


**COVER PAGE**

***Document downloaded by @DAEL***

***Tue Nov 24 01:08:14 2026***
***For personal use***

When automatic English translation is provided, only the original document is authentic.
The EAA cannot be held responsible of any translation error

Bibliographical reference

*Subjective and Objective Assessment of the Listening Quality of Customer Support Waiting Loops*, Peter Počta and G. Beerends, Acta Acustica **vol. 105** (Number 2), 2019, pp. 392-400

DOI
https://doi.org/10.3813/AAA.919322

**Acta Acustica united with Acustica**
**Vol. 105 (2019) 392 – 400**

DOI 10.3813/AAA.919322

# Subjective and Objective Assessment of the Listening Quality of Customer Support Waiting Loops

Peter Počta[1)], John G. Beerends[2)]

[1)] Dept. of Multimedia and Information-Communication Technology, FEEIT, University of Žilina, 01026 Žilina, Slovakia. pocta@fel.uniza.sk

[2)] TNO, P. O. Box 96800, 2509 JE The Hague, The Netherlands

**Summary**

This paper assesses the perceived listening quality of customer support waiting loops using both a subjective experiment as well as an objective perceptual assessment with POLQA (ITU-T Rec. P.863) and VISQOL. A modified version of the methodology defined in the ITU-T Rec. P.835 was derived and used in this paper to subjectively assess the listening speech quality, the listening audio quality as well as the overall listening quality of the waiting loop perceived by the user. It is expected that the perceived listening quality of the customer support waiting loop can influence the overall quality of the customer support communication service. The waiting loop consisted of a speech announcement and a music fragment. Both are assessed separately as well as overall. The results show that the degradations introduced by the investigated codecs and packet loss, representing typical degradations commonly seen in voice communication over telecommunication networks, seriously impact the subjectively perceived listening quality of the speech announcement and music fragment as well as the overall subjectively perceived listening quality of the waiting loop. As expected the listening quality of the music fragment is, in most of the cases, significantly lower than the listening speech quality of the announcement. The POLQA results show a high correlation between subjective and objective measurements for the listening speech and audio quality as well as the overall listening quality of the waiting loop. On the other hand, it is shown that the VISQOL model provides too low correlations between predicted objective and subjective scores for accurate prediction of the listening speech and audio quality as well as the overall listening quality of the waiting loop.

**PACS no. 43.71.Gv, 43.72.Kb**

## 1. Introduction

A large part of current speech communication over telecommunication networks is represented by customer support calls. Companies running support lines do their best to optimize their operational costs and consequently limit their capacity. Therefore, customers/callers will experience waiting loops that typically consist of a short "busy line" announcement and a music excerpt, which is supposed to kill the waiting time. The waiting loop is repeated a couple of times till a connection with an operator is established. It is worth noting here that telecommunication networks are usually optimized for transmitting speech signals. So music signals will be degraded severely when transmitted over telecommunication channels. This is especially the case when it comes to channels involving older parametric speech codecs such as the G.729 [1] and AMR-NB [2] codec. Newer ones, like EVS [3], are also optimized for music signals. As the waiting loop represents an integral part of the customer support communication it is expected that its listening quality has an influence on the overall quality of the communication service. Therefore, it is important to measure and optimize the perceived listening quality of waiting loops as much as possible to provide the end user with the best possible quality. The importance of this issue is also shown in the corresponding study item of Q9 of ITU-T SG12 dealing specifically with an objective assessment of audio signals such as music transmitted over telecommunication links like WCDMA and LTE with modern codecs and terminals. Besides the influence of the waiting loop listening quality on the overall quality of the communication service, this kind of communication can be also impacted, in terms of conversational quality, by a performance of Spoken Dialogue System (SDS), usually included in a transmission chain in the context of the customer support lines, or better to say its modules, like a speech recognizer, a semantic analyser, a dialogue manager, etc. As this study is purely limited to the perceived listening quality of customer support waiting loops, the interested reader is referred to [4],

Received 15 January 2018,
accepted 18 December 2018.



© S. Hirzel Verlag · EAA

*Document downloaded by @DAEL, 2026-11-24 01:08:17 - personal use only*

when it comes to the SDS performance impact on the overall quality experienced by the end user.

A well-established subjective methodology defined in ITU-T Rec. P.800 [5] is usually used to assess the listening quality of speech communication over telecommunication channels. Objective models such as PESQ (ITU-T Rec. P.862) [6, 7, 8], POLQA (ITU-T Rec. P.863) [9, 10, 11] and VISQOL [12, 13] (non-standardized) have been developed to predict the listening-only speech quality perceived by the end user. On the other hand, neither subjective nor objective methodologies are available when it comes to a quality assessment of music signals transmitted over telecommunication networks. For the high and intermediate quality broadcasting context well-established subjective methodologies like ITU-R Rec. BS.1116-3 [14] and ITU-R Rec. BS.1534-3 [15] and objective models like PEAQ (ITU-R Rec. BS. 1387-1) [16, 17], PEMO-Q (non-standardized) [18], VISQOL Audio (non-standardized) [19] and POLQA Music (currently under development) [20] exist to assess audio quality perceived by the end user.

