INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION
STANDARDIZATION SECTOR

STUDY PERIOD 2001-2004

COM 12 – C 55 – E

July 2003

English only

Original: English

Question: 7/12

STUDY GROUP 12 – CONTRIBUTION 55

Source: KPN, Netherlands

Title: A subjective/objective test protocol for determining the conversational quality of a voice link

Abstract

Currently Question 7 is working on a number of subjective assessment methods (P.ASPD, P.NSA, P.PAC, P.VOAD [1]) that allow accurate assessment of different aspects of the overall conversational quality of a voice link. An integrated method for assessing overall conversational quality, that takes into account listening quality (how do I perceive the other), talking quality (how do I perceive myself) and interaction quality (how easy can we interrupt each other, double talk distortions) is however outside the scope of these new recommendations. This contribution proposes a subjective/objective test protocol for determining this overall conversational quality of a voice link. Expert listeners judge the quality of pre-recorded speech and background noise under different circumstances. The protocol assesses all three components of conversational quality: listening, talking and interaction quality. It uses predefined anchoring conditions and an objective end-to-end delay measurement that is mapped to subjective quality. The method is focused on pinpointing the weak parts of a voice link.

Contact: John G. Beerends
TNO Telecom, Netherlands

Tel: +31 70 446 2644
Fax: +33 70 446 3477
Email: j.g.beerends@telecom.tno.nl

1 Introduction

In the last decades several new technologies have been introduced that degrade the conversational quality of a voice link in a complicated manner. A general trend is that the end-to-end delay is increasing due to signal processing (noise suppression, echo-cancellation), speech encoding/decoding, packetisation and error protection. Especially for mobile links the end-to-end delay is such that even weak echo's may become audible. Furthermore mobile networks sometimes employ echo/noise suppressors that may interact with the echo/noise suppressors that are found in mobile handsets. This may cause audible echo artifacts and audible background noise switching, sometimes leading to unacceptable low end-to-end conversational quality. This has urged the development of special tests to assess the overall conversational quality in an efficient, reproducible manner that allows pinpointing of the major problems. Currently Question 7 is working on a number of subjective assessment methods (P.ASPD, P.NSA, P.PAC, P.VOAD [1]) that allow accurate assessment of the different aspects of the overall conversational quality of a voice link. However, an integrated method for assessing overall conversational quality, that takes into account listening quality (how do I perceive the other), talking quality (how do I perceive myself) and interaction quality (how easy can we interrupt each other, double talk distortions) is outside the scope of these new recommendations.

This contribution proposes a subjective/objective test protocol for determining the overall conversational quality of a voice link. It separately assesses the three different aspects that attribute to the conversational quality for both the A and B side of the link [2]:

- One-way listening quality, how does A perceive the voice and background noise of the B-side (and vice versa)
- One-way talking quality, how does A perceive his own voice and background noise of the B-side during talking (and vice versa)
- Two-way interaction quality, how easy can A interrupt B (and vice versa) and are there disturbing artifacts audible during double talk

The first one is related to distortions as introduced by the coding and transport of the speech signal and background noise. It can be assessed in a listening only experiment. The second one is related to echo's, background noise switching and sidetone distortion and can be assessed in a talking only experiment. The third one is dominated by the end-to-end delay and the double talk capabilities of the system under test. Often echo's and background noise switching become audible under double talk conditions.

For each of three modalities of a conversation, listening, talking and interacting, objective assessments can be carried out. For one-way listening ITU-T recommendation P.862 was developed for end-to-end measurements between electrical interfaces [3], [4], [5]. With the signal processing that takes place in the mobile handset electric domain end-to-end quality measurements are becoming more difficult, because the correct coupling points are seldom available. Currently P.862 is extended towards the acoustic domain to cope with this problem [6], [7]. For one-way talking and two-way interaction work has been carried out but for most distortions the relations between objective and subjective measurements are not clear yet [8], [9], [10], [11]. For two-way interaction the influence of delay is important. It can be measured objectively and taken into account in the calculation of an overall conversational speech quality. Recommendation G.107 [9] and Annex A of P.562 [10] describe methods that use this approach. However with these approaches subjective data has to be gathered when new types of distortion occur in the network, e.g. when a new speech codec and/or noise suppression system is introduced. Taking into account these new types of distortion requires a continues update of these models and currently none of them take into account double talk distortions and background noise switching during talking. In the design of subjective experiments the main difficulty is that the subjective score is highly dependent on the experimental context, especially the amount of switching between partners significantly influences the final outcome of the assessment.

