5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)
29-31 August 2016, Berlin, Germany

Instrumental Assessment of Near-end Perceived Listening Effort

Jan Reimes

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

1. Introduction

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

So far no measurement technique is available to evaluate the impact of signal processing techniques such as "near-end listening enhancements" [1] (NELE), artificial bandwidth extension (BWE) or additional noise reduction (NR). For mobile phones, acoustic testing in downlink direction is always carried out in silent condition. However, in several state-of-the-art devices the aforementioned algorithms are already included. This implies that a device may behave differently under noisy conditions than in silence: e.g. NELE algorithms may be triggered by a certain noise level and/or spectrum.

Whenever speech processing is inserted into a conversation, quality aspects must be considered, too. A satisfactory balance between speech quality and listening effort is desirable from the user's point of view. Currently, no reliable objective or instrumental methods are available to evaluate speech quality and listening effort of a device under test (DUT) in downlink in the presence of background noise. Any possible metrics should take into account ongoing trends in acoustic telecommunication measurement standards, i.e.:

- Usage of real speech instead of artifical test signals.
- Realistic playback of background noise scenarios (e.g. according to [2] or [3]).
- "Black-Box-Approach": no internals of a DUT are known, only outer measurements are available.

Due to these requirements, several existing assessment measures targeting to intelligibility and/or speech quality aspects prove to be unfavorable:

- STITEL, STIPA, RASTI according to [4]: shaped noise signals are used for measurement.
- ITU-T Recommendations P.862 [5], [6] and P.863 [7]: noise or near-end noise is explicitly excluded in scope.
- ETSI EG 202 396-3 [8] and TS 103 106 [9]: methods are specified for noise reduction scenarios and only for uplink direction.

Another widely used measure for the instrumental intelligibility assessment is the speech intelligibility index (SII) [10]. Several drawbacks of this measurement algorithm should be considered, too:

- Pure ⅓ octave level-based measure, no real psychoacoustical model (except frequency weighting)
- Noise-free degraded speech signal is needed as input (not available in acoustic testing)

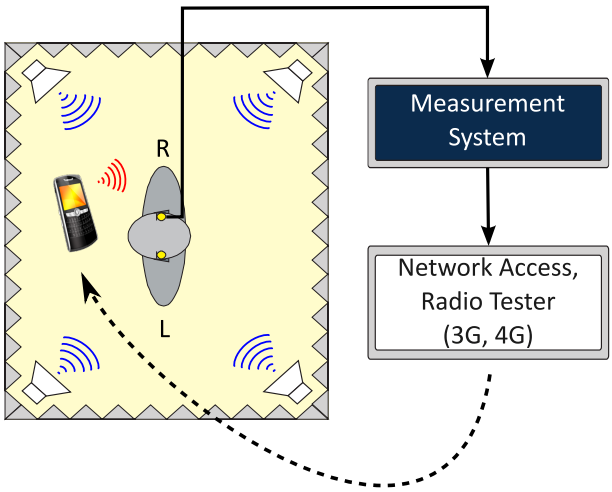


Figure 1: Recording setup for (binaural) signal assessment

- Does not consider speech distortions which may also decrease intelligibility

In overall, the SII method is also not applicable as a "black box" approach for devices with unknown and inaccessible signal processing components.

Auditory experiments addressing the trade-off between speech quality and listening effort (e.g. like presented in [11]) can be used to develop a new instrumental method for the evaluation of downlink signal processing. To address all concerns described above, a new method for the instrumental assessment of listening effort for mobile phones is introduced. Based on these auditory tests, a new prediction model can be developed.

2. 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 DUT is mounted at right ear of head and torso simulator (HATS) according to [12] 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 [13]. Then the HATS is placed into a measurement chamber. Inside this room, a realistic background noise playback system according to [2] or [3] is arranged.

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

10.21437/PQS.2016-1

as $d(k)$ in the following. The left ear signal is recorded and used for the 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 $n(k)$.

Obviously, the usage of playback systems according to [2] or [3] 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.

