# Overall End-to-End Conversational Quality Prediction

Frank Kettler, Stefan Bleiholder, Silvia Poschen, Nils Rohrer
[1] HEAD acoustics GmbH, 52134 Herzogenrath, E-Mail: frank.kettler@head-acoustics.com

## Abstract

Results of an auditory conversational test conducted with experts and naïve subjects using commercial wireless phones were presented in [1]. The phones have been specifically manipulated to cover a wide range of possible impairments. Thus, the auditory results reflect overall end-to-end conversational quality for single disturbances on one side, on both sides, the combination of multiple disturbances and the mutual interaction between the terminals.

Instrumental conversational speech quality tests were carried out on the same phones covering a wide range of test conditions to characterize the performance. These tests focused on perception-based parameters (instrumental MOS scores). The auditory results were then used to motivate an approach to predict end-to-end conversational quality from these instrumental results. The complexity of this task arises from the mutual influence of disturbances from the phones used on both ends, the interaction between phones and network and the perceptual weighting of disturbances by subjects in the conversational context. This motivates a two-stage approach for the prediction model. The correlation of auditory and predicted end-to-end conversational quality results justifies the approach, shows promising results and motivates further work.

## 1. Introduction

Laboratory tests allow an accurate, detailed, and comprehensive characterisation of terminals in terms of speech transmission quality. There are various test methods available today to determine technical parameters (Key Performance Indicators, KPI) and perception-based speech quality scores (Key Quality Indicators, KQI). The instrumental analyses cover the different transmission quality aspects of terminals, i.e

- MOS-LQO for listening speech quality according to ITU-T Recommendation P.863 [2],
- E-MOS for echo disturbance according to [3],
- S-MOS, N-MOS, G-MOS for communication in noisy environments according to ETSI TS 103 281 [4],
- double talk capability (DT-MOS, [5]) and
- Listening Effort (MOS-LE) according to ETSI TS 103 558 [6], when the listener is exposed to noisy environments.

All these conversational aspects can be quantified. However, it is not possible to measure, quantify or predict the end-to-end overall quality as perceived by the user in the conversational context, i.e. via a communication network and including the terminals on both sides. From a network planning perspective, terminals are only rudimentarily considered. The E-model [7], [8] assumes standard characteristics for terminals, which does not represent the complexity of modern terminals by far (see also [9]). Network monitoring or quality prediction models often limit the analysis to one-way transmission, e.g. the analysis of the IP data stream in terms of media loss, packet loss and silence periods (Barret, [10]). Other models additionally consider the temporal structure of disturbances (Clark [11], Telchemy [12], [13]), but do not consider terminals with their electroacoustic characteristics.

Terminals are considered in [14], but this is limited to the implemented speech codec in the terminal with its specific influence on one-way transmission quality.

Thus, electroacoustic characteristics of terminals, the typical artefacts of todays' signal processing in terminals (noise reduction, echo cancellation, listening enhancement, …) and the mutual interaction between both terminals on both ends is not appropriately considered in todays' prediction and monitoring models.

The structure of such a prediction or monitoring model must be motivated by and derived from conversational tests (CT). A CT is the most valid and reliable method to assess overall quality in the conversational context. Test subjects can rate overall quality potentially impaired by single or multiple disturbances in the conversational context. The individual weighting of different kinds of disturbances is only covered by CTs. On the other hand, the multitude of influencing parameters in CTs (test subjects, motivation of discussion, tasks, naturalness of conversation, user behavior, duration, test conditions, terminal characteristics, …) needs to be controlled very carefully.

## Conversational Tests – Wrap up

CTs were conducted using the Kandinsky test method [15]. Details about the test procedure, the test conditions and the auditory results can be found in [1]. Test persons used cordless handsets, six out of eight devices have been intentionally modified (impaired) in terms of signal processing and acoustic properties. The communication network was impaired by additional delay and packet loss. Single disturbances were tested as well as a combination of multiple disturbances. Subjects rated the overall communication quality in the end-to-end scenario for each connection they used. The results are designated as $MOS_{E2E}$.

At first, the CT results are used for the following analyses:

- The $MOS_{E2E}$ results for the single disturbances indicate the individual contribution of each quality aspect in the conversational context.
- In the same way, the $MOS_{E2E}$ results derived from tests with multiple disturbances are used to estimate and combine the individual contributions to overall quality in the conversational context.