This contribution proposes a combined subjective/objective test protocol that determines the conversational quality of a voice link by using expert listeners that run a number of tightly controlled experiments. The protocol uses a HATS (Head And Torso Simulator) / Loudspeaker combination for defining an exact controlled B-side of the voice link.

The complete subjective/objective test protocol is split in six steps:

a) One-way listening, speech quality (this part can be replaced by using the new objective method currently under development within ITU-T SG 12 and known as P.AAM, the Acoustic Assessment Model [6], [7])

b) One-way talking, side tone/echo quality

c) One-way talking, echo quality, maximum criticality

d) One-way talking, background noise quality

e) Two-way interaction, full/semi-full/half duplex quality

f) Two-way interaction, delay, double talk capabilities

The test protocol uses no background noise listening only test. During the development of the protocol it turned out to be better to combine this with a talking test. It should also be noted that the room in which the telephone sets are placed may have a significant influence on the final quality judgment, especially if one of the sets uses hands-free operation at a large microphone distance. Furthermore the method uses experts that give opinion scores that are anchored by predefined distortions. A minimum of two experts is required that show a correlation above 0.9 on a conversational quality test having a wide range of distortions. Each of the six tests is carried out by at least two experts at both the A and B-side giving a MOS for each of the six tests. The end-to-end conversational quality of the speech link is defined as the minimum over the twelve MOS scores in the six different tests (a, b, c, d, e, f of the A- and B side of the connection).

2 The subjective/objective test protocol for determining the conversational quality of a voice link

In the test as described below the quality of a connection between A and B is measured at the A side of the connection by the expert listener using a HATS on the B-side. The HATS plays speech recordings that are made in a dry acoustic environment at close distance from the mouth (about 10 cm) using a microphone that delivers a natural spectral balance of the voice at this distance when played back over the artificial mouth of the HATS. The background noise level in the recording environment should be below 30 dBA, the reverberation time below 0.5 seconds for frequencies above 300 Hz and below 1 second for frequencies above 50 Hz. If no HATS with high quality speech recordings are available the test can be carried out using a real voice on the B-side.

Speech material that is played over the HATS should be recorded at two levels, normal and loud, without adjustment of the recording level. If voice dependency may play a role at least five different voices should be used, two female, two male and one child for both the expert and the played back speech. Normal meaning full sentences are used.

For the background noise simulation at the B-side a set of minimal four loudspeakers should be used in the play back of a mono or stereo recording of background noises.

When the test is completed the same test must be repeated at the B-side of the connection with the role of A- and B-side interchanged.

When no HATS / loudspeaker set up is available the protocol can be used with a "live" voice at the B-side using natural background noises.

The following six tests are defined:

a) One-way listening, speech quality. Natural speech is played back over the HATS at the B-side. The two levels that are recorded should be played back such that the natural sound pressure level is reproduced for the voice. Linear (timbre), non-linear and level distortions should all be taken into account. The ITU-T P.800 [11] ACR listening quality scale (Absolute Category Rating) is used. The P.800 listening quality scale is anchored in the following way:

5 = excellent = the speech quality is essentially the same as the natural voice at the B-side

4 = good = PSTN quality, the speech signal is only distorted by a linear narrow band filter using the combined modified ITU-T IRS send and receive filtered speech [13]

3 = fair = speech quality that is equivalent to modified ITU-T IRS send filtered speech degraded by modulated noise at a level of 20 dB [14] and played back over a modified IRS receive handset

