ETSI TR 103 121 V1.1.1 (2013-03)

Speech and multimedia Transmission Quality (STQ);
Adaptation of the ETSI QoS Model
to better consider results from field testing

Reference
DTR/STQ-189

Keywords
delay, E-Model, QoS, quality

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Important notice

Individual copies of the present document can be downloaded from:
http://www.etsi.org

The present document may be made available in more than one electronic version or in print. In any case of existing or perceived difference in contents between such versions, the reference version is the Portable Document Format (PDF). In case of dispute, the reference shall be the printing on ETSI printers of the PDF version kept on a specific network drive within ETSI Secretariat.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at
http://portal.etsi.org/tb/status/status.asp

If you find errors in the present document, please send your comment to one of the following services:
http://portal.etsi.org/chaircor/ETSI_support.asp

Contents

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (http://ipr.etsi.org).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Speech and multimedia Transmission Quality (STQ).

Introduction

ETSI has developed a Transmission Planning Model for predicting QoS - also known as the E-Model; this model is originally described in ETR 250 [i.11] - which has been further developed and has gained global recognition.

TR 102 356 [i.12] summarizes global activities on improving the E-model.

In addition, popular field testing in modern technologies, such as UMTS, NGN and in future LTE typically reveals only one quality component of the QoS. Therefore, it is highly desirable for ETSI to develop an adapted version of the E-model which - on a reliable and on a proofed basis - can combine results from field trials with other impairments, such as one-way delay, etc.

The present document investigates to which extent parameters, other than one-way delay, were considered in this context. The verification of this approach by subjective tests of conversational QoS was carried out.

1 Scope

The present document addresses a new approach to assess or anticipate the conversational quality of end-to-end transmissions. It is based on the adaptation of the ETSI QoS Model (hereafter referred to as E-Model) in order to better consider results from field testing.

The present document defines the principles of this new approach, the test conditions including test equipment test set-up, the conversational subjective test plan and the results of the tests conducted for this new approach.

The model takes into account the variable parameters such as end-to-end delay, talker echo, degree of interactivity between the subjects (expressed as Talker Alternation Rate) and listening quality.

Comparisons between the new model and other approaches such as E-Model are also made available.

2 References

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

Referenced documents which are not found to be publicly available in the expected location might be found at http://docbox.etsi.org/Reference.

2.1 Normative references

Not applicable.

2.2 Informative references

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] Recommendation ITU-T G.711: "Pulse code modulation (PCM) of voice frequencies".

[i.2] Recommendation ITU-T G.729: "Coding of speech at 8 kbit/s using conjugate-structure algebraic code-excited linear-prediction (CS-ACELP)".

[i.3] ETSI TS 126 071: "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Mandatory speech CODEC speech processing functions; AMR speech Codec; General description (3GPP TS 26.071)".

[i.4] Recommendation ITU-T P.800: "Methods for Subjective Determination of Transmission Quality".

[i.5] Recommendation ITU-T P.805: "Subjective evaluation of conversational quality".

[i.6] ETSI SR 002 959: "Electronic Working Tools; Roadmap including recommendations for the deployment and usage of electronic working tools in the ETSI standardization process".

[i.7] 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".

[i.8] Recommendation ITU-T G.107: "The E-model: a computational model for use in transmission planning".

[i.9] Recommendation ITU-T 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".

[i.10] ETSI ES 202 396-1: "Speech and multimedia Transmission Quality (STQ);Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database".

[i.11] ETSI ETR 250: "Transmission and Multiplexing (TM); Speech communication quality from mouth to ear for 3,1 kHz handset telephony across networks".

[i.12] ETSI TR 102 356: "Speech Processing, Transmission and Quality Aspects (STQ); Application and enhancements of the E-Model (ETR 250); Overview of available documentation and ongoing work".

[i.13] 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.

[i.14] F. Hammer: "Quality Aspects of Packet-Based Interactive Speech Communication", Ph.D. Thesis. TU Graz 2006.

[i.15] F. Hammer, P. Reichl, A. Raake: "The Well-Tempered Conversation. Interactivity, Delay and Perceptual VoIP Quality", in Proceedings of IEEE ICC 2005, Seoul (South Korea), May 2005.

[i.16] Recommendation ITU-T P.59: "Artificial conversational speech".

[i.17] Recommendation ITU-T P.57: "Artificial ears".

[i.18] ETSI TR 126 935: "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Packet Switched (PS) conversational multimedia applications; Performance characterization of default codecs (3GPP TR 26.935)".

[i.19] Recommendation ITU-T G.113: "Transmission impairments due to speech processing".

[i.20] Recommendation ITU-T P.56: "Objective measurement of active speech level".

[i.21] Recommendation ITU-T COM 12-35-E (1997): "Development of scenarios for short a conversation test".

[i.22] Handbook on Telephonometry (1992): "Measurement methods: telephonometry".

[i.23] RICHARDS (D.L.): "The transmission performance of telephone networks", The Butterworth Group, pp. 199-203, London 1973.

[i.24] HAMMER (F.): "Quality Aspects of Packet-Based Interactive Speech Communication", PhD Thesis, University of Technology at Graz 2006.

[i.25] KITAWAKI (N.) and ITOH (K.): "Pure Delay Effects on Speech Quality in Telecommunications", IEEE Journal on Selected Areas in Communications, vol. 9 (4).

[i.26] RAAKE (A.): "Speech Quality of VoIP: Assessment and prediction", John Wiley and Sons Ltd., Chichester 2006.

[i.27] Recommendation ITU-T P.834: "Methodology for the derivation of equipment impairment factors from instrumental models".

3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

AMR-NB	Adaptative Multi-rate Narrowband
MCQP	Management Conversational Quality Predictor
MOS	Mean Opinion Score
MOS-CQE	Mean Opinion Score – Communication Quality Estimated
MOS-CQS	Mean Opinion Score – Communication Quality Subjective
MOS-LQO	Mean Opinion Score – Listening-only Quality Objective
PESQ	Perceptual Evaluation of Speech Quality

QoS	Quality of Service
RLR	Receive Loudness Rating
RMSE	Root Mean Square Error
SLR	Send Loudness Rating
TAR	Talker-Alternation Rate
TELR	Talker echo loudness rating

4 Development and review of the approach

The modelling is to be done in the subjective MOS domain and only the final result is converted into the E-Model domain as R-Value.

A user interface gives the choices of the MOS-LQO value, one-way delay and the additional parameters as outlined in clause 4.1.

With a similar user interface calculations can be made using the same parameter, but purely based on the E-Model and related documents.

Finally, graphs were derived to show the differences between both approaches.

Verification of this approach was done by subjective tests of conversational QoS.

The subjective conversation tests are covering the following characteristics:

- different coders
  - 3 coders, G.711 [i.1] A-law, G.729AB [i.2] (@ 8kbit/s), AMR-NB [i.3] (@ 12,2kbit/s)
- different delay values
  - 3 values, 100, 300, 600 ms one-way delay
- different echo situations
  - 2 situations, weak echo, strong echo, TELR= 46 dB, 32 dB
- different conversational scenarios
  - 3 levels of temperature i.e. different categories

The exact test scenarios can be found in annexes B and C:

- minimum number of 50 conditions
  - equal to 54 conditions in English, 18 conditions in Czech, total 72 conditions
- minimum of 40 votes per condition

48 votes, equals to 3 456 votes:

- the equivalent of a reference terminal
  - real-time adaptation to ES 202 737 [i.7] with diffuse field correction as per Recommendation ITU-T P.57 [i.17] in send and receive direction.
- different languages

The majority of tests are conducted in English language. There is a number of tests in Czech language, although limited so that a third coder can be used.

When possible, the E-Model default settings are used. However, for some parameters (e.g. Noise) the actual parameter values are used as default settings (as long as they do not change the E-Model results when using the actual or theoretical default setting values).

The four signals (send and receive for the two electrical ends) can be recorded in order to assess *a posteriori* the "temperature" of the conversation (TAR, as defined in the thesis of Florian Hammer [i.14] and in articles, e.g. The Well-Tempered Conversation [i.15]). It was made sure that the four recordings take into account the delay in the simulators. The method to determine this factor is also reported in the present document.

In order to test listening quality with Recommendation ITU-T P.862 (PESQ) [i.9] on electrical ends, the recordings from signals were kept from the "four ends" with the speech sequences of PESQ and also with one pair of subjects for all the scenarios. Results are available in annex D.

The recordings for the "TAR calculation" were done in the middle of the networks (two ways).

As the recordings are done for each way after the send part of the chain, the signals were time-shifted according to the delay necessary to compute the TAR value.

For Echo attenuation, a mask as defined in ES 202 737 [i.7], clause 7.2.2.2 was used, also when the requirement addresses the case in which echo cancellation is used.

Table 1: Echo attenuation limits

Frequency	Limit
100 Hz	-20 dB
200 Hz	-30 dB
300 Hz	-38 dB
800 Hz	-34 dB
1 500 Hz	-33 dB
2 600 Hz	-24 dB
4 000 Hz	-24 dB
NOTE 1: All sensitivity values are expressed in dB on an arbitrary scale. NOTE 2: The limit at intermediate frequencies lies on a straight line drawn between the given values on a log (frequency) - linear (dB) scale.	

During the measurement it should be ensured that the measured signal is the echo signal and not the Comfort Noise which potentially may be inserted in send direction in order to mask the echo signal.

An informal conversation needs to be done as trial for the first conversation for each pair of subjects in order to ensure that the instructions were well understood.

The conversation scenarios from Recommendation ITU-T P.805 [i.5] were used:

- Appendix V (18 potential scenarios), (one third of the tests in English language used such scenarios).
- A modified Appendix VII: the names of the figures were replaced by numbers and the table of figures was split in two or three parts, in order to reduce the potential time to reach the solution and to reduce the brain load. (one third of the English test conditions).
- Appendix VIII, to provide very high interactivity(one third of the English test conditions).

Czech test conditions are a sample selection of these conditions.

Each type of scenarios corresponds to one of the three interactivity categories, appendix VIII provides the "highest temperature", while the lower and medium are considered in appendices V and VII. This is determined by TAR computation.

Questions from SR 002 959 [i.6], Recommendation ITU-T P.805 [i.5] and additional proposals were considered. Only two questions were kept:

"How do you assess the conversation interactivity with the other person"				
No special effort required	Minimal effort required	Moderate effort required	Considerable effort required	Severe effort required
"What is your opinion of the connection you have just been using?"				
Excellent quality	Good quality	Fair quality	Poor quality	Bad quality

4.1 Market requirements and testability aspects of approach

The approach is to provide a "Management Conversational Quality Predictor (MCQP)".

The modelling is to be done in the subjective MOS domain and only the final result are converted into the E-Model domain as R-value.

A conversational quality predictor tool for technical management level is needed because the one way quality is not the quality really experienced by the users.

The principles retained for such a tools are:

Principle 1: to provide a decision support tool for the management level.

Principle 2: to hide parameters which are not needed by transmission planners or not accessible/monitored and which may create confusions for technical managers instead of helping them.

Many parameters are either not known to the technical decision makers, or they could have a wide range of values, e.g. the real terminal quality, the user's speech level, the local and distant noise levels.

The current E-model is rarely used to support decisions before changes are implemented in a network. Management needs to know how much impact deployment of a new technology will have on user perceived quality. So, the tool will implement 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 more appropriate 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.

As a consequence, several graphs will be provided as results of this project, comparing subjective results, the new predictor outputs, the E-model values for a number of variable parameters. If the technical managers are currently using E-model, they will be able to use these graphs to move to the new predictor without losing the historical evolution of the networks.

Finally, graphs will be derived to show the differences between the E-model and the new approach.

4.2 Development and Review of a test plan for a subjective conversational test

4.2.1 Requirements

As described in Recommendation ITU-T P.805 [i.5] in more detail subjective conversational tests allow the subjects involved to be in a more realistic situation simulating the actual service conditions experienced by telephone customers. In addition, subjective conversational tests are designed to assess the effects of impairments that can cause difficulty during conversation (such as delay, packet loss, echo, interruptions, noise, clipping, etc.). They can be used to study overall system effects or specific degradations, such as delay.

Subjects participate in the test as paired sets of communicators. They are seated in separate sound-proof rooms and asked to hold a conversation through the transmission chain (i.e. network simulator plus telephone sets) and then to give their opinion of the quality on a pre-selected quality scale. In the present tests acoustic noise environment were not simulated in both rooms.

Depending on the purpose of the test, expert, experienced or untrained (naive) subjects may participate. Such tests can be useful to manufacturers, operators and customers, and are an important assessment tool because they provide the closest simulation of real telephony interactions between subscribers. Untrained subjects are involved when it is important to get an indication of how the general telephone-using population would rate the overall quality and difficulty in using the connection with the system under test. This can be used to give a global evaluation of the performance in a range of conditions. However, untrained subjects are unable to describe and identify accurately the types of degradation associated with the system under test.

The main characteristics of a conversation-opinion test are:

- To be very close to a real conversation where people are required to interact and may adapt their behaviour to accommodate the system under test.
- The use of a task to stimulate a conversation with equal participation of both parties.
- Different subjects may have variable behaviour in a conversation (due to culture, personality, etc.), which could create greater variability in subjects' responses in the assessment of speech quality.
- Since subjects have to concentrate on participating in the conversation, and are not specifically involved in assessing the quality performance during the conversation, their final measures may be less sensitive than in listening-only tests.
- Conversation tests are the most valid method for measuring the effect on acceptability of certain system impairments, such as delay.
- Devices under test and simulation tools needs to be available at the testing lab and need to run in real time.
- This conversation test methodology can be adapted to field testing; however, it is foreseen that the control of some experimental variables (e.g. delay, packet loss, acoustic noise, etc.) would be limited.

4.2.1.1 Requirements regarding test facilities

A conversational test has to provide as realistic a communication environment as possible. All processes in the communication link are required to be real time.

Switching between conditions that involve different coders and/or different networks parameters has to be transparent to the subjects. This may require specialized instrumentation and procedures.

Asymmetry between two subjects in a communication is typical of many actual speech communication scenarios; an asymmetric scenario may be defined by different acoustic noise environments or different transmission conditions. Special consideration may be needed to ensure accurate simulation of acoustic noise environments.

Each subject sits in a separate sound-proof room, as defined in Recommendation ITU-T P.800 [i.4] where a variety of acoustic noise environments can be simulated. The environment in both rooms can be the same or different. Examples of different environments are quiet room, office, car, railway station, train and cafeteria. A quiet room might be simulated by the introduction of a suitable level of Hoth noise to fix the recommended floor noise. Certain chambers also allow reverberation to be considered as an experimental variable.

In addition, the send and receive sensors used by the subjects may be the same or different. For example, handset, headset with microphone or microphone and loudspeaker may be used; the choice of the equipment depends on the use case.

4.2.1.2 Requirements regarding test design

Most of the test design issues relevant to listening-only tests are also relevant to conversation tests, for example, reference conditions and presentation order effects. A major limitation to conversational test design is the duration of each individual task, or trial, required to exercise each experimental condition. Properly exercising a communication system requires conversations lasting a minimum of 2 minutes. Typical trials require 4 to 5 minutes duration where the conversation period takes 2 to 3 minutes and the response period another 2 minutes. This would limit the total number of conditions in a subject's session to about 24 conditions which would take about 3 hours including instructions, preliminaries and breaks. Tasks designed to measure some system degradations may require conversations longer than 2 to 3 minutes. Compromises have to be made between the test duration and the choice of conditions. If more conditions are to be tested, the test has to be separated into several sessions/experiments and may require different subject panels.

