Conversational quality assessment of advanced video conferencing systems

John Beerends, Niels Neumann
TNO
Netherlands

Presented at the 4th International Conference of the Acoustical Society of Nigeria
October 2020

1 Abstract

With the covid-19 pandemic outbreak the number of users of advanced video conferencing systems such as Zoom, Microsoft Teams, Skype, FaceTime etc., has increased significantly. The end-to-end conversational speech quality of these systems is in many cases below expectation, mainly because of large delays and the underestimated effect of degradations caused by the interaction of the microphone and loudspeaker with the send and receive room. This paper proposes a practical quality assessment approach based on ITU recommendations.

2 Introduction

In the last decades several advanced video conference technologies have been introduced that degrade the conversational quality of a voice link in a complicated manner. With the Covid-19 pandemic outbreak, the number of users of these conferencing systems, such as Zoom, Microsoft Teams, Skype, FaceTime etc., has increased significantly. Despite significant advances in signal processing, that allow high quality conversations over these systems, the end-to-end conversational speech quality is in many cases below expectation. This is mainly caused by large delays and the underestimated effect of degradations caused by the interaction of the microphone and loudspeaker with the send and receive room. In order to have a high quality experience, the video and audio have to be in sync leading to an increased delay in these systems due to the extra processing time needed for the video coding. This increased delay in turn leads to a lower conversational quality, especially in cases where partners in the conversation often interrupt each other.

Furthermore, users will try to get the video image correctly framed by using a self-view image and thus taking distance from the camera whereby in most cases the microphone distance will also increase beyond the normal close distance used in telephony. This increased distance degrades the conversational quality more than one would expect because the binaural decolorization is not able to compensate the room acoustic degradations. In “live” situations, this binaural decolorization maintains high quality even at large mouth/loudspeaker to ear distances. This leads to a hollow sound quality and decreased intelligibility in online conferencing. This problem can be solved by using a close coupled microphone in combination with a headphone. A disadvantage of the headphone is that it decreases the natural side tone at one’s ear and that it blocks the natural back ground sound from one’s own environment, leading to an uncomfortable talking quality. A correct, direct, feedback from microphone to headphone can restore the side tone degradation, but is rarely correctly implemented.

When the video and audio signal are correctly synchronized, the intelligibility degradation in the speech path can be partly compensated by the lip-sync information.

A problem with modern video conferencing is that it uses the Internet, forcing the use of advanced coding schemes, packetization, buffering and error protection technologies that increase the end-to-end delay and sometimes lead to short signal interruptions. Packet loss can cause short signal interruptions that can lead to situations where the overall perceived end-to-end quality is high, but the speech is temporarily not correctly perceived and the intelligibility is lower than one expects based on the speech quality.

The increased usage of audiovisual links in communication has urged the development of special tests to assess the overall conversational quality of these links in an efficient, reproducible manner that allows pinpointing of the major problems. This method should take into account all aspects that contribute to the overall perceived conversational quality. These aspects include the listening quality and intelligibility (how do I perceive the other), talking quality (how do I perceive myself), the interaction quality (how easy can we interrupt each other, double talk distortions) and the video quality (lip sync).

If an audiovisual presentation is multi/broadcasted and participants do not use interruptions, the impact of delay is marginal and large buffers can be used, thus allowing to have a significantly higher audiovisual streaming quality. However, as soon as interactions between meeting participants are foreseen, a low delay is a necessity in order to have an acceptable conversational quality. This paper is focused on the conversational quality of video conference links between two participants, however the protocol naturally extends towards group communication.

One can assess the conversational audio visual quality with subjective tests [1], [2], however the resulting subjective score of such tests is highly dependent on the experimental context. Especially the amount of switching between partners significantly influences the final outcome of the assessment. Furthermore, if insight is needed into the underlying causes of a low conversational quality, a better approach is to break down the overall quality into the different main components. In this paper a subjective/objective test protocol is given that separately assesses the four different main quality aspects that contribute to the overall conversational quality for both the A and B side of the link [3]:

- One-way listening quality and intelligibility, 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 (side tone/echo's) and background noise switching 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 artefacts audible during double talk
- One-way viewing quality, how does A perceive the video quality of the B-side (and vice versa), including lip-sync assessment

The first quality aspect is related to distortions introduced by the microphone/room, the coding and transport of the speech signal and by the play back loudspeaker/room. It can be assessed in a listening-only experiment.

The second quality aspect involves the side tone path from one's own mouth to ear and becomes especially important when headphones are used. With headphones, the natural side tone path should be simulated by feeding the correct signal from one's microphone back to the headphone. In the case of handsfree we have a natural acoustic side-tone path but echo's may

degrade the talking quality. It also involves the perceived switching of the background noise of the B-side which may occur during voice onsets on the A-side. Talking quality can be assessed in a talking-only experiment.

