Instrumental Assessment of Near-end Speech Quality

Jan Reimes

HEAD acoustics GmbH, 52134 Herzogenrath, Germany, Email: telecom@head-acoustics.de

Introduction

Communication in noisy situations may be extremely stressful for the person located at the near-end side. Since the background noise originates from their natural environment, it cannot be reduced for the listener. Thus, the only possibility to improve this scenario by means of digital signal processing is the insertion of speech enhancement algorithms in the receiving terminal.

In previous work, a large auditory database was presented for evaluating the trade-off between speech quality and listening effort, which is correlated with speech intelligibility. A balance between speech quality and listening effort is desirable from the user's point of view. While recent developments already indicate that the instrumental assessment of listening effort is possible, quality aspects have not yet been considered.

This contribution presents possible approaches and results for the instrumental assessment of perceived near-end speech quality in noisy scenarios. The previously developed auditory database is used for the evaluation of the proposed solution.

Measurement Setup

The test setup is motivated by the requirement that all signals can be measured outside the device, i.e. can be assessed by state-of-the-art measurement front-ends. For this purpose, the mobile device-under-test (DUT) is mounted at the right ear of head and torso simulator (HATS) according to [1] with an application force of 8 N. The artificial head is equipped with diffuse-field equalized type 3.3 ear simulators according to ITU-T P.57 [2]. Inside the measurement chamber, a realistic background noise playback system according to [3] or [4] is arranged.

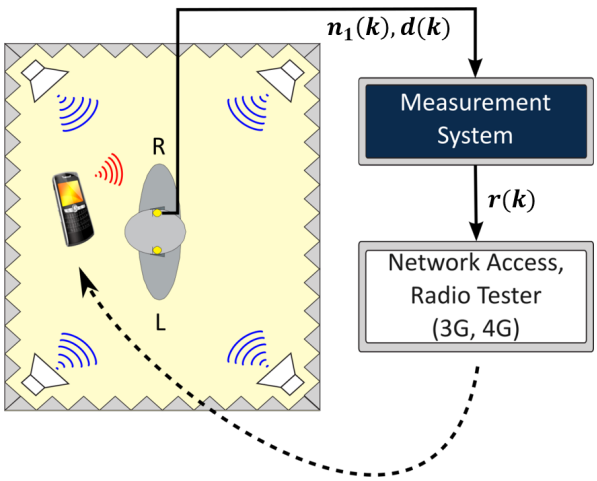


Figure 1: Recording setup for (binaural) signal assessment

Figure 1 illustrates the overall measurement setup. The recording procedure is conducted in two stages:

1. Transmission of speech in receiving direction and noise playback are started at the beginning of the recording. Simultaneously, degraded speech and near-end noise are recorded by the right artificial ear. This signal is denoted as $d(k)$ in the following. The left ear signal could additionally be recorded and used for an auditory evaluation (binaural presentation).

2. Transmission of speech is deactivated, only the near-end noise (with the phone still active and positioned at the artificial ear) is recorded, which is denoted as $n_1(k)$.

Obviously, the usage of playback systems according to [3] or [4] are crucial here for the further analysis. The sample-accurate playback precision allows time-synchronous recordings for multiple measurements, which is necessary for the proper time alignment between noisy speech signal and noise-only signal.

The degraded signal $d(k)$ can be considered as a superposition of a processed signal $s(k)$ and a noise signal $n_2(k)$ according to equation 1.

$$d(k) = s(k) + n_2(k) \tag{1}$$

As mentioned before, the background noise system provides an accurate playback precision, so that

$$n_1(k) \approx n_2(k) \tag{2}$$

can be assumed. With this knowledge, a simple estimate $\hat{s}(k)$ of the processed and noise-free signal $s(k)$ is provided by subtracting $d(k)$ and $n_1(k)$. This technique was already introduced in [5] as "time-synchroneous noise compensation" (TNC) for similar measurement applications. However, in order to additionally minimize the residual error $n_1(k) - n_2(k)$, a modified method is proposed and illustrated in figure 2.