An example is shown in table 2.

Table 2: Timetable for a 24 condition test

	Visit 1				Visit 2		
	Instruction	**Session 1**	**Break**	**Session 2**	**Session 3**	**Break**	**Session 4**
Number of conversations		7 (incl. practice)		6	6		6
Time	15 min	35 min	10 min	30 min	30 min	10 min	30 min

Conditions that are identical in both directions and that use the same sensors and same acoustic noise are called symmetric conditions. Any other case is considered asymmetric. For asymmetric conditions, subject pairs should be required to swap location for each condition. This limits the total number to 12 asymmetric conditions.

In order to achieve a sufficient resolution between conditions, it is recommended that the minimum number of subject pairs should in general be 16. It is also recognized that this number may have to be relaxed in some circumstances in order to reduce the available time for the test, however this will reduce the reliability of results.

4.2.1.3 Requirements regarding test conditions

Some conditions, including transmission channel and environmental noise, may vary with time. In order to take this into account, the trial time needs to be increased to adapt to the conditions. Care should be taken by the experimenter/analyst in order not to overestimate the impact of impairments of non-linear and/or time-variant systems occurring infrequently during the conversation.

Certain types of environmental noise may require sophisticated sound reproduction systems. ES 202 396-1 [i.10] describes methodologies to create appropriate noise conditions. It also provides a noise database for several environmental conditions, including car simulations.

Examples of test condition variables are:

- Environmental noise (street, car, cafeteria, etc.).
- Room reverberation (none to highly reverberant).
- Transducer (hands free, headset, handset, noise canceller, microphone array, etc.).
- Frequency bandwidth (narrow-band, wideband, audio band, etc.).
- Transmission channel/network characteristics (delay, packet loss, fading, etc.).
- Terminal (mobile phone, soft phone, POTS, etc.).
- Coder.

The test environment for each test room need to be defined with the following parameters:

- Room characteristics (size, reverberation time, etc.), see Recommendation ITU-T P.800 [i.4].
- Background noise:
  - level of noise;

- type of noise (car, babble, etc.);
- frequency spectrum;
- dynamic characteristics of the noise field.

4.2.1.4 Requirements regarding Subjects

The choice of naive (untrained), experienced or expert subjects depends on the questions and the required degree of precision in the results.

In general, the advice given in Recommendation ITU-T P.800 [i.4] should be taken into account when selecting test subjects.

Some care should be taken when selecting subjects for conversation tests. As with any speech signal processing equipment, some potential subjects will be more experienced than others. It is recognized that the levels of experience with specific equipment or technology is a continuum, ranging from those who are completely unfamiliar with technical behaviour of the equipment under test (non-experts) to those who are thoroughly competent in the operation and maintenance of this equipment (experts).

The age and gender of all types of subject, together with their partners, should be recorded for all types of tests, but especially for any formal conversation test as opposed to informal expert evaluations.

Unless gender, age and other socio-economic characteristics are design factors of the test, then a formal conversation test should be populated (on a best-endeavour basis) with a random mix of subjects.

4.2.1.4.1 Untrained subjects (naive)

Untrained subjects are accustomed to daily use of a telephone. However, they are neither experienced in subjective testing methodology, nor are they experts in technical implementations of the equipment under test. Ideally, they have no specific knowledge about the device that they will be evaluating. Consistent with Recommendation ITU-T P.800 [i.4], the subjects have not participated in any subjective test in the previous 6 months. Each subject pair is given the opportunity to become familiar with each other in a controlled period of time. Time should be allowed for instructing the subjects about the procedure of the test and the task they have to perform. Practice conditions (the result of which is not included in the result analysis) should be used at the start of the test to ensure that the subjects are comfortable with the test procedure and understand the task. The subject pool should be representative of the telecommunication user pool and the application that the experiment is designed to measure.

4.2.1.4.2 Experienced subjects

Experienced subjects are experienced in subjective testing including subjects who participate routinely in subjective testing but does not include individuals who routinely administer, design or run subjective evaluations. Experienced subjects are able to describe an auditory event in detail and are able to separate different events based on specific impairments. They are also able to describe their subjective impressions in detail. However, experienced subjects neither have a background in technical implementations of the equipment under test, nor do they have detailed knowledge of the influence of these implementations on subjective quality.

4.2.1.4.3 Experts

Experts are experienced in subjective testing. Experts are able to describe an auditory event in detail and are able to separate different events based on specific impairments. They are able to describe their subjective impressions in detail. They have a background in technical implementations of the equipment under test and do have detailed knowledge of the influence of particular implementations on subjective quality. Individuals directly involved in the design or development of the specific system under test has to be excluded from that particular test.

4.2.1.5 Requirements regarding Tasks

In addition to the descriptions for full conversation tests in Recommendations ITU-T P.800 [i.4] and P.805 [i.5], the following consideration may be taken into account. Conversational tests were carried out with observers (operators) present in the test room together with the subjects, but this is generally not recommended. Instead, an audio/visual link should be used to observe or communicate with the subjects. It is the task of the observers (operators) to document all comments which subjects mention during or after the test. This documentation can be useful for further analysis. In addition, audio/video recordings of the conversations can be made.

4.2.1.5.1 Requirements for tasks to be used for untrained subjects

A task should be selected that best fits the requirements of the specific objective of the experiment and the cultural factors of the subject pool. The characteristics required for selecting a task are that:

- it should allow for the generation of a sufficient number of equivalent versions. Each version should stimulate an equivalent level of conversation and interaction;
- it should stimulate semi-structured conversations (too 'open' conversations make it impossible to measure communication efficiency, but too structured communications do not leave room for the subjects to develop a balanced opinion of the channel);
- it should be easily learned;
- it should be intrinsically motivating;
- it should allow for interruptions from the subjects;
- it should be insensitive to changes in subjects' task strategy or skill in performing the task;
- it should represent a cooperative effort between the communicators rather than a competitive effort;
- it should induce the subjects to make use of a rich, varying vocabulary with sufficient two-way interaction;
- it should induce discussion that is phonetically rich and temporally widely distributed (short *and* long utterances and interruptions).

4.2.1.5.2 Examples of conversational tasks

The following conversational tasks meet the requirements given in clause 6.6.1 of Recommendation ITU-T P.805 [i.5]:

- Subjects are asked to reach an agreement on an order of preference or time for a set of picture postcards as described in Handbook on Telephonometry [i.22].
- In the so-called "Kandinsky test" the subjects are asked to describe to their partner the position of a set of numbers on a picture. Both subjects have similar pictures, but with some of the numbers in different positions. It is recommended that the picture should be designed for the task and that both the picture and the numbers are easy to describe. This can be achieved by using pictures consisting of coloured, geometrical figures (e.g. Kandinsky or others).
- In the so-called "short conversational tests" proposed by the Ruhr University (Bochum, Germany) in [i.21], scenarios developed by them are derived from typical situations of everyday life: railway enquiries, rental of a car or an apartment, etc. These scenarios were elaborated to allow a well-balanced conversation between both participants, to stimulate the discussion between persons and to facilitate the naturalness of the conversation. These conversations are approximately 2,5 to 3 minutes in duration. Examples of such scenarios are presented in Appendices IV (German), V (English) and VI (French); of Recommendation ITU-T P.805 [i.5].
- Handbook on Telephonometry [i.22] also gives some guidance on "simplified conversation tests", where shortcuts are suggested to reduce the time taken or to increase the number of treatments in one experiment. Subjects are asked to rate a number of individual degradations after they have given their opinions on quality and difficulty.

- In the task taken from [i.23], random shapes are presented to the subject on a paper sheet or screen. Twenty-four shapes is a typical number on one sheet. There are no meaningful relationships between shapes and their names. The detail and concrete method of how to generate the shapes can be found in [i.23]. The operator prepares the same set of sheets for both subjects, but with the shapes in a different order. During the conversation, each subject arbitrarily chooses one shape on the sheet and describes one of its features to his/her partner. His/her partner either guesses the name of the shape based on the information provided or requests additional information from their partner until the shape is identified. Then partners swap their role and continue with another shape. Example shapes are given in Appendix VII of Recommendation ITU-T P.805 [i.5].
- A "game" where subjects work with their partner to complete a cooperative task or solve a problem. This approach can be used effectively to control the trial-to-trial variability. Care has to be taken to ensure that the game does not limit the conversational vocabulary.

In addition to such conversational tasks, specific tasks may be used which stress the interactivity of the conversation, however at the expense of being less realistic and more competitive. Such tasks may be:

- The mutual reading of random numbers or other items as fast as possible, see e.g. [i.25].
- The mutual verification of numbers or other items as fast as possible, see, e.g. [i.25] or [i.24]. An example for such a task is given in Appendix VIII of Recommendation ITU-T P.805 [i.5].
- More interactive versions of the short conversation test tasks, called "interactive short conversation tests", see [i.26] and [i.24]. The task consists of the fast exchange of data. Two subjects are described to be colleagues working in two different sections in one big company, exchanging, e.g. telephone numbers and email-addresses. In order to speed up the conversations, tasks are presented in terms of tabulated data which were iteratively optimized based on a series of informal tests. These showed that the tabulated data underlying the conversations should not be too different for the two subjects, in order to avoid natural delay in the responses due to the necessity of searching for items in the tables. On the other hand, it was found that too identical list-orders lead to a training effect so that the subjects started to develop a "walkie-talkie" speaking style. As a compromise, one item in the list of each subject is chosen so that it cannot be found in the list of the other subject, with changing positions. This way, fast conversations can be achieved without a strong effect of a "walkie-talkie" style. An example for such a more interactive scenario can be found in Appendix IX of Recommendation ITU-T P.805 [i.5].

It should be noted that the impact of, e.g. transmission delay in situations provoked by such interactive tasks may be more severe than in situations provoked by the tasks which are in accordance with clause 6.6.1 of Recommendation ITU-T P.805 [i.5]. This may be due to the structure of the conversation being changed, see e.g. [i.24] for a discussion.

4.2.1.6 Requirements regarding Questions

Recommendation ITU-T P.800 [i.4] and Handbook on Telephonometry [i.22] recommend both a "quality" question using a five-point scale and a "difficulty" question using a binary scale. Some organizations felt that subjects were confused by the "difficulty" question, while other organizations would still prefer to continue using it. As a result, both these scales are reproduced here but new scales are also provided. These new scales may help the subjects to formulate an overall quality judgement by initially focusing their attention on different quality dimensions.

In Recommendation ITU-T P.800 [i.4] and Handbook on Telephonometry [i.22], the scales are as follows:

"*What is your opinion of the connection you have just been using?*"

- Excellent
- Good
- Fair
- Poor
- Bad

The experimenter allocates the following values to the categories: Excellent = 5; Good = 4; Fair = 3; Poor = 2; Bad = 1.

All further statistical processing is performed in terms of these numbers.

"*Did you or your partner have any difficulty in talking or hearing over the connection?*"

- Yes
- No

The experimenter allocates the following values to the responses: Yes = 1; No = 0.

The new scales are given below and the intention is that after each trial (corresponding to one specific condition) the subjects have to evaluate multiple aspects of the communication. The following questions are provided as examples and are representative of the multiple aspects to be considered. Several five-point category scales are provided as well as a binary response scale. The cognitive load on the subjects and therefore the number of questions asked should be minimized to reduce subject fatigue and any possible confusion.

"*How would you assess the sound quality of the other person's voice?*"

The five-point scale descriptors are:

- No distortion at all, natural
- Minimal distortion
- Moderate distortion
- Considerable distortion
- Severe distortion

"*How well did you understand what the other person was telling you?*"

The five-point scale descriptors are:

- No loss of understanding
- Minimal loss of understanding
- Moderate loss of understanding
- Considerable loss of understanding
- Severe loss of understanding

"*What level of effort did you need to understand what the other person was telling you?*"

The five-point scale descriptors are:

- No special effort required
- Minimal effort required
- Moderate effort required
- Considerable effort required
- Severe effort required

"*How would you assess your level of effort to converse back and forth during the conversation?*"

The five-point scale descriptors are:

- No special effort required
- Minimal effort required
- Moderate effort required
- Considerable effort required
- Severe effort required

"*Did you detect (insert distortion of interest here)?*"

- Yes
- No

"If yes, how annoying was it?"

The five-point scale descriptors are:

- No annoyance
- Minimal annoyance
- Moderate annoyance
- Considerable annoyance
- Severe annoyance

"*What is your opinion of the connection you have just been using?*"

The five-point scale descriptors are:

- Excellent quality
- Good quality
- Fair quality
- Poor quality
- Bad quality

The previous examples should be supplemented by the experimenter to address the needs of the specific experiment. When using multiple scales for assessing the multi-dimensional aspect of quality, care should be taken to ensure that the previous responses are not available to the subjects.

4.2.2 Test set-up

Based on the requirements described in clause 4.2.1 the following options were chosen.

The conversational scenarios can be found in annex B for the tests in English language and in Annex C for the test in Czech language. The instructions of the subjects and the quality question to be answered by the subjects after each test can be found in annex D.

An example of the detailed session plan is in annex D.

The conversational tests were conducted under the control of a supervisor with the two test persons sitting in two different rooms following the requirements defined in clause 4.2.1.

The technology needed for experiment (network simulator) is located in separated room where also experiment operators are seated. Each conversational room uses table and chair, fixed telephone terminal as described in the next chapters and microphone pre-amplifier. Further details of the test environment can be found at the MESAQUIN website where one picture shows one subject seated in one of the two conversational rooms, using the handset telephone.