2 = poor = speech quality that is equivalent to modified ITU-T IRS send filtered speech degraded by modulated noise at a level of 10 dB and played back over a modified IRS receive handset

1 = bad = speech quality that is equivalent to modified ITU-T IRS send filtered speech degraded by modulated noise at a level of 0 dB and played back over a modified IRS receive handset

b) One-way talking, side tone/echo quality. The expert talks at the A-side and listens for echo and side tone distortion. The reference is the natural side tone. Linear, non-linear, level distortions and echo should be taken into account. Speech should be produced at two levels, normal and loud. The ITU-T P.800 [12] DCR opinion scale (Degradation Category Rating) is used anchored in the following way:

5 = no echo and/or side tone distortion audible

4 = audible but not annoying; anchor delay 20 ms (mean one-way delay), TELR 25 dB

3 = slightly annoying; anchor delay 50 ms, TELR 25 dB

2 = annoying; anchor delay 150 ms, TELR 30 dB

1 = very annoying; anchor delay 300 ms, TELR 30 dB

c) One-way talking, echo quality, maximum criticality. The expert talks at the A-side and listens for echo while at the B-side the telephone horn position is changed from default to a hard surface with the loudspeaker side facing this surface. Speech should be produced at two levels, normal and loud. The ITU-T P.800 [12] DCR opinion scale is used, but only the top three categories are allowed using relaxed scaling criteria, see **b)**:

5 = audible echo but not annoying

4 = slightly annoying

3 = annoying

d) One-way talking, background noise quality. The expert talks at the A-side while at the B-side a set of background noise audio files (office, babble, car noise) is played. The recordings should contain a slowly decreasing and/or slowly increasing background noise level in order to assess the system under test under a wide range of background noise conditions. The rate of change should lie in the order of about 20 dB/s and have a duration of about 20 seconds. Speech should be produced at two levels, normal and loud. Most important distortions that occur are time clipping resulting in background noise switching and changes in noise suppression resulting in unnatural level and timbre variations especially during the starts/stop of a speech spurt at the A-side. The ITU-T P.800 [12] DCR opinion scale (Degradation Category Rating) is used in the assessment with 5 = no suppression of the background noise and other annoying changes audible. The same anchoring is used as under **b)**.

e) Two-way interaction, full/semi-full/half duplex quality. Over the HATS at the B-side a recording is played of a voice that counts fast and continuously: *1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5…etc (note the two different speech levels),* the expert at the A-side speaks with pauses using single letters and short consonant vowel consonant (cvc) words*: a, pause, <cvc>, pause, b, pause, <cvc>, pause, …etc.* In the assessment the quality of the continuous counting voice from the B-side (*1, 2, 3, 4, 5*) is judged together with the quality with which the own voice is perceived. During double talk no distortion and echo from one's own voice should be audible. Speech should be produced at two levels, normal and loud. The ITU-T P.800 [12] DCR opinion scale (Degradation Category Rating) is used in the assessment with 5 = no semi-full/half duplex degradation audible. The same anchoring is used as under **b)**.

f) Two-way interaction, delay. Objective measurement preferably with speech, DCR rating = 5 – 0.01* mean one-way delay [ms]. This quality rating is more stringent towards delay than the one given in recommendation G.107 [9] and P.562 [10] and is based on the subjective results as given in recommendation G.114 [15]. For delays above 400 ms the service is no longer considered to be a telephony service. If no objective measurements can be made an alternative procedure can be used that is based on an interactive counting protocol. In this test two experts take turn using the following test protocol: Expert A starts the procedure with the counting word "one" while at the same time he starts a timer, next expert B counts "two" after receiving "one", etc until expert B counts "ten" after which expert A stops the timing. This procedure is calibrated in a face-to-face test until the START STOP time ***T*** is about 4.5 seconds (4500 ms). Over the voice link the mean one-way delay is estimated by ***0.1*(T - 4500) ms***.

Each of the six tests is carried out by at least two experts at both the A and B-side giving a MOS for each of the six tests. The end-to-end conversational quality of the speech link is defined as the minimum over the twelve MOS scores in the six different tests (a, b, c, d, e, f of the A- and B side of the connection).

