PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Prediction of speech and noise quality for super-wideband and fullband transmission

Jan Reimes

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

Abstract

ITU-T Rec. P.835 provides a well-established listening test procedure for the auditory performance evaluation of devices containing signal processing components for noise reduction. In contrast to "classical" listening tests according to ITU-T Rec. P.800, test subjects are asked for independent votes for speech quality (S-MOS), noise intrusiveness (N-MOS) and overall/global quality (G-MOS).

However, since auditory testing is time-consuming and expensive, several instrumental models for quality prediction according to ITU-T Rec. P.835 were developed and standardized in the past. These models are already commercially available and widely used in industry for many years. However, so far only narrowband and wideband speech signals were supported. As devices with a higher acoustic bandwidth (e.g., terminals supporting the EVS-codec) are currently entering the market, an extension to super-wideband and fullband mode was necessary.

The new prediction model was developed in close interaction with several standardization committees. A common auditory testing framework was specified and a considerable number of auditory databases for training and validation was created. This contribution provides an overview of the final prediction model, which was recently introduced and standardized as ETSI TS 103 281. In addition, several prediction results from validation databases are presented.

Keywords: speech quality, noise intrusiveness, prediction, model

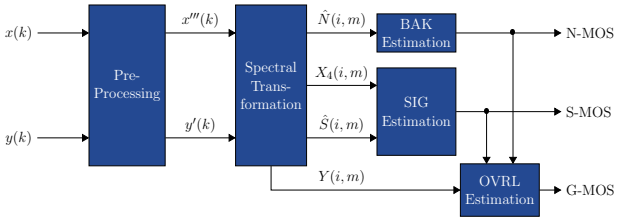
1 INTRODUCTION

Speech communication terminals including noise reduction techniques, like e.g., mobile phones, usually have to make a trade-off between a comfortable removal of noise and an intelligible, non-distorted residual speech signal. Existing quality prediction methods are not applicable for several reasons: For example, the commonly used method ITU-T Rec. P.863 [1], models a listening test design according to ITU-T Rec. P.800 and does not provide information on the three quality attributes according to ITU-T Rec. P.835 [2].

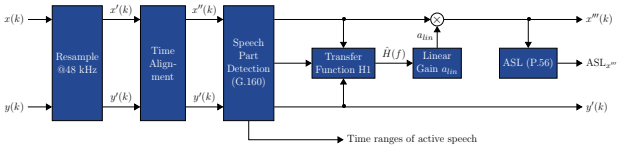
The obsolete, but still commonly used prediction models ITU-T Rec. P.862 [3] and ITU-T Rec. P.862 [4] explicitly exclude use cases of acoustic insertion paths and noise reduction – and are only specified for narrowband (NB) or wideband (WB) scenarios.

Instrumental methods capable of predicting these different quality dimensions are already available for several applications in NB or WB, e.g. [5]. However, as it is desired to accurately assess speech signals coded and processed with state-of-the-art technologies (e.g., EVS-SWB or -FB, containing audio bandwidth up to 20 kHz), a new development was started by standardization bodies.

A preliminary version of the model was already introduced in previous work [6]. This contribution provides an overview about the final prediction model for S-MOS, N-MOS and G-MOS, which was recently standardized as ETSI TS 103 281 (model A).



(a) Overall view



(b) Pre-processing

Figure 1

2 MODEL DESCRIPTION

2.1 Overview

Due to complexity of the whole model, only a brief overview of the prediction algorithm is introduced in the following sections. Functionality of the different blocks and sub-blocks is mainly provided as high-level descriptions. These focus on some key stages, which are unique compared to other speech quality prediction algorithms. For a complete and more detailed description, see clause 6.3 of [7].

