České vysoké učení technické v Praze
Fakulta elektrotechnická

Czech Technical University in Prague
Faculty of Electrical Engineering

doc. Ing. Jan Holub, Ph.D.

Pokroky v měření a optimalizaci kvality služeb a kvality uživatelské zkušenosti

Advances in Quality of Service and Quality of Experience Measurements and Optimization

Summary

Quality of Service (QoS) and recently enhanced concept of definition and measurement of customer experience (QoE) is a widely used concept of service assessment. This theses focus to QoE measurements and optimization in telecommunications where rapid development of digital technologies in the area of mobile telephony has led to an increased need for efficient resources deployment while improving the customer's subjective impression about the service provided.

Important results have been achieved recently in both subjective and objective domains for voice and multimedia transmission QoS measurements. In the area of subjective testing several experiments bringing new ways of understanding of QoE in voice and multimedia transmissions are described, e.g. experiments with non-native listeners or listeners originating from various socio-cultural environments. Also extensive conversational experiments challenging the paradigm of 30+ years old results preserved in ITU-T G.107 and G.108 are shown.

In the objective domain, the new application of ITU-T P.863 (POLQA) for e.g. military low-bit rate connections is described, including the non-native user emulation capability. Also an algorithm for conversational quality prediction, standardized afterwards by ETSI as TR 103 121 is briefly mentioned.

Finally, future trends and research areas are sketched.

Souhrn

Kvalita služby (QoS) a v nedávné době nově definované měření kvality zákaznické zkušenosti (QoE) jsou široce rozšířenou koncepcí posuzování služeb. Tyto teze se zaměřují na hodnocení kvality zákaznické zkušenosti v telekomunikacích, kde rychlý rozvoj číslicových technologií v oblasti mobilních komunikací vedl ke zvýšené potřebě optimalizace využívání zdrojů při současném zlepšení subjektivního vnímání služby zákazníkem.

V oblasti měření kvality přenosu hlasu a multimédií bylo v nedávné době dosaženo významného pokroku, a to jak v oblasti subjektivních tak v oblasti objektivních metod. V oblasti subjektivních měření jsou popsány experimenty, přinášející nové způsoby uchopení QoE v oblasti přenosu hlasu, např. experimenty s nerodilými posluchači či s posluchači z odlišných sociokulturních vrstev. Dále jsou představeny obsáhlé konverzační experimenty, rozporující paradigmata stanovená před více jak 30 lety a zakonzervovaná v doporučeních ITU-T G.107 a G.108.

V oblasti objektivních měření je popsáno nové využití algoritmu ITU-T P.863 (POLQA) pro např. vojenské komunikace s nízkými přenosovými rychlostmi s možností emulace uživatele, nepoužívajícího ke komunikaci svůj rodný jazyk. Dále je zmíněn predikční algoritmus pro odhad kvality konverzace, který byl následně standardizován organizací ETSI jako doporučení TR 103 121.

Závěrem jsou naznačeny trendy a oblasti budoucího výzkumu.

Klíčová slova:

Kvalita služby, kvalita uživatelské zkušenosti, přenos hlasu, objektivní testy, subjektivní testy, poslechový test, konverzační test

Keywords:

Quality of Service, Quality of Experience, voice transmission, objective testing, subjective testing, listening test, conversation test

Contents:

1. Introduction

Multiple definitions of Quality of Service (QoS) and Quality of Experience (QoE) exist and moreover these keep changing with time. Currently, two QoE definitions are recognized widely: one drafted by ITU and another one by ETSI:

[1] ITU-T P.10/G.100: QoE is the overall acceptability of an application or service, as perceived subjectively by the end user. Quality of experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.). Overall acceptability may be influenced by user expectations and context.

[2] ETSI STF 354, (2010): QoE is a measure of user performance based on objective and subjective psychological measures of using a service or product.

Quality of Service (QoS) is defined e.g. in [3] ITU-T E.800 (1994): QoS is the collective effect of service performance which determine the degree of satisfaction of a user of the service. Thus, QoS is rather technical term or set of parameters – as highlighted in alternative definition by [4] IETF RFC 2386: QoS is a set of service requirements to be met by the network while transporting a flow.

The following text is focused mostly to QoS and QoE measurements and innovations for voice transmission in telecommunications. However, similar ideas are applicable for most of other domains where QoS and QoE concepts are used, e.g. transportation or civil engineering.

Speech transmission during any call in the telecommunication network is affected by many impairments; including delay, echo, various kinds of noise, speech (de)coding distortions and artefacts, temporal and amplitude clipping etc. [25] Each transmission impairment has a certain perceptual impact on the speech transmission quality and in case of multiple impairment types these can mask each other. The overall quality can be evaluated and expressed in terms of a Mean Opinion Score (MOS) [5], covering the range from 1 (bad) to 5 (excellent). Speech transmission quality measurements are widely used to compare different coding and transmission technologies, or to monitor the network performance. The traditionally proven but expensive subjective methods [5], involving human listeners assessing many speech samples, have been partially replaced by objective digital signal processing algorithm based measurements that either compare the original undistorted signal to the received one [6] (so called intrusive or double-sided algorithms) or process only the received version [7]. All these methods have been designed and tested on past and contemporary telecommunication transmission standards that are widely used in common mobile and fixed telecommunication networks, e.g. those using 'toll quality' voice encoding.

Within the last 25 years, several aspects and models of signal processing in the auditory system have been applied to methods for objective speech quality measurement. In general, most quality assessment methods, subjective and objective, aim to quantify the quality (or the quality degradation) of a transmitted speech sample relatively to a non-degraded reference situation.

2. Speech and Multimedia Transmission Assessment Methods

2.1 Subjective methods

Subjective speech quality tests seek to quantify the range of opinions that listeners express when they listen to speech transmission systems that are under test. For the evaluation of these systems, many subjective assessment procedures have been developed and standardised over the past decades [25]. The different methods may be distinguished by many aspects. They can, e.g., involve conversational tests or listening-only tests [5]. Conversational tests (i.e., tests where two subjects have to listen and talk interactively via a transmission system) will achieve a more realistic test environment for the assessment of speech quality. On the other hand, they are much more time consuming to perform and are often subject to lower reproducibility [8].

For most cases the method recommended by [5] is therefore a listening-only test. Commonly used subjective test methods are Absolute Category Methods, Degradation Category Methods, Detectability Methods, Comparison Category Methods, and Threshold Methods. One widely used method is the direct evaluation of the speech quality by an Absolute Category Rating (ACR). The subject is presented with short groups of unrelated sentences which were passed through a system under test. Typically, the subject's task is to rate his/her impression on a five-point scale with absolute categories. An estimate of the quality is then the arithmetic mean of the responses of all subjects which is called the mean opinion score (MOS). Other currently recommended assessment methods are described e.g. in [9].

Subjective speech quality data acquired with good reliability and reproducibility generally require large investments in terms of technical equipment and manpower. Such efforts are necessary and accepted for standardisation or specification tests, that have been performed e.g. to establish the GSM [10], AMR [11] or recently EVS [12] codec standards. The costs of these tests are, however, unacceptable during the development of algorithms and devices. Therefore non-auditive instrumental methods for a quality judgement have been of great interest for a long time already.

2.2 Objective methods

The goal of objective speech transmission QoS measurement is to predict speech transmission quality based on objective measures of physical parameters and properties of the speech signal waveform [25]. An automated implementation of an objective test requires significantly less effort, time and expense than the corresponding subjective tests. On the other hand, often a considerable difference between objective and subjective test results is observed [13]. In this case subjective results are generally considered to hold the "correct answer". The performance of objective measures is therefore judged by their respective ability to approximate the subjective speech quality results as closely as possible.

The most recent achievement of ITU-T is [6] P.863 POLQA, an objective method for predicting overall listening speech quality from narrowband (300 to 3 400 Hz) to super-wideband (50 to 14 000 Hz). It defines a single algorithm for assessing the speech quality of current and near future telephony systems utilizing a broad variety of coding, transport and enhancement technologies. The measurement algorithm is a full reference model which operates by performing a comparison between a known reference signal and a captured degraded signal. This is consistent with the algorithms described in previous recommendations ITU-T P.861 and P.862.

3 Advances in subjective testing methods

3.1 Conversation tests to detect delay and echo sensitivity [8], [21]

To confirm or challenge the knowledge of human sensitivity to delay and talker echo as experienced during telephone conversation reported in [14], a set of conversational tests according to [15] ITU-T P.805 in English and Czech were run. The author observed already in his past experiments [16] severe differences in human perception compared to [14]. The test scenarios were defined in order to create several conversational interactivity levels between the two subjects involved in each conversation. To take into account the interactivity between the talkers, a new parameter called Talker Alternation Rate (TAR) is introduced. In some previous works (e.g. by F. Hammer in [17]), Speaker Alternation Rate (SAR) was used to denote the same, however, SAR can make confusion having alternative meaning of Specific Absorption Rate. Therefore, authors of this contribution propose to use TAR instead.

The subjective conversation tests covered the following characteristics:

- 3 coders: G.711 [18] A-law, G.729AB [19] (@ 8kbit/s), AMR-NB [11] (@ 12.2kbit/s)
- 3 delay values: 100, 300, 600 ms one-way delay
- 2 echo situations: weak echo, strong echo, TELR= 46dB, 32dB [20]
- 3 levels of interactivity i.e. different categories. The exact test scenarios can be found in Annexes B and C of ETSI TR 103 121 [21] and are mostly based on the scenarios defined in ITU-T P.805 [15].
- 54 conditions in English and 18 conditions in Czech, in total 72 conditions
- 48 votes per condition (equals to 3456 votes in total)
- The equivalent of a reference terminal - real-time adaptation to ES 202 737 [20] in send and receive direction, with diffuse field correction as per ITU-T Recommendation P.57 [22].

The subjective conversational tests have been performed on 24 English and 8 Czech native talker pairs. The test environment conformed to ITU-T P.800 [5] requirements. A proprietary DSP-based real-time network simulator has been designed, assembled, calibrated and used for the tests. Its terminals have been calibrated on Head and Torso Simulator as specified in [20].

