selecting present for friend, every participant have different budget and preferences
- Work on weekend – unexpected emergency work on weekend, participants already had plans for
- Party – participants are organizing party,
- Sport – participants have to decide which sport they will play, they have different preferences
- Culture event – participants have to decide which culture event they will attend, they have different preferences

It is clear from scenarios description that instructions for participants consist of two parts – common for all 3 participants and individual for each of them.

3 Results

In this part of paper we will evaluate whether there is difference between our two socio-economic groups. We will compare them directly and with results from objective tests.

3.1 Result Evaluation

Results from subjective test were processed and confidence intervals CI95 were computed using following formula:

$$u = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}\left(MOS_i - \overline{MOS}\right)^2}$$

where N is number of scores for conversation and set delay
MOS_i is score from participants
$\overline{MOS}$ is average MOS

For CI95 for difference between Managers and students we used formula for indirect measurement:

$$u_{undirect} = \sqrt{\sum_{j=1}^{m}\left(\frac{\partial f}{\partial X_j}u_{x\_j}\right)^2}$$

where f is function used for computation of result (in our case simple difference)
X_j is one of the values (in our case *MOS* from one of our groups)
u_{x_j} is uncertainty of X_j

For comparison of subjective and objective tests we used several criteria.

Pearson's correlation coefficient:

$$r = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2(Y_i - \overline{Y})^2}},$$

where $\overline{X}$, $\overline{Y}$ are mean values of measured samples
X_i, Y_i are the values of the i-th samples

Root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(MOS\_CQS_i - MOSC\_CQO_i)^2},$$

where N is number of tested delays
MOS_CQS_i is *MOS* result obtained by means of subjective testing
MOS_CQO_i is *MOS* result obtained by means of objective testing

Modified RMSE (RMSE*):

$$RMSE^* = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\Delta_i)^2},$$

where N is number of tested delays
Δ_i is 0 if difference between objective and subjective results is lesser than CI95 for subjective result and in other case it is given by following formula:

$$\Delta_i = \left|MOS\_CQS_i - MOSC\_CQO_i\right| - u_{k=2}$$

Maximum absolute difference:

$$D_{MAX} = MAX\left(\left|MOS\_CQS_i - MOS\_CQO_i\right|\right),$$

where MOS_CQS_i is *MOS* result obtained by means of subjective testing
MOS_CQO_i is *MOS* result obtained by means of objective testing

3.2 Subjective Tests

As we can see in Figs. 3 and 4 there is difference between our two groups. In case of both groups the quality drops with bigger delay. It is also clear that confidence intervals are bigger in case of Managers due to limited number of participants in this group.

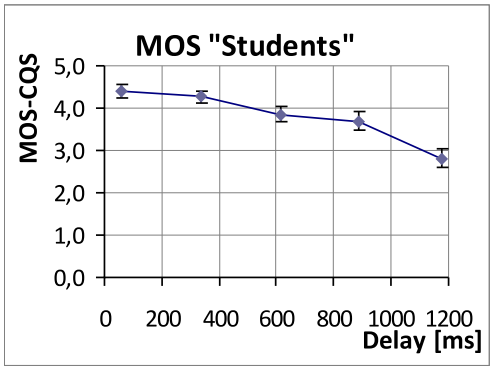

Fig. 3 MOS values for the group Students

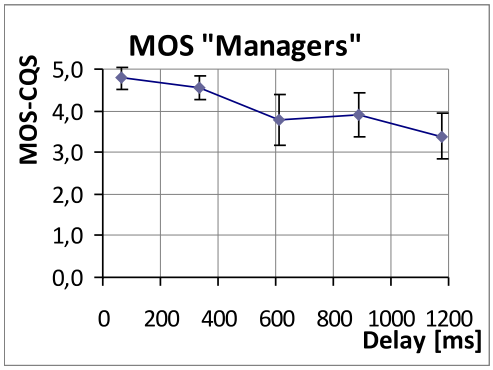

Fig. 4 MOS values for the group Mangers

Surprisingly from Fig. 5 is clear that the Managers are more tolerant to delay than the Students. This finding is directly opposite to our presumption that the Managers should be more demanding.