4.2.2.1 MESAQIN.com real-time network simulator description

The block scheme is depicted in figures 1 (overview) and 2 (DSP block):

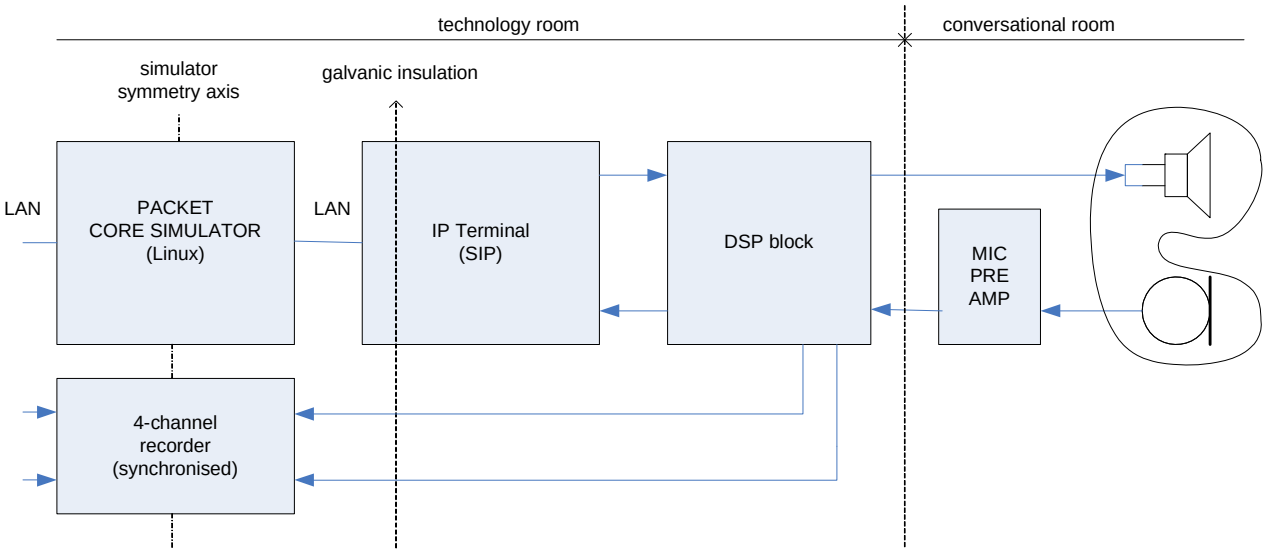


Figure 1: Network simulator (one half is shown, the other is symmetrical)

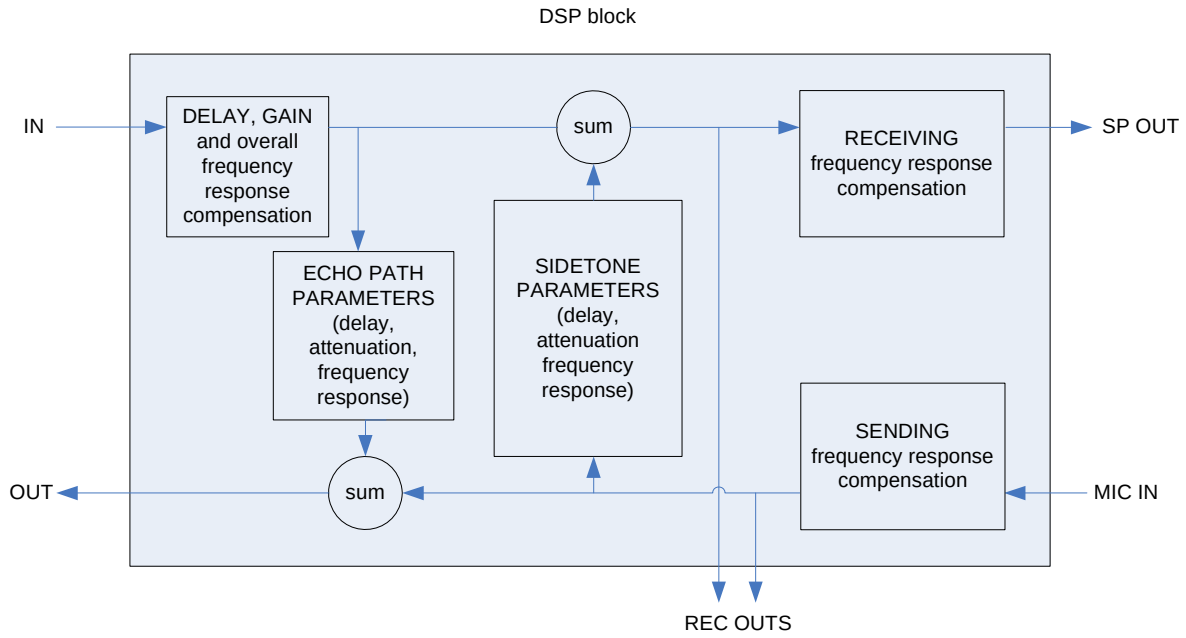


Figure 2: Detailed structure of DSP block of the network simulator (two such blocks are needed for real-time call simulations)

The selected simulator parameters are in bold characters through the list of available characteristics:

- Audio coder support:
  - G.711 A-law
  - G.711 μ-law
  - Speex-NB
  - Speex-WB
  - GSM
  - AMR-NB
  - AMR-WB

- G.729AB
- G.723 5.3k
- G.723 6.3k

- Delays:
  - up to 2 400 ms.
  - 100 ms
  - 300 ms
  - 600 ms

NOTE: Only one half of the simulator is shown at the picture, the other is symmetrical.
Packet core is galvanicly insulated from the DSP end parts.
Packet core simulator was not used for the experiment (no parameters to be varied there).
For the experiment, symmetrical setup is considered (ED=2*TD, equal TERL for sides A and B).

4.2.2.2 Terminal calibration and equalization to ES 202 737 in send and receive direction

The hardware used for conversational tests in the laboratory consists of analogue Panasonic™ phones, with external microphone preamplifiers added, and completely removed electronics.

All output/input signals are analogue, common levels (1,7 V peak max).

All frequency responses were measured on Brüel & Kjær Head and Torso Simulator 4128C, S/N 27891552, using positioner 4606, S/N 2768519 and artificial ear 4158_C type 3.3 and verified by measurement on HeadAcoustics Artificial Head MFE VI. "TC6199" and measurement system ACQUA 3.1.100. The latter was also used for application force sensitivity analysis.

The real-time compensation is used to equalize the responses to conform to ES 202 737 [i.7] responses. It deploys 24 bits, 96 kSa/s DSP. The delay introduced by the DSP compensation block is < 1 ms.

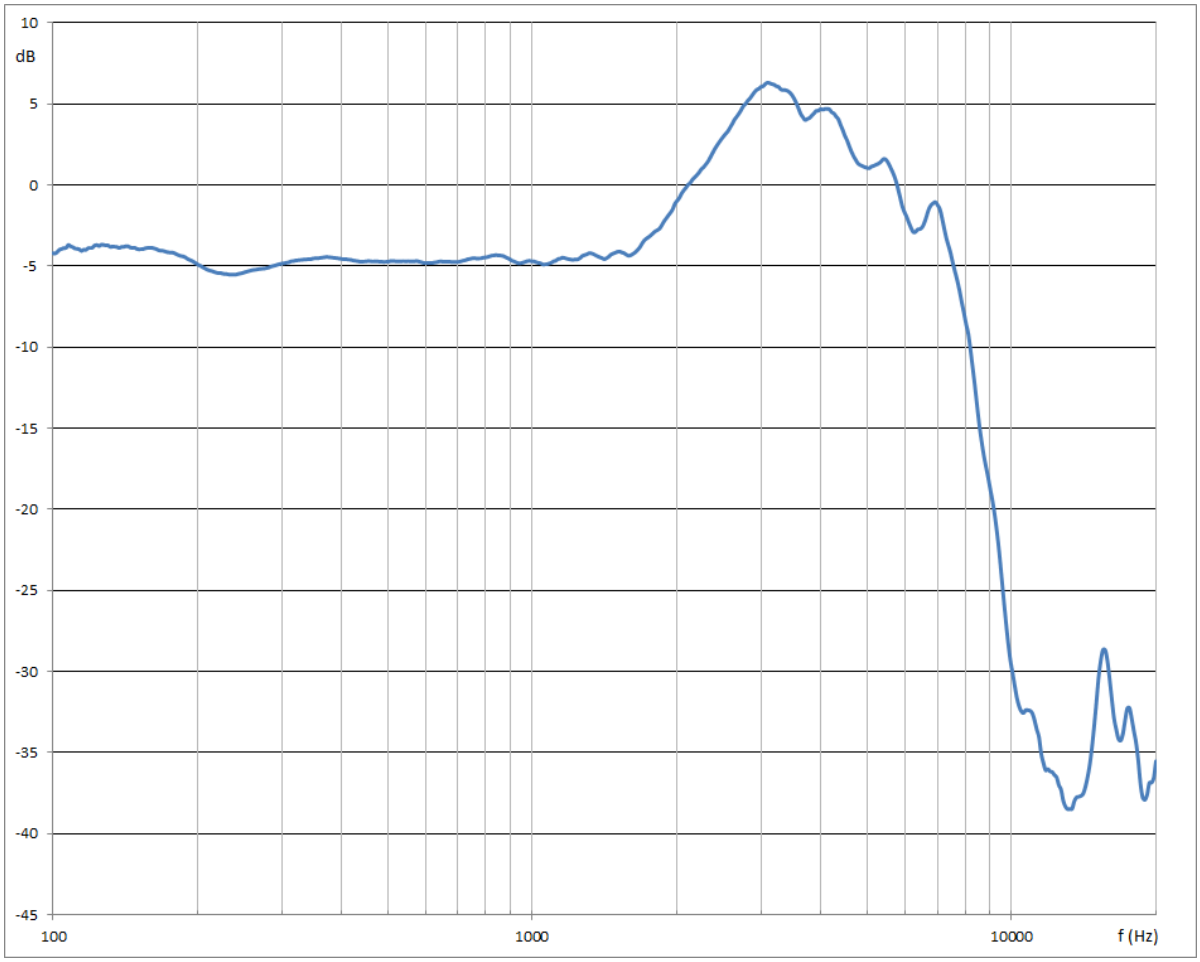


Figure 3: Original (uncompensated) frequency response in SEND direction dBV/Pa

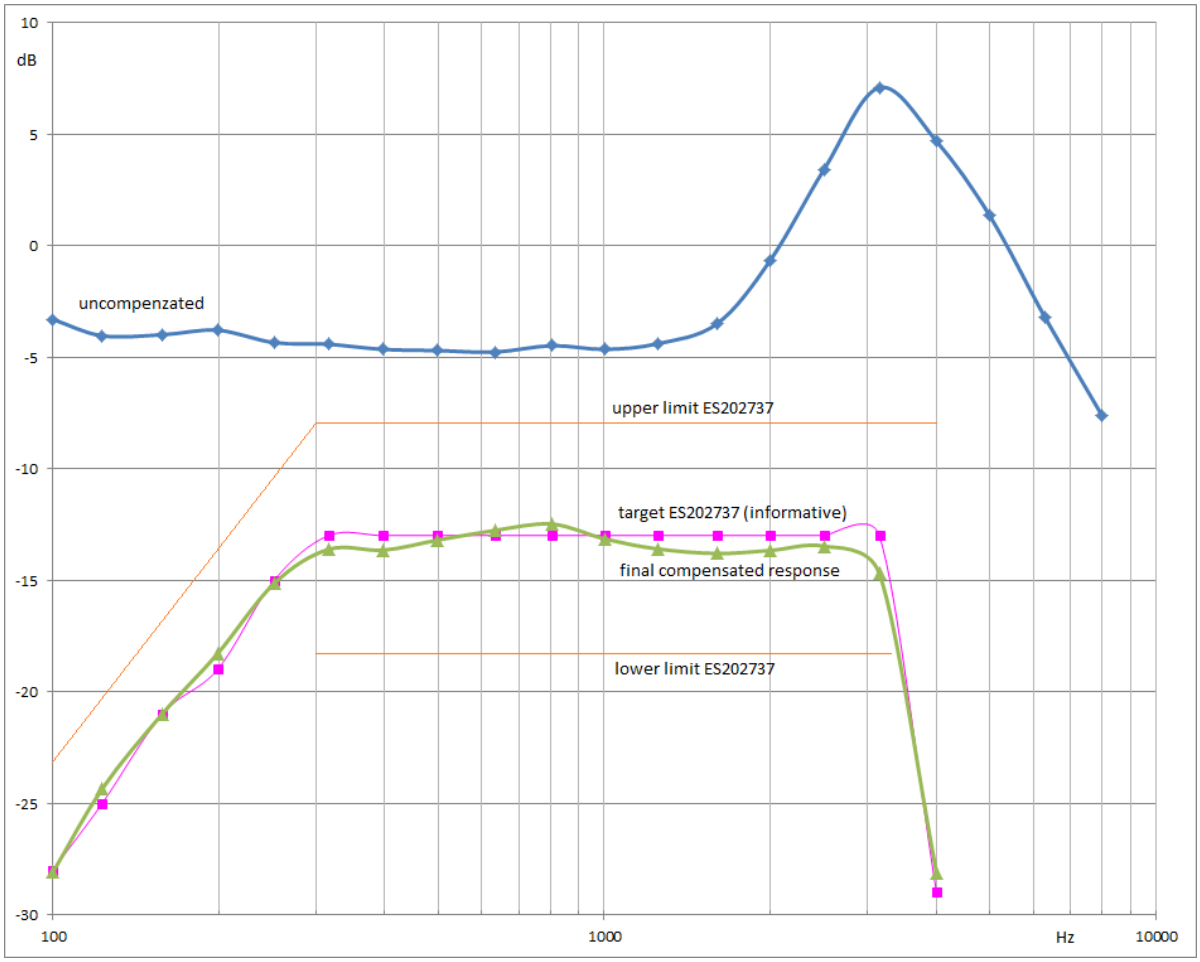


Figure 4: Final real-time compensated frequency response in SEND direction (dBV/Pa, green)

NOTE 1: Also shown in figure 4: informative target as per ES 202 737 [i.7] (pink), upper and lower limit as per ES 202 737 [i.7] (orange), and original uncompensated response (blue).

NOTE 2: Valid measurement points are indicated by marks, connecting lines are for informative purposes only.

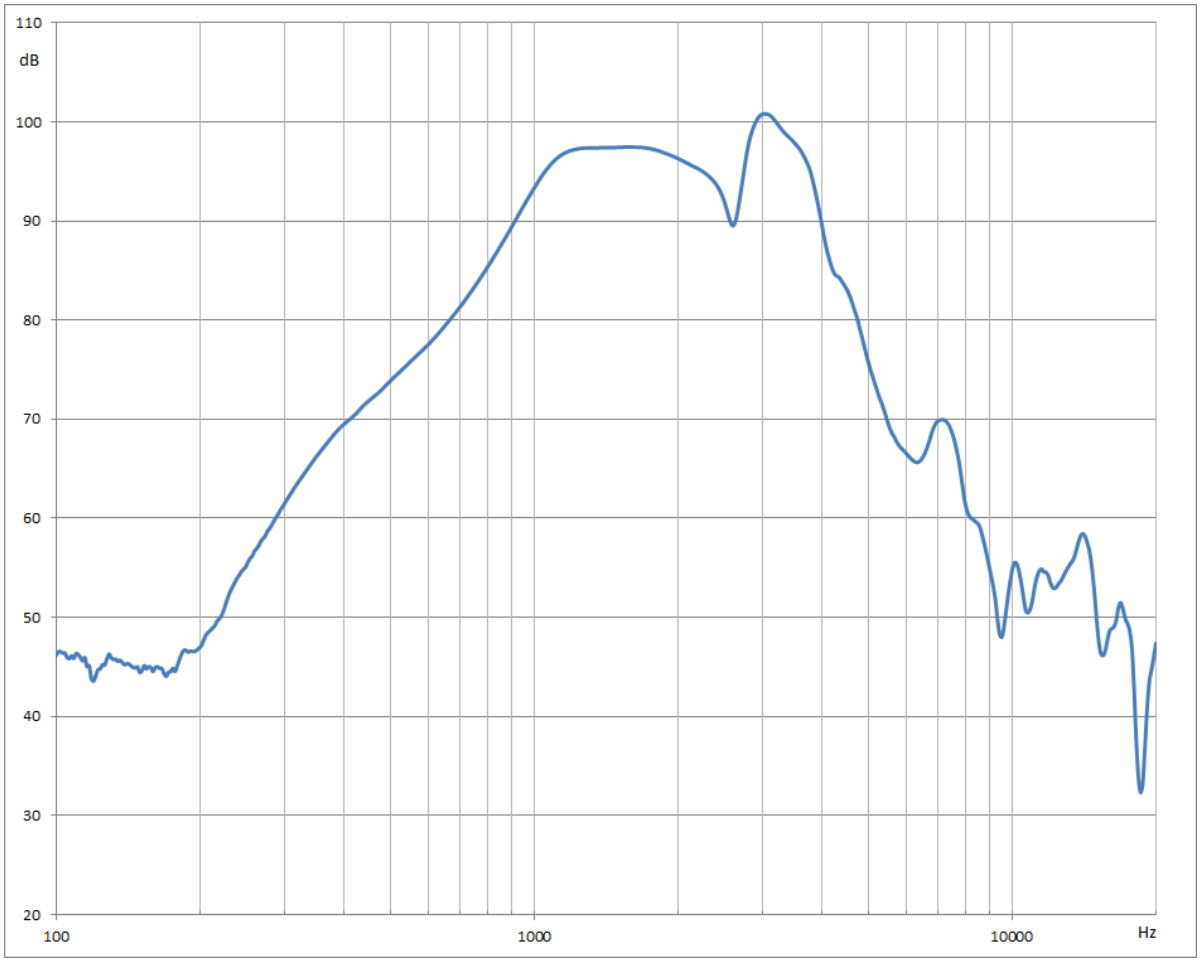


Figure 5: Original (uncompensated) frequency response in RECEIVE direction, dB SPL, before correction (Recommendation ITU-T P.57 [i.17], Paragraph 5.2)