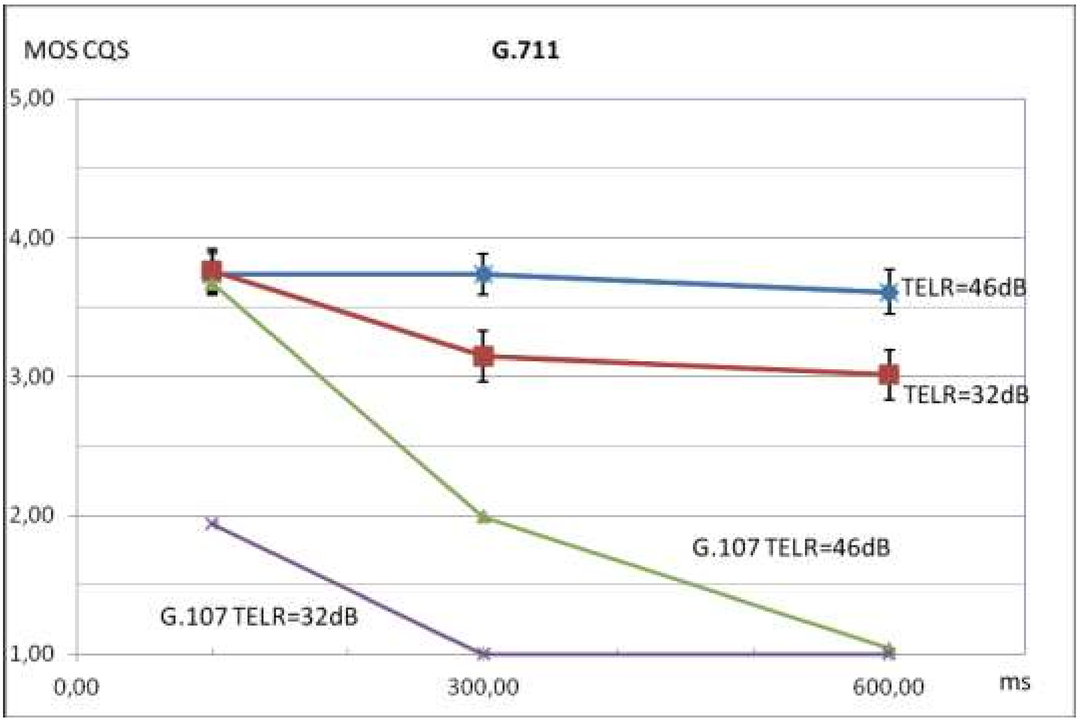

Fig. 3.1: Subjective test results example: G.711 coder and two tested TELR values (32dB, 46dB) including CI95% uncertainty intervals. Corresponding E-model (G.107) results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600ms, the connecting lines are shown for informative purposes only [21].

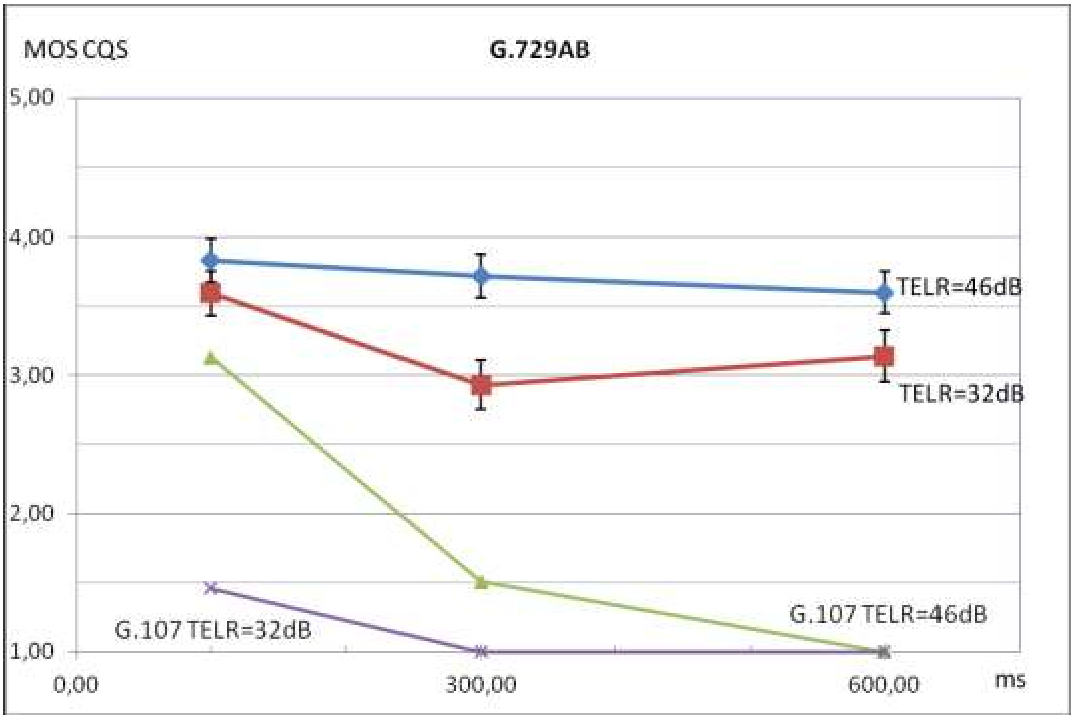

Fig. 3.2: Subjective test results example: G.729AB coder and two tested TELR values (32dB, 46dB) including CI95% uncertainty intervals. Corresponding E-model (G.107) results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600ms, the connecting lines are shown for informative purposes only [21].

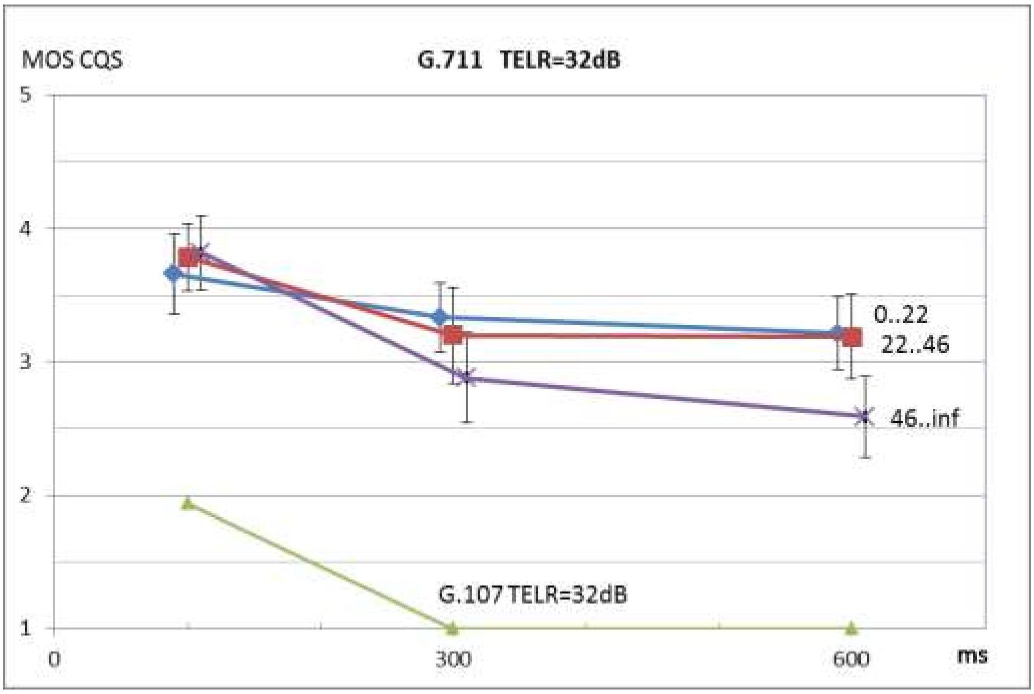

Fig. 3.3: Subjective test results for G.711 coder and TELR = 32dB, split for 3 different interactivity levels based on TAR analysis, including CI95% uncertainty intervals. Corresponding E-model (G.107) results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600ms, the connecting lines are shown for informative purposes only [21].

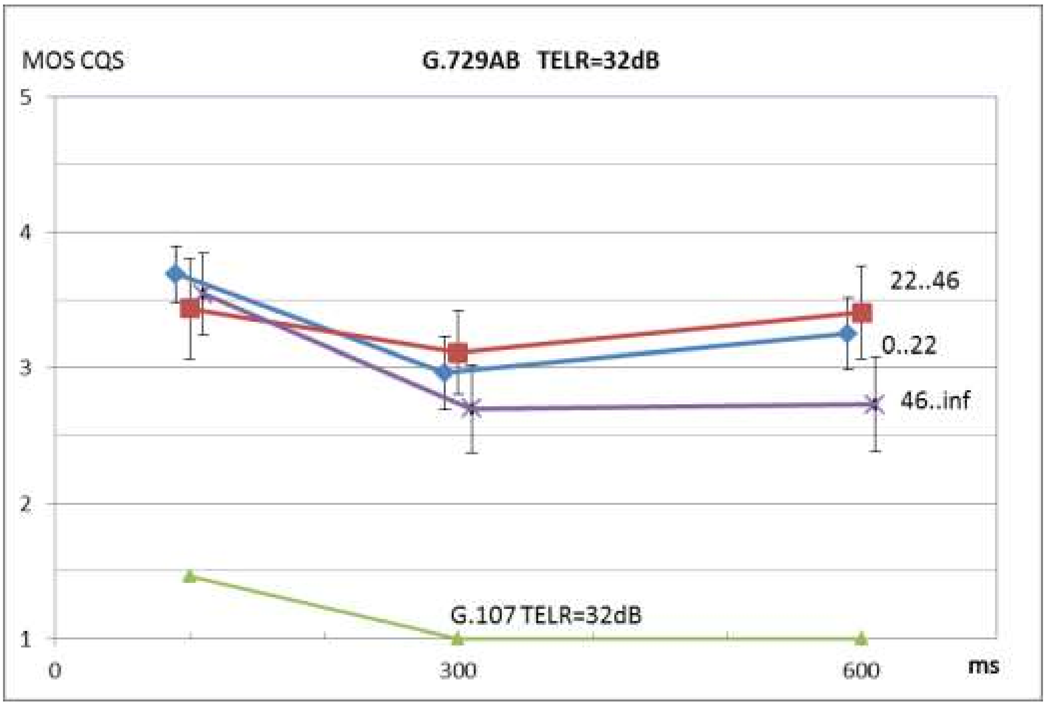

Fig. 3.4: Subjective test results for G.729AB coder and TELR = 32dB, split for 3 different interactivity levels based on TAR analysis, including CI95% uncertainty intervals. Corresponding E-model (G.107) results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600ms, the connecting lines are shown for informative purposes only [21].