3.3 Objective Tests

From collected subjective date we created correction function. As we can see from Table 2, difference between group of Students and results obtained from PESQ are after proper regressions very similar. RMSE* was zero for both groups. Before correction it was 0.025 for Managers and 0 for Students.

4 Conclusion and Future Work

We have proven that there actually is difference between groups of users with different socio-economic background. It is surprising

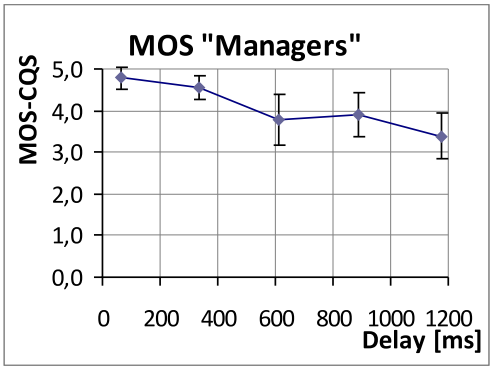

Fig. 5 Difference between the groups Managers and Students and its CI95. Important CI95 non-crossings with zero are obvious for the first two and last measured points.

Comparison between groups of managers and students Tab. 2

	Objective – "Managers"	Objective – "Students"
Criteria	Subjective – "Managers"	Subjective – "Students"
Pearson's correlation	0.952	0.990
RMSE	0.157	0.078
Maximum absolute difference	0.252	0.124

that the Managers, who we assumed should be used to higher standard, are less demanding. We suppose that this can be due to their higher experience with teleconference calls, or even due to the fact that they are usually older than participants from group of the Students, so they remember older technologies with less quality and are used to communication for longer distances. This was proven in [7]. On the other hand in [8] was proven that MOS scale shift during time even for same group of listeners.

For the future work we consider the validation of our results with higher number of participants from group of the Managers in order to reduce CI95.

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". Authors would like to thank to subjective test participants and to Ing. Michal Toula for their invaluable help and assistance.

References

[1] ITU-T Rec. P. 800: *Methods for Subjective Determination of Transmission Quality,* International Telecommunication Union, Geneva, 1996ITU-T Rec. P.863, Perceptual Objective Listening Quality Assessment (POLQA), Intern. Telecommunication Union, Geneva, 2011

[2] 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,* Intern. Telecommunication Union, Geneva

[3] ITU-T Rec. P.863, *Perceptual Objective Listening Quality Assessment (POLQA),* Intern. Telecommunication Union, Geneva, 2011

[4] ITU-T Rec. P.563, *Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications,* Intern. Telecommunication Union, Geneva, 2004

[5] HOLUB, J., KASTNER, M., TOMISKA, O.: *Delay Effect on Conversational Quality in Telecommunication Networks: Do We Mind?* In: Wireless Telecommunications Symposium 2007 [CD-ROM]. Pomona, California: IEEE Communications Society, 2007

[6] KITAWAKI, N., ITOH, K.: Pure Delay Effect on Speech Quality in Telecommunications. *IEEE J. Sel. Areas Comm.*, 9(4): 586–593, 1991 ITU-T Rec. P.863, Perceptual Objective Listening Quality Assessment (POLQA), Intern. Telecommunication Union, Geneva, 2011

[7] HOLUB, J., SMID, R., BACHTIK, M.: Child Listeners as the Test Subject – Comparison with Adults and P.862, *MESAQIN*, Prague, CTU, 2003

[8] SOUCEK, P., HOLUB, J.: *How do Perceived Speech Quality and Acceptability Level Shifts During Time,* MESAQIN, Prague, CTU, 2011.