Speech files according to ITU-T P.501 [14] are used for the evaluation. The eight sentences (two sentences of two male and two female talkers) should be centered in a grid of $4.0\,\mathrm{s}$ as exemplarily shown in figure 2 for the German speech corpus.

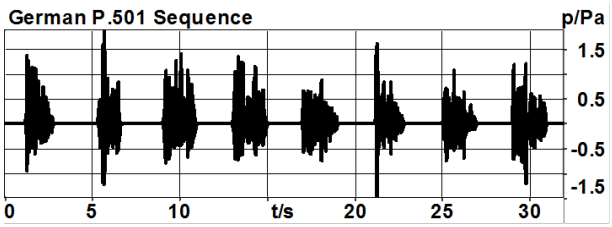


Figure 2: Example for German source signal

For the electrial insertion to the DUT, a subsequent pre-filtering according to the current application case (e.g. NB or WB) is applied. The active speech level according to [15] of this signal is calibrated to $-16.0\,\mathrm{dBm0}$, which refers to a default electrical input level for the DUT. Several volume control settings could be selected in order to investigate impacts on the listening effort. However, at least one condition including nominal receiving loudness rating (e.g. according to [16]) should be evaluated.

3. Auditory Base

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 range and aspects, 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 [11], 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 [17] and P.835 [18] 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 [17] and are provided in table 1.

For the assessment of stimuli of the listening test, the measurement setup as described in section 2 was used, but in conjunction with a mockup device. A background noise playback

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

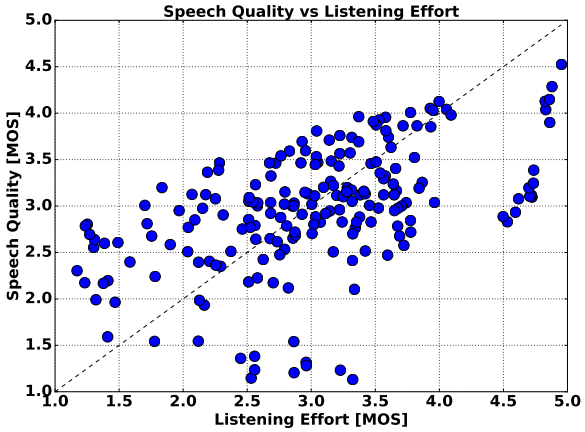


Figure 3: Speech Quality vs. Listening Effort

system according to [3] 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 *SNR(A)* $\approx -7 \ldots +15\,\mathrm{dB(A)}$. Additionally, a silence condition (noise $\leqslant 30\,\mathrm{dB(A)}$) 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 $8.0\,\mathrm{s}$ 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 participant, 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 can be regarded as almost orthogonal. 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.

4. Instrumental Testing

The structure of the new method is similar to other speech quality and/or intelligibility measures, e.g. blocks like time-

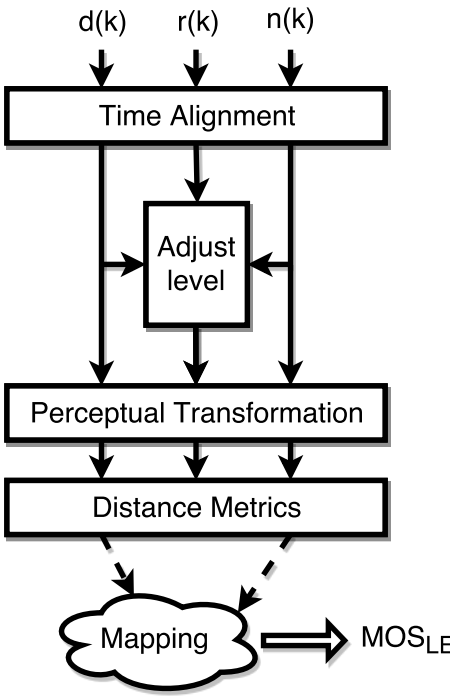


Figure 4: Block diagram of instrumental assessment

alignment and level adjustment are also present here. Unlike in other metrics like e.g. ITU-T P.863 [7], the noisy and degraded speech signal $d(k)$ must not be level-scaled, since it is an acoustically captured ear signal. It should be evaluated exactly with the real level with respect to perceived loudness. Figure 4 illustrates the general structure of the proposed assessment algorithm which expects three input signals:

- Degraded signal $d(k)$ as described in section 2.
- Noise-only signal $n(k)$ as described in section 2.
- The reference signal $r(k)$ is the speech signal which is electrically inserted to the DUT.