The third quality aspect, the two-way interaction quality, 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. This two-way interaction quality aspect is difficult to quantify and requires two-way speech activity on the link.

Finally, the video-related quality aspect quantifies the contribution of the video signal to the overall conversational quality. This aspect is of minor importance as one can achieve high conversational quality without a video signal. In the case of low video quality where the audio and video are not in lip-sync, a still picture of high quality may be preferred instead, allowing for a lower end-to-end audio delay.

For each of the four quality aspects, objective assessments can be carried out. For one-way listening ITU-T recommendation P.863 was developed for end-to-end speech quality measurements [4], [5], [6]. For one-way talking and two-way interaction, development work has been carried out, but for this class of distortions, the relations between objective and subjective measurements are not clear and highly dependent on the experimental context [7], [8], [9]. 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 [8], [9], [10]. However, especially double talk distortions and background noise switching during talking are difficult to assess objectively. Video quality can be assessed using ITU-T recommendations J.247 and J.341 ([11], 12]), but objective lip-sync quality is difficult to measure.

This paper proposes a combined subjective/objective test protocol that determines the conversational quality of a voice link by using trained listeners that run a number of controlled experiments. For an exact controlled B-side voice link, the protocol uses a HATS (Head And Torso Simulator) or loudspeaker at the B-side. If no HATS or loudspeaker is available, a second trained subject can be used in the test. The complete subjective/objective test protocol is split in five steps:

a) One-way listening, speech quality

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

c) Two-way interaction, full/semi-full/half duplex quality, double talk capabilities

d) Two-way interaction, impact of delay

e) Video impact assessment

Note that the rooms in which the audio sets are placed have a significant influence on the final quality judgment, especially if one of the sets uses hands-free operation at a large microphone distance. The method uses experts that give opinion scores that are anchored by the judgement of predefined distortions. Each of the five tests is carried out by two experts at both the A and B-side giving a MOS for the first four aspects and a MOS correction for the video impact. The end-to-end conversational quality of the videoconference link is determined by the worst contributing factor.

Note that the proposed test protocol only provides a rough estimate of the conversational quality. Exact measurements are almost impossible and only the listening quality can be

measured with high accuracy using ITU-T recommendation P.863 [4], [5], [6]. The philosophy of this paper is that it is better to be roughly correct than exactly wrong.

3 The subjective/objective test protocol for determining the conversational quality of a videoconferencing link

In the test as described in this paper, the quality of a connection between A and B is assessed at the A side of the connection by the expert listener using a HATS artificial mouth or loudspeaker on the B-side. The HATS/loudspeaker 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/loudspeaker. 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 two HATS are available, the artificial ear of the second HATS at the A-side can be used to make a recording that can be processed by ITU-T recommendation P.863 [4], [5], [6], allowing for an objective listening quality assessment that can be reproduced with high accuracy. Speech material that is played over the HATS/loudspeaker should be recorded and played back at the levels normally used in the audio visual link.

When no HATS or loudspeaker is available the protocol can be carried out with a "live" voice at the B-side using the natural background noise present at the B-side.

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.

The protocol consists of the following five tests:

a) One-way listening, speech quality. Examples are available for calibrating the expert opinion in two languages, one female Igbo talker and one male Dutch talker. Natural, high quality speech is played back over the HATS/loudspeaker on the B-side (female and male test signals are available) or short, known, sentences are spoken by a subject on the B-side. Linear (timbre), non-linear and level distortions should all be taken into account. The ITU-T P.800 [13] ACR listening quality scale (Absolute Category Rating, see Table 1) 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 (female and male examples)

4 = good = the speech signal is only marginally distorted, female and male anchor examples distorted by a linear band filter 50-4000 Hz

3 = fair = speech quality that is clearly distorted, female and male anchor examples distorted by a linear narrow band filter 300-3400 Hz; female and male anchor examples distorted by packet loss

2 = poor = speech quality that is severely distorted, female and male anchor examples distorted by a linear narrow band filter 300-2000 Hz; female and male anchor examples distorted by packet loss

1 = bad = unacceptable low speech quality showing severe intelligibility problems, female and male anchor examples distorted by severe packet loss

More degradation examples, including room reverberation degradations, can be found in: http://beesikk.nl/JohnBeerends/SpeechQualityExamples.htm

Note that a high quality reproduction is necessary for correct assessment of the anchor speech files. The authors suggest to download the anchor speech files and use a high quality headphone for the optimal assessment.