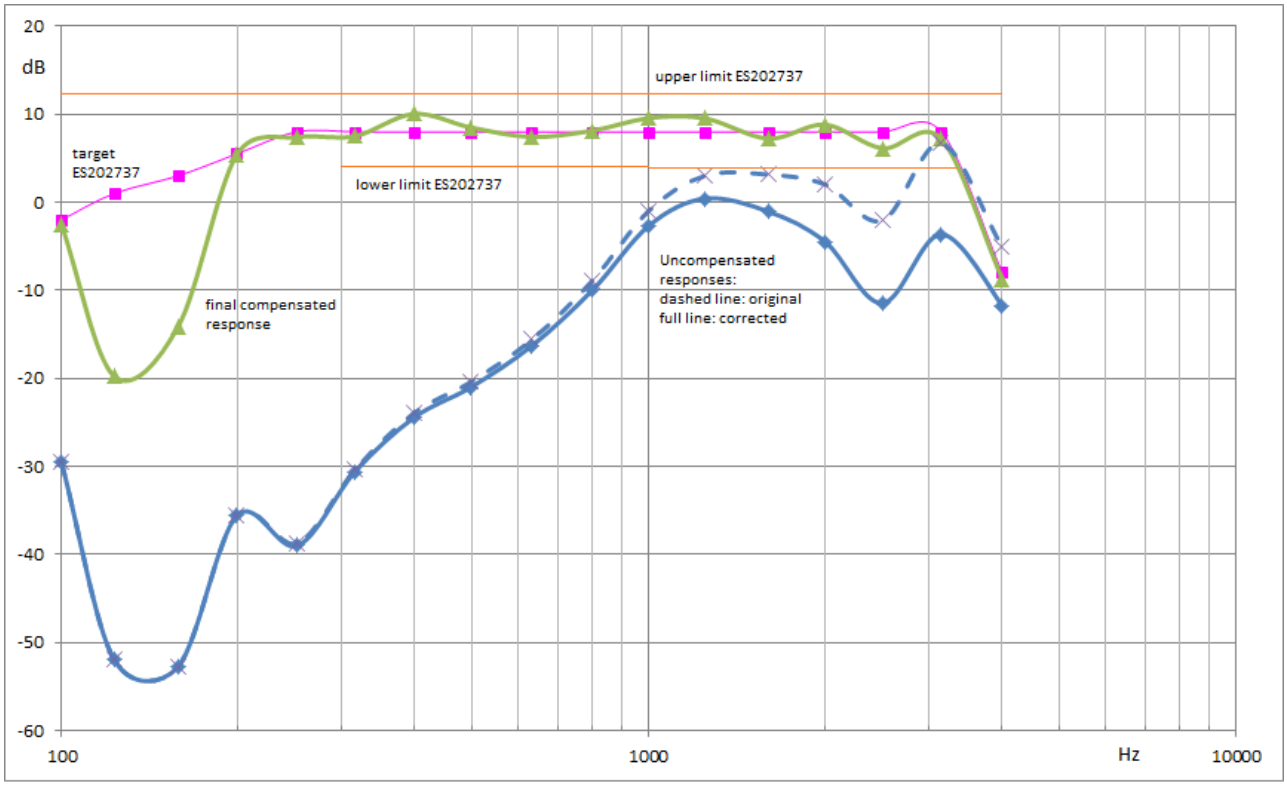


Figure 6: Final real-time compensated frequency response in RECEIVE direction (dBPa/V, green) after correction as per Recommendation ITU-T P.57 [i.17] (clause 5.2, Table 2a)

NOTE 3: Also shown in figure 6: informative target as per ES 202 737 [i.7] (pink), upper and lower limit as per ES 202 737 (orange), and original response (blue), corrected as per Recommendation ITU-T P.57 [i.17] (full line, dashed line shows the response before the correction).

NOTE 4: Valid measurement points are indicated by marks, connecting lines are for informative purposes only.

NOTE 5: Due to low sensitivity of the used (narrow-band) terminal loudspeaker below 200 Hz, measured results are masked there by HATS measurement setup noise (compare figure 5). Thus, they are not relevant for the real-time equalization (even though they also fully conform to ES 202 737 [i.7]).

With the responses given above, the following loudness ratings were evaluated:

- SLR=8,1 dB
- RLR=2,2 dB

Sensitivity of receiving frequency response to application force analysis.

Three measurements of frequency response were made using application forces 13N, 8N (nominal) and 2N, respectively. The measured responses are shown in figures 7 to 9. It is obvious that the dominant frequency peak is shifted towards higher frequencies while decreasing its magnitude for decreasing application force, due to the impedance of the transducer. Based on these measurements, the detailed instructions concerning how to carry the handset were given to the test subjects prior each test session, to assure constant and stable application force during the subjective tests.

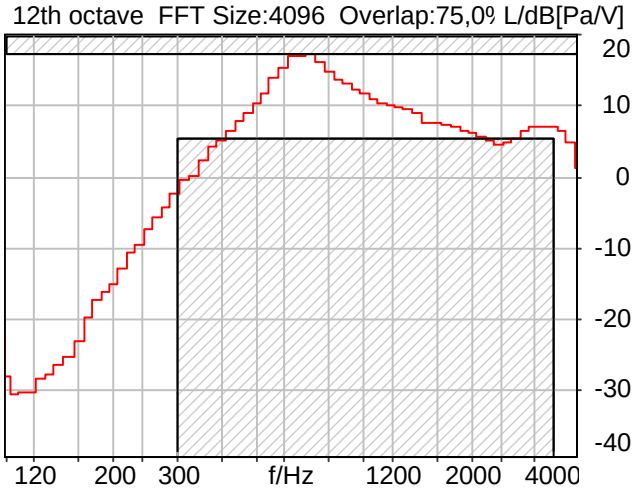


Figure 7: Frequency response in RECEIVE direction for 13 N application force

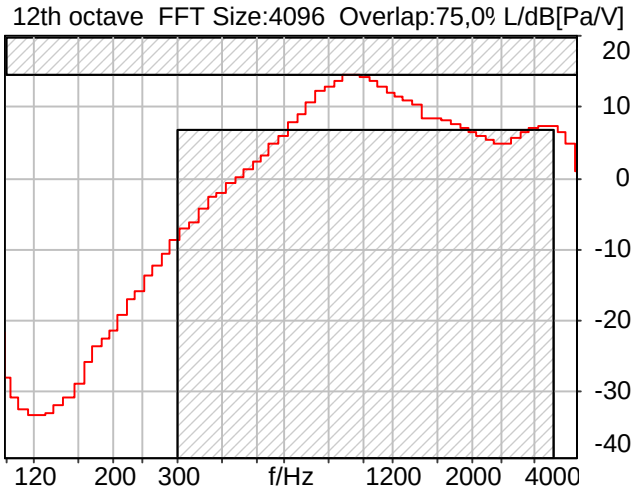


Figure 8: Frequency response in RECEIVE direction for 8 N application force

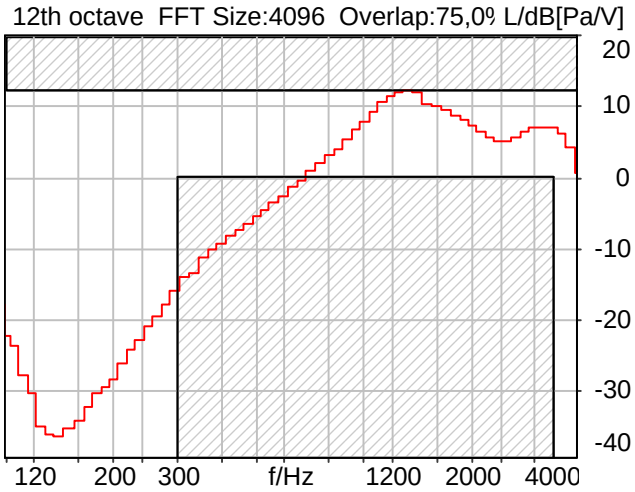


Figure 9: Frequency response in RECEIVE direction for 2 N application force

4.2.2.3 Conversational scenarios

The conversational scenarios are intended to stimulate different degrees of interactivity between the two subjects in the test. However, it is not sufficient just to define three scenarios, like one would define three different codecs or three different values for end-to-end delay. Repetition of "exactly" the same conversational task has to be avoided in the course of test conducted with one pair of subjects.

Annex B provides the conversational scenarios including the instruction for the test persons in English language. Annex C provides the same scenarios in Czech language.

The quality question to be answered by the subjects after each test can be found in annex D.

The conversational tests will be conducted under the control of a supervisor with the two test persons sitting in two different rooms with the properties as described in clause 4.2.2.

All conversational scenarios were developed based on the guidance provided by Recommendation ITU-T P.805 [i.5].

4.2.3 Subjective test plan

The subjective test plan is sub-divided in the matrix of parameter combination and the detailed session plan for the subjective test lab.

The matrix of parameter combinations is given in table 3.

Table 3: Subjective test plan

#	coder	echo	delay	interactivity l	English	Czech
01	G.711	46dB	100ms	lo	Yes	Yes
02	G.711	46dB	100ms	mi	Yes	
03	G.711	46dB	100ms	hi	Yes	
04	G.711	46dB	300ms	lo	Yes	
05	G.711	46dB	300ms	mi	Yes	Yes
06	G.711	46dB	300ms	hi	Yes	
07	G.711	46dB	600ms	lo	Yes	Yes
08	G.711	46dB	600ms	mi	Yes	
09	G.711	46dB	600ms	hi	Yes	
10	G.711	32dB	100ms	lo	Yes	
11	G.711	32dB	100ms	mi	Yes	Yes
12	G.711	32dB	100ms	hi	Yes	
13	G.711	32dB	300ms	lo	Yes	
14	G.711	32dB	300ms	mi	Yes	Yes
15	G.711	32dB	300ms	hi	Yes	
16	G.711	32dB	600ms	lo	Yes	
17	G.711	32dB	600ms	mi	Yes	
18	G.711	32dB	600ms	hi	Yes	Yes
19	AMR-NB	46dB	100ms	lo	Yes	
20	AMR-NB	46dB	100ms	mi	Yes	
21	AMR-NB	46dB	100ms	hi	Yes	Yes
22	AMR-NB	46dB	300ms	lo	Yes	Yes
23	AMR-NB	46dB	300ms	mi	Yes	
24	AMR-NB	46dB	300ms	hi	Yes	
25	AMR-NB	46dB	600ms	lo	Yes	
26	AMR-NB	46dB	600ms	mi	Yes	
27	AMR-NB	46dB	600ms	hi	Yes	
28	AMR-NB	32dB	100ms	lo	Yes	
29	AMR-NB	32dB	100ms	mi	Yes	
30	AMR-NB	32dB	100ms	hi	Yes	
31	AMR-NB	32dB	300ms	lo	Yes	Yes
32	AMR-NB	32dB	300ms	mi	Yes	
33	AMR-NB	32dB	300ms	hi	Yes	Yes
34	AMR-NB	32dB	600ms	lo	Yes	
35	AMR-NB	32dB	600ms	mi	Yes	Yes
36	AMR-NB	32dB	600ms	hi	Yes	Yes
37	G.729	46dB	100ms	lo	Yes	Yes
38	G.729	46dB	100ms	mi	Yes	
39	G.729	46dB	100ms	hi	Yes	
40	G.729	46dB	300ms	lo	Yes	
41	G.729	46dB	300ms	mi	Yes	
42	G.729	46dB	300ms	hi	Yes	Yes
43	G.729	46dB	600ms	lo	Yes	Yes
44	G.729	46dB	600ms	mi	Yes	
45	G.729	46dB	600ms	hi	Yes	
46	G.729	32dB	100ms	lo	Yes	
47	G.729	32dB	100ms	mi	Yes	Yes
48	G.729	32dB	100ms	hi	Yes	
49	G.729	32dB	300ms	lo	Yes	
50	G.729	32dB	300ms	mi	Yes	Yes
51	G.729	32dB	300ms	hi	Yes	
52	G.729	32dB	600ms	lo	Yes	
53	G.729	32dB	600ms	mi	Yes	
54	G.729	32dB	600ms	hi	Yes	Yes

Two TELR values were agreed, namely TELR= 46 dB TELR= 32 dB; and three one-way delay values: 100 ms, 300 ms and 600 ms; for the interactivity, lo refers to Appendix V, mi refers to Appendix VII and hi refers to Appendix VIII of Recommendation ITU-T P.805 [i.5] with the provision that this classification has to be resorted after the tests as per TAR measured.

18 Czech samples were selected randomly to exercise various conditions (including interactivity) are shown in the right column of table 3.

Annex E provides the objective quality measured from end to end using PESQ according to Recommendation ITU-T P.862 [i.9], showing that the transmission chain is correctly implemented.

4.3 Conducting the subjective tests and creation of report describing results obtained

4.3.1 Conducting the subjective tests

The subjective tests were conducted from July 2012 till December 2012 with 16 Czech native subjects and 48 English native subjects. The information about the age of subjects are available in annex D. Three different randomizations were used in both cases.

The conversational opinion scores were obtained and Mean Opinion Score (MOS-CQs) were calculated. Also root mean square errors (RMSE) were calculated for each MOS value. The 95 % confidence interval CI95 is then calculated as follows:

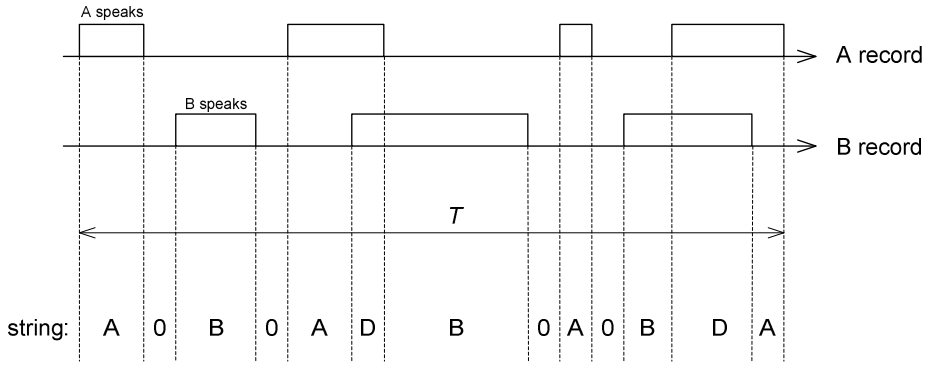
$$CI95=2*RMSE$$

For each converation, the Talker Alternation Rate (TAR) parameter was calculated prior further data processing. More information about TAR are also available in clause 5.

TAR analysis:

During the tests the speech signals are recorded electrically after being amplified by microphone pre-amplifier at each conversation room. It should be noted these signals are to be considered as time synchronized related to (middle) reference simulator point and cannot be considered as representing the subjective conversational situation at either side (each conversation participant perceives the other side with certain delay that is not reflected in the recording. For studies of subjective situation at each side, the delay can be artificially introduced to one or the other recording channel; however, this is not a purpose of TAR calculation).