As outlined above, the graphs are derived to show the differences between the E-model [14] and the new approach. In fact, two different values of MOS-CQ are obtained for each combination of input parameters (codec, delay, echo level, etc.):

-The E-model (G.107) output, recalculated from R to MOS scale (referred further as „E-model“)

-MOS-CQS as obtained by subjective tests with appropriate 95% confidence intervals (CI95%)

Results obtained for other coder, TELR and TAR combinations can be found in [21]. For low echo condition of TELR=46dB, the subjective sensitivity to delay is significantly lower than as predicted by E-model. The typical difference between MOC-CQS for 100ms and 600ms is for low echo condition approximately 0.5 MOS. For coders deploying higher perceptual compression (G.729AB) affecting the listening quality the MOS-CQS becomes for stronger echo (TELR=32dB) non-monotonic with new local minima located (in our case) at 300ms. Similar effects have been accidentaly reported by various labs in previous experiment but usually neglected and rendered out by data interpolation. This feature was first identified in [16].

3.2 Listening tests with subjects originating from different socio-cultural environments [26]

There are many impairment types that can affect conversation experience in telecommunications. One of the most significant is delay, because it directly influences QoE. As we said earlier effects of delay can be assessed by means of conversational tests. We focused on teleconference calls, which are often for business calls or online lessons. Because there are recommendations only for conversation between two users, the methods from [5] and [15] have been extended to conversation of three participants.

The relationship between delay and resulting MOS score is known (i.e. [14]) and is reflected by many algorithms. The way how delay affects users depends on many factors. We aim to prove that among others it is socio-economic background of user. Therefore, the following tests in Czech language have been designed. Experiment was conducted in four separate rooms (Fig. 3.5) to avoid direct contact between the respondents. One of the rooms fully meets the requirements of the recommendation P.800 - reverberation 182ms, noise below 30 dB SPL(A). The second and third rooms meet the requirements of reverberation time <500ms, the other parameters have not been measured in the room. In the fourth room the technical background of the experiment, the network simulator and seat for experiment supervisor are located.

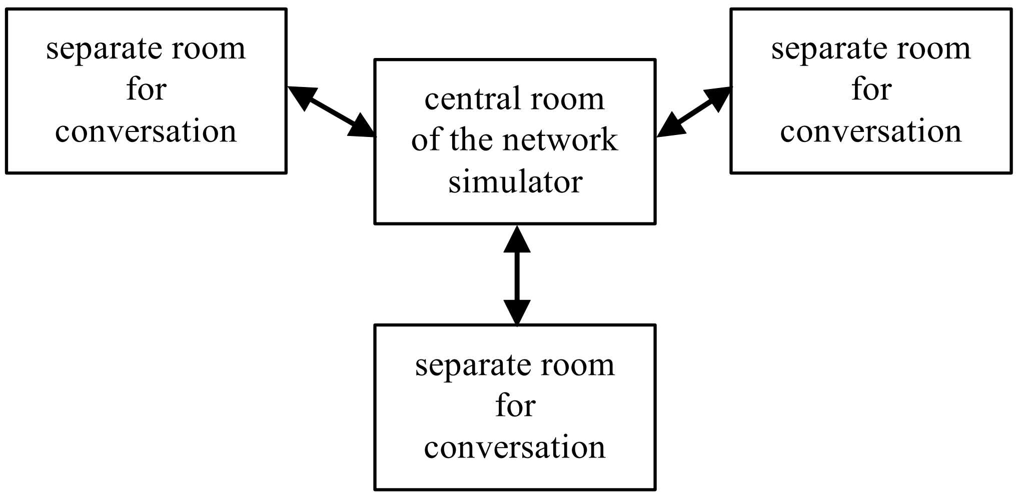

Fig. 3.5: Test-bed [26]

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. 3.6). 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 48^{th} level, 303Hz and low-pass filter Bessel 24^{th} 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.

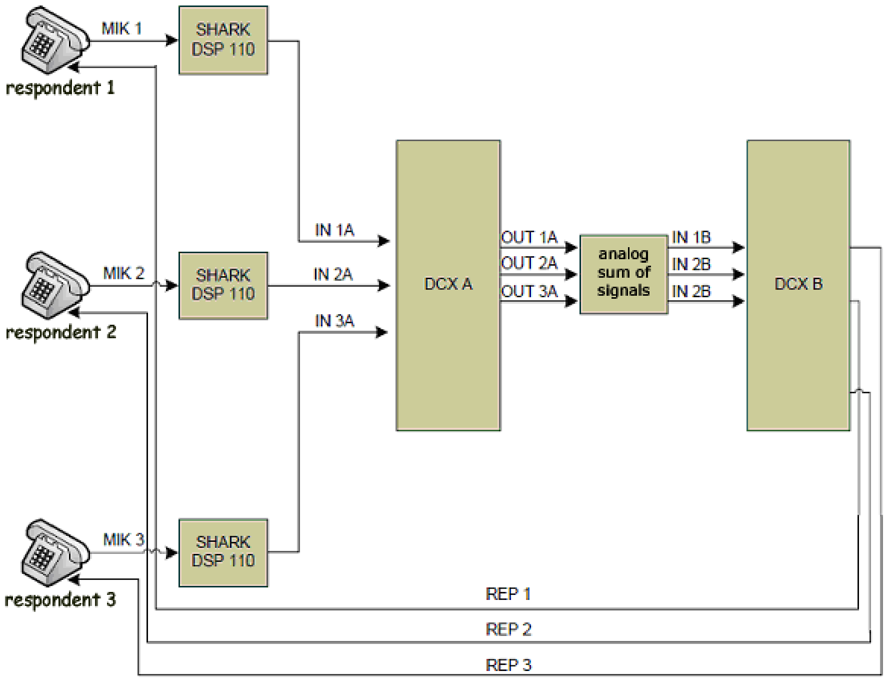

Fig. 3.6: Block diagram of network simulator [26]

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 speaker‘s side 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 is led via satellite. In this experiment the following values of delay were adjusted: 62, 337, 612, 887 and 1176ms.

Certain criteria must be met in selection of participants. They are described in detail in [6]. 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. [23]), 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 the experiment participants with different socio-economic background were needed. 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 got used to certain standards and they are willing to pay for quality. Also they expect to get quality they paid for. On the contrary, „Students“ have lower income than average and are willing to easier accept cheaper services. It was not necessary for participants from this group to be actually studying at the time when tests took place.

55 people participated in our experiment: 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 of session instruction. The numbers of participants in both groups clearly show that Managers are much more difficult to acquire for participation in subjective tests.

There are several ways to conduct conversational tests. In this project so called weakly defined scenarios [15] based on real everyday life situations, 2 to 3 minutes long, were used. We tried to find scenarios which will be interactive enough and symmetric if possible:

-Selecting gift – selecting a 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

Instructions for participants consisted of two parts – common for all 3 participants and individual for each of them.

As seen from Figs. 3.7 and 3.8 there is difference between our two groups. In case of both groups the quality drops with increased delay. It is also clear that confidence intervals are bigger in case of „Managers“ due to limited number of participants in this group.

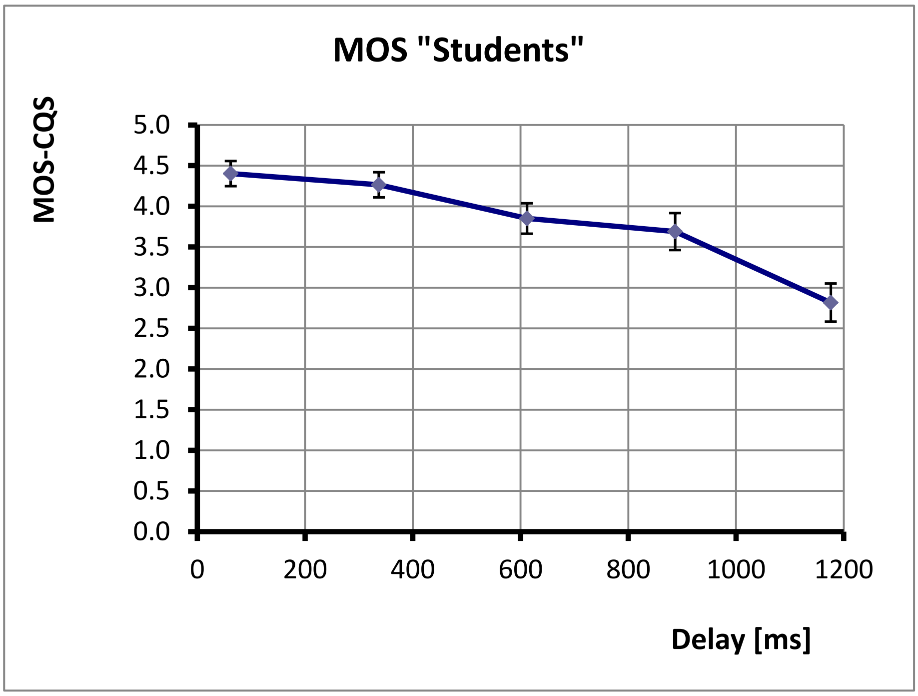

Fig. 3.7: MOS values for the group “Students” [26]

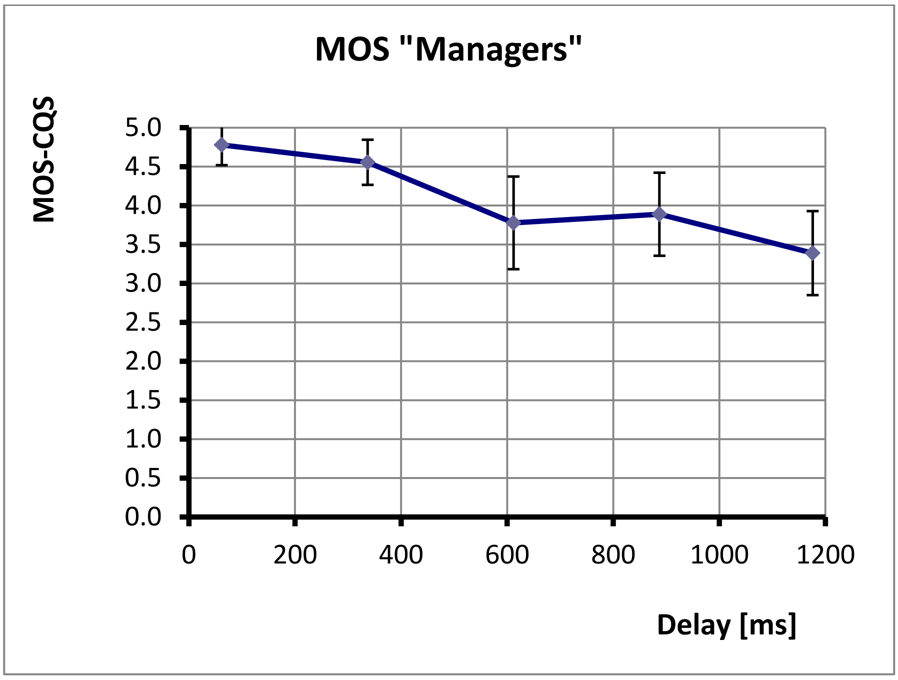

Fig. 3.8: MOS values for the group "Mangers" [26]

Surprisingly from Fig. 3.9 follows that „Managers“ are more tolerant to delay than the „Students“. This finding is directly opposite to original presumption that the „Managers“ should be more demanding.