This test results in a **MOS-LISTEN(A) and (B)** for both sides of the connection.

Opinion	Absolute Category Rating ACR	Degradation Category Rating DCR
5	Excellent	Degradation is inaudible
4	Good	Degradation is audible but not annoying
3	Fair	Degradation is slightly annoying
2	Poor	Degradation is annoying
1	Bad	Degradation is very annoying

Table 1. Definition of the ACR and DCR MOS scales used in the test protocol, taken from ITU-T P.800 [13].

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 and level distortions and echo's and background noise switching should be taken into account. Speech should be produced at varying levels between soft and loud and degradations in the background noise of the B-side, such as noise switching at voice onsets, should be taken into account. The ITU-T P.800 [13] DCR opinion scale (Degradation Category Rating, see Table 1) is used, anchored in the following way:

5 = no echo, no side tone distortion or audible background noise switching, equivalent to the natural acoustic situation

4 = audible but not annoying echo, side tone distortion or background noise switching

3 = slightly annoying echo, side tone distortion or background noise switching

2 = annoying echo, side tone distortion or background noise switching

1 = very annoying echo making normal conversations difficult

This test results in a **MOS-TALK(A) and (B)** for both sides of the connection.

c) Two-way interaction, full/semi-full/half duplex quality, double talk capabilities. This test requires two experts, one on each side of the connection. The expert at the B-side 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 (use varying speech levels),* the expert at the A-side speaks with pauses using short consonant vowel consonant (cvc) words*: pause, <cvc1>, pause, <cvc2>, pause, …etc.* In the assessment, the quality of the continuous counting voice from the B-side (*1, 2, 3, 4, 5*) is judged at the A-side and the *cvc* speech quality produced at the A-side is

judged at the B-side, together with the quality with which A perceives his own voice. During double talk no distortion and echo from one's own voice should be audible. Speech should be produced at varying levels between soft and loud. The ITU-T P.800 [13] 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)**.

This test results in a **MOS-INTERACTION(A) and (B)** for both sides of the connection.

d) Two-way interaction, delay. Objective measurement of the end-to-end delay, preferably with speech. For mean one-way delays up to 72 ms, the DCR MOS is 5.0 (excellent), for larger delays the DCR MOS rating = 11.5 – 3.5*LOG(mean one-way delay [ms]) (see Figure 1). This quality rating is roughly in line with recommendations G.107 [8], [9] and G.114 [10]. For delays above 1,000 ms the service is no longer considered to be a conversational 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 subjects take turn using the following test protocol: subject A starts the procedure with the counting word "one", while at the same time he starts a timer, next B counts "two" after receiving "one", etc until expert A receives "ten" from the B-side, at which point he 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***.

This test results in a **MOS-DELAY** by applying the transformation as given in Fig.1.