Figure 1a shows the overall flow chart diagram of the proposed prediction model. The *degraded signal* $y(k)$ contains processed speech and residual noise. The model assumes an input provided in the unit *Pascal*, i.e. that $y(k)$ is already scaled to a suitable presentation level of 73 $\mathrm{dB_{SPL}}$ (see also listening test description in Annex D of [7]). Beside $y(k)$, also the clean speech reference $x(k)$ is needed as an input. Similar to related quality prediction models, a "full-reference" (or sometimes also called "double-ended") approach is chosen here.

2.2 Pre-processing

Pre-processing of input signals is an indispensable step in all common speech quality prediction models. It ensures that systematic differences between $y(k)$ and $x(k)$ (regarding e.g., delay or level) are removed and that subsequent analyses (in e.g., time-frequency representation) only detect audible differences.

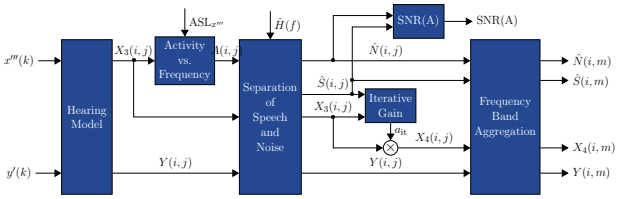
Figure 1a depicts the different sub-stages of the pre-processing. First, both input signals are resampled to 48 kHz and then time-aligned by a cross-correlation analysis. With the activity classification algorithm described in [8], active speech and pauses are determined. In order to quantify possible band limitations, level differences and/or non-correlated noise components between both signals, the transfer function $\hat{H}(f)$ is calculated by the *H1 methodology* (cross-power spectral density) on a short-term basis. Based on this intermediate result, the gain a between both signals can easily be obtained. The reference $x(k)$ is scaled by this gain, i.e. the pre-processed signals $y'(k)$ and $x'''(k)$ provide approximately the same active speech level $\mathrm{ASL}_{x'''}$, which is calculated according to [9].

2.3 Spectral Transformation

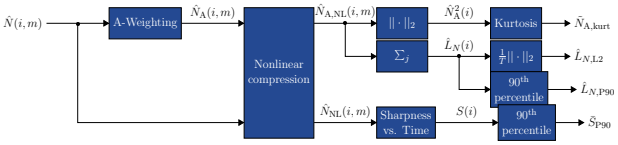
In the next block, the pre-processed signals are transformed into the time-frequency domain. Similar to other speech quality prediction models, a hearing model representation close to human perception is chosen. Figure 2a provides an overview of this sub-block for the spectral transformation.

2.3.1 Hearing model of Sottek

The hearing model according to Sottek [10] is utilized for the time-frequency representation of the input signals. This transformation includes an auditory filter bank representation of the signal and a hearing-adequate envelope



(a) Spectral Transformation



(b) Metric calculation N-MOS

Figure 2

determination which leads to a frame resolution of about 8 ms. The frequency bandwidth $\Delta f(f_m)$ for the m-th frequency band is given by equation 1.

$$\Delta f(f_m) = \Delta f(f_0) + c \cdot f_m. \tag{1}$$

The initial bandwidth $\Delta f(f_0)$ is set to 70 Hz and the factor c equals 0.14, which results in $M = 27$ bands up to 20 kHz. In contrast to other hearing-adequate frequency scales, the proposed method includes the whole fullband (FB) range.

In addition, each band is divided into $Q = 3$ equidistant sub-bands. This leads to $J = Q \cdot M = 81$ bands, which are denoted with index j. This high resolution in frequency is necessary for the separation of speech and noise (see section 2.3.2). At a later stage, both spectra versus time are aggregated back to $M = 27$ by applying a squared average, as exemplarily shown in equation 2 for the degraded spectrum.

$$Y(i,m) = \sqrt{\frac{1}{Q} \sum_{j=m\cdot Q}^{Q\cdot(m+1)-1} Y(i,j)^2} \tag{2}$$

This time-frequency representation is calculated for the pre-processed signals $y'(k)$ and $x'''(k)$. At this sub-block, these spectra $Y(i,j)$ and $X_4(i,j)$ exclude the non-linear compression (or *loudness transformation*) described in [10], which is applied at later stages.

