Speech Quality Testing - Part II: Performance of Instrumental Assessment

Jan Reimes[1], Magnus Schäfer[1], Alexander Kling[2]
[1] *HEAD acoustics GmbH, 52134 Herzogenrath, Germany*
[2] *Technische Universität Köln, Germany*

Abstract
Instrumental assessment methodologies facilitate efficient testing and developing modern communication systems, and ensure that high speech quality irrespective of use case and acoustic environment is provided. Nowadays, these methodologies need to be able to deal with a wide range of degradations like reverberation or low signal-to-noise ratios. These acoustic impairments are particularly common for hands-free devices, which are increasingly prevalent in today's communication.

For the evaluation of speech quality in telecommunication scenarios, Recommendation ITU-T P.800 has been providing guidelines for subjective testing for decades and is well-established in the industry. However, such auditory experiments are extremely time consuming and therefore costly. For many years, quality prediction models have been developed to provide a comfortable remedy for this purpose.

Based on two comprehensive auditory experiments presented in a companion study, the capabilities of several standardized and commonly used speech quality prediction methods were investigated. The realistic telecommunication scenarios comprise the usage of smartphones in different positions as well as vehicle and desktop hands-free devices. In particular, the prediction results for adverse acoustic conditions are analyzed in detail. Even though some models claim to accurately predict such degradations according to their scope, several inconsistent results were observed.

1 Introduction
Recommendation ITU-T P.863 [1] is currently regarded as the "Gold Standard" when it comes to the prediction of speech quality for telecommunication. The method models listening tests according to ITU-T P.800 [2] and has been widely used in the telecommunication industry for many years.

However, especially when predicted scores are quite far away from the expectations, it is often useful to investigate the model in detail. One of the most important impacts on the prediction accuracy are the underlying auditory databases, which were used during development. In case certain applications or scenarios were not covered at that time and/or were also not validated sufficiently, it is unlikely that the model provides accurate predictions in these cases.

In the development phase of ITU-T P.863 for example, only very few and not well documented databases were included, which consider the various possibilities of acoustical interfaces. In recent years, it appeared

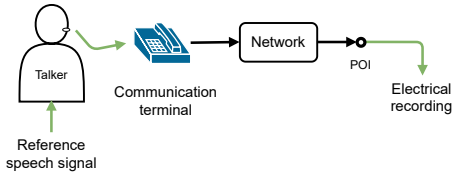


Figure 1: Electrical recordings in send direction

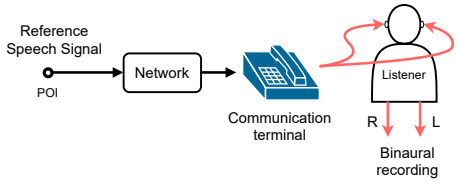


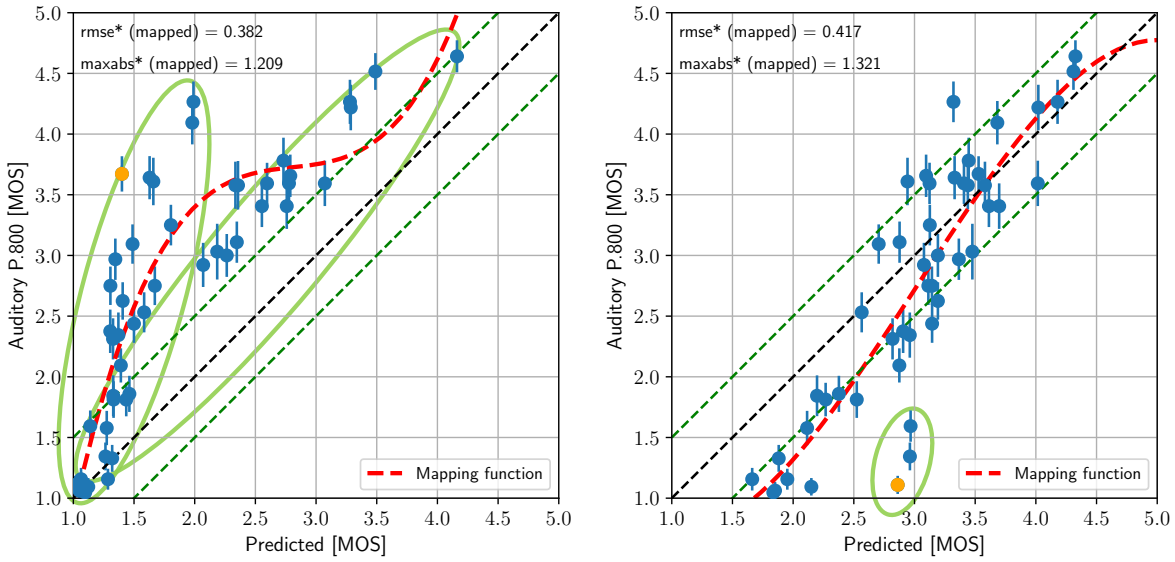
Figure 2: Acoustical recordings in receive direction