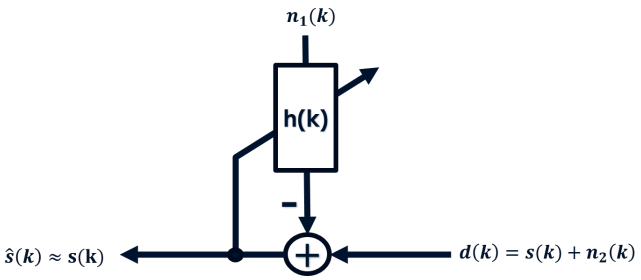


Figure 2: Estimation of $\hat{s}(k)$ by AdvTNC

A time-variant adaptive filter $h^{(i)}(k)$ based on normalized least-mean-square (NLMS) optimization is used here. The initialization according to equation 3 with a dirac impulse response corresponds to the previously mentioned simple subtraction and provides a fast convergence of the filter.

$$h^{(0)}(k) = \delta(k) \tag{3}$$

This pre-processing step is called "Advanced TNC" (AdvTNC) in the following sections.

Auditory Database

In general, perceptually-motivated instrumental methods predict quality indexes based on a specific experimental setup. These listening test databases typically include audio samples and corresponding results for certain auditory attributes. Providing that such a database includes a wide range of quality degrees, an instrumental measure can be trained based on these samples. Usually this is realized by calculating metrics of difference between the measured and the (known) reference signal. In [6], a suitable database for the current work based on simulated mobile devices was introduced, thus only a brief summary will be given in the following.

The auditory evaluation included a new procedure for the combined assessment of speech quality and listening effort on the well-known 5-point scale. The average over all participants per attribute is reported as mean opinion score (MOS). A kind of mixture between ITU-T P.800 [7] and P.835 [8] listening test was used. Here test participants vote each presented sample twice. A rating for listening effort (LE) is given after the first playback, then after a second trial the speech quality (SQ) was assessed. The scales of both attributes were taken from ITU-T P.800 [7] and are provided in table 1.

Score	Listening Effort	Speech Quality
5	No effort required	Excellent
4	No appreciable effort required	Good
3	Moderate effort required	Fair
2	Considerable effort required	Poor
1	No meaning understood with any feasible effort	Bad

Table 1: Auditory scales for combined assessment

For the assessment of stimuli of the listening test, the measurement setup as described in the previous section was used, but in conjunction with a mockup device. A background noise playback system according to [4] with an 8-speaker-setup was used to reproduce a realistic and level-correct sound field around the HATS. The standardized noises *Full-size car 130 km/h*, *Cafeteria*, *Road* and *Train station* were evaluated. Two additional gains of $-6\,\mathrm{dB}$ and $+6\,\mathrm{dB}$ for the background noise level were applied to each scenario. This step was conducted to obtain an overall noise level range of $\mathrm{SNR(A)} \approx -7 \ldots +15\,\mathrm{dB(A)}$ due to the individual level of each noise. Additionally, a silence condition (noise $\leq 30\,\mathrm{dB(A)}$) was used.

Several NELE, BWE and combinations of both algorithms were simulated in NB and WB mode instead utilizing real devices. All processed samples were calibrated to a monaural active speech level of $79.0\,\mathrm{dB_{SPL}}$. Bad as well as good conditions could be generated for both LE and SQ scales with this procedure.

In overall, 197 conditions with 8 sentences each were evaluated. A listening sample of duration 8.0 s included two sentences of a certain talker, which results in 788 different samples. One random sample per condition was selected for each of the 56 participants, which obtained 14 pairs of LE/SQ votes per sample, respectively 56 votes per condition.

Figure 3 shows one important finding of this experiment, i.e. that both assessed dimensions - speech quality and listening effort - can be regarded as orthogonal to some degree. The correlation coefficient according to Pearson is determined to $r_{\mathrm{Pearson}} = 0.52$, which indicates at least a minor correlation. This can be explained by the fact that good speech quality ratings (i.e. $\mathrm{MOS_{SQ}} > 4.5$) cannot be expected for very low listening effort scores (i.e. $\mathrm{MOS_{LE}} < 1.5$). On the other hand, even in silent or noise-free situations (i.e. $\mathrm{MOS_{LE}} > 4.5$), poor speech quality (i.e. $\mathrm{MOS_{SQ}} < 1.5$) affects also the perceived listening effort.