TAR definition:



A: A speaks, B is silent
B: B speaks, A is silent
0: both A and B is silent
D: both A and B speaks (doubletalk)

Figure 10: TAR definition

Any call progress can be then translated into a string containing the above characters. As a role swap, the following cases are considered: A0B, B0A, ADB, BDA, and also theoretically AB and BA, even though those combinations are quite rare in real scenarios. For the example in figure 10, six role swaps measured during T = 20 s measurement interval mean TAR is 18 min^{-1}.

TAR measurement details:

The TAR measurement is performed on 5 ms energy packets of the original speech recording with adaptive threshold of active speech. The detection adaptation algorithm is based on Recommendation ITU-T P.56 [i.20]. Any silent periods shorter than 350 ms are considered to be inter-syllabic pauses and thus neglected. Recording parts before the first role swap and after the last one are not considered to be a part of the measurement time T.

The TAR analysis results are shown in table 4.

Table 4

	English test conversations	Czech test conversations
Average TAR	34,4	32,6
Minimum TAR	3,7	5,7
Maximum TAR	83,1	74,8

4.3.2 Test results

Test results are presented in detail in tables 5 and 6 and in graphs for different coders (figures 11 to 13) and for combinations of coder and TELR, each split for 3 different interactivity levels based either on conversational scenario (left) or on TAR analysis (right).

Table 5

SCENARIO	ENGLISH LANGUAGE									
	IQ MOS	**CQ MOS**	IQSTD MOS	**CQ STD MOS**	TARaver	IQ MOS	**CQ MOS**	IQSTD MOS	**CQ STD MOS**	TARclass
	according to scenario type					according to TAR				
1	3,96	**3,82**	0,13	**0,10**	19,08	4,10	**3,80**	0,13	**0,11**	0...22
2	4,28	**3,80**	0,11	**0,11**	24,85	4,15	**3,79**	0,11	**0,10**	22...46
3	4,26	**3,60**	0,12	**0,16**	59,06	4,25	**3,63**	0,13	**0,17**	46...
4	4,24	**3,80**	0,13	**0,11**	19,13	4,36	**3,86**	0,10	**0,10**	0...22
5	4,38	**3,82**	0,10	**0,13**	24,14	4,24	**3,64**	0,12	**0,13**	22...46
6	4,14	**3,58**	0,12	**0,14**	57,37	4,12	**3,67**	0,14	**0,16**	46...
7	3,84	**3,72**	0,12	**0,14**	18,95	3,94	**3,72**	0,11	**0,12**	0...22
8	4,14	**3,64**	0,11	**0,14**	23,98	4,08	**3,58**	0,12	**0,15**	22...46
9	4,24	**3,46**	0,11	**0,14**	59,02	4,22	**3,50**	0,12	**0,15**	46...
10	4,14	**3,60**	0,12	**0,15**	22,02	4,27	**3,66**	0,13	**0,15**	0...22
11	4,06	**3,86**	0,14	**0,12**	21,51	3,96	**3,79**	0,12	**0,13**	22...46
12	4,20	**3,82**	0,11	**0,14**	58,27	4,20	**3,82**	0,11	**0,14**	46...
13	3,80	**3,28**	0,13	**0,15**	19,24	3,98	**3,33**	0,11	**0,13**	0...22
14	3,96	**3,28**	0,11	**0,15**	23,29	3,73	**3,20**	0,14	**0,18**	22...46
15	3,94	**2,88**	0,14	**0,17**	58,62	3,94	**2,88**	0,14	**0,17**	46...
16	3,98	**3,24**	0,12	**0,15**	20,02	4,07	**3,21**	0,12	**0,14**	0...22
17	3,92	**3,16**	0,12	**0,14**	24,13	3,83	**3,19**	0,10	**0,16**	22...46
18	3,52	**2,64**	0,14	**0,16**	55,87	3,46	**2,59**	0,15	**0,15**	46...
19	4,04	**3,72**	0,13	**0,15**	20,40	4,14	**3,74**	0,14	**0,17**	0...22
20	4,16	**3,62**	0,11	**0,15**	24,97	4,06	**3,60**	0,10	**0,12**	22...46
21	4,42	**3,94**	0,10	**0,12**	59,75	4,42	**3,94**	0,10	**0,12**	46...
22	3,96	**3,78**	0,12	**0,13**	20,83	4,06	**3,89**	0,10	**0,11**	0...22
23	4,00	**4,02**	0,10	**0,12**	21,44	3,88	**3,78**	0,13	**0,15**	22...46
24	4,04	**3,66**	0,12	**0,15**	59,34	4,02	**3,76**	0,12	**0,15**	46...
25	4,18	**3,98**	0,10	**0,12**	21,01	4,28	**3,91**	0,10	**0,11**	0...22
26	4,20	**3,82**	0,11	**0,13**	25,87	4,09	**3,88**	0,11	**0,13**	22...46
27	4,00	**3,44**	0,12	**0,15**	60,62	4,02	**3,44**	0,12	**0,15**	46...
28	4,06	**3,60**	0,11	**0,11**	19,97	4,10	**3,65**	0,11	**0,12**	0...22
29	3,98	**3,66**	0,10	**0,13**	23,75	3,96	**3,60**	0,11	**0,12**	22...46
30	4,16	**3,74**	0,12	**0,13**	60,44	4,15	**3,75**	0,13	**0,13**	46...
31	3,80	**3,34**	0,15	**0,17**	19,94	4,02	**3,25**	0,15	**0,17**	0...22
32	4,00	**3,26**	0,14	**0,15**	26,45	3,80	**3,34**	0,13	**0,15**	22...46
33	3,64	**2,86**	0,13	**0,16**	57,68	3,60	**2,85**	0,13	**0,17**	46...
34	3,78	**2,98**	0,14	**0,16**	20,75	3,94	**3,13**	0,15	**0,17**	0...22
35	3,96	**3,10**	0,15	**0,16**	24,82	3,85	**3,00**	0,14	**0,15**	22...46
36	3,78	**2,84**	0,13	**0,14**	59,90	3,73	**2,79**	0,13	**0,14**	46...
37	4,00	**3,74**	0,12	**0,13**	23,42	4,00	**3,81**	0,17	**0,18**	0...22
38	4,26	**3,90**	0,12	**0,14**	24,62	4,19	**3,82**	0,10	**0,11**	22...46
39	4,22	**3,86**	0,11	**0,14**	58,93	4,22	**3,86**	0,11	**0,14**	46...
40	3,94	**3,78**	0,12	**0,14**	20,46	4,02	**3,58**	0,11	**0,13**	0...22
41	3,96	**3,40**	0,10	**0,12**	23,36	3,92	**3,63**	0,11	**0,13**	22...46
42	4,26	**3,98**	0,11	**0,13**	61,49	4,23	**3,96**	0,11	**0,14**	46...
43	4,00	**3,84**	0,12	**0,10**	20,75	3,98	**3,76**	0,12	**0,11**	0...22
44	3,92	**3,38**	0,12	**0,14**	24,76	3,94	**3,48**	0,12	**0,13**	22...46
45	4,04	**3,58**	0,14	**0,15**	58,88	4,04	**3,58**	0,14	**0,15**	46...
46	4,10	**3,50**	0,12	**0,14**	17,38	4,15	**3,69**	0,10	**0,11**	0...22
47	4,24	**3,80**	0,11	**0,11**	24,13	4,23	**3,43**	0,15	**0,18**	22...46
48	4,08	**3,48**	0,14	**0,15**	58,55	4,07	**3,54**	0,15	**0,15**	46...
49	3,56	**2,88**	0,12	**0,14**	20,41	3,74	**2,96**	0,13	**0,13**	0...22
50	3,88	**3,16**	0,13	**0,15**	24,48	3,78	**3,11**	0,12	**0,15**	22...46
51	3,78	**2,76**	0,14	**0,16**	58,66	3,70	**2,70**	0,15	**0,16**	46...
52	3,92	**3,28**	0,14	**0,16**	19,12	4,11	**3,25**	0,12	**0,13**	0...22
53	4,04	**3,28**	0,13	**0,15**	24,62	3,83	**3,40**	0,15	**0,17**	22...46
54	3,68	**2,86**	0,14	**0,17**	55,53	3,59	**2,73**	0,14	**0,18**	46...

Table 6: Results including experiment in Czech langage

SCENARIO	ENGLISH LANGUAGE					CZECH LANGUAGE				
	IQ MOS	**CQ MOS**	IQSTD MOS	**CQ STD MOS**	TARaver	IQ MOS	**CQ MOS**	IQSTD MOS	**CQ STD MOS**	TARaver
	according to scenario type					according to scenario type				
1	3,96	**3,82**	0,13	**0,10**	19,08	4,04	**3,76**	0,12	**0,11**	19,72
2	4,28	**3,80**	0,11	**0,11**	24,85					
3	4,26	**3,60**	0,12	**0,16**	59,06					
4	4,24	**3,80**	0,13	**0,11**	19,13					
5	4,38	**3,82**	0,10	**0,13**	24,14	4,19	**3,72**	0,11	**0,11**	22,31
6	4,14	**3,58**	0,12	**0,14**	57,37					
7	3,84	**3,72**	0,12	**0,14**	18,95	3,70	**3,44**	0,13	**0,14**	18,68
8	4,14	**3,64**	0,11	**0,14**	23,98					
9	4,24	**3,46**	0,11	**0,14**	59,02					
10	4,14	**3,60**	0,12	**0,15**	22,02					
11	4,06	**3,86**	0,14	**0,12**	21,51	4,07	**3,65**	0,14	**0,12**	23,42
12	4,20	**3,82**	0,11	**0,14**	58,27					
13	3,80	**3,28**	0,13	**0,15**	19,24					
14	3,96	**3,28**	0,11	**0,15**	23,29	3,96	**3,07**	0,13	**0,13**	22,56
15	3,94	**2,88**	0,14	**0,17**	58,62					
16	3,98	**3,24**	0,12	**0,15**	20,02					
17	3,92	**3,16**	0,12	**0,14**	24,13					
18	3,52	**2,64**	0,14	**0,16**	55,87	3,75	**2,50**	0,15	**0,14**	51,26
19	4,04	**3,72**	0,13	**0,15**	20,40					
20	4,16	**3,62**	0,11	**0,15**	24,97					
21	4,42	**3,94**	0,10	**0,12**	59,75	4,37	**3,76**	0,11	**0,14**	54,84
22	3,96	**3,78**	0,12	**0,13**	20,83	3,90	**3,56**	0,13	**0,12**	21,11
23	4,00	**4,02**	0,10	**0,12**	21,44					
24	4,04	**3,66**	0,12	**0,15**	59,34					
25	4,18	**3,98**	0,10	**0,12**	21,01					
26	4,20	**3,82**	0,11	**0,13**	25,87					
27	4,00	**3,44**	0,12	**0,15**	60,62					
28	4,06	**3,60**	0,11	**0,11**	19,97					
29	3,98	**3,66**	0,10	**0,13**	23,75					
30	4,16	**3,74**	0,12	**0,13**	60,44					
31	3,80	**3,34**	0,15	**0,17**	19,94	3,63	**2,96**	0,14	**0,14**	18,07
32	4,00	**3,26**	0,14	**0,15**	26,45					
33	3,64	**2,86**	0,13	**0,16**	57,68	3,94	**2,83**	0,13	**0,13**	53,79
34	3,78	**2,98**	0,14	**0,16**	20,75					
35	3,96	**3,10**	0,15	**0,16**	24,82	3,83	**2,83**	0,13	**0,14**	21,70
36	3,78	**2,84**	0,13	**0,14**	59,90	3,83	**2,50**	0,15	**0,15**	54,92
37	4,00	**3,74**	0,12	**0,13**	23,42	4,06	**3,83**	0,10	**0,11**	19,57
38	4,26	**3,90**	0,12	**0,14**	24,62					
39	4,22	**3,86**	0,11	**0,14**	58,93					
40	3,94	**3,78**	0,12	**0,14**	20,46					
41	3,96	**3,40**	0,10	**0,12**	23,36					
42	4,26	**3,98**	0,11	**0,13**	61,49	4,27	**3,52**	0,12	**0,12**	53,95
43	4,00	**3,84**	0,12	**0,10**	20,75	4,07	**3,61**	0,12	**0,13**	17,74
44	3,92	**3,38**	0,12	**0,14**	24,76					
45	4,04	**3,58**	0,14	**0,15**	58,88					
46	4,10	**3,50**	0,12	**0,14**	17,38					
47	4,24	**3,80**	0,11	**0,11**	24,13	4,19	**3,56**	0,12	**0,12**	23,32
48	4,08	**3,48**	0,14	**0,15**	58,55					
49	3,56	**2,88**	0,12	**0,14**	20,41					
50	3,88	**3,16**	0,13	**0,15**	24,48	3,91	**3,19**	0,12	**0,14**	23,13
51	3,78	**2,76**	0,14	**0,16**	58,66					
52	3,92	**3,28**	0,14	**0,16**	19,12					
53	4,04	**3,28**	0,13	**0,15**	24,62					
54	3,68	**2,86**	0,14	**0,17**	55,53	3,61	**2,22**	0,16	**0,15**	51,47

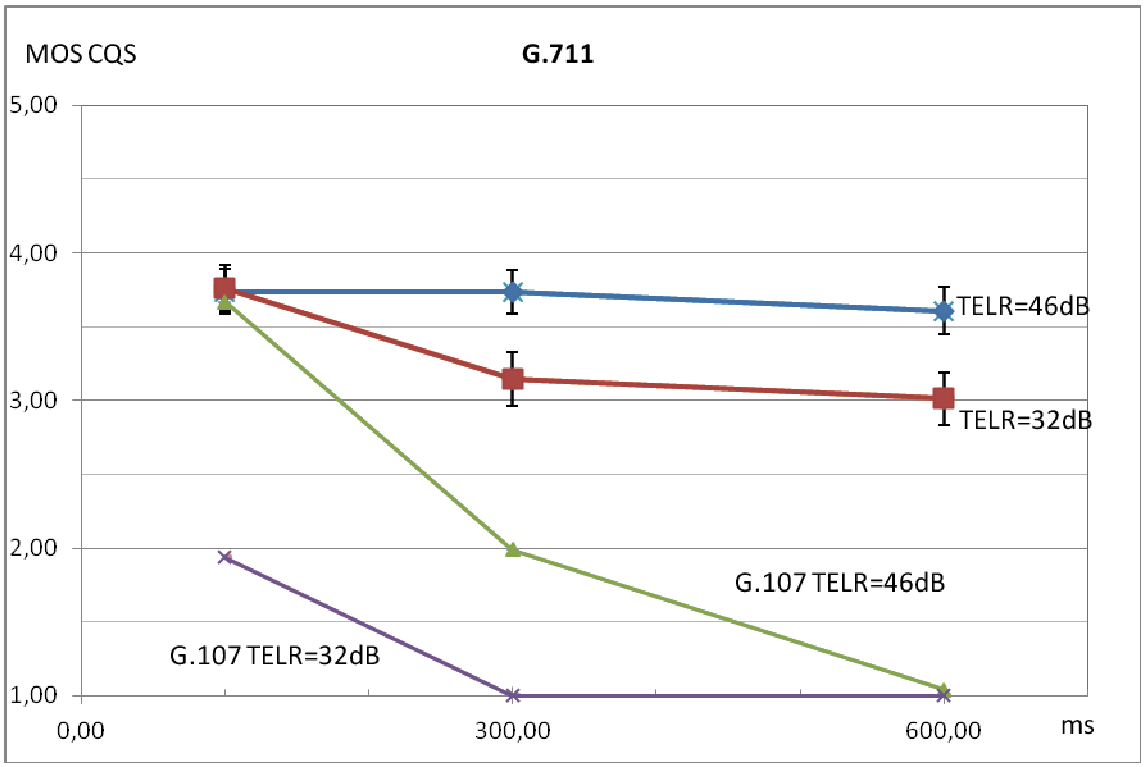


Figure 11: Subjective test results for G.711 coder and two tested TELR values (32 dB, 46 dB) including CI95% uncertainty intervals

NOTE 1: Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600 ms, the connecting lines are shown for informative purposes only.