that already moderate acoustic impairments like e.g., background noise or reverberation, lead to decreased prediction accuracy. However, these issues were not formally reported so far and needed further investigation. To assess if a certain predicted score really corresponds to an average auditory quality impression or not, new listening tests and comparison to prediction results are necessary.

2 Auditory Databases
In the present work, the performance of two standardized prediction models are investigated for the evaluation of speech communication terminals in typical acoustic scenarios. For the instrumental assessment of speech quality, the two auditory databases introduced in previous work [3] are used. Therefore, only a brief overview is provided in the following. In these experiments, acoustic interfaces and paths in send (SND) and receive (RCV) were considered, as well as three different device classes:

- vehicle hands-free (VHF)
- desktop hands-free (DHF)
- smartphones in handset mode (HS)

The measurement setup for the recordings in SND is illustrated in Figure 1, where the source signal is played back via an artificial mouth simulator of an head-and-torso simulator (HATS) according to ITU-T P.58 (talking situation). The acoustical path is located between mouth reference-point (MRP) and device under test (DUT) microphone. After transmission via DUT and test network, the impaired signals are recorded electrically at the point of interconnect (POI). In overall, 48 test conditions were recorded and selected (22 from VHF, 10 from DHF, 16 from HS).



(a) ITU-T P.863 **(b)** ETSI TS 103 281

Figure 3: Prediction results for database SND (all conditions)

The measurement setup for the recordings in RCV is illustrated in Figure 2, where the source signal is inserted electrically at the POI into the test network. The acoustical path is located between DUT loudspeaker and the HATS, which is equipped with artificial ears according to ITU-T P.57 (listening situation). The impaired signals are recorded binaurally at the drum reference-point (DRP). In overall, 48 test conditions were recorded and selected (26 from VHF, 6 from DHF, 16 from HS).

Since "acoustic noise in receiving environment" is excluded in the official scope of ITU-T P.863, actually only a few test conditions of the RCV database are slightly exceeding the usage of ITU-T P.863: the car ventilation fan was used in some cases as a moderate noise source in VHF. Otherwise, all residual conditions and the whole database SND are within the scope.

Additionally, the test conditions in both databases were compiled according to the guidelines in Appendix II of ITU-T P.863 [1]. These specify so-called *full-scale subjective tests*, where the distribution of, e.g., speech levels, bandwidths, usage of anchor conditions, amount of noisy conditions and other parameters are specified. Auditory experiments following these design rules can be used to validate the prediction model.

Both listening tests were conducted following the principles of ITU-T P.800 [2]. More details on the measurement setup, sample generation and listening test can be found in [3].

3 Prediction Models

Several standardized as well as proprietary models for the prediction of speech quality are available and may be used for the qualification of speech communication terminals. However, typically only standardized and well-established methods are clearly preferred in measurement specifications. For this reason, the evaluation of prediction performance is limited in the present work to the following two models:

- Recommendation ITU-T P.863 [1]: the most relevant prediction model, which is well known and widely used in the industry, but also in academia. The main application of this method is in the area of network-related processing, like, e.g., codec impairments or time-varying processing, where it is known to work very accurately.