Some work has been carried out to evaluate the quality of super-wideband speech and audio codecs in the context of telephone communication and study the performance of the POLQA model when it comes to digital audio broadcasting services. In [21], Feiten *et al.* compared the quality provided by popular low-delay super-wideband speech and audio codecs in the context of telecommunication applications. The MUSHRA test approach was deployed in this experiment as a subjective methodology. 26 naïve listeners were involved in the subjective test. The test was divided into two parts, one dedicated to mono test conditions and one dedicated to stereo test conditions. When it comes to the mono test conditions the following codecs were used; AAC-ELD operating at 24, 32 and 48 kbps, AAC-LC operating at 32, 48 and 64 kbps, CELT operating at 32, 48 and 64 kbps, G.718 operating at 12, 24, 32, 48 and 64 kbps, G.719 operating at 32, 48 and 64 kbps, G.722 at 64 kbps, G.722.1 Annex C operating at 24, 32 and 48 kbps, G.722.2 (AMR-WB) operating at 12.65 and 23.05 kbps, SILK operating at 12, 24, 40 and 64 kbps and Speex operating at 24, 32 and 44 kbps. The stereo test conditions involved the following codecs AAC-ELD operating at 48 and 64 kbps, AAC-LD at 64 kbps, CELT operating at 48, 64 and 96 kbps, G.718 at 64 kbps, G.719 operating at 64 kbps and Speex operating at 32 and 44 kbps. Test signals deployed in this experiment covered speech, music, speech over/with music and speech with background noise. It is worth noting here that different test signals were deployed for the mono and stereo test conditions. The results obtained for the mono test conditions showed that an excellent quality can be achieved with a bit rate of 48 kbps, independently of the source signal deployed. When it comes to the AAC-LC this quality level can be even obtained for 32 kbps. Regarding the results obtained for the stereo test conditions, excellent quality was achieved for AAC-ELD operating at 64 and even 48 kbps very closely followed by CELT operating at 96 kbps and G.719 at 64 kbps.

In [22], Sloan *et al.* investigated the performance of four quality prediction models, i.e. VISQOLAudio, PEAQ, POLQA and PEMO-Q, on three datasets containing full band audio signals encoded with a variety of mostly audio codecs; HE-AACv2, MP3, AAC-LC, Opus and Ogg Vorbis, namely TCDAudio14, AACvOpus15 and CoreSV14 (bit rates ranging from 24 to 256 kbps). The results showed that POLQA performed well for high quality audio clips in TCDAudio14 and AACvOpus15 but not for CoreSV14. Moreover, POLQA performed poorly for all other treatments, when compared to the predictions of the other models. The VISQOL Audio model achieved the best performance on all metrics for two of the three datasets and was just short of the best accuracy for the third one.

It is worth noting here that POLQA is a speech quality model using the concept of idealization [9, 10] and can thus in general not be used for assessing audio quality. The POLQA idealization algorithm uses deviations from the optimal voice timbre in the calculation of the MOS score. For audio quality assessment that uses music fragments such timbre deviations are not expected to be possible. Furthermore the POLQA idealization algorithm uses the concept of noise suppression in the reference signal, a concept that is expected to fail in the assessment of music signals.

To the best of our knowledge, there is no work dealing specifically with the listening quality of waiting loops as used in customer support calls. Therefore, we study the impact of different speech codecs, representing typical codecs deployed in current telecommunication networks, including packet loss on the listening speech and audio quality as well as the overall listening quality perceived by the end user. A subjective test is carried out that uses a modified version of subjective methodology defined in the ITU-T Rec. P.835 [23] originally developed for evaluating the quality provided by speech communication systems that include noise suppression algorithms. Moreover, we also check the performance of the POLQA 2014 model and the VISQOL model, i.e. version 238, for the investigated conditions. It should be noted here that for both, POLQA and VISQOL, the speech model was used for predicting the listening speech and audio quality as well as the overall listening quality because the degradations introduced by telecommunication networks are more severe than targeted in the objective modelling approach for music. The performance of the POLQA and VISQOL models is assessed by comparing the predictions with subjective quality scores obtained from the test described in this paper. The aim of this study is two-fold: firstly, we would like to know to what extent degradations introduced by the investigated codecs and packet loss have an impact on the listening speech and audio quality as well as the overall listening quality perceived by the end user in the context of the waiting loops. Secondly, we would like to see whether the POLQA and VISQOL models are able to provide valid predictions of the perceived listening speech and audio quality as well as the overall perceived listening



Document downloaded by @DAEL, 2026-11-24 01:08:17 - personal use only

quality for the given application domain using a speech, music and combined speech and music signal respectively.

The remaining of the paper is organized as follows. Section 2 describes the subjective test carried out within this study and its results. In Section 3, the experimental results obtained from the prediction models are compared with the subjective data presented in this paper and discussed. Section 4 provides the final conclusions.

## 2. Subjective test

The first aim of this study is to quantify the impact of degradations introduced by the investigated codecs, including packet loss, on the quality perceived by the end user. To this aim, a subjective listening-only test is carried out in which subjects judged the quality of waiting loops that are composed of a speech announcement (in Slovak) and music fragment. The following subsections provide a description of this test and the results that are obtained.

### 2.1. Experiment description

A modified version of subjective methodology defined in the ITU-T Rec. P.835 was used to perform the subjective listening-only test. The P.835 methodology was selected for the subjective test because it, in our view, represents the best candidate among the considered methodologies, i.e. ITU-T Rec. P.835, ITU-T Rec. P.806 [24] and ITU-T Rec.1301 [25], as it asks for three different aspects of a stimulus and is well established in the telephone communication quality assessment community. When it comes to the P.835 methodology, it is important to mention here that it recommends to present each stimulus thrice, and to collect a rating on one of the three assessed aspects, i.e. speech distortion, background noise intrusiveness and overall quality, using different rating scales after each presentation. In comparison to the original P.835 methodology, speech distortion and background noise intrusiveness aspects were replaced by speech quality (S-MOS) and audio quality (A-MOS) respectively. Moreover, when it comes to speech quality and audio quality evaluation, the subjects listened to and assessed only the corresponding part of the samples, i.e. the short announcement (speech quality) and the music excerpt (audio quality). Note that in P.835 noise impact assessment the background noise is also present in the speech and thus one has to use the complete noisy speech file in the assessment of both the speech distortion and the noise intrusiveness. In the waiting loop experiment the speech and audio quality can be assessed independently. Similarly as in the P.835 case, an overall quality (O-MOS) was also assessed.