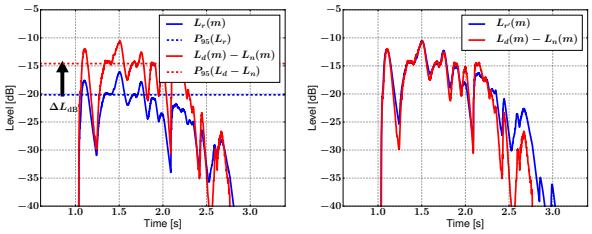
4.1. Time Alignment

For the proper time alignment, first the envelope of the cross-correlation between $d(k)$ and $r(k)$ is calculated. The delay between both signals is determined by the position of the maximum peak in the envelope function. Since $d(k)$ and $n(k)$ are already time-aligned against each other (see section 2), $n(k)$ is compensated in the same way as $d(k)$.

4.2. Reference Calibration

When feeding the reference signal $r(k)$ into the prediction model, it may have any arbitrary active speech level relative to the degraded signal $d(k)$. For the comparison between both signals, it is necessary to compensate possible bias between them. For this purpose, level vs. time according to [19] is calculated for all three input signals with a time constant of 35 ms. The resulting level signals are denoted as $L_r(m)$, $L_d(m)$ and $L_n(m)$. The estimated level vs. time of the pure degraded speech without noise $L_{d-n}(m)$ is determined in the level domain according to equation 1.

$$L_{d-n}(m) = \max\left(0, L_d(m) - L_n(m)\right) \tag{1}$$



(a) Uncompensated level **(b)** Level after compensation

Figure 5: Percentile-based reference level alignment

The level difference ΔL (on a linear scale) is determined by the ratio of 95th percentiles between estimated pure degraded speech and reference level vs. time according to equation 2.

$$\Delta L = \frac{P_{95}(L_{d-n}(m))}{P_{95}(L_r(m))} \tag{2}$$

Finally, the scaled reference signal $r'(k)$ can be determined according to equation 3. The principle of the level calibration method is exemplarily illustrated in 5.

$$r'(k) = \Delta L \cdot r(k) \tag{3}$$

Based on the level vs. time of the reference signal $L_{r'}(m)$, a speech frame classification according to ITU-T G.160 Appendix II is performed [20]. For each time frame, an indicator for high (H), mid (M) and low (L) speech activity is provided. Additionally, pause (P) and silent frames (S) are reported. Finally, all active time frames are combined in a meta class A as defined by equation 4.

$$A = \{H, M, L, P\} \tag{4}$$

4.3. Psychoacoustic Core Model

For the perceptual modeling, the algorithm known as *Relative Approach* is employed as a hearing-adequate time-frequency transformation. The algorithm introduced in [21] and [22] models a major characteristic of human hearing: the much stronger subjective response to distinct patterns (tones and/or relatively rapid time-varying structure) than to slowly changing levels and loudnesses. Thus this representation detects noticeable patterns of audio signals in the time-frequency domain.

The algorithm is already used in several other applications, e.g. for the evaluation of packet loss scenarios [23] and speech quality assessment according to [8] and [9].