2.3.2 Separation of speech and noise

For the calculation of metrics correlating with auditorily assessed noise intrusiveness and speech distortion, it is useful to separate the degraded spectrum $Y(i,j)$ into a speech- ($S(i,j)$) and noise-only ($N(i,j)$) component, as per equation 3. A (pseudo-)Wiener filter is used for this purpose, as shown in equation 4.

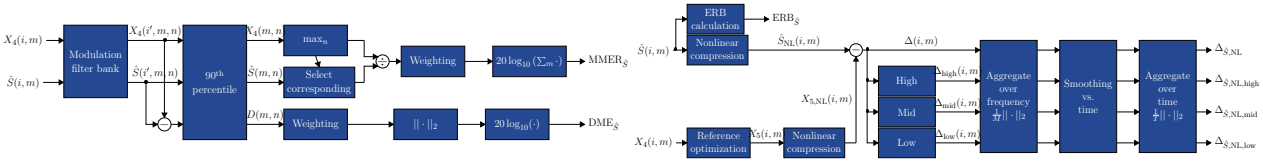
$$Y(i,j) = S(i,j) + N(i,j) \tag{3}$$

$$S(i,j) \approx \hat{S}(i,j) = Y(i,j) \cdot W(i,j) \tag{4}$$

The Wiener gain $W(i,j)$ is obtained according to equation 5, utilizing a noise estimation $\hat{N}(i,j)$ and a compensated reference $\tilde{X}_3(i,j)$, which can be regarded as rough estimate for $\hat{S}(i,j)$. The latter spectrum is determined as $X_3(i,j)$ multiplied (filtered) by the magnitude of $\left|\hat{H}'(i,j)\right|$ ($\hat{H}(f)$ interpolated to hearing model bands).

$$W(i,j) = \sqrt{\frac{\tilde{X}_3(i,j)^2}{\tilde{X}_3(i,j)^2 + \hat{N}(i,j)^2}} \quad \text{with} \quad \tilde{X}_3(i,j) = X_3(i,j) \cdot \left|\hat{H}'(i,j)\right| \tag{5}$$

For an estimation of the separated noise component $\hat{N}(i,j)$, first a binary mask of active time-frequency bins is calculated based on $\tilde{X}_3(i,j)$. The methodology described in [8] is applied on the level-vs-time for each frequency



(a) Modulation-based

(b) Spectral distance-based

Figure 3. S-MOS metric calculations

band j, in order to classify active or inactive speech frames. The resulting activity mask $A(i,j)$ is multiplied by the degraded spectrum $Y(i,j)$ and suppresses parts of active speech, noise-only bins remain. Each single frequency band j of $A(i,j) \cdot Y(i,j)$ can be regarded as a windowed time signal, which contains multiple zeros with unknown (noise) content. In order to reconstruct these missing parts, the deconvolution algorithm described in e.g., [11], is applied to obtain the noise estimate $\hat{N}(i,j)$.

As described in section 2.2, the input signal $x'''(k)$ includes already a level calibration in order to provide a similar speech level as $y'(k)$. In order to refine the with regard to temporal clipping and/or decreased frequency bandwidth, an additional iterative scaling procedure is applied on the reference spectrum $X_3(i,j)$, resulting in $X_4(i,j)$. This procedure is already extensively described in [6] and is left out here for sake of brevity.

2.4 Assessment of N-MOS

For the instrumental assessment of N-MOS, several metrics are calculated based on the noise-only spectrum $\hat{N}(i,m)$. Since this quality attribute mainly depends on the perceived annoyance, psycho-acoustic parameters like loudness or sharpness are addressed by these metrics. In contrast to common and/or standardized methods, measures for the prediction model were developed and optimized with regard to the underlying auditory databases.