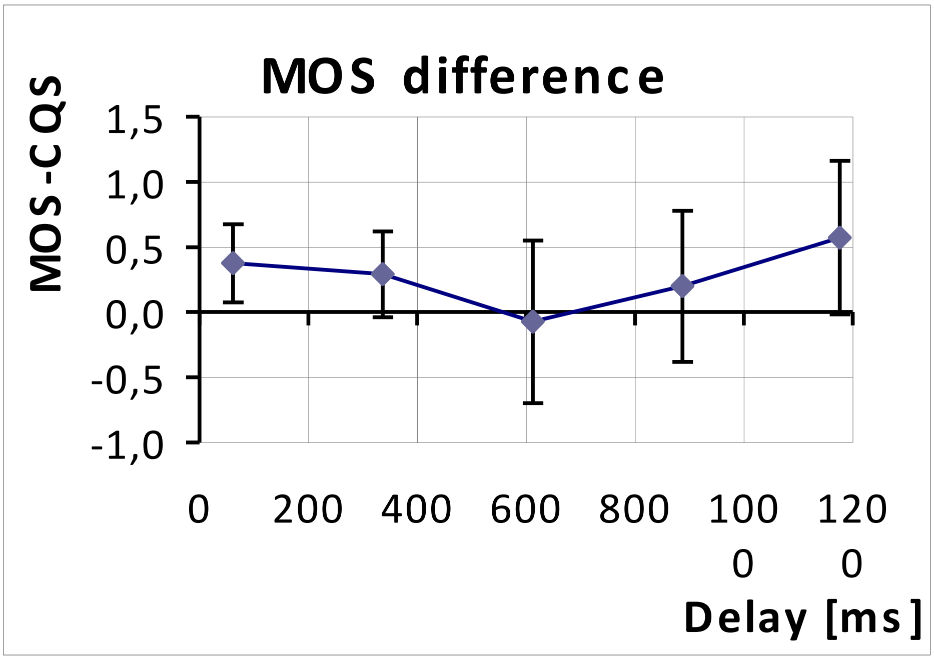

Fig. 3.9: Difference between the groups "Managers" and "Students" and its CI95.
Important CI95 non-crossings with zero are obvious for the first two and last measured points [26]

It was proved that there actually is difference between groups of users with different socio-economic background. It may be considered surprising 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 [24]. On the other hand in [25] was proven that MOS scale shifts during time even for same group of listeners.

3.3 Listening tests with subjects of different language proficiency [27], [28]

In many practical cases, the communication in the telecommunication network is carrying a non-native language for one or more conversation participants. Typical examples are e.g. international and/or roaming calls in today's public fixed and mobile telecommunication networks or communications in military radio telecommunication networks during multi-national tactical operations, international governmental organisations or multi-national companies. There are procedures to automatically estimate perceived quality of the transmitted speech [6] and their results correlate well with subjective experiments carried on native speakers and native listeners. However, it is not clear if the effect of listener non-nativity can affect the quality perception. This work examined methods to quantify such effects by listening test results performed on non-native listeners, pre-sorted according their English proficiency.

Unfortunately, there are contradictory hypotheses about such an influence:

- Non-native listeners have higher difficulties to understand the contents even for less distorted samples as native listeners, thus they should assess quality worse (=giving generally lower scores) than native listeners.
- Non-native listener's brain is more occupied by message content decoding than in case of native listener, thus the quality assessment should not be so detailed, so some impairments can be missed, thus the final scores should be higher than for native listeners.

A speech database fulfilling P.800 requirements and containing two background noise conditions (no noise / Hoth noise +10dB SNR) has been recorded on selected coders (PCM 8 bit [18], GSM 06.10 [10], MELPe 2.4 kbit/s [29]). The final database contained 120 different sentences spoken by native English speakers. Voices of more than 2 female and 2 male speakers, recorded in studio environment, have been used. In each case, 15 sentences per condition (noise+coder, see Table 1) have been prepared. The active speech level as per ITU-T P.56 has been equalized to -26 dBoV that corresponded then to 79 dB SPL (A) during the listening tests.

Subjective tests have been carried out on naive subjects as required by P.800. Their age was in the range between 20 and 30. None of them was a native English speaker, the nationalities represented in the group were: Czech, Slovak, Italian. The English proficiency of each subject has been verified by short quiz, composed by played-out English sentences/articles and followed by set of questions to be answered in a written using multiple-choice principle. The language test lasted 7 minutes and was always performed right before the quality testing. The maximum achievable number of points in the language test was 21. Based on the language test results, the subjects were assigned to one of 3 categories:

- "Beginners" (0-3 points)
- "Intermediate" (4-10 points)
- "Advanced" (11-21 points)

The subjects were not informed about their results after the language tests.

Subjective tests as per ITU-T P.800 [5] have been performed on the above mentioned 120 sample database. The subjective listening-only tests have been performed in a critical listening room where up to 8 listeners can be seated. The reverberation time of the room is 185 ms and natural background noise less than 25dB SPL (A). The samples have been played-back in random order, by means of digital playback system with SNR higher than 90 dB. The loudspeakers were actively compensated to achieve transition ripple less than 0,8 dB in audible frequency range. Multiple sessions have been run always with different listeners. In total, 36 votes per sample have been obtained, 13 per “Beginners”, 11 per “Intermediate” and 12 per “Advanced” groups.

For verification purposes, similar subjective tests as described above have been performed using the same database but on native listeners. This experiment was carried out by the author in Los Angeles, California, in April 2008. The test subjects were students of Cal Poly Pomona. The purpose of the test was to compare influence of (non-) nativity and different expectations of both groups of subjects, coming at the same time from different continents. Test subjects were seated in standard class room with only basic anechoic measures (plasterboard lining). The play-out system used non-compensated loudspeakers with transition ripple up to 9 dB in audible frequency range.

Due to different expectations driven by different communication technologies used in different countries and also due to different environmental conditions (room and equipment) the experimental results achieved on native and non-native listeners can not be directly compared. This is also well noticeable from Table 3.3 where Pearson correlation coefficients are reported. The results coming from native listeners provide significantly lower correlations with all other listener groups than in the remaining rows (where results between two non-native listener groups are reported). Note that correlation calculation is invariant to offset and gain changes so the systematic offset identified between “Advanced“ and other non-native groups is not influencing the results in Table 3.3.

Test results are given in the following tables and figures. A special attention has been paid to differences in quality perception between Advanced group and the remaining two non-native (Intermediate and Beginners) groups. Per-condition results are listed in Table 3.2 and shown in Figure 3.10. Figure 3.11 shows results per sample. Both per-condition and per-sample results shown clear shift in subjective scoring of non-native listeners and the difference between Advanced and both other groups (Beginners and Intermediate) is about 0.5 MOS for the entire MOS scale. The difference between Intermediate and Beginners is not so evident (not shown in the pictures) and fits within confidence intervals of subjective experiments. The results of native listeners testing are reported in Fig. 3.12.

Table 3.1. Subjective test results (per condition)

Condition	Noise type	Coder	MOS-LQSn Advanced	MOS-LQSn Intermediate	MOS-LQSn Beginners
1	no noise	clean	4,38	3,78	3,88
2	no noise	PCM 8bit lin.	3,32	2,87	3,09
3	no noise	GSM 06.10	2,51	1,81	1,75
4	no noise	MELPe 2.4	2,87	2,58	2,44
5	10 dB Hoth	clean	3,55	2,74	2,61
6	10 dB Hoth	PCM 8bit lin.	2,75	2,35	2,36
7	10 dB Hoth	GSM 06.10	1,94	1,48	1,33
8	10 dB Hoth	MELPe 2.4	2,16	1,91	1,85

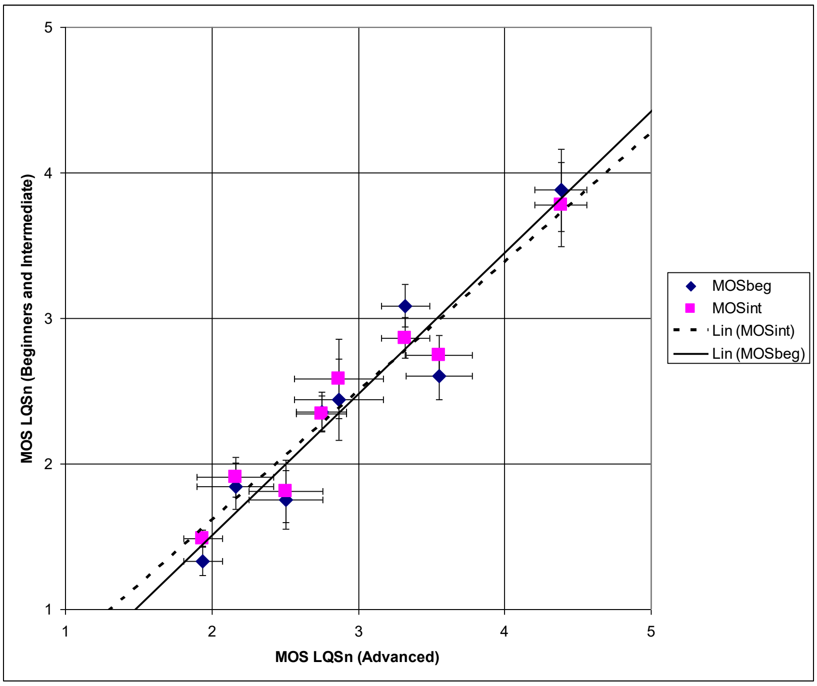

Fig.3.10: Subjective test results per condition, comparison between "Advanced" and other ("Intermediate" and "Beginners") groups. 95% confidence intervals (CI95) are reported [27]