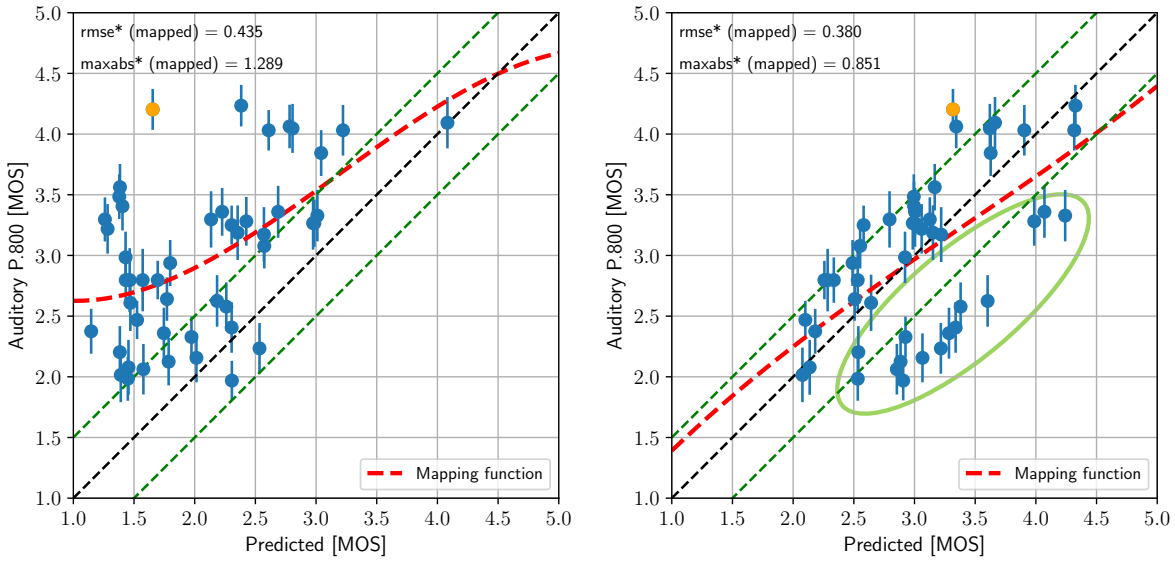
- ETSI TS 103 281 (model A) [4]: provides a similar usage as ITU-T P.863, with a degraded and reference signals as inputs and operates in a super-wideband (SWB)/fullband (FB) context. The scope is quite different, i.e., the evaluation of noise suppression techniques in terminals. The model predicts three separate attributes for speech/signal distortion (SIG), background noise intrusiveness (BAK) and overall quality (OVRL) according to the listening test design of ITU-T P.835 [5]. It was so far only validated for the send direction. For the comparison with auditory data, the prediction output *global MOS* (G-MOS) is utilized.

The prediction results of the two methods are presented in the present work. Note that also other standardized as well as proprietary prediction models were investigated. However, only the most relevant (ITU-T P.863) and (w.r.t. performance metrics) the best-performing (ETSI TS 103 281) models were considered.

4 Prediction Performance

4.1 Overview

The comparison between auditory and predicted quality is always conducted on a per-condition basis, which means



(a) ITU-T P.863 **(b)** ETSI TS 103 281

Figure 4: Prediction results for database RCV (all conditions)

that per-sample results are averaged first. Then, it is a commonly used procedure to apply a third order mapping on these averaged prediction results, in order to compensate for any bias in the auditory data. These mapped results are used for the calculation of the following performance metrics:

- *rmse**- a modified root-mean-square error according to Recommendation ITU-T P.1401 [6], which considers the uncertainty of the auditory data by means of the 95% confidence interval (CI95).

- *maxabs**- the absolute maximum error observed within one database, again considering CI95 to compensate for the uncertainty of the auditory data. This metric can be interpreted as a kind of "worst-case estimate" of a model per database.

In the following, only the results of the whole databases are provided as scatter plots. For sake of clarity and simplicity, performance metrics are reported for each device classes individually in tabular format. Note that due to the low amount of test conditions in some classes, very high or low performance values have to be interpreted carefully. For this reason, the tabular values also provide the performance metrics without mapping.

4.2 Database SND

Prediction results for the database SND and both models are provided in Figure 3. The predictions as per ITU-T P.863 reveal two clearly visible clusters, as indicated with green shapes in Figure 3a. The first cluster on the left seems to heavily underestimate the auditory data, but somehow shows at least a correct ranking. The second cluster also follows a linear trend and has approximately a constant bias. Despite a rather steep third mapping function has to be applied, at least *rmse** obtains an acceptable accuracy here.

The predictions for ETSI TS 103 281 are in general more consistent. No clustering can be observed and the third order mapping function to be applied is rather linear across the whole range. However, both performance metrics are impacted by three clear outliers, as indicated with a green shape in Figure 3b.

When investigating the prediction performance for each device category individually as per Table 1, it can be observed that ITU-T P.863 does not perform well without applying a third order mapping. The strong gain in performance values between unmapped and mapped predictions indicates a strong/steep mapping function, especially for VHF and even more for DHF. This observation is consistent with Figure 3a, where a strong underestimation is visible.

The prediction model ETSI TS 103 281 shows comparable, but slightly worse performance values for *rmse**. The consistently low for *maxabs** in all classes can be explained by the aforementioned outliers, since each class contains one of them.

4.3 Database RCV

Prediction results for the database RCV and both models are provided in Figure 4. In contrast to SND, the prediction accuracy significantly decreases for ITU-T P.863 and a lot of outliers are observed in Figure 4a.