The MOS values for PSTN are 4.0 for tests a, c, d and 5.0 for tests b, e, f, for both sides of the connection, resulting in a conversational MOS score of 4.0.

The MOS values for GSM show more variance, in The Netherlands all six tests gave a score of about 3.0; resulting in a conversational quality of about 3.0.

3 Conclusion

A fast and simple method for the measurement of the conversational quality of a voice link is presented that provides stable, reproducible, results. It uses a combined subjective/objective test protocol with prerecorded speech that is played over a HATS and prerecorded background noise that is played over loudspeakers. Expert listeners assess listening quality (how do I perceive the other), talking quality (how do I perceive myself) and interaction quality (how easy can we interrupt each other, double talk distortions). The method uses predefined anchoring conditions and an objective end-to-end delay measurement that is mapped to subjective quality. The final result of the assessment is a single number that represents the overall conversational quality of a voice link. The method is focused on pinpointing the weak parts of a voice link and can be carried out in less than half an hour. For special purposes a limited subset of the test can be used.

4 References

[1] P. Usai, "Report of Question 7/12 Rapporteur Meeting (Paris, 19-20 May 2003)," Contribution COM 53 to ITU-T Study Group 12, May 2003.

[2] D.L. Richards, *Telecommunication by speech,* London Butterworths, 1973.

[3] ITU-T Rec. P.862, "Perceptual Evaluation Of Speech Quality (PESQ), An Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," International Telecommunication Union, Geneva, Switzerland (2001 February.).

[4] A. W. Rix, M. P. Hollier, A. P. Hekstra and J. G. Beerends, *"PESQ, the new ITU standard for objective measurement of perceived speech quality, Part 1 - Time alignment,"* J. Audio Eng. Soc., vol. 50, pp. 755-764 (2002 Oct.).

[5] J. G. Beerends, A. P. Hekstra, A. W. Rix and M. P. Hollier, *"PESQ, the new ITU standard for objective measurement of perceived speech quality, Part II - Perceptual model,"* J. Audio Eng. Soc., vol. 50, pp. 765-778 (2002 Oct.).

[6] J. G. Beerends, J. Berger, A. W. Rix, "Preliminary results for the ITU-T acoustic speech quality models," Delayed Contribution D.109 to ITU-T Study Group 12, January 2003.

[7] A. W. Rix, J. Berger and J. G. Beerends, "*Perceptual quality assessment of telecommunications systems including terminals,"* presented at the 114^{t}th Convention of the Audio Engineering Society*, J. Audio Eng. Soc. (Abstracts),* (2003), convention paper 5724 (equivalent to KPN Research report 33124).

[8] S. R. Appel and J. G. Beerends, *"On the quality of hearing ones own voice,"* J. Audio Eng. Soc., vol. 50, pp. 237-248 (2002 April) (equivalent to KPN Research publication 00-32300).

[9] ITU-T Rec. G.107, "The E-model, a Computational Model for Use in Transmission Planning," International Telecommunication Union, Geneva, Switzerland (1998 Dec.).

[10] ITU-T Rec. P.562, "Analysis and interpretation of INMD voice-service measurements," International Telecommunication Union, Geneva, Switzerland (2000 May).

[11] ITU-T Rec. P.340, "Transmission characteristics and speech quality parameters of hands-free terminals," International Telecommunication Union, Geneva, Switzerland (2000 May).

[12] ITU-T recommendation P.800*, Methods for subjective determination of transmission quality,* August 1996.

[13] ITU-T Rec. P.830, "Subjective performance assessment of telephone-band and wideband digital codecs," International Telecommunication Union, Geneva, Switzerland (1996 Feb.).

[14] ITU-T Rec. P.810, Modulated Noice Reference Unit (MNRU), February 1996.

[15] ITU-T Rec. G.114, "One-Way Transmission Time", International Telecommunication Union, Geneva, Switzerland (1996 Feb.).