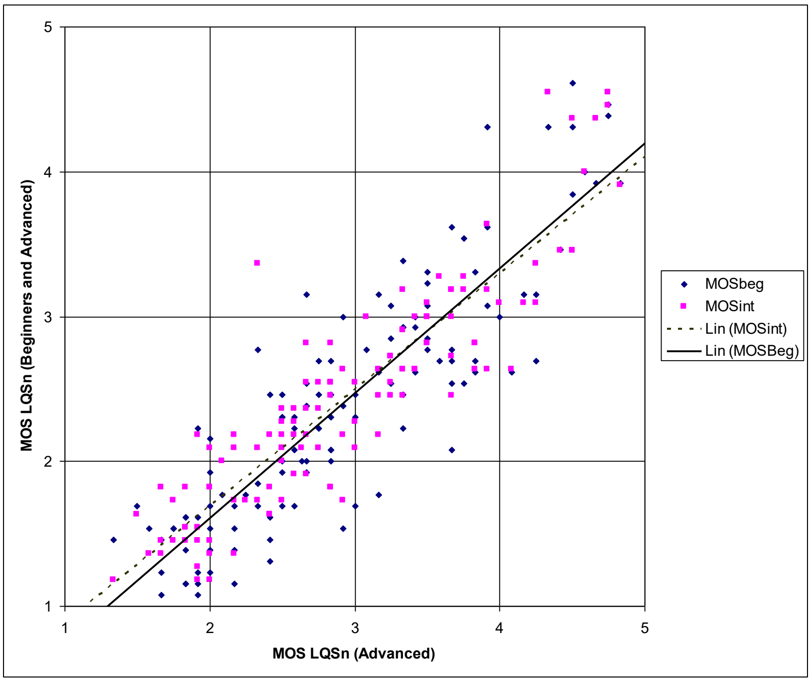

Fig. 3.11: Subjective test results per sample, comparison between "Advanced" and other ("Intermediate" and "Beginners") groups [27]

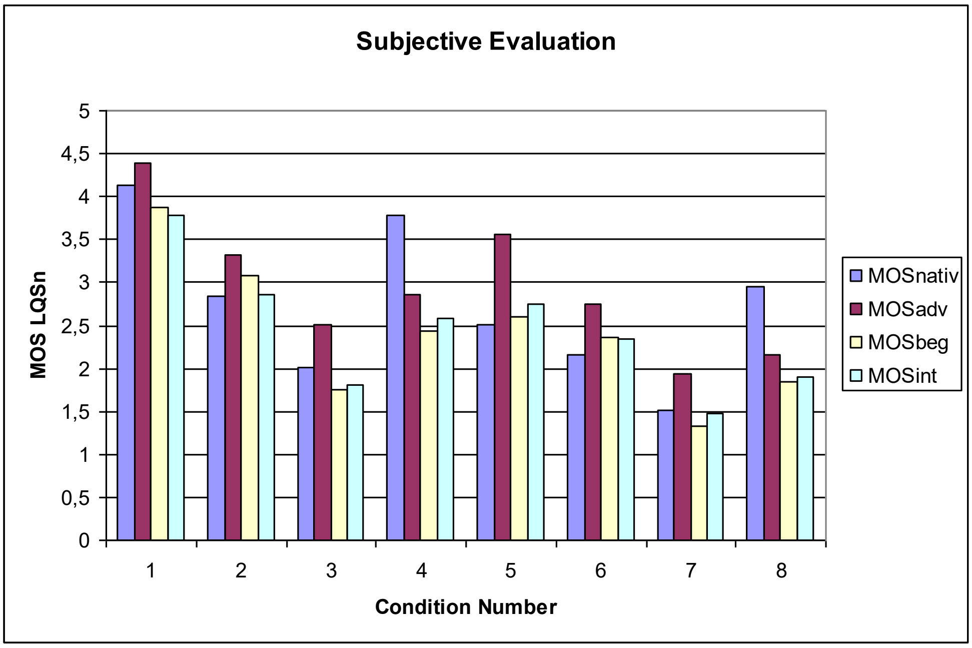

Fig. 3.12: Comparison between non-native and native listeners with different expectation factors [27]

Table 3.2. Pearson correlation coefficients between different listener groups ("per condition" results) [27]

	Native	**Advanced**	**Intermediate**	**Beginners**
Native	1,000	**0,670**	**0,789**	**0,758**
Advanced		1,000	**0,972**	**0,957**
Intermediate			1,000	**0,991**
Beginners				1,000

It is evident from the results that both non-advanced groups of non-native listeners (means "Beginners" and "Intermediate") scored the samples systematically lower than "Advanced" listeners. It means that the first hypothesis mentioned before was confirmed. The offset is approximately 0.5 MOS along the entire MOS scale.

This systematic offset can be conveniently used to re-map objective algorithm output to bring the algorithm result closer to "conventionally correct" (meaning subjective) results in case the communication in the telecommunication network is carrying a non-native language for one or more conversation. Such correction can impact significantly e.g. threshold-based decisions on link quality acceptability in automatic measurements performed by network monitoring systems or drive-test systems.

4 Advances in objective testing methods

4.1 P.863 Aplication for low bit-rate and high packet-loss connections [28]

To provide the ability to measure voice transmission quality, objective methods like ITU-T P.863 [6] are widely deployed. Such methods are widely used to compare different coding and transmission technologies, or to monitor the network performance. All these methods have been designed and tested on telecommunication transmission standards that are widely used in common mobile and fixed telecommunication networks, e.g. those using rather good voice encoding. The application of objective digital signal processing based methods to any other area, such as special radio communication networks [30] that deploy low bit-rate speech coding and transcoding must be carefully verified by proper testing and result comparison with subjective assessment. This can widen algorithm potential applicability area beyond its scope, declared by its authors in [6].

To verify POLQA [6] for low bit-rate coded speech, the following tasks had to be performed:

-Selection of coders and conditions, sample database recording

-Subjective testing

-Objective testing

The following coders have been selected:

-PCM 8bit, 8kSa/s [18]

-GSM 6.10 (Full Rate) [10]

-MELPe 2.4 kbit/s [29]

Two background noise conditions (no noise / Hoth noise +10dB SNR) have been defined. For each coder and condition combination, 11 different sentences have been recorded, 5 male and 6 female voices, using 2 male and 2 female native speakers. The test language was French as one of the NATO commanding language. Including original (undistorted) samples, 88 samples have been generated in total, 25% of them being affected by low bit-rate MELPe coder.

Subjective testing complying to ITU-T P.800 [5] have been performed on the entire database. Each sample has been evaluated at least by 24 listeners as required by [5] using Opinion Score 1..5. Afterwards, the per-sample and per-condition MOS values have been calculated by averaging and confidence intervals CI95 have been calculated. Per condition MOS values are shown in Table 4.1. below.

Table 4.1. Subjective test results

Condition No.	**Technology**	**MOS-LQSn**
1	Original sample	4,62
2	PCM 8kSa/s	3,22
3	GSM 6.10	2,55
4	MELPe 2.4	2,92
5	Original + Hoth	3,55
6	PCM 8kSa/s + Hoth	3,06
7	GSM 6.10 + Hoth	1,72
8	MELPe 2.4 + Hoth	2,11

For comparison between subjective and objective results, the following algorithms have been selected for analysis [6]:

ITU-T P.862 – PESQ

ITU-T P.863 – POLQA narrowband

ITU-T P.863 – POLQA super-wideband

Similarly to subjective test analysis, the objective per-condition MOS values have been obtained by averaging always 11 results coming from each speech sample. The objective scores are listed in Table 4.2, see below. Also Pearson correlation coefficient showing the correlation between particular objective method results and subjective results is reported.

Table 4.2. Objective test results [28].

Condition No.	**PESQ**	**POLQA NB**	**POLQA SWB**
1	4,55	4,48	4,72
2	2,91	3,48	3,11
3	2,63	3,07	1,56
4	2,67	3,40	2,70
5	1,83	2,87	2,83
6	1,79	2,72	2,48
7	1,58	2,5	1,61
8	2,00	2,43	1,87
Correlation with subjective tests	**0,71**	**0,80**	**0,92**

The results are also shown in Fig. 4.1, combining all 3 objective algorithms and the subjective results into one graph.

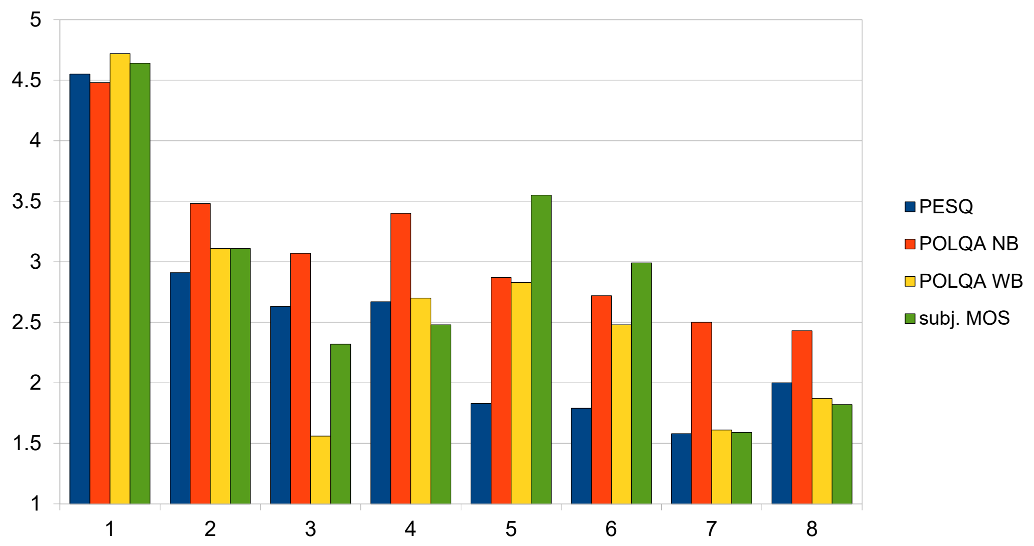

Fig.4.1 Results of PESQ, POLQA NB and POLQA algorithm and subjective test results comparison [28]

The comparison between objective and subjective results shows high correlation between subjective MOS and POLQA-SWB scores even for speech sample databases including significant portion of low bit-rate samples. Despite the fact the POLQA algorithm is presented by its authors as unsuitable for coding technologies below 4 kbit/s, it can be stated that POLQA-SWB can be used even for low bit-rate coded speech measurements, even though more samples and more languages should be tested to validate this statement.