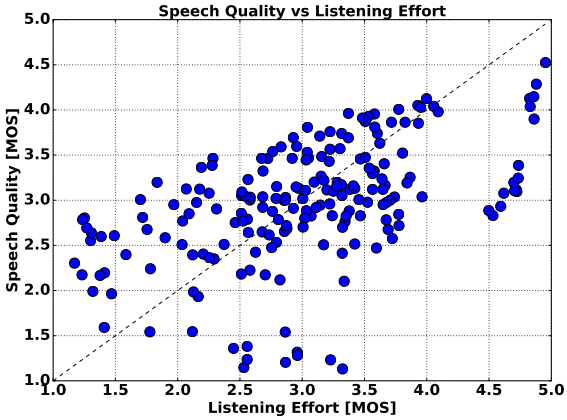


Figure 3: Speech Quality vs. Listening Effort

Instrumental Speech Quality

Similar to the already introduced instrumental assessment of listening effort in [9], it is desired to perform any speech quality analysis only on the basis of external recordings. No internal states or signals (like the processed signal $s(k)$) are known, only the obtained signals as specified in the previous sections are available.

A common method for the assessment of instrumental speech quality is described in ITU-T P.863 [10]. The method evaluates a disturbed signal against a given optimal reference. For all calculations with this predictor

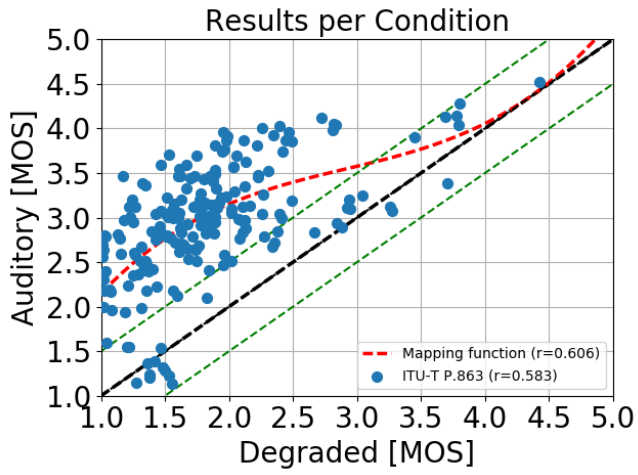
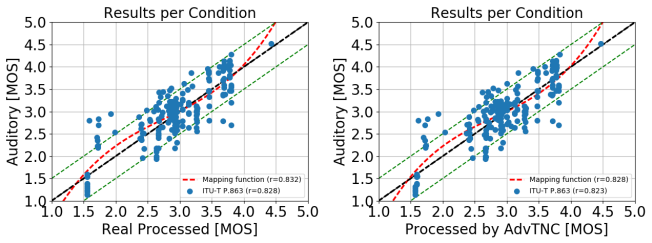


Figure 4: ITU-T P.863 results obtained by signal $d(k)$



(a) Real signal $s(k)$ **(b)** Estimated $\hat{s}(k)$

Figure 5: ITU-T P.863 carried out with speech-only signals

below, the signal $r(k)$ is always used as the reference signal.

Applying and comparing the method to the signal $d(k)$ of each condition of the auditory test database, a poor correlation can be observed as shown in the scatter plot of figure 4. Since the specification [10] explicitly excludes acoustic near-end noise, this is not a surprising result. In consequence, another approach has to be chosen.

For the analysis of the noise-free but processed speech signal $s(k)$, the calculation according to ITU-T P.863 is carried out with the estimation $\hat{s}(k)$ obtained by the AdvTNC method described in the aforementioned section. Since in this specific study also the real signal $s(k)$ is known, a comparison to the noise-compensated signals can be conducted. Figure 5 provides the results for both signal types used an input for the predictor according to ITU-T P.863. Since both signal types obtain almost identical scores, the application of the AdvTNC method seems to be a valid pre-processing for the instrumental quality assessment.