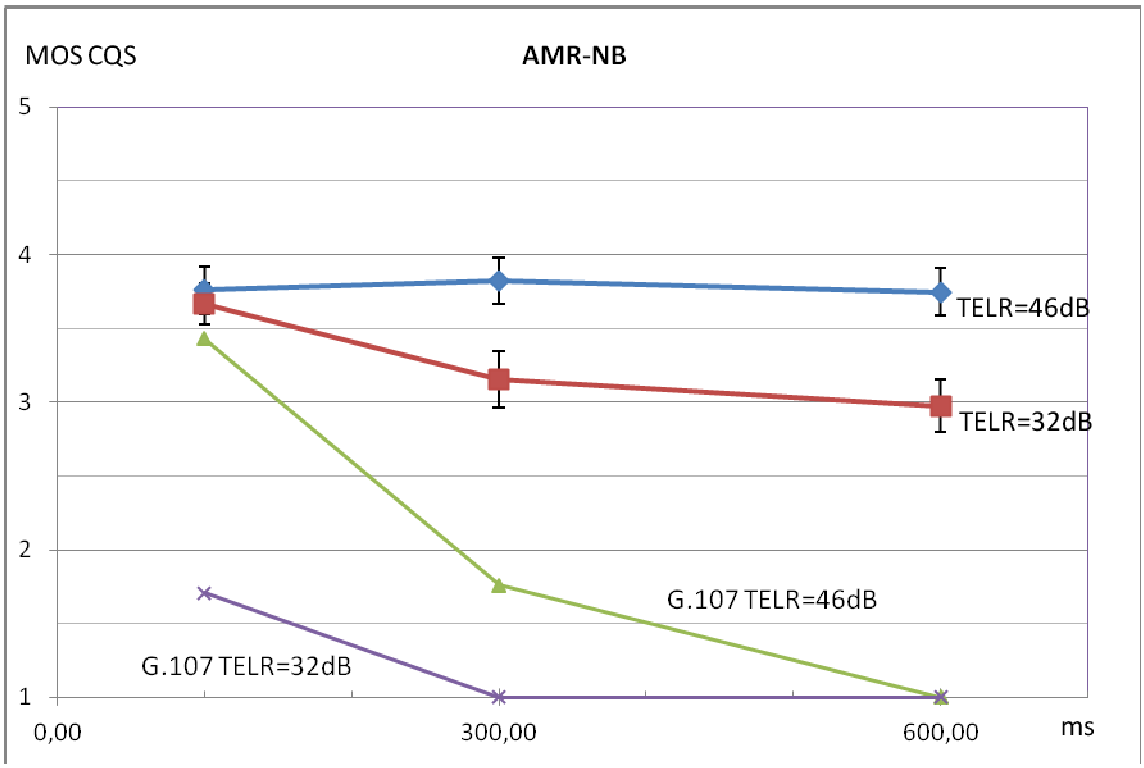


Figure 12: Subjective test results for AMR-NB coder and two tested TELR values (32 dB, 46 dB) including CI95% uncertainty intervals

NOTE 2: Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600 ms, the connecting lines are shown for informative purposes only.

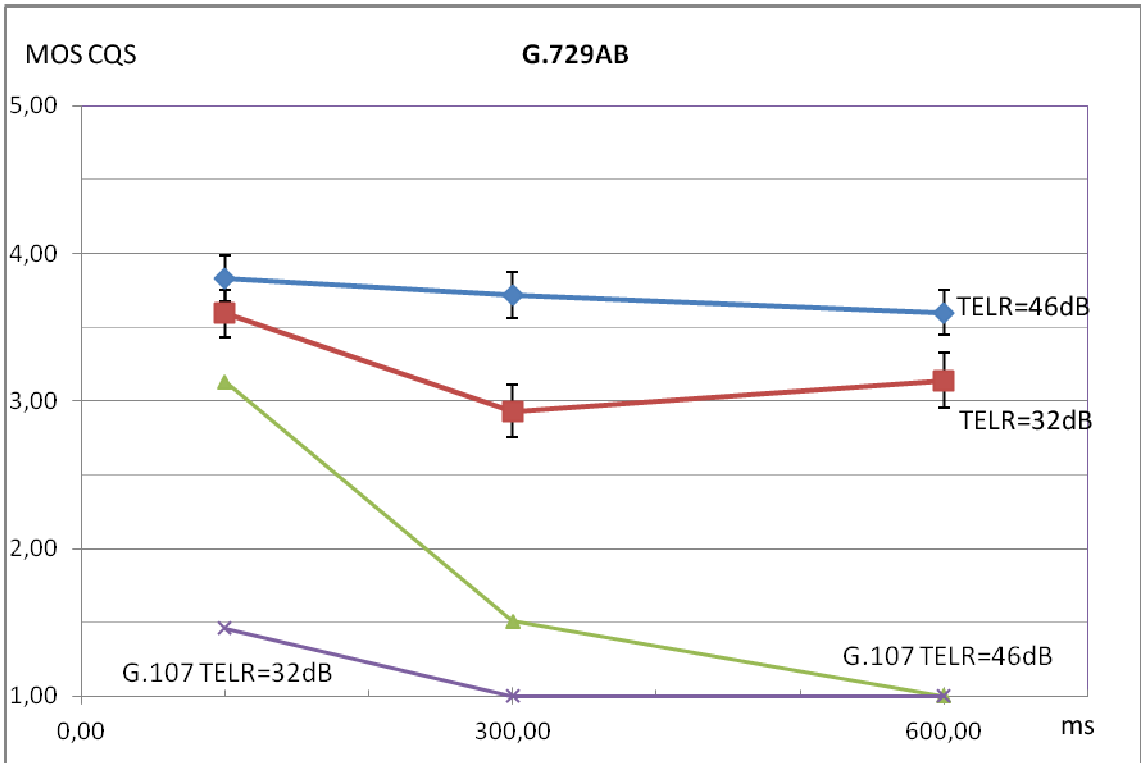


Figure 13: Subjective test results for G.729AB coder and two tested TELR values (32 dB, 46 dB) including CI95% uncertainty intervals

NOTE 3: Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600 ms, the connecting lines are shown for informative purposes only.

From figures 11-13 the following conclusions follows:

- For low echo condition of TELR = 46 dB, the subjective sensitivity to delay is significantly lower than as predicted by E-model. The typical difference between MOC-CQS for 100 ms and 600 ms is approximatively 0,5 MOS for low echo condition.

- For coders deploying higher perceptual compression (G.729AB) affecting the listening quality the MOS-CQS becomes non-monotonic with new local minima located (in our case) at 300 ms. Similar effects were reported by previous experiments by various labs, see [i.11].

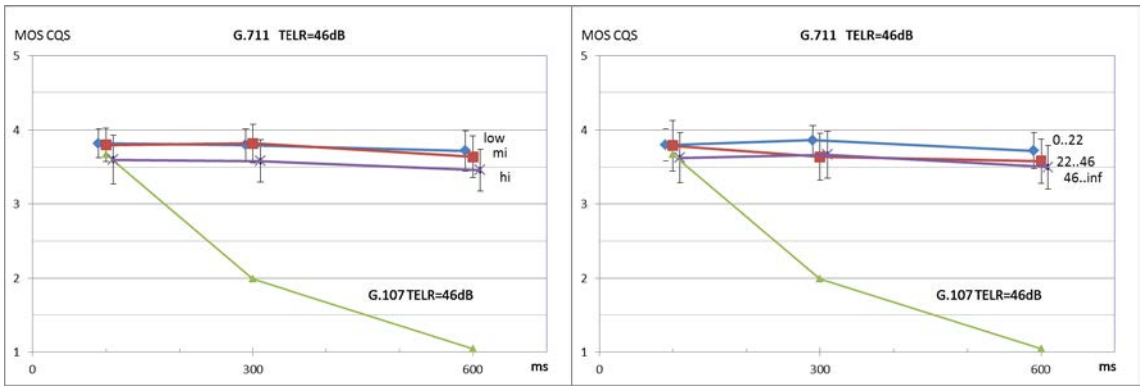


Figure 14: Subjective test results for G.711 coder and TELR = 46 dB

Figure 14 split for 3 different interactivity levels based either on conversational scenario (left) or on TAR analysis (right), including CI95% uncertainty intervals. Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600 ms, the connecting lines are shown for informative purposes only.

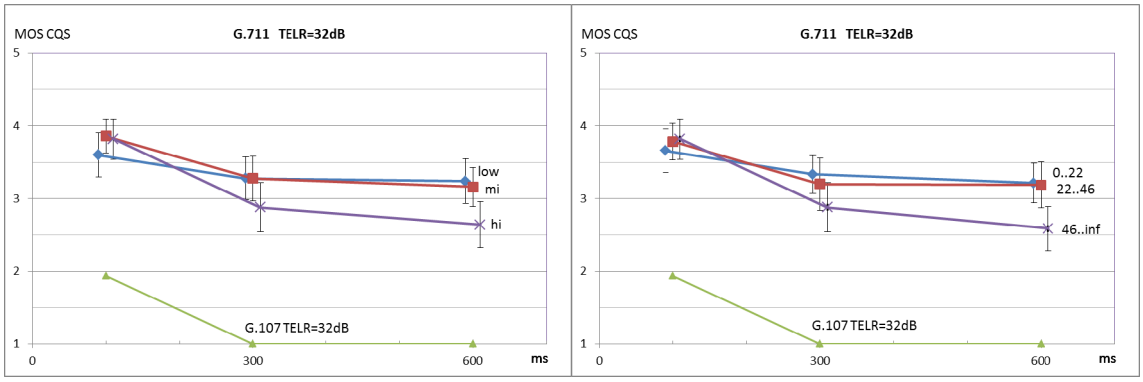


Figure 15: Subjective test results for G.711 coder and TELR = 32 dB

Figure 15 split for 3 different interactivity levels based either on conversational scenario (left) or on TAR analysis (right), including CI95% uncertainty intervals. Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600 ms, the connecting lines are shown for informative purposes only.

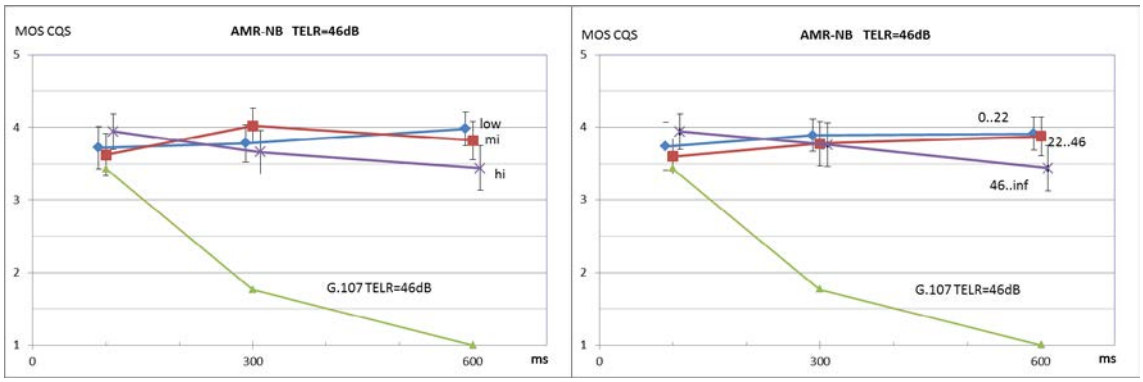


Figure 16: Subjective test results for AMR-NB coder and TELR = 46 dB

Figure 16 split for 3 different interactivity levels based either on conversational scenario (left) or on TAR analysis (right), including CI95% uncertainty intervals. Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600 ms, the connecting lines are shown for informative purposes only.

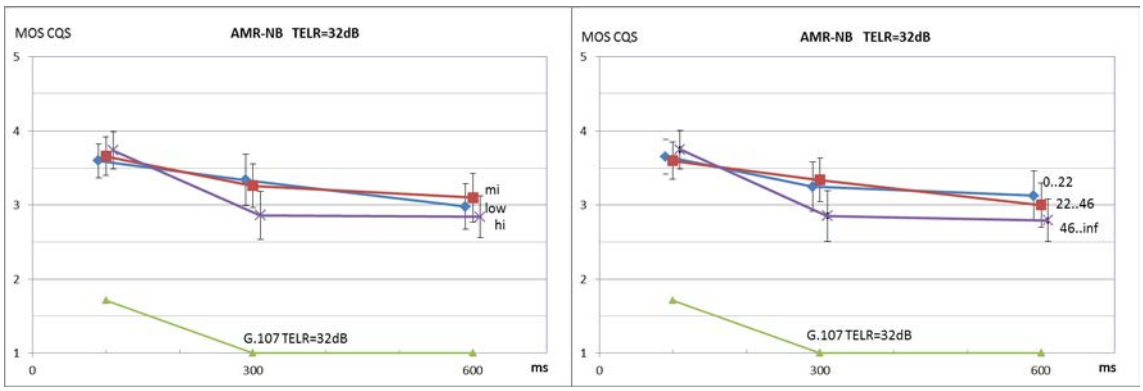


Figure 17: Subjective test results for AMR-NB coder and TELR = 32 dB

Figure 17 split for 3 different interactivity levels based either on conversational scenario (left) or on TAR analysis (right), including CI95% uncertainty intervals. Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600 ms, the connecting lines are shown for informative purposes only.

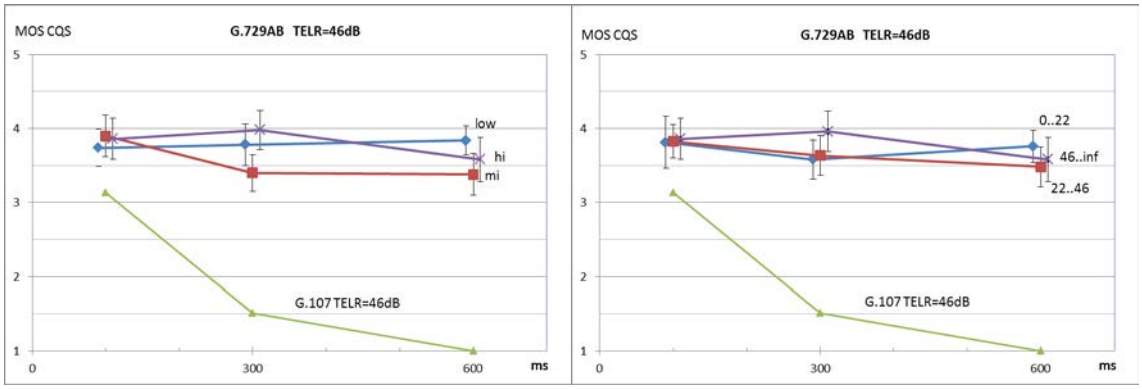


Figure 18: Subjective test results for G.729AB coder and TELR = 46 dB

Figure 18 split for 3 different interactivity levels based either on conversational scenario (left) or on TAR analysis (right), including CI95% uncertainty intervals. Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600 ms, the connecting lines are shown for informative purposes only.