For all the quality aspects investigated in this study, i.e. speech quality, audio quality and overall quality, a five-point ACR (Absolute Category Rating) scale was deployed. The rest of the P.835 test protocol was fully followed in this subjective test. In all experiments, up to 2 listeners were seated in a small listening room (acoustically treated) with a background noise below 20 dB SPL (A). All subjects were Slovak Nationals whose first language was Slovak. All together, 34 listeners (17 male, 17 female, 20–58 years, mean 36.33 years) participated in the test. The subjects were remunerated for their efforts. The samples were played out using high quality studio equipment in a random order and diotically presented (presentation level: 73 dB SPL (A)) to the test subjects.

Four different 12 seconds long samples replicating a typical waiting loop, i.e. containing the short announcement in Slovak language and music excerpt both lasting around 5 seconds, were used in this test. As, according to our experience, male voices, either in a synthetic or human form, dominate when it comes to the announcements, a male voice was deployed in this case. One of the reasons for the male human and synthetic voice dominance is its higher accurateness (human and synthetic voice) and persuasiveness (synthetic voice) than that of a female voice, see [26] for more detail. As the focus of this paper is on the impact of telecom degradations on music signals, only one male human voice was used for the announcement. On the other hand, 4 different music excerpts covering a wide range of genres (music mood according to the Thayer's arousal-valence emotion plane, see [27] for more detail) from an instrumental jazz (calm), a psychedelic music (peaceful) and to a popular instrumental (excited) and electronic (happy) music, representing typical waiting loop music, were used. It is worth noting here that the music was, in all the cases, without a singing voice.

Twenty-seven test conditions representing typical degradations commonly seen in voice communication over telecommunication networks were investigated in this test, see Table I. In order to emulate a typical telecommunication channel, the samples were filtered by the following filters:

- a PCM filter (test conditions involving ITU-T G.711 and ITU-T G.729 codecs)
- an MSIN and LP35 filter (test conditions involving AMR-NB and EVS-NB codecs)
- a P.341 filter (test conditions involving ITU-T G.711.1, AMR-WB, EVS-WB, ITU-T G.729.1 codecs)
- a 14KBP filter (test conditions involving EVS-SWB codec).

More details of these filters can be found in ITU-T Rec. G.191 [28].

Regarding the packet loss, four different loss locations were simulated in the samples in order to take into account the location of the loss in the waiting loop. Both random (Bernoulli loss model) and bursty (Gilbert model) loss distributions were covered. Packet loss was generated by deleting short segments of speech and music (after applying the codecs and filters), corresponding to an actual length of the speech codec frame (varying depending on the codec), to ensure the same loss distribution for all the test conditions involving packet loss and a derogation of both speech and music parts of the samples. This approach has not allowed us to deploy a packet loss concealment technique and thus represents the worst case scenario. The packet loss rates deployed in the experiment were selected



Document downloaded by @DAEL, 2026-11-24 01:08:17 - personal use only

Table I. Description of the test conditions used in the subjective test.

| No. | Description of test condition |
|---|---|
| 1 | ITU-T G.711.1 codec at 96 kbps (Wideband speech codec) [29] |
| 2 | ITU-T G.711.1 codec at 96 kbps + 10% Packet loss |
| 3 | ITU-T G.711.1 codec at 96 kbps + 20% Packet loss |
| 4 | ITU-T G.729.1 codec at 32 kbps (Wideband speech codec) [30] |
| 5 | ITU-T G.729.1 codec at 32 kbps + 10% Packet loss |
| 6 | ITU-T G.729.1 codec at 32 kbps + 20% Packet loss |
| 7 | EVS-SWB codec at 5.9 kbps (Multiband speech codec operating in Super-wideband (SWB) mode) [3] |
| 8 | EVS-SWB codec at 5.9 kbps + 10% Packet loss |
| 9 | EVS-SWB codec at 5.9 kbps + 20% Packet loss |
| 10 | EVS-WB codec at 5.9 kbps (Multiband speech codec operating in Wideband (WB) mode) [3] |
| 11 | EVS-WB codec at 5.9 kbps + 10% Packet loss |
| 12 | EVS-WB codec at 5.9 kbps + 20% Packet loss |
| 13 | ITU-T G.711 codec (Narrowband speech codec) [31] |
| 14 | ITU-T G.711 codec + 10% Packet loss |
| 15 | ITU-T G.711 codec + 20% Packet loss |
| 16 | EVS-NB codec at 5.9 kbps (Multiband speech codec operating in Narrowband (NB) mode) [3] |
| 17 | EVS-NB codec at 5.9 kbps + 10% Packet loss |
| 18 | EVS-NB codec at 5.9 kbps + 20% Packet loss |
| 19 | AMR-WB codec at 6.6 kbps (Wideband speech codec) [32] |
| 20 | AMR-WB codec at 6.6 kbps + 10% Packet loss |
| 21 | AMR-WB codec at 6.6 kbps + 20% Packet loss |
| 22 | ITU-T G.729 codec (Narrowband speech codec) [1] |
| 23 | ITU-T G.729 codec + 10% Packet loss |
| 24 | ITU-T G.729 codec + 20% Packet loss |
| 25 | AMR-NB codec at 4.75 kbps (Narrowband speech codec) [2] |
| 26 | AMR-NB codec at 4.75 kbps + 10% Packet loss |
| 27 | AMR-NB codec at 4.75 kbps + 20% Packet loss |

to cover a broad range of packet loss rates experienced by the end user in the current telecommunication networks, especially in the case of Over-The-Top Voice over IP services.

The four samples were processed by the test conditions listed in Table I to yield to 108 test items. The 108 test items were divided into two test sessions balanced from a size and degradation perspective in order to avoid subject fatigue as advised in ITU-T Rec. P.835.