Figure 2b depicts the different sub-blocks for the metric calculations. Functionality like e.g., A-weighting, non-linear compression (see section 2.3.1) and different methods for aggregation versus time and/or frequency are used here. More detailed descriptions of the specific calculations of metrics can be found in [7]. Finally, a Random Forest Regression (RFR) [12] with four features is used for the estimation of N-MOS.

2.5 Assessment of S-MOS

The auditory assessment of S-MOS is typically the hardest task for subjects in a listening test according to ITU-T Rec. P.835. The term *speech distortion* may address several aspects like e.g., decrease of bandwidth, temporal clipping, added artifacts, changes in modulation, level/loudness and/or signal-to-noise ratio (SNR) (cf. [13, 14]). In consequence, any instrumental model has to analyze multiple possible impacts of degradation.

In the prediction algorithm for S-MOS, the separated speech component $\hat{S}(i,m)$ and the reference $X_4(i,m)$ are used the calculation of comparison metrics, which can be divided into three classes of disturbances described in the following subsections. Similar as for N-MOS, a RFR is used again for the combination of nine features to the prediction of S-MOS.

2.5.1 Basic Metrics

Several basic metrics can be calculated with the separated components $\hat{S}(i,m)$ and $\hat{N}(i,m)$:

- Speech level $\mathrm{ASL}_{\hat{S}}$ (only active time ranges, aggregate energy of $\hat{S}(i,m)$ vs. frequency).
- SNR: ratio of active speech level and averaged A-weighted noise level:
  $\mathrm{SNR(A)} = \mathrm{ASL}_{\hat{S}} - 10 \cdot \log_{10}\left(\frac{1}{T_i}\sum_i \sum_{m=1}^{M} \hat{N}(i,m)^2\right)$
- Equivalent Rectangular Bandwidth $\mathrm{ERB}_{\hat{S}}$ according to [13].

More detailed descriptions of the specific calculations of metrics can be found in [7].

2.5.2 Modulation-based Metrics

Figure 3a provides a flow chart of the two modulation-based metrics. The spectral representations of the speech component $\hat{S}(i,m)$ and the reference $X_4(i,m)$ are used as the inputs of this sub-block.

In a first analysis, both spectra are transformed by a modulation filter bank according to [15], which utilizes $N = 8$ sub-bands per frequency index m with center frequencies $f_{c,mod}(n)$, ranging from 4 Hz to 128 Hz. The analysis described in [15] calculates modulation energy based on sub-frames of 256 ms, denoted as index i'.

The resulting 4D-representations $\hat{S}(i',m,n)$, $X_4(i',m,n)$ as well as the difference $D(i',m,n) = \left|\hat{S}(i',m,n) - X_4(i',m,n)\right|$ are then first aggregated versus time. To address only the active frames, the 90 % percentile is used. As an intermediate result, three average spectra versus frequency index m and modulation index n are obtained.

To emphasize critical frequency bands on the one hand and slightly depreciate low and high frequencies, a modified A-weighting function [16] is derived according to equation 6.

$$A_{\text{clip}}(m) = \max\left(-10\,\text{dB}, A(m)\right) \tag{6}$$

For the metric *Difference in Modulation Energy* (DME), $D(m,n)$ is aggregated versus bands and versus modulation frequencies, as shown in equation 7.

$$\text{DME}_{\hat{S}} = 20\log_{10}\left(\frac{1}{N}\sum_{n=0}^{N-1}\sqrt{\sum_{m=0}^{M-1}\left(A_{\text{clip}}(m)\cdot D(m,n)\right)^2}\right) \tag{7}$$

In another branch, the second metric is derived from the principle described in [15]. For each frequency band m in the reference $X_4(m,n)$, the maximum magnitude across all modulation bands is determined. These corresponding modulation indices $n_{\max}(m)$ are then selected in $\hat{S}(m,n)$ to obtain the *maximum modulation energy ratio* (MMER), as given by equation 8.

$$\text{MMER}_{\hat{S}} = 20\log_{10}\left(A_{\text{clip}}(m)\cdot\frac{\hat{S}(n_{\max},m)}{X_4(n_{\max},m)}\right) \tag{8}$$

2.5.3 Spectral Distance Metrics

Similar to in other speech quality prediction models, a processing step called *reference optimization* is applied on $X_4(i,m)$. The goal of this procedure is to compensate the reference spectrum for inaudible differences especially for the spectral distance metrics, i.e. to adjust it towards the degraded spectrum. This procedure is already extensively described in [6] and is left out here for sake of brevity. The output of this optimization step is denoted as $X_5(i,m)$ and is only used for the spectral distance metrics (cf. 3b).