The same phones, that were used in the CT, have been instrumentally measured in laboratory tests to determine the KPIs for the different communication aspects (MOS-LQO, E-MOS, S-, N, G-MOS, DT-MOS and MOS-LE). These laboratory tests were carried out following common standards for terminal testing but extended by a wide range of test signal levels and network impairments. The relevant KPI of a terminal for the quality prediction in the end-to-end scenario with another phone is the one, which best represents the characteristic of the phone on the opposite side and the given network condition.

## Prediction Model Based on a 2-Stage Approach

The complexity of the task to derive an overall end-to-end communication quality prediction model motivates a two-stage approach, i.e. gaining as much information as possible from laboratory terminal tests and combining these KQI to the overall quality in the conversational context.

The 2-stage approach is shown in figure 1. In a first step, terminals are instrumentally measured and characterized in terms of technical parameters (KPI) and perception-based parameters (KQI, MOS scores). The relevant MOS in an end-to-end scenario for user A and B is determined not only by the instrumental MOS score of one individual phone, but also by the mutual interaction of the phone on the other side.

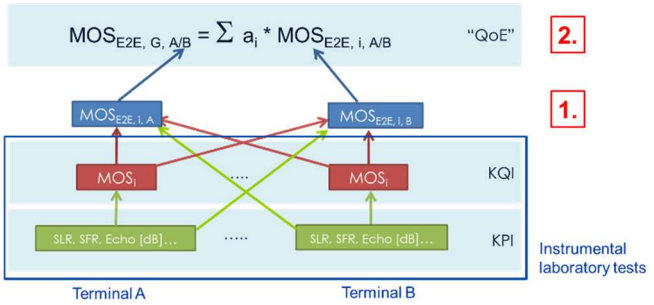


**Figure 1:** 2-Stage Approach for the Overall End-to-End Conversational Quality Prediction Model. Measured KPIs and KQIs for both terminals are combined to instrumental MOS scores in the end-to-end scenario ($MOS_{E2E,\ I,\ A}$ or $MOS_{E2E,\ I,\ B}$, quality aspect *i*, end-to-end scenario, relevant for user A or B)

Example listening quality: Listening speech quality relevant for user A is not only determined by the receiving direction of terminal A but also by the sending quality of terminal B (in addition to network impairments). Thus, the sending and receiving characteristics of both phones need to be considered, sending and receiving MOS-LQO need to be combined. Moreover, the sending sensitivity of terminal B is used to identify the best suited MOS-LQO for the receiving direction of terminal A.

Example echo: The echo disturbance and echo perception for user A is caused by the echo performance of terminal B (E-MOS). But the microphone sensitivity of terminal A significantly influences the echo performance of terminal B. A high microphone sensitivity leads to a high speech level in receiving direction of phone B, potentially causing stronger distortions and echoes. The E-MOS for terminal B is therefore determined for a wide range of test signal levels and the best suited E-MOS score is selected for further calculations in the prediction model.

In the next step 2, these individual $MOS_{E2E,i}$ need to be combined to an overall MOS score, designated here as $MOS_{E2E,G}$ on side A or side B.

A weighted sum of the individual contributions is assumed, leading to the challenge to define appropriate weighting functions. This also needs to reflect the fact, that an individual impairment may dominate overall quality. A first set of weighting functions was derived from interviews with expert listeners participating in the CT and comments given by test subjects.

This leads to non-linear weighting functions with different gradients for individual impairments. Figure 2 shows the principle of these weighting function, exemplarily for 3 of 4 quality aspects ($a_{LQO}$, $a_{echo}$, $a_{DT}$). The gradient in each of the curves indicates the importance of the individual aspect on overall conversational quality.

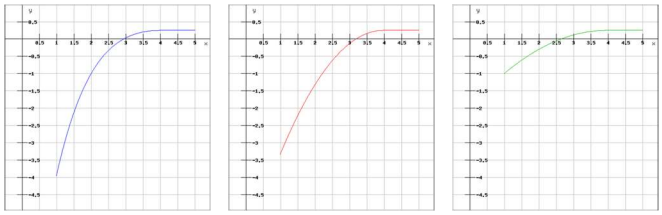


**Figure 2:** Examples of non-linear weighting functions $a_i$ for quality aspect *i* listening quality, echo and double talk performance (left to right). $a_i$ for listening quality and echo show a stronger gradient than $a_i$ for double talk capability.