In order to ensure that each participant of the test properly understood his/her task a familiarization session consisting of two parts preceded each test. In the first part test subjects listened to 4 items representing the typical waiting loop degraded by 4 different conditions corresponding to those to be used in the main test. These conditions covered the whole degradation range and types, namely the test conditions No. 4, 7, 21 and 25 listed in Table I. The aim of the first part was to put the subjects into the context of the experiment. The second part of the familiarization consisted of 16 items (8 test conditions, i.e. No. 1, 7, 9, 10, 16, 17, 19 and 25, and 2 samples) covering again all

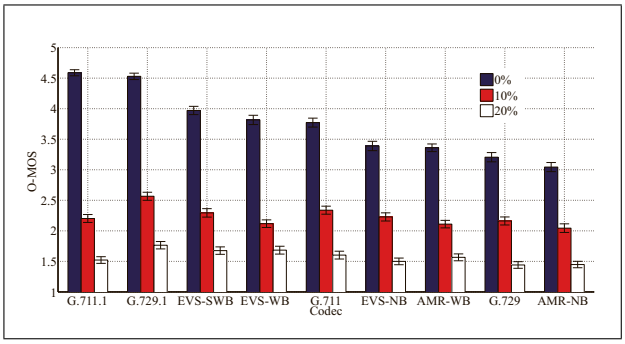


Figure 1. Effect of codec and packet loss on average O-MOS. The vertical bars show 95% CI computed over 136 O-MOS values (34 subjective scores per sample for 4 samples).

the range of the degradations included in the main test. The aim of this part was to familiarize the subjects with the test procedure and rating scale using a different speaker and different typical waiting loop music fragments as those to be used in the main test in order to prevent boredom. This was carried out by running a short part of the main test. The different speaker and music fragments were used in both parts of the familiarization session. The overall test time per subject was 79.5 minutes on average including an instruction session, the familiarization session, 2 test sessions and 10 minutes long rest-break.

### 2.2. Experimental results

Figure 1 presents the results of the subjective test for the test conditions, i.e. codec and packet loss impact, averaged over 136 values (34 scores per sample for 4 samples) for the overall evaluation (O-MOS, the overall MOS over speech and audio). The results are downwardly ordered on the basis of the 0% packet loss values and show a lower than expected quality for the AMR-WB condition. In general a WB condition will have a higher quality than a NB condition while the results show that the AMR-WB condition has about the same quality as the EVS-NB condition. As expected the results show a rapid decrease in quality as a function of packet loss, see red and white bars for more detail.

A three-way analysis of variance (ANOVA) test was conducted on the O-MOS values using codec, sample and packet loss as fixed factors (Table II). The effect of packet loss was found to be highly statistically significant (the highest F-ratio of 2760.1, $p < 0.001$). Furthermore, the effect of codec was also highly statistically significant with $F = 47.6$, $p < 0.001$. The last factor investigated in the ANOVA test was the sample factor and it turned out to have a weaker effect on the O-MOS values than the previous factors on their own ($F = 46.13$, $p < 0.001$) but was also highly statistically significant. Regarding interactions of all the involved factors, the results show that all of them were highly statistically significant. Moreover, the highest F-ratio was reported for an interaction of the codec and packet loss ($F = 18.81$, $p < 0.001$), followed by an interaction of the sample and packet loss ($F = 4.38$, $p <$



Document downloaded by @DAEL, 2026-11-24 01:08:17 - personal use only

Table II. Summary of ANOVA test conducted on the O-MOS values.

| Effect | SS | df | MS | F | p |
|---|---|---|---|---|---|
| Codec | 207.80 | 8 | 25.97 | 47.60 | 0.0000 |
| Sample | 75.51 | 3 | 25.17 | 46.13 | 0.0000 |
| Packet loss | 3012.18 | 2 | 1506.09 | 2760.10 | 0.0000 |
| Codec*Sample | 31.77 | 24 | 1.32 | 2.43 | 0.0001 |
| Codec*Packet loss | 164.23 | 16 | 10.26 | 18.81 | 0.0000 |
| Sample*Packet loss | 14.34 | 6 | 2.39 | 4.38 | 0.0002 |
| Error | 1970.94 | 3612 | 0.55 | | |
| Total | 5476.78 | 3671 | | | |

0.001) and the codec and sample ($F = 2.43$, $p < 0.001$). To summarize, the results of the ANOVA test reveal that subjects are much more sensitive to the packet loss than to the codec and sample, and highly statistically significant interactions between all the investigated factors are found.

The average S-MOS (speech MOS) values obtained from the subjective test are presented in Figure 2. Regarding the results reported for the packet loss of 0%, a trend is more or less the same as that reported for O-MOS values above although the AMR-WB codec now provides a slightly better performance than the EVS-NB codec. When it comes to the impact of packet loss we see the same rapid decrease in quality as a function of the loss as reported above for the O-MOS values (see red and white bars for more detail). The reported S-MOS values are mostly higher than those reported for O-MOS with the exception of the test condition involving the AMR-WB codec with 10% packet loss.

A three-way analysis of variance (ANOVA) test was conducted on the S-MOS values using codec, sample and packet loss as fixed factors (Table III). The highest F-ratio ($F = 2818.14$) was determined for the packet loss factor. The effect of packet loss was found to be highly statistically significant ($p < 0.001$). Moreover, the codec factor appeared to have a weaker effect on the S-MOS values than the packet loss factor with $F = 37.89$. Furthermore, the effect of codec was again highly statistically significant ($p < 0.001$). The last factor investigated in the ANOVA test was the sample factor and it turned out to have a weaker effect on the S-MOS values than the packet loss and codec factors on their own even though it is also highly statistically significant, ($F = 25.05$, $p < 0.001$). Regarding interactions of all the involved factors, the results show that all factors are highly statistically significant. To summarize, the results of the ANOVA test reveal that listeners are much more sensitive to the packet loss than to the codec and sample, and highly statistically significant interactions between all the investigated factors are found.