Metrics based on the difference (or ratio) between degraded and (optimized) reference spectra are often used in speech and audio quality prediction. Spectral differences usually correlate with the perceived disturbance and are then aggregated versus time. In the current prediction model, this approach is also used in several ways. Figure 3b illustrates the flow chart for the determination of distance-based metrics.

First, both hearing model spectra $\hat{S}(i,m)$ and $X_5(i,m)$ are transformed by the non-linear loudness function (see section 2.3.1) to $\hat{S}_{\text{NL}}(i,m)$ and $X_{5,\text{NL}}(i,m)$, respectively. Then the delta-spectrum $\Delta(i,m)$ is determined by subtracting each time-frequency bin of speech and clean reference spectra. In order to obtain a single value of this analysis, an aggregate versus frequency is performed by an L2-norm and averaging across frequency. Temporal smoothing of the resulting curve vs. time is then applied by a median filter (window size 40 ms). Finally, the single value $\Delta_{\hat{S},\text{NL}}$ obtained by again applying an L2-norm versus time, taking only active speech parts into account.

In a similar way, three related metrics are calculated. In three additional branches, $\Delta(i,m)$ is separated into three frequency bands for *low* (50-3800 Hz), *mid* (3.8-7.8 kHz) and *high* (7.8-20 kHz) as shown in Figure 3b. These subsets $\Delta_{\text{low}}(i,m)$, $\Delta_{\text{mid}}(i,m)$ and $\Delta_{\text{high}}(i,m)$ contribute in different ways to speech quality perception. In the same way as described above, the single values $\Delta_{\hat{S},\text{NL,low}}$, $\Delta_{\hat{S},\text{NL,mid}}$ and $\Delta_{\hat{S},\text{NL,high}}$ are calculated.

2.6 Assessment of G-MOS

In several ITU-T Rec. P.835 prediction models, the overall quality G-MOS is usually obtained by combining the outputs for S-MOS and N-MOS (by e.g. linear regression). In the proposed model, this approach is extended by an additional similarity measure $\Delta_{Y,\mathrm{NL}}$ between degraded and reference spectra. It is calculated exactly in the same way as $\Delta_{\hat{s},\mathrm{NL}}$ described in section 2.5.3, but utilizing the spectrum $Y(i,m)$ (includes speech plus residual noise). The resulting three features for the estimation of G-MOS are again combined with a RFR.

3 TRAINING AND TEST DATABASES

Obviously, the introduced prediction model has to be trained and tested with data originated from auditory tests. In order to reduce variance and ensure comparability, the 3GPP work item DESUDAPS [17] was initiated. Within this work item in standardization, more than 20 comprehensive auditory databases were contributed by different parties to the pool.

Each database followed a framework extending the existing guidelines according to ITU-T Rec. P.835 and consists of 32-48 conditions each. 8-16 samples and at least four talkers per condition were used. By providing databases with American English, German and Mandarin speech sequences, it was made possible to develop a model robust to language, talkers and/or specific samples.

Each database contains recordings of real and simulated terminals in conjunction with realistic background noise field simulation according to e.g., [18]. Several applications and use cases like e.g., handset, headset and hands-free modes are represented in this comprehensive pool of databases.

Annex B of [7] provides a complete list of databases used for training and testing[1]. For each database declared as training, the metrics as described in section 2 were calculated and the three RFR estimators were fitted to the auditory data on a per-sample basis.