By removing the near-end noise from the recordings, the correlation significantly improves. However, still multiple larger over- and under-predictions of the quality can be observed.

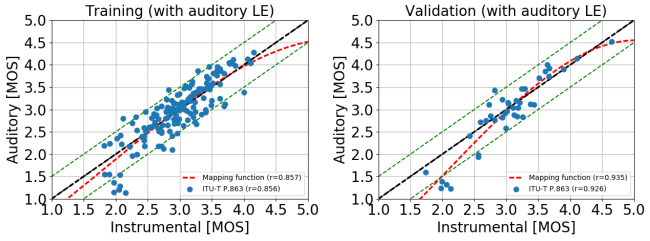
Proposed Prediction Algorithm

The results of the previous section indicate that the quality perception in silence and in noisy conditions may be different. The comparison shown in figure 3 suggests that LE and SQ are not fully orthogonal and may affect each other. Thus an obvious step is to evaluate this assumption, the previously determined speech quality score according to ITU-T P.863 is combined with the auditorily determined listening effort score. Equation 4 describes a linear regression model taking this relation into account.

$$\begin{aligned}\widehat{\mathrm{MOS}}_{\mathrm{SQ}} = a_0 + a_1 \cdot \widehat{\mathrm{MOS}}_{\mathrm{P.863}} + \\ a_2 \cdot \mathrm{MOS}_{\mathrm{LE}} + a_3 \cdot \widehat{\mathrm{MOS}}_{\mathrm{P.863}} \cdot \mathrm{MOS}_{\mathrm{LE}}\end{aligned} \tag{4}$$

Since this method can be seen a new predictor, a division of the listening test database into training and validation sets is carried out. The identical partitioning as in [9] is used and includes 147 conditions (588 samples) for training and 50 conditions (200 samples) for validation. The linear regression is carried out on the per-sample results.

Figure 6 illustrates the prediction results of this approach for the auditory database of [6]. Here the results are shown on a per-condition basis. In contrast to figure 5, the prediction results are more accurate here.



(a) Training **(b)** Validation

Figure 6: Estimation of $\widehat{\mathrm{MOS}}_{\mathrm{SQ}}$ with auditory $\mathrm{MOS}_{\mathrm{LE}}$

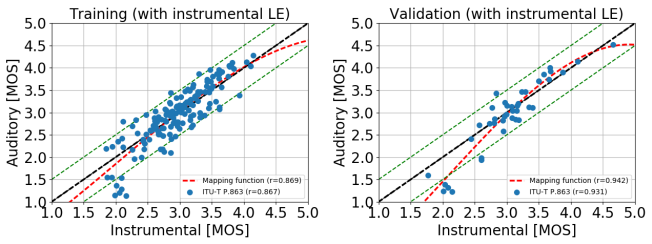
These results indicate that listening effort can possibly impact the speech quality perception. For a complete instrumental evaluation, also prediction of the listening effort is necessary. In [9], such a prediction algorithm was already introduced. To include this model, the linear regression algorithm is adapted according to equation 5.

$$\begin{aligned}\widehat{\mathrm{MOS}}_{\mathrm{SQ}} = a_0 + a_1 \cdot \widehat{\mathrm{MOS}}_{\mathrm{P.863}} + \\ a_2 \cdot \widehat{\mathrm{MOS}}_{\mathrm{LE}} + a_3 \cdot \widehat{\mathrm{MOS}}_{\mathrm{P.863}} \cdot \mathrm{MOS}_{\mathrm{LE}}\end{aligned} \tag{5}$$

Again, the coefficients of the linear regression are determined on a per-sample basis. Figure 7 shows the prediction results for speech quality determined purely with instrumental measures.

A first observation is that neither for the training nor the validation set the prediction results differ significantly from the ones shown in figure 6. Thus, the exchange of auditory to instrumental listening effort did not have a noticeable impact on this model.