Figure 3 reports the average A-MOS (audio MOS) values obtained from the subjective test. When it comes to the packet loss of 0%, a trend is again more or less the same as that reported above although the AMR-WB now provides a slightly lower performance than the EVS-NB codec. As expected the AMR codec is less suited for music signals than the EVS codec. Regarding the packet loss

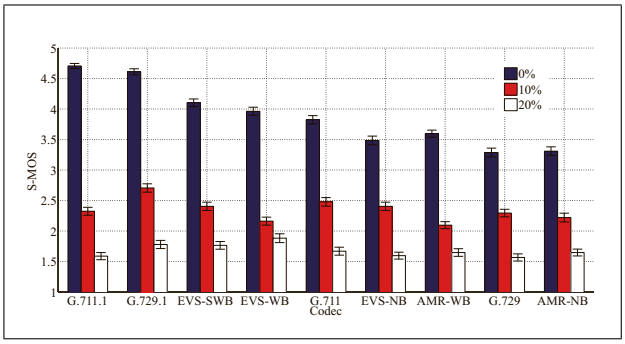


Figure 2. Effect of codec and packet loss on average S-MOS. The vertical bars show 95% CI computed over 136 S-MOS values (34 subjective scores per sample for 4 samples).

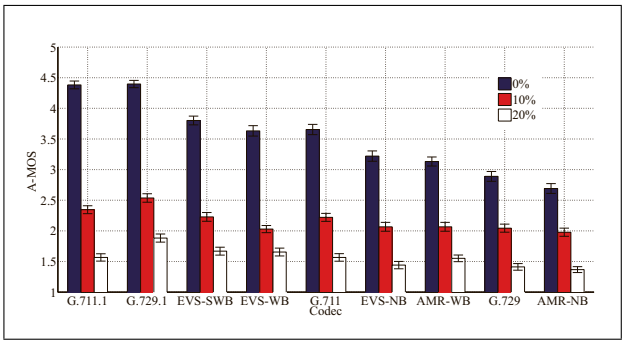


Figure 3. Effect of codec and packet loss on average A-MOS. The vertical bars show 95% CI computed over 136 A-MOS values (34 subjective scores per sample for 4 samples).

impact, the trend here follows the trend already reported for O-MOS and S-MOS values. It is worth noting here that the reported A-MOS values are mostly lower than those reported for O-MOS with the exception of the test conditions involving the G.711.1 codec and 10 and 20% packet loss and G.729.1 codec and 20% packet loss. Moreover, it should be noted here that the reported A-MOS values are also mostly lower than those reported for S-MOS with the exception of the test conditions involving the G.711.1 codec and 10% packet loss and G.729.1 codec and 20% packet loss. All the quality aspects investigated in this study, i.e. speech quality (S-MOS), audio quality (A-MOS) and overall quality (O-MOS) are excellently correlated. To be more precise, the following correlations were



Document downloaded by @DAEL, 2026-11-24 01:08:17 - personal use only

Table III. Summary of ANOVA test conducted on the S-MOS values.

| Effect | SS | df | MS | F | p |
|---|---|---|---|---|---|
| Codec | 166.71 | 8 | 20.84 | 37.89 | 0.0000 |
| Sample | 41.33 | 3 | 13.78 | 25.05 | 0.0000 |
| Packet loss | 3099.97 | 2 | 1549.98 | 2818.14 | 0.0000 |
| Codec*Sample | 39.99 | 24 | 1.67 | 3.03 | 0.0000 |
| Codec*Packet loss | 179.81 | 16 | 11.24 | 20.43 | 0.0000 |
| Sample*Packet loss | 13.93 | 6 | 2.32 | 4.22 | 0.0003 |
| Error | 1986.61 | 3612 | 0.55 | | |
| Total | 5528.35 | 3671 | | | |

Table IV. Summary of ANOVA test conducted on the A-MOS values.

| Effect | SS | df | MS | F | p |
|---|---|---|---|---|---|
| Codec | 298.94 | 8 | 37.37 | 62.89 | 0.0000 |
| Sample | 135.94 | 3 | 45.31 | 76.26 | 0.0000 |
| Packet loss | 2484.01 | 2 | 1242.01 | 2090.32 | 0.0000 |
| Codec*Sample | 51.27 | 24 | 2.14 | 3.60 | 0.0000 |
| Codec*Packet loss | 164.55 | 16 | 10.28 | 17.31 | 0.0000 |
| Sample*Packet loss | 17.48 | 6 | 2.91 | 4.90 | 0.0001 |
| Error | 2146.15 | 3612 | 0.59 | | |
| Total | 5298.34 | 3671 | | | |

obtained 0.9979 (O-MOS vs S-MOS), 0.9957 (O-MOS vs A-MOS) and 0.9902 (S-MOS vs A-MOS).

A three-way analysis of variance (ANOVA) test was conducted on the A-MOS values using codec, sample and packet loss as fixed factors (Table IV). The effect of packet loss was again found to be highly statistically significant (the highest F-ratio of 2090.32, $p < 0.001$). Furthermore, the effect of sample was also highly statistically significant with $F = 76.26$, $p < 0.001$. The last factor investigated in the ANOVA test was the codec factor and it turned out to have a weaker effect on the A-MOS values than the previous factors on their own ($F = 62.89$, $p < 0.001$) but was also highly statistically significant. Regarding interactions of all the involved factors, the results show that all of them were highly statistically significant. Moreover, the highest F-ratio was reported for an interaction of the codec and packet loss ($F = 17.31$, $p < 0.001$), followed by an interaction of the sample and packet loss ($F = 4.9$, $p < 0.001$) and the codec and sample ($F = 3.6$, $p < 0.001$). To summarize, the results of the ANOVA test reveal that subjects are much more sensitive to the packet loss than to the sample and codec, and highly statistically significant interactions between the investigated factors are found.

## 3. Objective test

In this section, the subjective results are compared to the predictions made with the POLQA 2014 model as well as the VISQOL model version 238. The comparison is performed for all the experimental conditions.