4.2 Conversation QoE predictor ETSI TR 103 121 [21]

Today, the current E-model, defined in the ITU-T Recommendation G. 107 [14], is rarely used to support decisions before changes are implemented in a network due to its complex and obsolete content. Network planners asking how much impact deployment of a new technology will have on a QoE. Therefore an algorithmic estimator has been developed that implements the parameters effectively impacted by these new technologies. Instead of providing instructions for many parameters, most of which finally are left at their default values, it is better to hide these parameters inside the tool, and make only most important network parameters available, such as delay, talker echo, listening quality and interaction level. In order to compare the developed approach with the E-model, several graphs are provided as a result of the project, comparing subjective results, the new predictor outputs and the E-model values for a number of variable parameters. The newmodel was designed and trained based on polynomial fit of subjective test data. Only English data have been used for the training, but the model was validated for Czech data, too [8]. Its input variable values are:

-end-to-end delay

-the talker echo (TELR)

-Talker Alternation Rate (TAR)

-Coder used (affecting listening quality)

It should be noted the model is quite simple as the number of parameters is currently limited. Further subjective data would be needed to properly consider other important parameters, e.g., the effect of background noise or other possible impairments.

However, for the given set of subjective data it achieves significantly higher correspondence with conversational subjective data than the E-model in the context of different call types. It also considers the influence of call interactivity and distorted echo that is not considered by E-model at all.

The following analyses have been performed and are reported in Table 4.3:

-Pearson correlation coefficient *R* between MOS-CQS and E-model output. This analysis shows the differences between existing standardized estimator and subjective test results.
-*RMSE** against E-model (root mean squared error with suppressed influence of subjective testing uncertainty). This analysis shows the differences between the nearest CI95% interval border and the standardized E-model result (zero if the E-model output is located within the CI95% interval).
-Pearson correlation coefficient *R* between MOS-CQS and the developed predictor output. This analysis shows the difference between the developed predictor and subjective test results.
-*RMSE** against the developed predictor (root mean squared error with suppressed influence of subjective testing uncertainty). This analysis shows the differences between the nearest CI95% interval border and the developed predictor (zero if the predictor output is located within the CI95% interval).

Table 4.3 Result analysis overview [8]

	MOS-CQS versus E-model	MOS-CQS versus new model
R	0,546	0,911
RMSE	1,984	0,148
*RMSE**	1,722	0,029

The comparison of results of tests performed in Czech language and in English language clearly indicates insignificant systematic offset (0.2 MOS in average) causing Czech testers being virtually more demanding (more critical), however, the reason of this systematic offset is not clear. It can be caused e.g. by slightly lower average TAR for Czech tests (32,6) than for English tests (34,4) or by different age distribution or by other unknown reason.

As follows from Table 4.3, the new estimator outperforms the E-model for the given set of test conditions in *R, RMSE* and *RMSE** parameters. However, due to the following significant differences, the estimations provided by the E-model and the new model can not be directly compared:

The E-model is a complex model taking into account a lot of different parameters and due to its rather pessimistic results (in particular linked with high delay figures) it delivers safe predictions during network planning phase or to guarantee a high quality e.g. for business calls. However, its results are questionable to use during the operational phase and are impaired by a lack of fundamental inputs like interactivity (characterized by TAR). Also the amount of distortion in echo caused by multiple coding of the echo signal is not reflected (only TELR and echo delays are considered).

The new model, on the contrary, provides a good match with MOS-CQS (conversation quality) because one of its major innovations - taking into account the interactivity between the talkers and to introduce the new parameter TAR which is very important for the overall quality of speech conversations. The results provided by the developed model (see graphs presented in the ETSI TR 103 121 [21]) and a reference implementation give the opportunity to determine the expected quality of communications, taking into account delay, talker echo, listening quality and TAR. However, to be wider applicable, it should be extended towards other parameters such as noise (effects of noisy environments and of noise cancellation), and bandwidth (considering wideband and super wideband speech).

The developed model applies in particular for new IP-based networks where the end-to-end delay may be high and could be seen as a model dedicated to NGN and new mobile networks (e.g. UMTS and LTE). The model has been approved as ETSI TR 103 121 [21] and is available on http://www.etsi.org/standards

5. Future research challenges

New areas for QoS and QoE research in telecommunications are emerging due to technology development and user experience expectations development. The new signal coding principles must be followed by properly working algorithms for their transmission testing (e.g. EVS [12].). As the future speech coding techniques might be based rather on speech recognition including emotion detection etc. and consequent speech synthesis (such approach would greatly reduce required bandwidth), completely new subjective and objective testing procedures might be needed, covering much wider area of research (emotion detectors, synthetic speech quality testing, etc.). Small steps towards this direction are already being made

Another obvious trend in subjective testing is an attempt to bring the test situations and conditions closer to the real situations where the tested technology will be used in practice. Thus, subjective testing might be moved from laboratories and anechoic chambers to real environments and/or parallel mental or physical tasks might be introduced [32]. The new methodologies will thus incorporate knowledge from field of usability testing, psychology, etc. The challenging task, of course, will be maintaining or even improving the test repeatability and robustness compared to current laboratory test procedures.

6. References:

[1] ITU-T P.10/G.100, Vocabulary for performance and quality of service, Geneva 1998, am. 2006

[2] ETSI STF 354, Guidelines and Tutorials for Improving the User Experience of Real-time Communication Services, 2010

[3] ITU-T E.800: Terms and definitions related to quality of service and network performance including dependability, Geneva, 1994, am. 2008

[4] IETF RFC 2386, A Framework for QoS-based Routing in the Internet, 1998, am. 2013

[5] ITU-T Rec. P.800 "Methods for Subjective Determination of Transmission Quality", Series P: Telephone Transmission Quality, ITU, Geneva, 1996, am. 1998

[6] ITU-T Rec. P. 863 "Perceptual Objective Listening Quality Assessment (POLQA)", International Telecommunication Union, Geneva, 2011, am 2014.

[7] ITU-T Rec. P. 563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications", International Telecommunication Union, Geneva, 2004, am. 2007

[8] Holub, J. - Počta, J. - Monfort, J.-F. - Pomy, J.: Management Conversational Quality Predictor. In PQS 2013. Vienna: Telecommunications Research Center Vienna FTW, 2013, p. 147-150

[9] ITU-R BS.1116-3 (02/2015), Methods for the subjective assessment of small impairments in audio systems, Geneva, 1994, am. 2015

[10] ETSI ETS 300 961, 06-10, GSM codec FR, Sophia Antipolis, 1997

[11] 3GPP TS 26 071: Mandatory speech CODEC speech processing functions; AMR speech Codec; General description

[12] 3GPP TR 26.952, Codec for Enhanced Voice Services (EVS); Performance Characterization, 2015

[13] Voran, S.: Techniques for Comparing Objective and Subjective Speech Quality Tests. In Proc. Workshop "Speech Quality Assessment", Ruhr-Uni Bochum, 1994

[14] ITU-T Recommendation G.107: The E-model: a computational model for use in transmission planning, Geneva, 1998, am. 2015

[15] ITU-T Recommendation P.805: Subjective evaluation of conversational quality, Geneva, 2007

[16] Holub, J. - Tomíška, O.: Non-monotonicity in Perceived Quality of Delayed Talker Echo. In: Measurement of Speech, Audio and Video Quality in Networks. Prague: Czech Technical University, 2007, p. 67-68. ISBN 978-80-01-03734-8

[17] F. Hammer: "Quality Aspects of Packet-Based Interactive Speech Communication", Ph.D. Thesis. TU Graz 2006

[18] ITU-T Recommendation G.711: Pulse code modulation (PCM) of voice frequencies, Geneva 1988

[19] ITU-T Recommendation G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic code-excited linear-prediction (CS-ACELP), Geneva 1996 am. 2012

[20] ETSI ES 202 737: Speech and multimedia Transmission Quality (STQ); Transmission requirements for narrowband VoIP terminals (handset and headset) from a QoS perspective as perceived by the user, Sophia Antipolis 2010

[21] ETSI TR 103 121: Speech and multimedia Transmission Quality (STQ); Adaptation of the ETSI QoS Model to better consider results from field testing, 2013

[22] ITU-T Recommendation P.57: Artificial ears, Geneva 2009, am. 2012

[23] KITAWAKI, N., ITOH, K.: Pure Delay Effect on Speech Quality in Telecommunications. IEEE J. Sel. Areas Comm., 9(4): 586-593, 1991

[24] Holub, J., Smid, R., Bachtik, M.: Child Listeners as the Test Subject – Comparison with Adults and P.862, MESAQIN, Prague, CTU, 2003

[25] Holub, J., New Methods for Speech Transmission Quality Measurements, Habilitation Thesis, FEE CTU, Prague, 2004

[26] Holub, J., Souček, P.: Subjective Testing and Objective Modelling of Influence of Different Social Classes to Voice Call Quality Perception. Communications. 2012, vol. 14, no. 2, art. no. 3, p. 17-21. ISSN 1335-4205

[27] Blašková, L. - Holub, J.: How Do Non-native Listeners Perceive Quality of Transmitted Voice?. Communications. 2008, vol. 10, no. 4, p. 11-15. ISSN 1335-4205

[28] Holub, J.: Objective Quality Measurements of Low Bit-Rate Coded Speech. In MCC 2011 - Military Communications and Information Technology: A Comprehensive Approach Enabler. Warsaw: University of Warsaw, 2011, p. 441-445. ISBN 978-83-62954-20-9

[29] Street, M. and Collura, J.: "Interoperable voice communications: test and selection of STANAG 4591", RTO-IST conf. on 'Military communications', Warsaw, Poland, 2001

[30] Holub, J., Street, M., Šmíd, R.: Intrusive Speech Transmission Quality Measurements for Low Bit Rate Coded Audio Signals, AES115 Convention, New York, October 2003

[31] ETSI STF 504, DTS/STQ-236 TS 103 296 (draft), Detection of Emotions in Telecommunication Measurement Applications, Sophia Antipolis, 2016

[32] Holub, J., Avetisyan, H.: Low Bit-rate Coded Speech Intelligibility – Comparison of Laboratory Test Results and Results of Test with Parallel Task, submitted to PQS 2016, Berlin, August 2016