This first model calibration was then applied to calculate and predict the overall end-to-end conversational quality for the same test conditions as used in the CT.

## Discussion of Results

The scatter plot in figure 3 show the comparison of instrumental MOS given on the x axis and auditory MOS from the conversational tests on the vertical axis. The results represent the overall conversation quality on both sides, i.e. for user A and user B. The following conclusion can be drawn:

- In principle, auditory and instrumental MOS scores cover the entire quality range.
- An accumulation of higher quality ratings can be observed. This goes for the auditory results and for the instrumentally predicted scores.
- The plot shows a good correlation for a majority of test condition, applying a moderate higher order mapping function.
- There are some obvious and interesting outliners, numbered 1 to 5 and high-lighted by the red circles, where the model prediction overestimates the overall conversational quality.

The corresponding test condition are listed in table 1. Interestingly, they can all be identified as listening speech quality related.

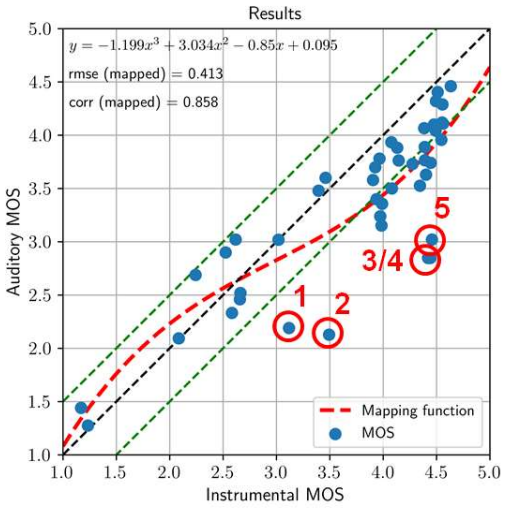


**Figure 3:** Scatterplot with model prediction scores (instrumental MOS) and conversational test results (auditory MOS) for both subscribers A and B. The red marked outliers 1 to 5 can clearly be identified as impairments caused by listening degradations.

**Table 1:** Outlier Conditions and Devices, Subset of [1]

| Outlier | Label 1 | Label 2 | Device 1 | Device 2 |
|---|---|---|---|---|
| 1 | nominal | Rx HighPass | #1 | #3 |
| 2 | nominal | Rx HighPass, BandPass | #2 | #3 |
| 3 | nominal | BandPass | #2 | #6 |
| 4 | DT12 | L+9 | #10 | #4 |
| 5 | nominal | L+9 | #10 | #6 |

The two outliers 1 and 2 are affected by an highpass filter limitation (outlier 1) and a combination of the same highpass in conjunction with a bandpass filter in the microphone path on the opposite side (outlier 2). The outliers 3, 4 and 5 indicate similar disturbances, no. 3 is a microphone bandpass filter, whereas no. 4 and 5 are affected by a 9 dB increased microphone sensitivity on the opposite side. This introduced audible distortions as also annotated by the test subjects. Note that connection no. 4 is additionally impaired by a DT attenuation of 12 dB in the send path, but this is audible only for the user on the other end.

The prediction model was calculated using 2 approaches for the $MOS_{E2E,LQO}$:

- Use the lower $MOS_{LQO}$ from sending and receiving direction, weighted by a factor, scaled by the difference between the higher $MOS_{LQO}$ and the optimum $MOS_{LQO}$ according to common standards
- Calculation of the $MOS_{E2E,LQO}$ as a function of MOS-LQO in sending and receiving direction. This formular was derived by means of a linear prediction based on measurements for MOS-LQO in sending direction, receiving direction and end-to-end, with different phones (VoIP devices, handset).

Obviously, this does not adequately consider listening speech quality, at least not in the conversational context. This is surprising as instrumental listening speech quality analysis of terminals has been intensively addressed in standardization and laboratory tests. Further work and model verifications are needed here.

Figure 4 exemplarily shows the scatterplot without these outliers. The mapping leads to the following performance metrics:

- rmse* 0.242, a modified root-mean-square error according to ITU-T Recommendation P.1401 [16], which considers the uncertainty of the auditory data by means of the 95% confidence interval,
- Pearson correlation coefficient of 0.952.