Figures 4–9 compare the subjective scores obtained from the test described in Section 2 with the POLQA and

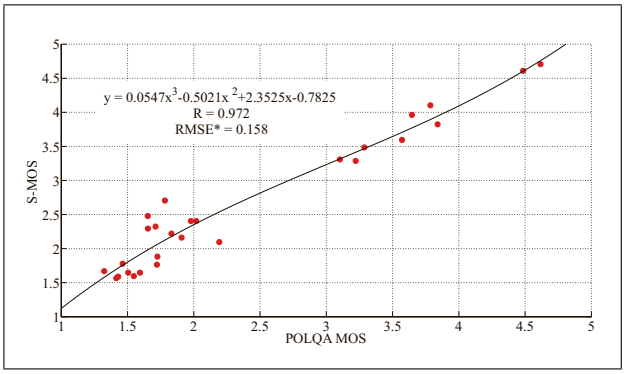


Figure 4. Correlation between the subjective results and POLQA predictions (POLQA MOS) for the subjectively perceived listening quality of the speech fragment (S-MOS).

VISQOL predictions. In order to model the experimental context of the particular experiment, 3rd order monotonic regressions are used in the calculation of the correlation and RMSE* between the corresponding subjective quality scores and POLQA and VISQOL MOS values. More details on the exact procedure can be found in [33]. A similar approach was also used in the standardization process of the PESQ and POLQA model and allows to ignore bias and context effects which are caused by the experimental context. More detail about the mapping procedure specifically designed for the POLQA model can be found in [34]. It can be observed from Figure 4 that the correlation results achieved for the POLQA model are excellent for the S-MOS scores representing the subjectively perceived listening quality of the speech fragment. Moreover, the reported RMSE* value is rather small. This does not come as a surprise as the POLQA model was de-



Document downloaded by @DAEL, 2026-11-24 01:08:17 - personal use only

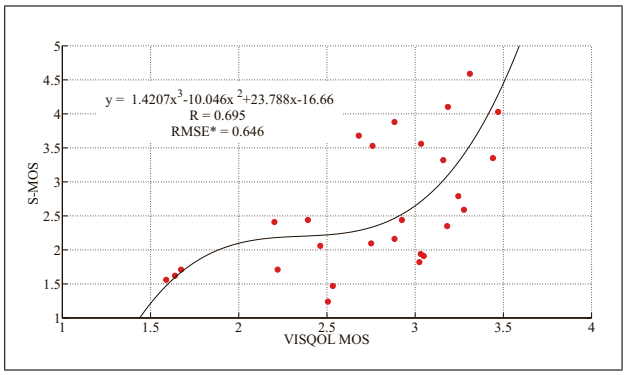


Figure 5. Correlation between the subjective results and VISQOL predictions (VISQOL MOS) for the subjectively perceived listening quality of the speech fragment (S-MOS).

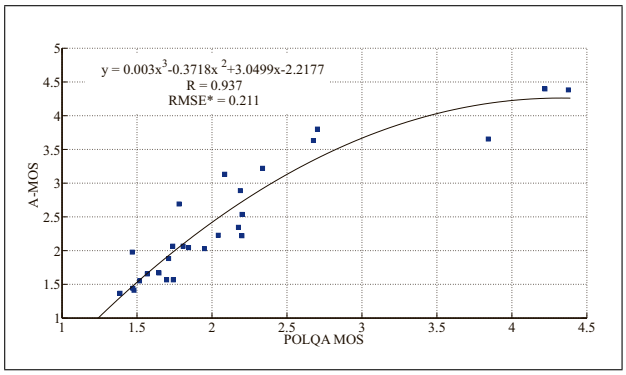


Figure 6. Correlation between the subjective results and POLQA predictions (POLQA MOS) for the subjectively perceived listening audio quality of the music fragment (A-MOS).

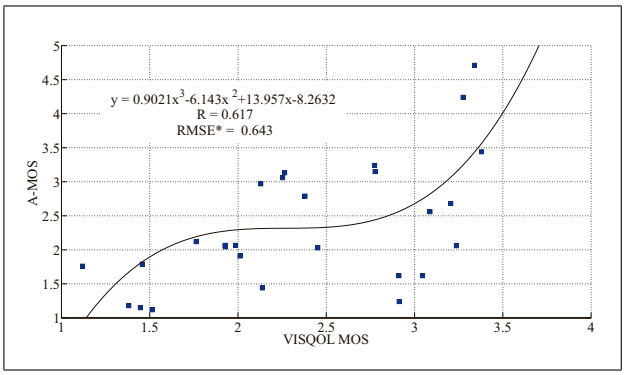


Figure 7. Correlation between the subjective results and VISQOL predictions (VISQOL MOS) for the subjectively perceived listening audio quality of the music fragment (A-MOS).

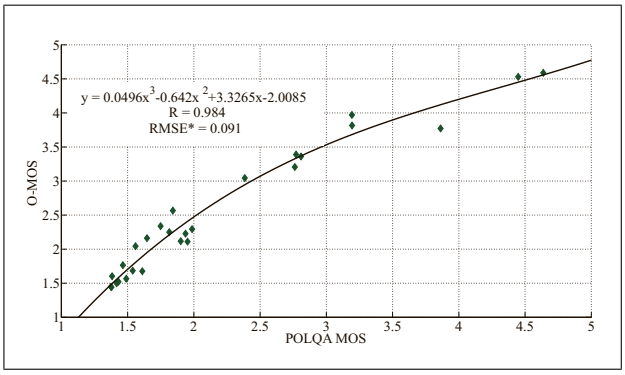


Figure 8. Correlation between the subjective results and POLQA predictions (POLQA MOS) for the overall subjectively perceived listening quality of the waiting loop (O-MOS).