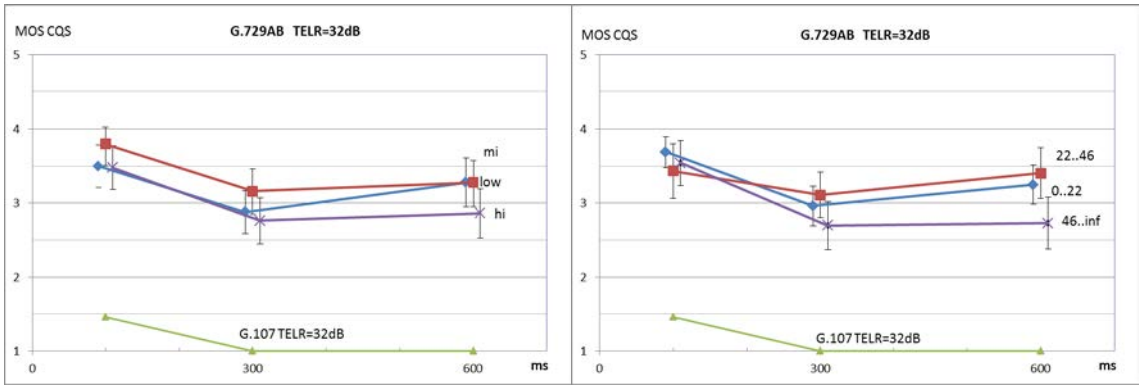


Figure 19: Subjective test results for G.729AB coder and TELR = 32 dB

Figure 19 split for 3 different interactivity levels based either on conversational scenario (left) or on TAR analysis (right), including CI95% uncertainty intervals. Corresponding E-model results are shown, too. The valid measurement points are highlighted by symbols and are located at positions 100, 300 and 600 ms, the connecting lines are shown for informative purposes only.

Language comparison

The comparison of results of tests performed in Czech language and in English language is depicted in figure 20. The results clearly indicate insignificant (0,2 MOS in average) systematic offset causing Czech tester 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 some other unknown reason.

Corresponding results of E-model, on the other hand, show significant differences for both language results, especially for delay values of 300 and 600 ms.

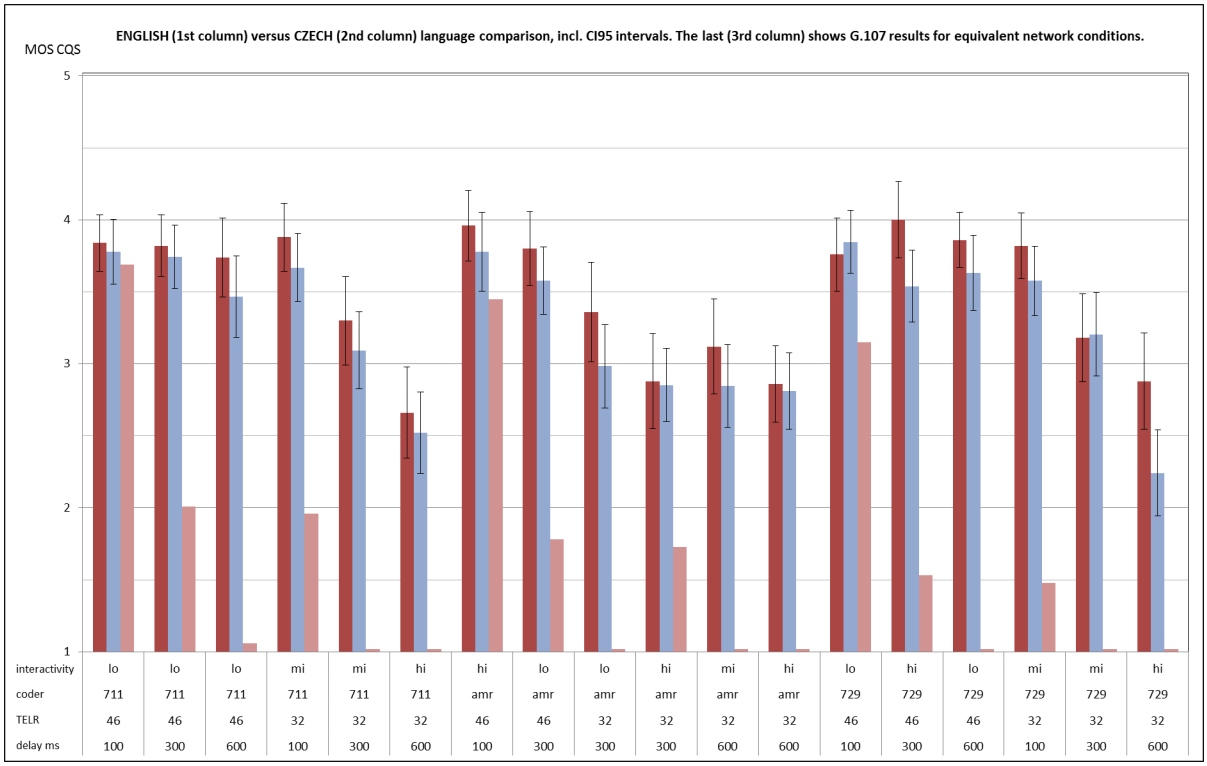


Figure 20: Language comparisons

First column (English results) and second column (Czech results) show systematic offset of cca 0,20 MOS indicating that Czech results are more critical (more demanding testers). However, CI95% intervals are overlapping for all tested cases. E-model result (third column) shows significant differences.

4.4 Computation and comparison of the different data resulting of the tests

Three different values of MOS-CQ are obtained for each combination of input parameters (codec, delay, echo level, etc.):

- E-model (G.107) output, recalculated for R to MOS scale (later referred to as "Emodel")
- MOS-CQS as obtained by subjective testing (later referred to as "MOS") with appropriate CI95% intervals
- Output of the predictor developed based on the MOS-CQS results as above (later referred to as "MCQP")

These different values are compared:

a) Comparison between MOS-CQS and E-model output:

- E-model versus MOS-CQS
- This analysis shows the differences between existing standardized estimator and subjective test results for each input parameter vector

b) RMSE* against E-model (root mean squared error with suppressed influence of subjective testing uncertainty). This analysis shows 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).

c) Comparisons between MOS-CQS and the developed predictor MCQP:

- = MCQP versus MOS-CQS
- This analysis shows the difference between the developed predictor and subjective test results

d) RMSE* against the developed predictor MCQP (root mean squared error with suppressed influence of subjective testing uncertainty). This analysis shows differences between the nearest CI95% interval border and the developed predictor MCQP (zero if the MCQP output is located within the CI95% interval).

e) Comparison between MCQP and E-model:

- = Emodel versus MCQP
- This analysis shows the differences between existing standardized estimator and the new developed predictor for each input parameter vector

5 The new model and the comparisons with other methods

This clause defines the model MCQP developed from the subjective tests and compares the graphs obtained with the model MCQP for variable values of the end-to-end delay, the talker echo, the conversation temperature (Talker Alternation Rate) and the listening quality with the results from other models such as E-Model, RMSE.

Since, the E-model as per Recommendation ITU-T G.107 [i.8] does not take into account effects of variable interactivity of conversations, it is important to consider the inter-relations as depicted in figure 21:

In principle, conversational interactivity can be defined as follows [i.1].

Conversational interactivity is a single scalar measure based on quantitative attributes of the participants' spoken contribution.

In order to fully understand a concept of conversational interactivity, we have to understand parametric conversation analysis, which is fully based on conversational model. According to [i.14] and [i.15], a two-way conversation can be divided into four different states, as illustrated in figure 21.

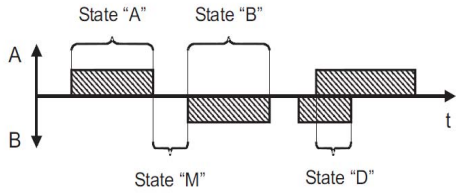


Figure 21: Illustration of conversation states (adopted by Hammer et al. [i.14])

States A and B denote that either person A or person B is talking and the other person does not speak. State M (mutual silence) reflects the situation that both persons are silent and state D (double talk) represents the case that both persons talk simultaneously.

There are three different models for conversational temperature, originally called speaker-alternation rate (called TAR in the present document), conversational temperature model and model based on the entropy of speech turns. The first one represents the number of speaker alternations, i.e. A-M-B, B-M-A, A-D-B and B-D-A per minute. A low TAR corresponds to low conversational interactivity and a high TAR corresponds to highly interactive conversation. A major advantage of TAR is that, given the conversation pattern (talk spurts); it can simply be calculated by counting the talker alternations and dividing them by duration of the call. On the other hand the second model, namely conversational temperature model is based on the conversational model described above. In fact, conversational temperature in this model is modeled by scalar parameter $\tau = \tau(t_A, t_B, t_M, t_D)$, which is a function of mean sojourn times, leading to a simple but efficient and intuitive one dimensional metric for describing conversational interactivity. The last model is called entropy model and is based on a speaker turn model and corresponds to the uncertainty about who of the participants is talking. It should be noted that this model was designed for multi-party conversations. It means that this model is not suitable for our purpose. Finally, it was shown in [i.14], [i.15] and [i.16] that TAR is most efficient metric providing meaningful representation of interactivity. That is a reason why this metric was taken into account.

5.1 Definition of the Model MCQP

The MCQP model is based on polynomial fit of subjective test data. Only English data were used for the derivation, and the model is valid for Czech data, too. Let us define the following parameters (E-model variables used where possible):

- *T*means one-way delay in ms
- *TELR* means Talker Echo Loudness Rating in dB
- *Ie* meansEquipment impairment factor (according to Recommendation ITU-T G.113 [i.19], appendix I) which can be derived from listening quality
- *TAR means*average Talker Alternation Rate of the call

Then, several products are derived:

- *P0*=1/(ABS((*T*/100)-3)+1)
- *P1=T*TAR*
- *P2*=(65-*TELR*)**T*
- *P3*=(65-*TELR*)**P0**(*Ie*+5)
- MCQP is then defined as
- ***MCQP*=A0+A1****T*+A2****Ie*+A3****TAR*+A4****P1*+A5****P2*+A6****P3*** *or **4.50**, whichever value is lower*

Should the previous formula provide value < 1, the estimator output is set to 1,00.

Where the constant values A0…A6 were obtained by polynomial regression and are listed in table 7:

A0=3,7368704882422000000
A1=0,0020530419466113700
A2=0,0112691465589648000
A3=0,0031369723762006100
A4=-0,0000220133980334889
A5=-0,0000809867387433772
A6=-0,0010251326366576700

It should be noted the model is quite simple as the number of parameters is currently limited. Further subjective data are needed to properly consider other important parameters e.g. 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.

5.2 Results from other Models and comparison with MCQP

5.2.1 Results from E-Model

The results delivered by E-model are reported in table 8 and in graphs shown in figures 11 to 19. Their comparison with subjective test data (MOS-CQS) is reported in the next clause.

5.2.2 Comparisons of E-Model with MOS-CQS and RMSE*

Comparisons of the E-Model and MOS-CQS results are shown in figure 22.

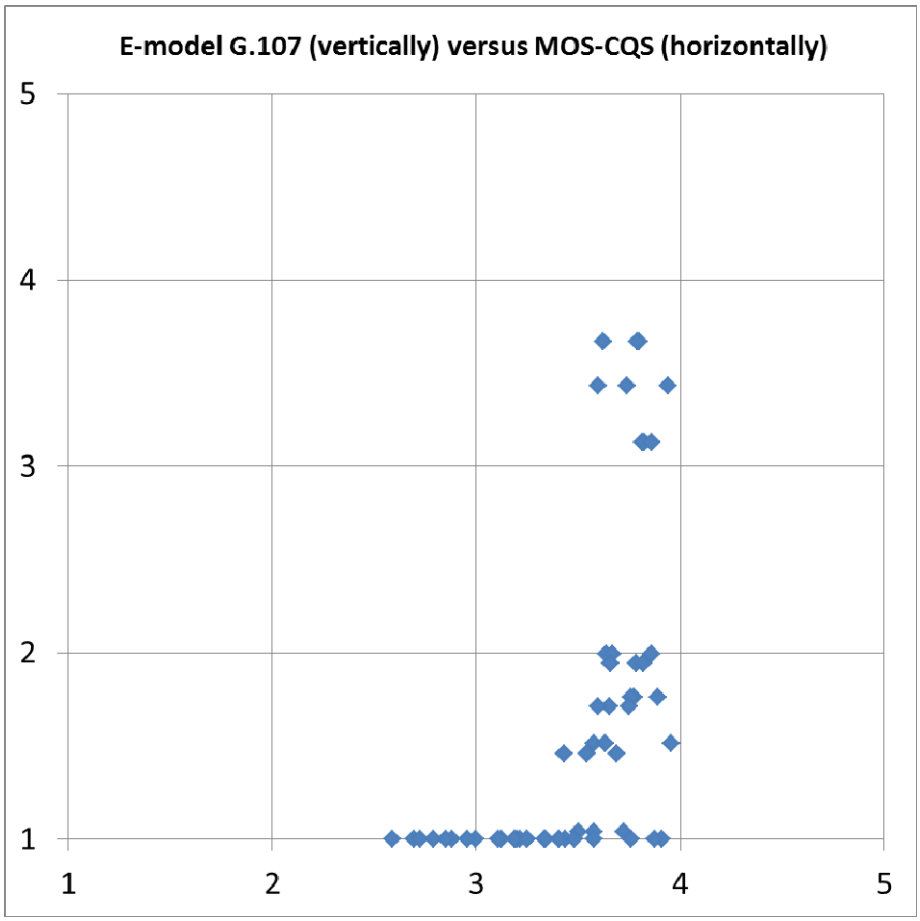


Figure 22: Comparison of E-model data (vertically) and subjective test data MOS-CQS (horizontally)

The comparison values (Pearson correlation coefficient, root mean square error rmse and root mean square error with suppresed influence of subjective test data uncertainty rmse* is reported in table 8 and is high.

Table 8: Comparison between E-model and subjective test data MOS-CQS

R	0,546
rmse	1,984
rmse*	1,722

5.3 Comparisons of the results from MCQP

The results of MCQP are shown in table 9. Also subjective test results and E-model results are shown.