The prediction model, even in its first implementation, seems suitable to structure such a complex task. The results are promising, disregarding the above mentioned listening speech quality impairments.

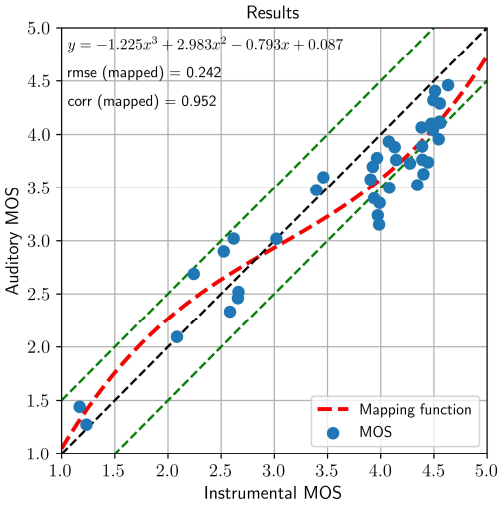


**Figure 4:** Scatterplot with model prediction scores (instrumental MOS) and conversational test results (auditory MOS) exemplarily disregarding the detected listening speech quality outliers.

## Conclusion

The results are very promising considering the high complexity of the given task. However it is acknowledged by the authors that more data from conversations tests are needed to further validate the model approach.

Other type of terminals, in particular

- mobile phones,
- vehicle mounted hands-free terminals,
- teleconferencing devices,
- and messenger services

need to be considered in conjunction with specific ambient noise scenarios.

## Acknowledgement

The work is the prosecution of developments initiated in research project (KF2485605MS4) funded as part of the program for "Joint Industrial Research (IGF)" by the German Federal Ministry of Economics and Technology (BMWi) via the AiF.

## Literature


[1] S. Bleiholder, N. Rohrer and F. Kettler, “Overall End-to-End Conversation Quality in Complex Scenarios” in Fortschritte der Akustik - DAGA 2020, Hannover, DEGA e.V., 2020.

[2] ITU-T Recommendation P.863, Perceptual objective listening quality prediction, March 2018

[3] S. Bleiholder, J. Reimes, F. Kettler, “Super-Wideband Extension of a Perceptual Based Echo Assessment Method for Aurally Adequate Evaluation of Residual Single Talk Echoes”, 13. ITG Fachtagung Sprachkommunikation, Oldenburg, 10.-12. Oktober 2018

[4] ETSI TS 103 281 V1.3.1, Speech and multimedia Transmission Quality (STQ); Speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals, May 2019.

[5] S. Bleiholder, J. Reimes, F. Kettler, “Auditory Assessment of Echo during Double Talk and Double Talk Distortions, in Fortschritte der Akustik - DAGA 2019, Rostock, DEGA e.V., 2019.

[6] ETSI TS 103 558 V1.3.1, Speech and multimedia Transmission Quality (STQ); Methods for objective assessment of listening effort, July 2021

[7] ITU-T Recommendation G.107, The E-model: a computational model for use in transmission planning, June 2015

[8] ITU-T Recommendation G.107.1, Wideband E-model, June 2015

[9] Möller et al., Extending the E-Model for Capturing Noise Reduction and Echo Canceller Impairments, JAES 2012

[10] Barret et al., System and Method for Diagnostic Modelling of Audio and Video Quality of Service, US Patent No. US 9,210,419 B2, Dec. 8, 2015

[11] Clark, Method for Estimating A Subjective Quality of a Signal, Patent No. US 7,869,377 B2 Jan.11, 2011

[12] Telchemy Inc., Quality of Service Monitor for Multimedia Communication System, US 6,741,569 B1, May 25, 2004

[13] Telchemy Inc., Per-Call Quality of Service Monitor for Multimedia Communication System, US 7,058,048 B2, June 6, 2006

[14] Towns-vonStauder, Quality Determination for Packetized Information, US 2006/0153174 A1, July 13, 2006

[15] ITU-T Recommendation P.805, “Subjective evaluation of conversational quality,” April 2007.

[16] ITU-T Recommendation P.1401, Statistical analysis, evaluation and reporting guidelines of quality measurements, Jan. 2020.