Table 2 summarizes all considered approaches presented in this work. The epsilon-sensitive performance metrics RMSE^* and e^*_{max} (maximum error) according to [11] are evaluated on a per-condition basis and after 3rd order mapping.



(a) Training **(b)** Validation

Figure 7: Estimation of $\widehat{\mathrm{MOS}}_{\mathrm{SQ}}$ with instrumental $\widehat{\mathrm{MOS}}_{\mathrm{LE}}$

Method	Figure	RMSE*	$e^*_{\max}$
ITU-T P.863 with $d(k)$	4	0.40	1.56
ITU-T P.863 with $s(k)$	5a	0.23	0.88
ITU-T P.863 with $\hat{s}(k)$	5b	0.23	0.86
Proposed (auditory LE, Training set)	6a	0.19	0.81
Proposed (auditory LE, Validation set)	6b	0.15	0.49
Proposed (instrumental LE, Training set)	7a	0.18	0.79
Proposed (instrumental LE, Validation set)	7b	0.14	0.48

Table 2: Performance metrics for all considered approaches

Conclusions

Speech enhancement algorithms in receiving direction are already present in state-of-the-art terminals. In addition, signal processing originated from the far-end device and network jitter and/or packet loss may further decrease perceived quality at the user's side. Thus, quality evaluation covering all these these impacts is strongly desired.

From the measurement perspective, a "black box approach" is desired for any evaluation, i.e. no internal signals and/or information can be expected to be known. The usage of speech signals, realistic background noise systems and measurements conducted with head and torso simulator are inevitable for an evaluation close to the perception of the user.

The only available quality assessment method ITU-T P.863 is carried out on a large auditory database, which was presented in earlier work. However, the application in conjunction with near-end noise is an invalid use case and provides only poor correlation with auditory data. Thus, a method called AdvTNC is introduced in order to compensate for the noise component. Prediction results improve with this method, but still contains unsatisfactory correlation and rank order shifts.

To check the impact of perceived listening effort on the speech quality, a simple regression model based on auditorily determined listening effort and scores predicted by ITU-T P.863 is proposed. Compared to the noise-free estimation, correlation and error metrics significantly improve for training and validation sets. However, still several rank order shifts are observed with this approach.

In order to obtain a complete instrumental assessment of speech quality, a previously developed model for the prediction of listening effort is used instead of the auditory data. With an updated regression model, the prediction accuracy is almost identical compared to the model with auditory data.

Based on these results, it seems inevitable that instrumental speech quality assessment cannot be seen independent from the near-end noise and/or the perceived listening effort.

Further work should also include other scenarios like e.g., in-car communication and hands-free terminals, where the trade-off between speech quality and intelligibility/listening effort is crucial. Since such terminals are not limited to monaural use case, aspects of binaural listening and perception must be taken into account. In consequence, future instrumental listening effort and speech quality measures should consider binaural aspects as well.

References

[1] *Use of head and torso simulator for hands-free and handset terminal testing*, ITU-T Recommendation P.581, Feb. 2014.

[2] *Artificial ears*, ITU-T Recommendation P.57, Dec. 2011.

[3] *Part 1: Background noise simulation technique and background noise database*, ETSI EG 202 396-1 V1.2.4, Feb. 2011.

[4] *A sound field reproduction method for terminal testing including a background noise database*, ETSI TS 103 224 V1.1.1, Aug. 2014.

[5] U. Müsch, "Tnc," in *Fortschritte der Akustik - DAGA 2017*. Berlin: DEGA e.V., 2017.

[6] J. Reimes, "Auditory evaluation of receive-side speech enhancement algorithms," in *Fortschritte der Akustik - DAGA 2016*. Berlin: DEGA e.V., 2016.

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

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

[9] J. Reimes, "Instrumental assessment of near-end perceived listening effort," in *Perceptual Quality of Systems Workshop*, Berlin, 2016.

[10] *Methods for objective and subjective assessment of speech quality*, ITU-T Recommendation P.863, Sep. 2014.

[11] *Statistical analysis, evaluation and reporting guidelines of quality measurements*, ITU-T Recommendation P.1401, Jul. 2012.