For the proposed prediction model, time frames of 10.0 ms and a filter-bank resolution of 1/12 octave are chosen. In the following, the time-frequency representations of the previously mentioned signals are denoted as $RA_x(m, j)$, with $x \in \{d, n, r'\}$. Here (m, j) refers to the mth time frame and the jth frequency band. As an intermediate representation, $RA_s(m, j)$ is calculated according to equation 5 and refers to an estimation of the spectral representation of the degraded speech signal without noise.

$$RA_s(m, j) = \max\left(0, RA_d(m, j) - RA_n(m, j)\right) \tag{5}$$

4.4. Distance Metrics

Based on the spectral representations of the signals, single value metrics correlating with the auditory results. For this purpose,

a correlation measure $\mathrm{Corr}(X,Y)$ for two arbitrary spectra X and Y according to equation 6 is introduced. Here the activity class A as described in section 4.2 is utilized, i.e. that the calculation is carried out only over the active and paused time frames. In the frequency domain, only the WB frequency range $\Delta F = 100 \ldots 7000\,\mathrm{Hz}$ is evaluated.

$$\mathrm{Corr}\,(X,Y) = \frac{\sum_{m\in A}\sum_{j\in\Delta F}(X(m,j)-\bar{X})\,(Y(m,j)-\bar{Y})}{\sqrt{\sum_{m\in A}\sum_{j\in\Delta F}(X(m,j)-\bar{X})^2 \cdot \sum_{m\in A}\sum_{j\in\Delta F}(Y(m,j)-\bar{Y})^2}} \tag{6}$$

The average values $\bar{X}$ and $\bar{Y}$ are provided in equation 7. Here N_A denotes the number of active time frames and $N_{\Delta F}$ the number of frequency bands included in ΔF.

$$\left[\bar{X},\bar{Y}\right] = \frac{1}{N_{\Delta F}\cdot N_A}\sum_{m\in A}\sum_{j\in\Delta F}\left[X,Y\right](m,j) \tag{7}$$

With this introduced correlation measure, the similarity between the estimated noise-free speech RA_s and $RA_{r'}$ can be calculated according to 8. This index $m_{SR'}$ provides a measure for the remaining structure of the degraded speech compared to the reference.

$$m_{SR'} = \mathrm{Corr}\,(RA_s(m,j), RA_{r'}(m,j)) \tag{8}$$

As a second measure $m_{DR'}$ is determined by 9 and employs the time-frequency representations RA_d and $RA_{r'}$. This metric takes the perceived noise into account by comparing noisy degraded speech and the clean reference.

$$m_{DR'} = \mathrm{Corr}\,(RA_d(m,j), RA_{r'}(m,j)) \tag{9}$$

4.5. Mapping

The two extracted features $m_{SR'}$ and $m_{DR'}$ are mapped with a simple linear regression against the auditory $\mathrm{MOS_{LE}}$.

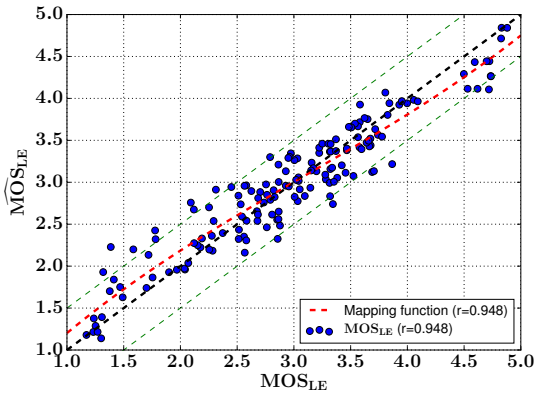
$$\widehat{\mathrm{MOS}}_{\mathrm{LE}} = a_0 + a_1\cdot m_{SR'} + a_2\cdot m_{DR'} \tag{10}$$

Other machine learning algorithms like support vector regression (SVR) or neural networks would also be possible here to achieve a better mapping. However, since the performance metrics are already located at the upper realistic range, any further improvement may lead to decreased generalization.

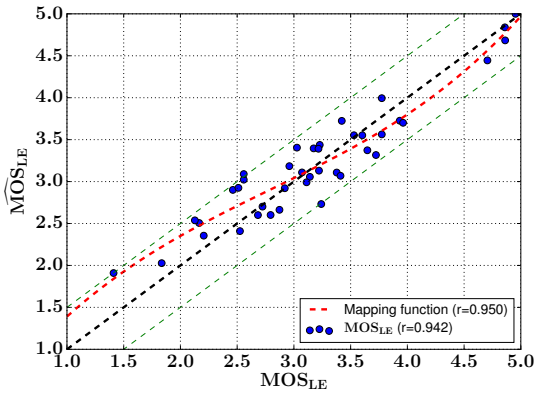
5. Results

For the training of the model 147 conditions (588 samples) are utilized. 50 conditions (200 samples) remain for validation check. Prediction results for instrumental listening effort $\widehat{\mathrm{MOS}}_{\mathrm{LE}}$ are evaluated graphically as shown in Figure 6. For training and validation, the proposed model performs adequately over the whole MOS range.