On the other hand, ETSI TS 103 281 provides much better performance metrics, and the overall distribution looks in general smoother. However, as indicated in Figure 4b, a too optimistically predicted cluster of conditions is visible (highlighted as green shape). These belong to the class VHF and contain moderate ventilation fan noise and thus lead to a decreased prediction performance for this class.

Table 1: Prediction performance for databases SND and RCV (per device group)

		SND				RCV			
		without mapping		after 3rd order mapping		without mapping		after 3rd order mapping	
Class	Metric	ITU-T P.863	ETSI TS 103 281	ITU-T P.863	ETSI TS 103 281	ITU-T P.863	ETSI TS 103 281	ITU-T P.863	ETSI TS 103 281
all	$rmse^*$	1.14	0.56	0.44	0.38	1.14	0.56	0.44	0.38
	$maxabs^*$	2.55	0.98	1.29	0.85	2.55	0.98	1.29	0.85
VHF	$rmse^*$	1.00	0.69	0.52	0.43	1.00	0.69	0.52	0.43
	$maxabs^*$	2.55	0.98	1.55	1.25	2.55	0.98	1.55	1.25
DHF	$rmse^*$	2.17	0.35	0.10	0.03	2.17	0.35	0.10	0.03
	$maxabs^*$	2.18	0.49	0.14	0.04	2.18	0.49	0.14	0.04
HS	$rmse^*$	0.92	0.38	0.12	0.06	0.92	0.38	0.12	0.06
	$maxabs^*$	1.36	0.67	0.36	0.17	1.36	0.67	0.36	0.17

When investigating again the prediction performance for each device category individually as per Table 1, a similar trend as for the database SND can be observed: ITU-T P.863 strongly underestimates the auditory data in all classes, especially for the hands-free scenarios. Most outliers are located in class VHF, which leads to very low performance also after third order mapping. Conditions of class DHF are predicted in the range of 1.0 to 1.5 MOS by ITU-T P.863, which explains the extremely high gain in performance before and after mapping.

Another similarity to the SND database is the performance of ETSI TS 103 281, which shows a more smoother and consistent trend than ITU-T P.863 and does not require a strong mapping function. Even though the number of conditions in class DHF are limited, they are predicted accurately by ETSI TS 103 281 already without any mapping.

5 Conclusion

Based on the auditory listening test databases available from previous work, the performance of two standardized speech quality prediction models was investigated. An analysis for different classes of terminals and corresponding acoustic scenarios revealed that the state-of-the-art model according to ITU-T P.863 lacks accuracy in many typical evaluation scenarios, especially for hands-free use cases. A much higher performance was expected, because most of the test conditions were clearly within the scope of the model [1] and the listening test databases were created according to the design guidelines from Annex II of ITU-T P.863.

Even though developed exclusively for the evaluation of noise reduction algorithms, the model according to ETSI TS 103 281 [4] showed a surprisingly good prediction performance in many cases, especially in RCV, where the model was not even intended to be used. However, even though it outperforms ITU-T P.863 in several cases, outliers and lower performance metrics were observed as well.

For both directions and considered models, the prediction is accurate for devices of class HS, where the acoustic path has minor impact. However, ITU-T P.863 is more pessimistic and requires a steeper mapping. By introducing slight inherent reverberation for VHF and DHF scenarios, ITU-T P.863 requires even steeper and more winding mapping functions to match the context of the listening test. It seems unlikely that the very few conditions, which are slightly out of scope of ITU-T P.863 in database RCV can explain these mispredictions.

Especially when tests and performance requirements in measurement specifications are based on such prediction models, the characteristics and steepness of a specific mapping function in a certain application is quite relevant. Based on the results of this study and due to the steep and winding mapping functions to be applied in particular for hands-free scenarios, ITU-T P.863 seems not suitable for the performance evaluation of speech communication devices under realistic acoustic conditions.

References

[1] ITU-T Recommendation P.863, *Perceptual objective listening quality prediction*, Mar. 2018.

[2] ITU-T Recommendation P.800, *Methods for subjective determination of transmission quality*, Aug. 1996.

[3] M. Schäfer, J. Reimes, and A. Kling, "Speech Quality Testing – Part I: Auditory Assessment of Communication Scenarios," in *Fortschritte der Akustik - DAGA 2021*, Berlin: DEGA e.V., 2021.

[4] ETSI TS 103 281 V1.2.1, *Speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals*, Jan. 2018.

[5] ITU-T Recommendation P.835, *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, Nov. 2003.

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