Two test databases (DES-25 and DES-26, cf. clause 8.4 and 8.5 of [7]) provided speech recordings that address typical use cases (handset, headset and hands-free mode of mobile phones) and noise scenarios. During the development of the prediction model, the auditory data of the test databases were unknown and could not be used for training/validation/testing purposes. Instrumental results were obtained by an independent third party laboratory. According to the approved performance requirements defined in standardization [19], the prediction model had to pass certain thresholds:

- $rmse^* < 0.30$ for S-MOS and $rmse^* < 0.25$ for N-MOS,
- $maxabs^* < 0.5$ for S-MOS and N-MOS.

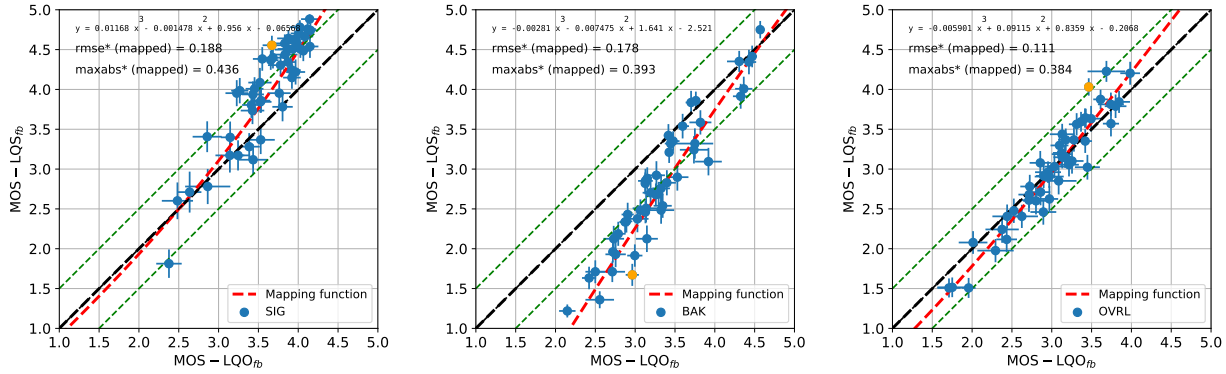
Figure 4 and 5 provides prediction results per condition of the two test databases in form of a scatter plot. In addition, prediction performance metrics $rmse^*$ (according to ITU-T Rec. P.1401 [20]) and $maxabs^*$ (cf. clause 8.1 of [7]) are provided for each attribute inside the plot.

For S-MOS (Figure 4a and 5a), the model provides a highly accurate prediction in the medium quality range. A slight underprediction for the upper range in both databases, which can be compensated by a 3rd-order mapping function. Performance metrics for S-MOS meet the requirements mentioned above.

For N-MOS (Figure 4b and 5b), the absolute prediction is high accurate across the whole quality attribute, but including an overprediction at the lower end. This effect is most likely due to the training of the model, where quite bad noise conditions were included. Also for N-MOS, the requirements regarding performance metrics are met after a 3rd-order mapping function.

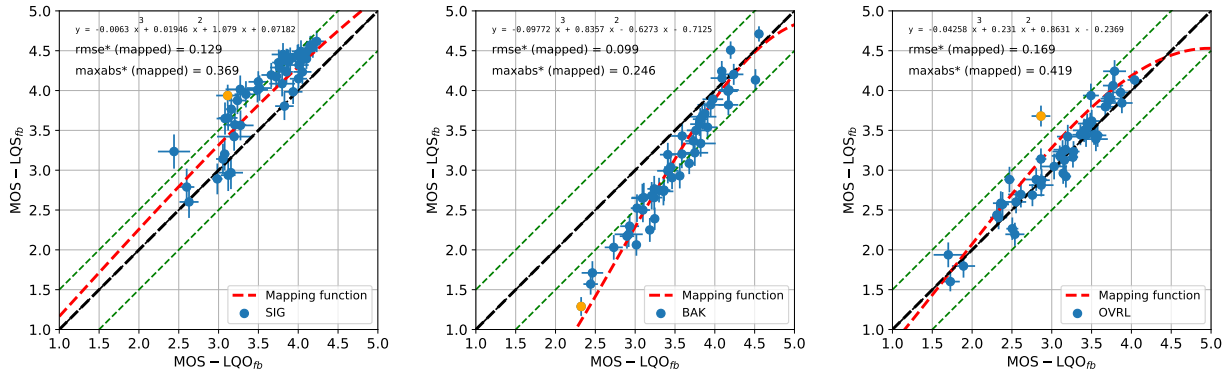
As already described in section 2.6, G-MOS is explained by the most part by S-MOS and N-MOS. Thus, no specific requirement was set here. However, the model also provides excellent prediction accuracy for this attribute (Figure 4c and 5c), even without any additional mapping function.