In order to qualify the performance of the model, several accuracy metrics are provided in table 2. Here the well-known correlation coefficients r_{Pearson} and r_{Spearman} are listed, as well as root-mean-square error (*RMSE*) according to [24]. Another widely used measure for the performance of prediction models is the so-called "epsilon-insensitive RMSE" as described in [24], which takes the 95% confidence intervals of the auditory data into account. All metrics are provided before and after third order mapping.



(a) Training



(b) Validation

Figure 6: Instrumental results for listening effort

Metric	Training		Validation	
	raw	3rd order	raw	3rd order
r_{Pearson}	0.936	0.948	0.942	0.950
r_{Spearman}	0.948	0.948	0.899	0.899
*RMSE**	0.140	0.145	0.150	0.151
RMSE	0.282	0.245	0.277	0.222

Table 2: Performance metrics for proposed model

6. Conclusions

In the presented work, a model for the instrumental assessment of perceived listening effort was presented. The corresponding measurement setup as well as a new auditory test was introduced. The prediction model which consists of several blocks for pre-processing, perceptual transformation and feature extraction was described.

For future work, several improvements and new considerations could be taken into account. The current auditory evaluation only included a fixed listening level of $79.0\,\mathrm{dB_{SPL}}$ and thus the model may be unconditioned for varying levels,

Another enhancement could be the extension to other receive-side applications (e.g. any kind of hands-free scenarios, public address systems). Here the model must also consider binaural perception effects.

Finally, an extended model for the combined assessment of listening effort and speech quality as introduced by the work in [11] would be desirable.

7. References

[1] B. Sauert, "Near-end listening enhancement: Theory and application," Ph.D. dissertation, RWTH Aachen, 2014.

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

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

[4] IEC 60268-16, *Objective rating of speech intelligibility by speech transmission index*, International Electrotechnical Commission, 2011.

[5] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-T Recommendation P.862, Feb. 2001.

[6] *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, ITU-T Recommendation P.862.2, Nov. 2007.

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

[8] *Part 3: Background noise transmission - Objective test methods*, ETSI EG 202 396-3 V1.5.1, Oct. 2015.

[9] *Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals-objective test methods*, ETSI TS 103 106 V1.3.1, Apr. 2014.

[10] ANSI S3.5, *Methods for the Calculation of the Speech Intelligibility Index*, American National Standards Institute, 1997.

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

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

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

[14] *Test signals for use in telephonometry*, ITU-T Recommendation P.501, Jan. 2012.

[15] *Objective measurement of active speech level*, ITU-T Recommendation P.56, Dec. 2011.

[16] *Calculation of loudness ratings for telephone sets*, ITU-T Recommendation P.79, Nov. 2007.

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

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

[19] IEC 61672-1, *Electroacoustics - Sound level meters*, International Electrotechnical Commission, 2013.

[20] *Voice enhancement devices - Appendix II*, ITU-T Recommendation G.160 Amendment 2, Mar. 2011.

[21] K. Genuit, "Objective evaluation of acoustic quality based on a relative approach," in *Internoise*, Liverpool, UK, Jul. 1996.

[22] R. Sottek and K. Genuit, "Models of signal processing in human hearing," *International Journal of Electronics and Communications*, vol. 59, pp. 157–165, 2005.

[23] F. Kettler, H.W. Gierlich, and F. Rosenberger, "Application of the relative approach to optimize packet loss concealment implementations," in *Fortschritte der Akustik - DAGA 2003*, Aachen, Germany, Mar. 2003.

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