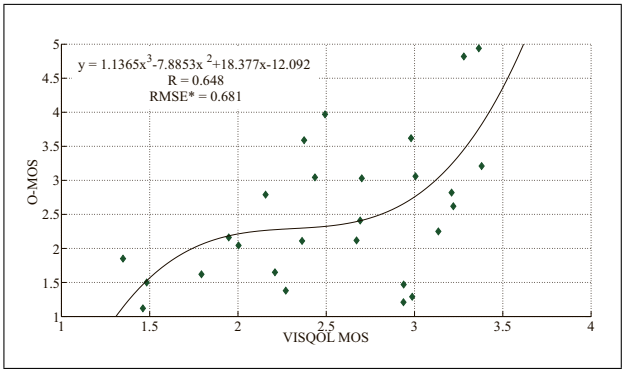


Figure 9. Correlation between the subjective results and VISQOL predictions (VISQOL MOS) for the overall subjectively perceived listening quality of the waiting loop (O-MOS).

signed to predict speech listening quality in a wide variety of conditions. On the other hand, when it comes to the VISQOL model performance, we can clearly see from Figure 5 that the VISQOL model performs worse than the POLQA model but the correlation is still rather good. In addition to that, a rather high value of the RMSE* is reported for the VISQOL model and the S-MOS scores. The lower performance of VISQOL was rather surprising for us as the VISQOL model has competed well with the POLQA model in the study published in [13] including the similar degradations as those deployed in this study.

Figure 6 shows that the A-MOS scores, representing the subjectively perceived audio quality of the music fragment, and the POLQA MOS values correlate also on an excellent level with a correlation coefficient of 0.937. The RMSE* value is a bit larger than that reported for the S-MOS scores but still excellent considering the fact that POLQA was not designed to predict audio quality of music fragments. When comparing these results with the results published in [22], we can see that the performance of the POLQA model here is much better than that reported in [22]. Figure 7 depicts the correlation between the subjective results and VISQOL predictions for the A-MOS scores. We can clearly see from this figure that the correlation is lower than for the POLQA model but still rather good considering the fact that VISQOL was not designed to predict audio quality of music fragments. The reported RMSE* value is again rather high. It should be noted here that the lowest RMSE* value was obtained in this case.

Figure 8 compares the O-MOS scores typifying the overall subjectively perceived listening quality of the waiting loop and POLQA predictions. The performance of the POLQA model is a bit better than that reported for the S-MOS and A-MOS scores above with a correlation of 0.984. Moreover, the lowest RMSE* value was obtained in this case. A comparison of the O-MOS scores and



*Document downloaded by @DAEL, 2026-11-24 01:08:17 - personal use only*

VISQOL predictions is presented in Figure 9. In comparison to the POLQA behavior, the performance in terms of the correlation coefficient was only a bit better than that reported for the A-MOS scores above. It is worth noting here that the best performance of the VISQOL model from the correlation coefficient perspective was reported for the S-MOS scores. Contrary to the POLQA behavior, the highest RMSE* value was achieved in this case. To sum up, in all the cases, the correlations obtained for the POLQA model are above 0.9, a level that allows to make reliable quality predictions. On the top of it, the reported low RMSE* values allow practical use. Regarding the VISQOL performance, the reported correlations are well below the level that allows to make reliable quality predictions. Moreover, as the obtained RMSE* values are rather high, they do not allow practical use in this context.

## 4. Conclusions

A large part of current speech communication over telecommunication networks is represented by customer support calls. In these calls the listening quality of the waiting loop is one of the important issues. These loops mostly use a speech announcement and a music fragment. The modified version of the methodology defined in the ITU-T Rec. P.835 was derived and used in this paper to subjectively assess a listening speech quality, a listening audio quality as well as an overall listening quality of the waiting loop perceived by the user. Moreover, the performance of the POLQA 2014 model and the VISQOL model (version 238) for the investigated conditions was evaluated in this study.

We show that the degradations introduced by the investigated codecs and packet loss, representing typical degradations commonly seen in voice communication over telecommunication networks, have a serious impact on the subjectively perceived listening quality of the speech announcement and music fragment as well as the overall subjectively perceived listening quality of the waiting loop. Moreover, as expected, the listening quality of the music fragment is, in most of the cases, significantly lower than the listening speech quality of the announcement. The overall listening quality of the waiting loop is mostly lower than the listening speech quality. All the quality aspects investigated in this study, i.e. the listening quality of the speech announcement, the listening quality of the music fragment and the overall quality of the waiting loop are excellently correlated. In our view, a different ratio between the speech announcement and music fragment would lead to the similar results.

Furthermore we show that the ITU-T standard for listening speech quality assessment (POLQA – ITU-T Rec. P.863) can be used to predict the listening speech quality of the speech announcement in customer support waiting loops. We also show that, contrary to the expectation, POLQA can be used to predict the listening audio quality of the music fragment and the overall listening quality of the waiting loop. The best accuracy was achieved for the overall listening quality of the waiting loop, very closely followed by the listening speech and audio quality. When the reported excellent accuracy of the POLQA model will be confirmed by other tests conducted within the corresponding study item of the Q9 of ITU-T SG12, the validated factors listed in an application guide for recommendation ITU-T P.863 [34] can be updated accordingly. In contrast to the information reported above for the POLQA model, the VISQOL model is not capable to provide valid predictions of neither the listening speech and audio quality nor the overall listening quality of the waiting loop.

The modified version of the methodology defined in the ITU-T Rec. P.835 has provided in this study rather stable results when it comes to all the investigated quality aspects. When this behavior will be confirmed by other tests, the modified version of the P.835 methodology can be proposed as a new method to assess a listening speech and audio quality as well as an overall listening quality perceived by the end user in the context of mixed scenarios involving a transmission of both speech and music signals over telecommunication networks, e.g. customer support waiting loops. As clearly expected, a music genre/music emotion has proved to be a rather influential factor in this study when it comes to the listening audio quality (the strongest effect) as well as the overall quality of the waiting loop.