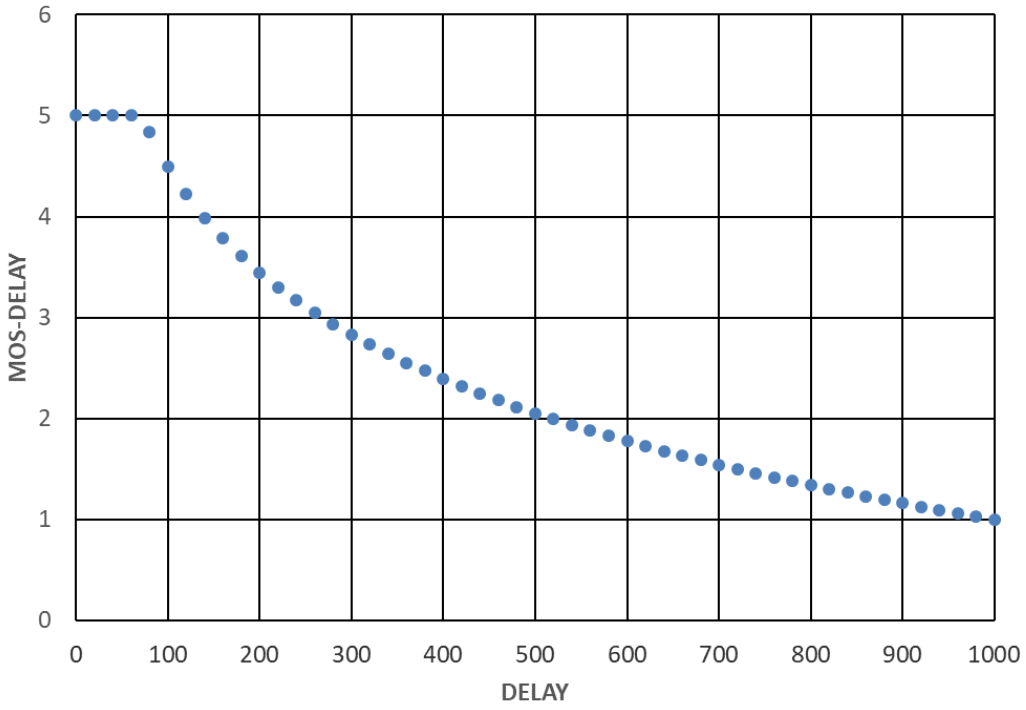

Figure 1. Impact of delay on the conversational quality, For mean one-way delays up to 72 ms, the DCR MOS is 5.0 (excellent), for larger delays the DCR MOS rating = 11.5 – 3.5*LOG(mean one-way delay [ms]).

e) Video quality. If the video and audio are clearly out of sync, the video quality impact is set to zero. If the video is of full High Definition TV quality (>1000 lines vertical resolution) with no clearly visible degradations and perfect lip-sync, the overall audio MOS value is increased by 0.5 MOS. For non-perfect lip-sync and lower video resolutions the overall MOS increase is lower than 0.5 and is estimated with Table 2. The video quality assessment can be determined for both the A- and the B-side.

Video assessment	**Increase in conversational MOS**
Ideal reference full HD resolution with perfect lip-sync	0.5
Difference with Ideal is visible but not annoying	0.4
Difference with Ideal is visible and slightly annoying	0.3
Difference with Ideal is visible and annoying	0.1
Difference with Ideal is visible and very annoying	0.0

Table 2. Impact of the video signal quality on the final conversational MOS, the maximum increase in MOS is only obtained for perfect lip-sync with high quality full HD video resolution (>1000 lines of vertical resolution).

Each of the five tests is carried out by two experts at both the A and B-side giving seven MOS for the first four tests (a, b, c, d), MOS-LISTEN(A), (B), MOS TALK(A), (B), MOS-DELAY and MOS INTERACTION(A), (B). The end-to-end conversational quality of the speech link is defined as the minimum over the seven MOS scores in the four different audio tests (a, b, c, d) increased by the video impact of the lowest quality of the A and B side, leading to a theoretical a maximum MOS of 5.5.

4 Conclusion

A fast and simple method for the measurement of the conversational quality of a video conferencing link is presented that provides stable results on a five point MOS scale. It uses a combined subjective/objective test protocol with expert listeners. Subjects can be trained in a short training session to become expert listeners.

In the protocol the three main contributing factors to conversational quality are assessed, the 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 listening quality experiment uses prerecorded speech that is played over a HATS/loudspeaker, or a live talking subject. The talking quality experiment uses a single expert. The interaction quality is determined by measuring the round trip delay and by an assessment of two experts, one on each side of the audiovisual link.

The method uses predefined anchoring conditions and an objective end-to-end delay measurement that is mapped to the subjective conversational quality impact. The final conversational quality is determined by the worst contributing factor, compensated with a correction factor of maximum 0.5 MOS for links with high definition video that is in exact lip-sync with the speech signal. The theoretical maximum MOS score in this approach is 5.5.

The conversational MOS values for high quality telephony links lie between 3.0 and 4.0.

5 References

[1] ITU-T Rec. P.805, “Subjective evaluation of conversational quality,” International Telecommunication Union, Geneva, Switzerland (2007 April).

[2] ITU-T Rec. P.920, “Interactive test methods for audiovisual communications,” International Telecommunication Union, Geneva, Switzerland (2000 May).

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

[4] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy and M. Keyhl, “Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I – Temporal Alignment,” J. Audio Eng. Soc., vol. 61, pp. 366-384 (2013 June).

[5] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy and M. Keyhl, “Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II – Perceptual Model,” J. Audio Eng. Soc., vol. 61, pp. 385-402 (2013 June).

[6] ITU-T Rec. P.863, “Perceptual Objective Listening Quality Assessment,” Geneva, Switzerland (2014 Sep.).

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

[8] ITU-T Rec. G.107, "The E-model, a computational model for use in transmission planning," International Telecommunication Union, Geneva, Switzerland (2015 June).

[9] ITU-T Rec. G.107.2, "Fullband E-model," International Telecommunication Union, Geneva, Switzerland (2019 June).

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

[11] ITU-R Rec. J.247, " Objective perceptual multimedia video quality measurement in the presence of a full reference," International Telecommunication Union, Geneva, Switzerland (2008 Aug.).

[12] ITU-R Rec. J.341, " Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference," International Telecommunication Union, Geneva, Switzerland (2016 March).

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