[1] In the field of standardization, the term *validation* is commonly used for the "blind" evaluation of new algorithmic methods like e.g., speech/audio codecs or quality prediction models. This terminology is contrary or even confusing with regard to recent developments in machine learning, where typically the term *test* is used for this purpose. *Validation* is used for an optimization step during the training process that also requires non-training related data. For sake of consistency, the term *test* is used in this contribution for the "blind" model evaluation.



(a) S-MOS (b) N-MOS (c) G-MOS

Figure 4. Prediction results of test database DES-25



(a) S-MOS (b) N-MOS (c) G-MOS

Figure 5. Prediction results of test database DES-26

4 CONCLUSION

Based on an excellent collaboration between different parties in standardization, a new prediction model for the instrumental assessment of speech distortion, noise intrusiveness and overall quality was developed and finally standardized as ETSI TS 103 281 (model A). This work summarized some of the algorithmic key features of the new model as well as results of test databases, which were evaluated by a third-party. Based on numerous training databases, the introduced model provides a highly accurate prediction performance for the evaluation of noise suppression capabilities of terminals.

In contrast to existing prediction methods [5], the new model does not require a second reference signal (*unprocessed reference*) as an input anymore. Thus, the new approach could be extended also for other use cases and applications, like e.g. vehicle hands-free for the usage in measurement specifications ITU-T Rec. P.1100/1110/1120. A possible update may also provide an implicit or explicit operational mode for NB or WB bandwidth.

REFERENCES

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

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

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

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

[5] ETSI TS 103 106 V1.5.1. *Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals-objective test methods*, March 2018.

[6] Jan Reimes and Hans W. Gierlich. Instrumental speech and noise quality assessment for super-wideband and fullband transmission. In *ITG-Fachtagung Sprachkommunikation*. VDE Verlag, Paderborn, Germany, September 2016.

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

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

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

[10] Roland Sottek. A hearing model approach to time-varying loudness. *Acta Acustica united with Acustica*, 102(4):725–744, Jul / Aug 2016.

[11] Til Aach and Volker Metzler. Defect interpolation in digital radiography - how object-oriented transform coding helps. *SPIE*, 4332, 2001.

[12] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[13] Nicolas Côté. *Integral and Diagnostic Intrusive Prediction of Speech Quality*. PhD thesis, TU Berlin, 2011.

[14] Marcel Wältermann. *Dimension-based Quality Modeling of Transmitted Speech*. PhD thesis, Fakultät für Elektrotechnik und Informatik, Technische Universität Berlin, Germany, 2012.

[15] Tiago Falk, Chenxi Zheng, and Wai-Yip Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18:1766 – 1774, 10 2010.

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

[17] 3GPP SA4. S4-160548 - DESUDAPS-1: Common subjective testing framework for training and validation of SWB and FB P.835 test predictors, Apr. 2016. Memphis, USA.

[18] ETSI TS 103 224 V1.3.1. *A sound field reproduction method for terminal testing including a background noise database*, July 2017.

[19] 3GPP S4-160747. *Requirements for SWB/FB P.835 objective predictor model(s)*, Jul. 2016.

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