Table 9: MCQP results

Test case	MOS CQS	CI95	E-model	MCQP
1	3,8	0,214	3,67	**3,8**
2	3,8	0,208	3,67	**3,8**
3	3,6	0,340	3,67	**3,8**
4	3,9	0,200	1,99	**3,7**
5	3,6	0,267	1,99	**3,7**
6	3,7	0,317	1,99	**3,6**
7	3,7	0,245	1,04	**3,8**
8	3,6	0,297	1,04	**3,8**
9	3,5	0,297	1,04	**3,4**
10	3,7	0,298	1,94	**3,6**
11	3,8	0,253	1,94	**3,6**
12	3,8	0,278	1,94	**3,7**
13	3,3	0,259	1,00	**3,3**
14	3,2	0,359	1,00	**3,3**
15	2,9	0,341	1,00	**3,2**
16	3,2	0,273	1,00	**3,1**
17	3,2	0,319	1,00	**3,1**
18	2,6	0,302	1,00	**2,8**
19	3,7	0,332	3,43	**3,8**
20	3,6	0,242	3,43	**3,8**
21	3,9	0,245	3,43	**3,8**
22	3,9	0,223	1,76	**3,7**
23	3,8	0,308	1,76	**3,7**
24	3,8	0,299	1,76	**3,5**
25	3,9	0,223	1,00	**3,8**
26	3,9	0,265	1,00	**3,8**
27	3,4	0,309	1,00	**3,4**
28	3,6	0,232	1,71	**3,6**
29	3,6	0,249	1,71	**3,6**
30	3,7	0,263	1,71	**3,7**
31	3,2	0,334	1,00	**3,2**
32	3,3	0,295	1,00	**3,2**
33	2,8	0,342	1,00	**3,1**
34	3,1	0,335	1,00	**3,1**
35	3,0	0,299	1,00	**3,1**
36	2,8	0,285	1,00	**2,7**
37	3,8	0,353	3,13	**3,8**
38	3,8	0,225	3,13	**3,8**
39	3,9	0,280	3,13	**3,9**
40	3,6	0,269	1,51	**3,6**
41	3,6	0,269	1,51	**3,6**
42	4,0	0,273	1,51	**3,5**
43	3,8	0,217	1,00	**3,9**
44	3,5	0,268	1,00	**3,8**
45	3,6	0,297	1,00	**3,5**
46	3,7	0,211	1,46	**3,6**
47	3,4	0,367	1,46	**3,6**
48	3,5	0,303	1,46	**3,7**
49	3,0	0,268	1,00	**3,1**
50	3,1	0,307	1,00	**3,0**
51	2,7	0,328	1,00	**2,9**
52	3,2	0,267	1,00	**3,2**
53	3,4	0,341	1,00	**3,1**
54	2,7	0,352	1,00	**2,8**

5.3.1 Comparisons of MCQP with MOS-CQS

Comparison of the MCQP and MOS-CQS results is shown in figure 23.

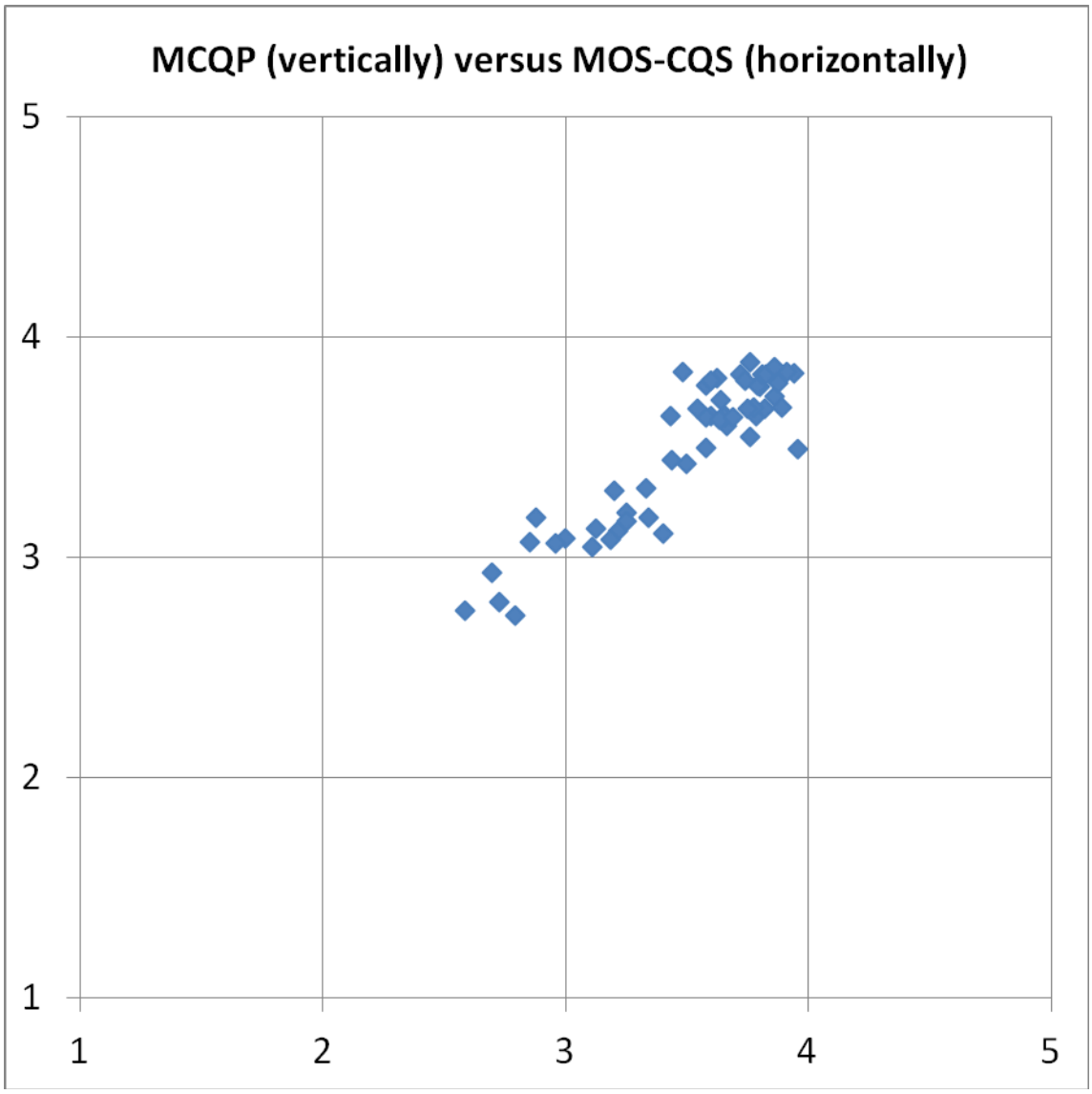


Figure 23: Comparison of MCQP (vertically) and subjective test data MOS-CQS (horizontally)

The comparison values (Pearson correlation coefficient, root mean square error rmse and root mean square error) with suppressed influence of subjective test data uncertainty rmse* is reported in table 10 and is close to zero.

Table 10: Comparison between MCQP and subjective test data MOS-CQS

R	0,911
rmse	0,148
rmse*	0,029

5.3.2 Comparisons of MCQP with E-Model

Due to the significant differences, the E-model and MCQP are models which results cannot be directly compared.

E-model is a much more 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 lack of fundamental inputs like interactivity (as 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 MCQP model, on the contrary, provides a good match with MOS-CQS (conversation quality). However, for this work it was rather simple and considers only parameters that were included in the subjective tests performed within the project.

6 Applications of MCQP

One of the major innovations of MCQP is to take into account the interactivity between the talkers and to introduce a new parameter TAR which is very important for the overall quality of speech conversations.

The graphs made available in the present document give the opportunity to technicians and managers to determine the expected quality of communications, taking into account delay, talker echo, listening quality and TAR.

Such a 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. However, to be wider applicable, it should take into account other parameters such as noise (effects of noisy environments and of noise cancellation).

6.1 Potential additional actions

This is based on the assumption that one major application area of the functions described here will be the inclusion of MOS-LQO values derived by drive testing (where no background noise is present) into the dimension of the E-Model.

However, since background noise is an important factor in quality perception by users, it is advisable to extend the present approach by background noise aspects.

The approach applies only for narrowband speech and should be extended at least for wideband and possibly to higher bandwidths.

The model can be expanded in order to become applicable to dynamic situations, considering IP-related impairments and impairments related to the radio links.

7 Conclusions

Since E-Model was developed, users are experiencing more and more communications, with increased delay, eg VoIP,

A low delay is needed for high quality conversations, especially for business calls that need a high interactivity, high intelligibility and significant comfort, in particular when their duration is long. For these types of calls the transmission planning with E-Model is still applicable.

The tests have also confirmed the need for high talker echo loss.

However for social calls, it appears that users are more tolerant to delay impairments (as long as there is no other type of impairment), as it was shown in previous tests such as described in TR 126 935 [i.18] or [i.13]. This is well taken into account by the new model MCQP.

Annex A:
Implementation Example of MCQP

This annex has an electronic attachment DTR_STQ-00189-v1.xls contained in archive tr_103121v010101p0.zip which accompanies the present document.

This spreadsheet provides conversion of Listening Quality scores MOS-LQO according to Recommendation ITU-T P.862.1 [i.9] into Equipment Impairment factors Ie for use in the E-model of Recommendation ITU-T G.107 [i.8]. The conversion is based on Appendix I of Recommendation ITU-T P.834 [i.27]. Furthermore two simple E-model calculations for different values of end-to-end delay are performed and the results are displayed as MOS-CQE scores.

Annex B:
Conversational scenarios in English

The conversational scenarios in English are contained in archive tr_103121v010101p0.zip which accompanies the present document.

Annex C:
Conversational scenarios in Czech

The conversational scenarios in Czech are contained in archive tr_103121v010101p0.zip which accompanies the present document.

Annex D:
Detailed session plans for subjective lab

The detailed session plan, which describes the actual procedures during the conduction of the conversational tests is intended to be the handbook for the personnel in the subjective lab. The detailed session plan will be produced after Milestone C was approved and all parameters and setting were approved by STQ. The following section contains an example of a session plan for illustrative purposes.

D.1 Session plans

The following tables contain the instructions for the subjects and one example of the randomized composition of tests.

INSTRUCTIONS TO SUBJECTS In this experiment we are evaluating systems that might be used for telecommunication services. You are going to have a conversation with another user. The test situation simulates communication between two pieces of equipment under test. During the test you have to talk with your distant partner, and to precisely follow the instructions given for the conversation scenario. The aim of the conversation is to fulfil the allocated task. However, if you do not succeed, the conversation will be ended after app. 2 minutes. After the completion of each call conversation, you will have to give your opinions on the quality by answering to two questions.
From then on you will have a break approximately every 45 minutes. The test will last a total of approximately 3 hours. Please do not discuss your opinions with other listeners participating in the experiment. Thank you!

At the end of each conversation, you will answer two questions:

"How do you assess the conversation interactivity with the other person"				
No special effort required	Minimal effort required	Moderate effort required	Considerable effort required	Severe effort required
"What is your opinion of the connection you have just been using?"				
Excellent quality	Good quality	Fair quality	Poor quality	Bad quality

INSTRUKCE PRO ÚČASTNÍKY TESTU V tomto experimentu vyhodnocujeme systémy, které mohou být použity pro telekomunikační účely. Budete konverzovat s druhým účastníkem experimentu. Testované situace představují různá použití testované technologie. Je zapotřebí, abyste během testu co nejvíce konverzoval se svým protějškem a co nejpřesněji následoval instrukce ve scénářích, které máte k dispozici. Cílem je splnit zadanou úlohu. Nicméně pokud se Vám to nepodaří, konverzace bude ukončena instruktorem po cca 2 minutách. Po skončení hovoru zodpovíte vždy dvě otázky, týkající se jeho kvality.
Po každých 45 minutách následuje přestávka nezbytné délky. Celý test včetně této instruktáže nezabere víc než 3 hodiny. Prosíme, během testu nediskutujte Váš názor s jinými účastníky experimentu. Děkujeme!

Po ukončení každé konverzace zodpovíte následující dvě otázky:

"Jak hodnotíte interaktivitu vašeho rozhovoru?"				
Nevyžadovalo žádné zvláštní úsilí	Vyžadovalo pouze minimální úsilí	Vyžadovalo střední úsilí	Vyžadovalo značné úsilí	Vyžadovalo maximální úsilí
"Jaký je váš názor na kvalitu spojení, které jste právě používali?"				
Excelentní	Dobrá	Střední	Špatná	Velmi špatná

Example of subjective test session of STF436 (compliant to P.805)

	Instruction	Session 1	Break	Session 2	Break	Session 4
Number of conversations		19 (incl. practice)		18		18
Test plan scenarios		practice (5min) 37,38,19,23,39,34 1,21,36,18,22,51 6,32,8,44,30,27		29,2,20,12,24,28 1,53,52,42,4,5 7,46,41,9,48,10		25,45,26,16,11,15 33,13,54,40,43,47 14,35,50,17,49,3
Time	15 min	50 min	10 min	45 min	10 min	45 min

Table D.1: summary of the test conditions

Subjects	48 English 16 Czech	Untrained
Groups	32	2 subjects/group
Rating scales	2	
Objective of the test		Delay, echo, coder, interactivity level, language
Communication system	Types	Handset audio terminal conform.to ES202737
Communication environment		Floor noise:Hoth 40dB SPL(A)

Figure D.1 providing the histogram of the age of the subjects having performed the tests.

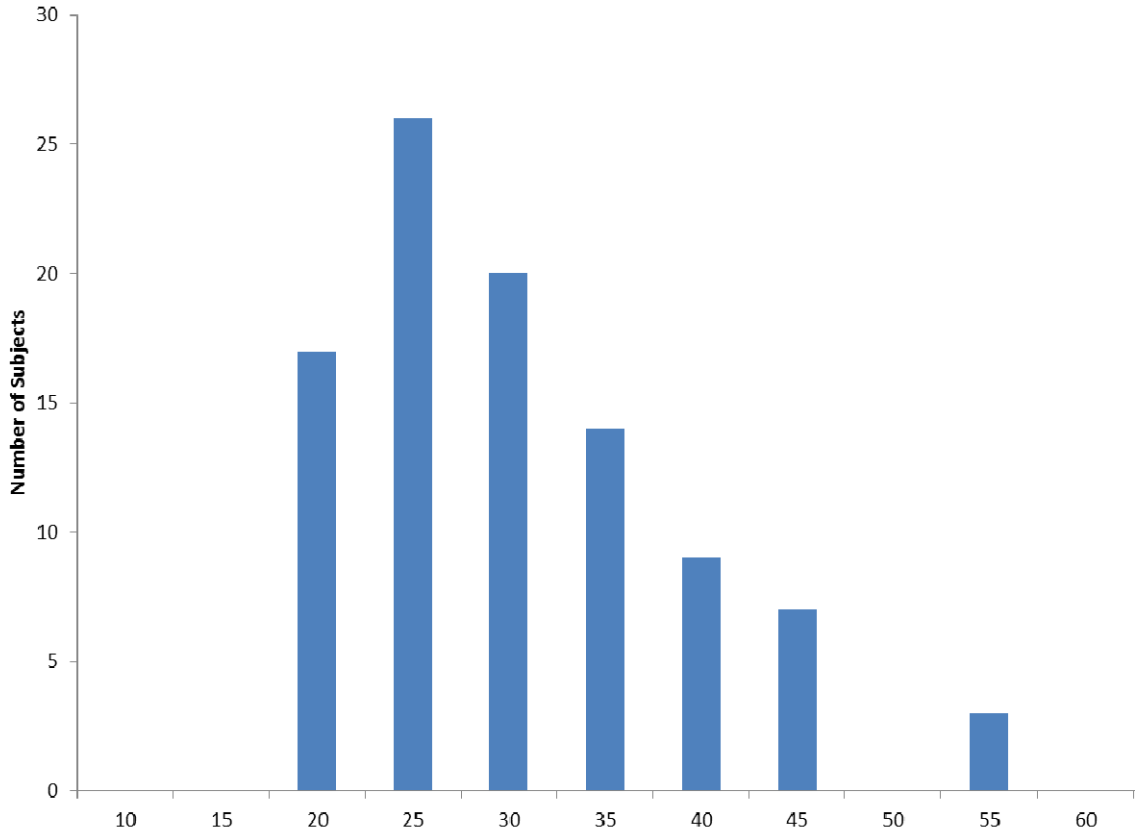


Figure D.1: providing the actual objective listening quality of the end-to end transmission chain

MOS-LQO were computed by PESQ (according to Recommendation ITU-T P.862) between the electrical ends of the transmission chain (just before the terminal transducers) and are reported in the table D.2.

Table D.2

Coder	G.711	G.729	AMR-NB
MOS-LQO	4.13	3.78	3.81

History

Document history		
V1.1.1	March 2013	Publication