As regards future work, one aspect is to validate the stable behavior of the modified version of the methodology defined in the ITU-T Rec. P.835. Another aspect is that the excellent accuracy of the POLQA model achieved in this study needs further verification. Finally, a study comparing a performance of speech codecs optimized also for music signals, e.g. EVS codec, with those optimized only for speech signals in terms of audio quality perceived by the end user in the context of telecommunication scenarios would be of great interest of research community.

### References

[1] ITU: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP). International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. G.729, 2007.

[2] 3GPP: Mandatory speech CODEC speech processing functions; AMR speech Codec; General description. Third Generation Partnership Project, 3GPP TS 26.071, 2012.

[3] ETSI: Universal Mobile Telecommunications System (UMTS), LTE, EVS Codec Detailed Algorithmic Description (3GPP TS 26.445 version 12.0.0 Release 12). European Telecommunications Standards Institute, Sophia-Antipolis, France, ETSI TS 126 445, 2014.

[4] S. Möller: Quality of telephone-based spoken dialogue systems. Springer Science & Business Media, 2004.

[5] ITU: Methods for subjective determination of transmission quality. International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. P.800, 1996.

[6] A. W. Rix, M. P. Hollier, A. P. Hekstra, J. G. Beerends: Perceptual evaluation of speech quality (PESQ) – The new ITU standard for objective measurement of perceived speech quality, Part I – Time-delay compensation. Journal of Audio Engineering Society **50** (2002) 755–764.



*Document downloaded by @DAEL, 2026-11-24 01:08:17 - personal use only*

[7] J. G. Beerends, A. P. Hekstra, A. W. Rix, M. P. Hollier: Perceptual evaluation of speech quality (PESQ) – The new ITU standard for objective measurement of perceived speech quality. Part II – Psychoacoustic model. Journal of Audio Engineering Society **50** (2002) 765–778.

[8] ITU: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. P.862, 2001.

[9] J. G. Beerends, CH. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, 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. Journal of Audio Engineering Society **61** (2013) 366–384.

[10] J. G. Beerends, CH. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, 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. Journal of Audio Engineering Society **61** (2013) 385–402.

[11] ITU: Perceptual objective listening quality assessment. International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. P.863, 2018.

[12] A. Hines, J. Skoglund, A. Kokaram, N. Hart: ViSQOL: The virtual speech quality objective listener. International Workshop on Acoustic Signal Enhancement, IWAENC 2012. VDE, 2012.

[13] A. Hines, J. Skoglund, A. C. Kokaram, N. Harte: ViSQOL: an objective speech quality model. EURASIP Journal on Audio, Speech, and Music Processing **2015** (2015).

[14] ITU: Methods for the subjective assessment of small impairments in audio systems. International Telecommunication Union, Geneva, Switzerland, ITU-R Rec. BS.1116-3, 2015.

[15] ITU: Method for the subjective assessment of intermediate quality levels of coding systems. International Telecommunication Union, Geneva, Switzerland, ITU-R Rec. BS.1534-3, 2015.

[16] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, B.Feiten: PEAQ-The ITU standard for objective measurement of perceived audio quality. Journal of Audio Engineering Society **48** (2000) 3–29.

[17] ITU: Method for objective measurements of perceived audio quality. International Telecommunication Union, Geneva, Switzerland, ITU-R Rec. BS. 1387-1, 1999.

[18] R. Huber, B. Kollmeier: PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception. IEEE Audio, Speech, Language Process. **14** (2006) 1902–1911.

[19] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, N. Harte: ViSQOLAudio: An objective audio quality metric for low bitrate codecs. J. Acoust. Soc. Am. **137** (2015) EL449–EL455.

[20] P. Počta, J. G. Beerends: Subjective and Objective Assessment of Perceived Audio Quality of Current Digital Audio Broadcasting Systems and Web-Casting Applications. IEEE Transactions on Broadcasting **61** 407–415.

[21] B. Feiten, J. Kroll, A. Raake, M. Wältermann, U. Wüstenhagen: Evaluation of super-wideband speech and audio codecs. Audio Engineering Society Convention 129. Audio Engineering Society, 2010.

[22] C. Sloan, N. Harte, D. Kelly, A. C. Kokaram, A. Hines: Objective Assessment of Perceptual Audio Quality Using ViSQOLAudio. IEEE Transactions on Broadcasting, 2017.

[23] ITU: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm" International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. P.835, 2003.

[24] ITU: A subjective quality test methodology using multiple rating scales. International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. P.806, 2014.

[25] ITU: Subjective quality evaluation of audio and audiovisual multiparty telemeetings. International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. P.1301, 2017.

[26] J. W. Mullennix, S. E. Stern, S. J. Wilson, C.-I. Dyson: Social perception of male and female computer synthesized speech. Computers in Human Behavior **19** (2003) 407–424.

[27] R. E. Thayer: The Biopsychology of Mood and Arousal. New York, Oxford University Press, 1989.

[28] ITU: ITU-T Software Tool Library 2009 User's manual. International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. G.191, 2009.

[29] ITU: Wideband embedded extension for ITU-T G.711 pulse code modulation. International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. G.711.1, 2012.

[30] ITU: G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729. International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. G.729.1, 2006.

[31] ITU: Pulse code modulation (PCM) of voice frequencies. International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. G.711, 1988.

[32] 3GPP: Speech codec speech processing functions. Adaptive Multi-Rate - Wideband (AMR-WB) speech codec. General description. Third Generation Partnership Project, 3GPP TS 26.171, 2012.

[33] ITU: Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. International Telecommunication Union, Geneva, Switzerland, ITU-T P.1401, 2012.

[34] ITU: Application guide for Recommendation ITU-T P.863. International Telecommunication Union, Geneva, Switzerland, ITU-T P.863.1, 2014.



*Document downloaded by @DAEL, 2026-11-24 01:08:17 - personal use only*