7 Assoc. Prof. Ing. Jan Holub, Ph.D.

EDUCATION

M.S. (measurement technology)	Czech Technical University	1996
Ph.D. (measurement technology)	Czech Technical University	1999

APPOINTMENTS

4/14-present	Chair, Department of Measurement, Faculty of Electrical Engineering, Czech Technical University in Prague
6/04-present	Associate Professor, Department of Measurement, Faculty of Electrical Engineering, Czech Technical University in Prague
6/99-6/04	Assistant Professor, Department of Measurement, Faculty of Electrical Engineering, Czech Technical University in Prague
5/96-8/12	Telecom Business Consultant at T&M Direct - TEKTRONIX

ACTIVITIES

- Curriculum Development at the Faculty of Electrical Engineering, Czech Technical University in Prague (Distributed Systems and Computer Networks (2010-), Circuits of Digital Instruments (2004-), Design of Medical Devices (2005-))
- IET (2004) and ISCA (2002) member
- involved in standardisation activities of ETSI and ITU-T via MESAQIN.com Ltd. (owner)
- chair of organizing and member of program committees of MESAQIN 2002-2012, chair of Dithering in ISDDMI 98 (10th International Symposium on Development in Digital Measuring Instrumentation, Naples, Faculty of Engineering), awarded by Antoio Menchetti prize for the best contribution of younger scientist there
- member of Program Committee of WTS 2006-2016 and Program Committee Chair of WTS 2009
- IMEKO member (2004). Czech delegate in IMEKO TC-1 (Education and Training in Measurement) 2008, chair of IMEKO TC-1 (2010)

GRADUATE STUDENTS TOTAL LAST 5 YEARS (ING-6, PHD-2)

Fexa, Pavel (Ph.D. 2015) Non-traditional ADC and DAC testing with poly-harmonic signals, Svatoš Jakub (Ph.D. 2015) Advanced Instrumentation for Polyharmonic Metal Detectors, Halcin Jakub (Ing. 2014) Sentence Design with Minimum Emotional Uncertainty, Nguyen Huu Chanh (Ing. 2014) Modern Methods of Layer Sound Transmission Analysis, Večeř Michael Pascal (Ing. 2013) Trend Analysis of Speech Quality Perception for Czech Listeners between 2008 and 2012, Blašková Ľubica (Ing. 2012) Economical efficiency of automatic quality monitoring system implementation in telecommunications, Toula Michal (Ing. 2011) Influence of Different Social Classes to Conference Call Quality Perception - Subjective Testing and Objective Modelling, Novák Jiří (Ing. 2011) Influence of Listener Emotional or Psychological Disorder to Conference Call Quality Perception - Subjective Testing and Objective Modelling, Šplíchal Jiří (Ing. 2011) Non-intrusive Sound Distribution Quality Measurement Algorithm

SELECTED PUBLICATIONS (last 10 years)

Journal publications

- Holub, J. - Slavata, O. - Souček, P. - Zisimopoulos, O. - Toumpakaris, D. - et al.: Towards Layer Adaptation for Audio Transmission. International Journal of Interdisciplinary Telecommunications and Networking. 2014, vol. 6, no. 4, p. 35-41. ISSN 1941-8663.

- Počta, P. - Holub, J.: Effect of speech activity parameter on PESQ's predictions in presence of independent and dependent loses. Computer Standards & Interfaces. 2013, vol. 1, no. 36, p. 143-153. ISSN 0920-5489. IF=0,879
- Slavata, O. - Holub, J.: Evaluation of Objective Speech Transmission Quality Measurements in Packet-based Networks. Computer Standards & Interfaces. 2013, vol. 36, no. 3, art. no. 20, p. 626-630. ISSN 0920-5489. IF=0,879
- Holub, J. - Souček, P.: Shift in Speech Quality and Acceptability Level between 2008-2012. International Journal of Interdisciplinary Telecommunications and Networking. 2013, vol. 5, no. 3, art. no. 7, p. 63-68. ISSN 1941-8663.
- Holub, J. - Souček, P.: Subjective Testing and Objective Modelling of Influence of Different Social Classes to Voice Call Quality Perception. Communications. 2012, vol. 14, no. 2, art. no. 3, p. 17-21. ISSN 1335-4205.
- Počta, P. - Holub, J.: Predicting the Quality of Synthesized and Natural Speech Impaired by Packet Loss and Coding Using PESQ and P.563 Models. ACTA ACUSTICA UNITED WITH ACUSTICA . 2011, vol. 97, no. 5, p. 852-868. ISSN 1610-1928. IF=0,783
- Počta, P. - Holub, J. - Vlčková, H. - Polková, Z.: Impact of Different Active-Speech-Ratios on PESQ`s Predictions in Case of Independent and Dependent Losses. Radioengineering. 2010, vol. 19, no. 1, p. 79-88. ISSN 1210-2512. IF=0,653
- Tučková, J. - Holub, J. - Duběda, T.: Technical and Phonetic Aspects of Speech Quality Assessment: The Case of Prosody Synthesis. Lecture Notes in Artificial Intelligence. 2009, vol. 5641, no. 2009931057, p. 126-132. ISSN 0302-9743.
- Počta, P. - Holub, J. - Mrvová, M.: Impact of Different Active-Speech-Ratios on PESQ's Predictions in Simulated VoIP Environment. ACTA ACUSTICA UNITED WITH ACUSTICA . 2009, vol. 95, no. 5, p. 950-958. ISSN 1610-1928. IF=0,783
- Blašková, L. - Holub, J.: How Do Non-native Listeners Perceive Quality of Transmitted Voice? Communications. 2008, vol. 10, no. 4, p. 11-15. ISSN 1335-4205.

Other Publications & Patents

- **Holub, J.: Portable device for measuring the acoustic reflectivity coefficient in situ. Patent Úřad průmyslového vlastnictví, (in Czech), 305173, 2015-04-15.**
- Slavata, O. - Holub, J.: Impact of the Codec and Various QoS Methods on the Final Quality of the Transferred Voice in an IP Network. In 2014 joint IMEKO TC1-TC7-TC13 Symposium: Measurement Science Behind Safety and Security. Madeira: IOPscience, 2015, art. no. 012011, ISSN 1742-6588.
- Souček, P. - Slavata, O. - Holub, J.: New Approach in Subjective and Objective Speech Transmission Quality Measurement in TCP/IP Networks. In 2014 joint IMEKOTC1-TC7-TC13 Symposium: Measurement Science Behind Safety and Security. Madeira: IOPscience, 2015, art. no. 012020, p. 1-4. ISSN 1742-6588.
- Drábek, T. - Holub, J.: The Innovation of the Autonomous System for Indoor Illuminance. In Proceedings of the 21st International Conference LIGHT SVĚTLO 2015. Brno: Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Electrical Power Engineering, 2015, p. 273-276. ISBN 978-80-214-5244-2.
- Holub, J. (ed.): XXI IMEKO WORLD CONGRESS - Full Papers. Prague: Czech Technical University in Prague, Faculty of Electrical Engineering, 2015. 2279 s. ISBN 978-80-01-05793-3.
- Holub, J. - Desbos, B. - Vacek, V. - Kolísko, J.: Determination of Material Acoustic Features Using Small Samples. In Local Mechanical Properties X. ZURICH: TRANS TECH PUBLICATIONS LTD, 2014, p. 111-114. ISSN 1013-9826. ISBN 978-3-03835-062-0.
- Vacek, V. - Holub, J.: Options of Assessment of Absorption Capacity of Noise Barrier. In Advanced Materials Research. Uetikon-Zurich: Trans Tech Publications Inc., 2014, p. 125-129. ISSN 1022-6680. ISBN 978-3-03835-083-5.

- Holub, J. - Slavata, O. - Souček, P. - Zisimopoulos, O. - Toumpakaris, D. - et al.: Towards Layer Adaptation for Audio Transmission. In 2014 Wireless Telecommunications Symposium (WTS). Piscataway: IEEE, 2014, p. 1-4. ISBN 978-1-4799-1297-1.
- Desbos, B. - Holub, J. - Vacek, V. - Kolísko, J.: Determination of Material Acoustic Features Using Small Samples. In Local Mechanical Properties 2013 - 10th International Conference - Book of Abstracts. Praha: CTU Publishing House, 2013, p. 7. ISBN 978-80-01-05374-4.
- Bruna, O. - Holub, J. - Pačes, P.: Experimental Stress Assessment in Biomedical Measurement Class. In IDAACS 2013 - Proceedings of the 2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems. Berlin: IEEE, 2013, p. 99-102. ISBN 978-1-4799-1426-5.
- Bruna, O. - Souček, P. - Holub, J.: Incorporating Human Stress Measurements into Biomedical Engineering Class. In 2013 Joint IMEKO (International Measurement Confederation) TC1-TC7-TC13 Symposium: Measurement Across Physical and Behavioural Sciences. Bristol: IOP Publishing Ltd, 2013, art. no. 012011, p. 1-6. ISSN 1742-6588.
- Souček, P. - Slavata, O. - Holub, J.: Innovation of Laboratory Exercises in Course "Distributed Systems and Computer Networks". In 2013 Joint IMEKO (International Measurement Confederation) TC1-TC7-TC13 Symposium: Measurement Across Physical and Behavioural Sciences. Bristol: IOP Publishing Ltd, 2013, p. 1-4. ISSN 1742-6588.
- Holub, J. - Počta, J. - Monfort, J.-F. - Pomy, J.: Management Conversational Quality Predictor. In PQS 2013. Vienna: Telecommunications Research Center Vienna FTW, 2013, p. 147-150.
- Vacek, V. - Holub, J.: THE POSSIBILITY OF EVALUATION OF ACOUSTIC ABSORPTION PANELS. In Zkoušení a jakost ve stavebnictví 2013. Brno: Vysoké učení technické v Brně, 2013, s. 135-140. ISBN 978-80-214-4777-6. (in Czech).
- Slavata, O. - Souček, P. - Holub, J.: New Concept of Laboratory Exercise on Temperature Measurements Using Thermocouple. In 2013 Joint IMEKO (International Measurement Confederation) TC1-TC7-TC13 Symposium: Measurement Across Physical and Behavioural Sciences. Bristol: IOP Publishing Ltd, 2013, art. no. 012059, p. 1-6. ISSN 1742-6588.
- Holub, J. - Souček, P.: Shift in Speech Quality and Acceptability Level between 2008-2012. In Wireless Telecommunications Symposium 2013 - Papers and Presentations. Piscataway: IEEE, 2013.
- **Holub, J.: System for Tariffication Control in Telecommunication Networks Based on Quality of Transmitted Call. Patent European Patent Office, EP2541883. 2013-08-16.**
- Holub, J.: System for billing control based on transmitted speech quality. Patent Úřad průmyslového vlastnictví, 303711. 2013-02-13. (in Czech).
- Bruna, O. - Holub, J. - Pačes, P. - Levora, T.: Small Aircraft Emergency Landing Decision Support System - Pilots' Performance Assessment. In XX IMEKO World Congress 2012 - Proceedings. Busan: IMEKO, 2012, art. no. TC18-O-1, p. 397-402. ISBN 978-1-62748-190-8.
- Slavata, O. - Holub, J. - Hübner, P.: Impact of Jitter and Jitter Buffer on the Final Quality of the Transferred Voice. In IDAACS-SWS'2012. Piscataway: IEEE, 2012, p. 120-123. ISBN 978-1-4673-4677-1.
- Souček, P. - Holub, J.: Evaluation of ITU-T P.863 POLQA in Chinese. In IDAACS-SWS\'2012. Piscataway: IEEE, 2012, p. 124-126. ISBN 978-1-4673-4677-1.
- Holub, J. - Slavata, O.: Impact of IP Channel Parameters on the Final Quality of the Transferred Voice. In Wireless Telecommunications Symposium 2012 Papers and Presentation. Pomona (CA): California State Polytechnic University, 2012, p. 1-5. ISBN 978-1-4577-0580-9.
- **Holub, J.: User-Centric Service Model in Wireless Networks: The Transition from Technical Excellence to Customer Experience Excellence in Wireless Networks. In Wireless Telecommunications Symposium 2012 Papers and Presentation. Pomona (CA): California State Polytechnic University, 2012, art. no. A, p. 1-37. ISBN 978-1-4577-0580-9 (invited lecture)**

- Holub, J. - Souček, P.: Subjective Testing and Objective Modelling of Influence of Different Social Classes to Voice Call Quality Perception. Communications. 2012, vol. 14, no. 2, art. no. 3, p. 17-21. ISSN 1335-4205.
- Bruna, O. - Levora, T. - Holub, J.: Evaluation of a Flight-guidance System for Ultra-Light Aircraft. In HCI-Aero 2012 Proceedings - "Transition of Science into Reality". Brussels: Honeywell (USA), Eurocontrols (France), 2012, art. no. PDMT4, p. 1-4.
- Holub, J.: Transmitted Voice Quality Controlled Billing System for Telecommunication Networks. Užitný vzor Úřad průmyslového vlastnictví, 23142. 2012-01-02. (in Czech).
- Holub, J.: Objective Quality Measurements of Low Bit-Rate Coded Speech. In MCC 2011 - Military Communications and Information Technology: A Comprehensive Approach Enabler. Warsaw: University of Warsaw, 2011, p. 441-445. ISBN 978-83-62954-20-9.
- Souček, P. - Holub, J.: How the Perceived Speech Quality and Acceptability Level Shift during Time. In Measurement of Speech, Audio and Video Quality in Networks 2011. Praha: CTU Publishing House, 2011, p. 31-33. ISBN 978-80-01-04848-1.
- Slavata, O. - Holub, J.: Subjective Measurement and Objective Modeling of Voice Quality in a Conference Call for a Stressed Listener. In Measurement of Speech, Audio and Video Quality in Networks 2011. Praha: CTU Publishing House, 2011, p. 14-19. ISBN 978-80-01-04848-1.
- Holub, J.: Validation of ITU-T P.863 for Low Bit-rate Coded Speech Quality Measurements. In WTS 2011 - Wireless Telecommunications Symposium 2011. New York: IEEE, 2011, p. 1-2. ISSN 1934-5070. ISBN 978-1-4577-0162-7.
- Holub, J. - Šmíd, R. (ed.): Measurement of Speech, Audio and Video Quality in Networks 2011. Praha: CTU Publishing House, 2011. 33 p. ISBN 978-80-01-04848-1.
- Holub, J. - Počta, P.: Transmitted Speech Quality versus Perceptual Annoyance and Service Acceptibility Thresholds. In Third International Workshop on Perceptual Quality of Systems. Bochum: ISCA - International Speech Communication Association, 2010, p. 100-102. ISSN 1680-8908.
- Holub, J. - Slavata, O.: Impact of IP Channel Parameters on the Final Quality of the Transferred Voice. In MESAQIN 2010. Prague: CTU, 2010, p. 18-22. ISBN 978-80-01-04569-5.
- Holub, J. - Souček, P.: Intelligibility in Case Non-native Listeneres and its Dependence on their Language Proficiency. In MESAQIN 2010. Prague: CTU, 2010, p. 23-25. ISBN 978-80-01-04569-5.
- Holub, J.: Voice Transmission Quality Measurement Algorithm with Non-native Listener Emulation Capability. In Wireless Telecommunications Symposium (WTS), 2010. Pomona, California: IEEE Communications Society, 2010, ISSN 1934-5070. ISBN 978-1-4244-6558-3.
- Holub, J. - Šmíd, R. (ed.): MESAQIN 2010. Prague: CTU, 2010. 58 s. ISBN 978-80-01-04569-5.
- Holub, J.: Objektívne a subjektívne hodnotenie kvality multimediálnych signálov. [Nepublikovaná přednáška]. Žilinská univerzita v Žiline. 2010-09-14.
- Blašková, L. - Holub, J. - Novák, J.: Speech Quality Aspects and Issues in NNEC. In MCC 2009 Military Communications and Information Systems Conference. Brno: Univerzita obrany, 2009, ISBN 978-80-7231-678-6.
- Holub, J. - Dušek, J.: Dependency of Subjective Echo Perception on Echo Quality. In MESAQIN 2009 Measurement of Speech, Audio and Video Quality in Networks. Praha: CTU Publishing House, 2009, ISBN 978-80-01-04361-5.
- Holub, J. - Mička, J.: End-to-End Network Simulator for Conversational Quality Measurements. In WTS 2009 Wireless Telecommunications Symposium 2009 Papers and Presentations. Praha: CTU Publishing House, 2009, ISSN 1934-5070. ISBN 978-1-4244-2588-4.
- Holub, J. - Šmíd, R. (ed.): Measurement of Speech, Audio and Video Quality in Networks. Prague: CTU, 2009. 50 p. ISBN 978-80-01-04361-5.

•Holub, J. - Tomíška, O.: Objective and Subjective Degradations of Transcoded Voice - Heterogeneous Radio Networks Interoperability. In DAGA 2008. Dresden: DAGA, 2008, p. 245-246. ISBN 978-3-9808659-4-4.
•Holub, J. - Blašková, L.: Transcoded Speech Contemporary Objective Quality Measurements Reliability. In WTS 2008 Wireless Telecommunications Symposium. Pomona, California: IEEE Communications Society, 2008, ISSN 1934-5070. ISBN 978-1-4244-1869-5.
•Holub, J. - Blašková, L.: How do Non-native Listeners Perceive Quality of Transmitted Voice?. In Measurement of Speech, Audio and Video Quality in Networks 2008. Praha: CTU Publishing House, 2008, p. 22-26. ISBN 978-80-01-04193-2.
•Holub, J. - Šmíd, R. (ed.): Measurement of Speech, Audio and Video Quality in Networks 2008. Praha: CTU Publishing House, 2008. ISBN 978-80-01-04193-2.
•Holub, J. - Kastner, M. - Tomíška, O.: Delay Effect on Conversational Quality in Telecommunication Networks: Do We Mind?. In Wireless Telecommunications Symposium 2007. Pomona, California: IEEE Communications Society, 2007,
•Holub, J.- Liu, J.J.: Intrusive Speech Transmission Quality Measurement in Chinese Environment. In Sixth International Conference on Information, Communications and Signal Processing ICICS 2007. Singapore: Nanyang Technological University, 2007, ISBN 1-4244-0983-7.
•Holub, J. - Šmíd, R. (ed.): Measurement of Speech, Audio and Video Quality in Networks. Prague: Czech Technical University, 2007. 70 p. ISBN 978-80-01-03734-8.
•Holub, J. - Tomíška, O.: Non-monotonicity in Perceived Quality of Delayed Talker Echo. In Measurement of Speech, Audio and Video Quality in Networks. Prague: Czech Technical University, 2007, p. 67-68. ISBN 978-80-01-03734-8.
•Holub, J. - Street, M. - Tomíška, O.: A Novel Non-Intrusive Voice Transmission Quality Measurement Algorithm. In WTS 2006 - Wireless Telecommunications Symposium 2006. Piscataway: IEEE, 2006, ISBN 1-4244-0046-5.
•Holub, J. - Street, M. - Tomíška, O.: Computationally Efficient Non-Intrusive Algorithm for Speech Transmission Quality Measurement. In IMEKO - XVIII World Congress and IV Brazilian Congress of Metrology. Rio de Janeiro: IMEKO, 2006, p. 162.
•Tomíška, O. - Holub, J.: Conversation Quality Assessment. In Nové smery v spracovaní signálov VIII. Medzinárodná vedecká konferencia. Tatranské Zruby: Akadémia ozbrojených síl gen. M. R. Štefánika, 2006, p. 305-309. ISBN 80-8040-294-9.
•Holub, J. - Šmíd, R. (ed.): Measurement of Speech and Audio Quality in Networks. Prague: Czech Technical University, 2006. ISBN 80-01-03262-0.
•Holub, J. - Tomíška, O.: Non-Intrusive Voice Transmission Quality Objective Estimator for Devices with Limited Computational Power. In Digital Signal Processing and its Applications. Moskva: IEEE, 2006, vol. VIII, p. 623-626.
•Holub, J. - Motlová, J. - Šmíd, R. - Tomíška, O.: Objective Assessment of Speech Distribution Quality by means of Unconventional PESQ Application - A Case Study. In Second ISCA/DEGA - Tutorial and Research Workshop on Perceptual Quality of System. Berlin: Niscom, TU Dresden et al., 2006, p. 160-163. ISSN 1680-8908.
•Holub, J. - Drozdová, A.: Proprietary Low Bit-rate Radio-communication Network - Objective and Subjective Speech Transmission Quality Assessment. In Measurement of Speech and Audio Quality in Networks. Prague: Czech Technical University, 2006, p. 63-66. ISBN 80-01-03262-0.