# Diagnostic Prediction of Transmitted Speech Quality: A New Framework for Signal-based and Parametric Models

*Sebastian Möller*[1], *Marcel Wältermann*[1], *Nicolas Côté*[2]

[1]Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
[2]Institute of Electronics, Microelectronics and Nanotechnology, UMR CNRS 8520, ISEN, Lille, France
sebastian.moeller@telekom.de, mar.wael@gmx.de, nicolas.cote@isen.fr

## Abstract

In this paper, we present a new framework for the diagnostic prediction of transmitted speech quality. The idea is to extract perceptually relevant feature estimations from the speech signal, and combine them with an overall quality metric in order to obtain more reliable as well as more diagnostic predictions of speech quality. We implement this framework in two complementary ways: In terms of a signal-based model which can be used for online and offline measurement, and in terms of a parametric model which can be used for network planning. The implementations are compared to standard state-of-the-art models and show a similar level of reliability, while providing additional diagnostic value.

**Index Terms**: speech quality assessment, speech transmission, quality prediction, Quality of Experience

## 1. Introduction

With the advent of mixed interconnected networks including Voice-over IP (VoIP) technology, speech quality prediction models have become more and more important to ensure proper network planning, set-up and monitoring. The aim is to provide an optimum quality to the user while limiting the technical resources, and thus expenses. Most prediction models which are used for this purpose provide an "overall quality index" in terms of a Mean Opinion Score (MOS), which characterizes the listening-only or conversational quality in an *integral* way. The prediction either refers to a particular part of the channel (e.g. a codec), or to the entire transmission path mouth-to-ear (sometimes also called "overall quality"). Predictions are based either on the signals which are available at the output and (mostly) also at the input of the transmission channel, and comparing them on a perceptual and cognitive level, or by describing the elements of the transmission channel via parameters and mapping these parameters to integral quality. Prominent examples of the former type of model are PESQ [1] and POLQA [2], and for the latter the E-model [3].

Unfortunately, the MOS alone provides little insight into *why* speech transmitted over a particular channel is perceived not with the highest (optimum) quality. In order to uncover the sources, it is necessary to find out about the perceptual reasons underlying the quality judgment. From psychophysics, methods are known that are able to identify the underlying reasons of a perceptual event (e.g. when listening to a degraded speech signal), in terms of perceptual dimensions. Because the judgment of quality is also based on a perceptual event, we can use such methods also to identify the underlying reasons of sub-optimum quality, i.e. perceptually-relevant speech quality dimensions [4].

More precisely, our approach is threefold: First, perceptual dimensions have to be identified. Then, algorithms have to be developed which estimate values for each perceptual dimension, in terms of its impact on the integral quality. Finally, algorithms have to be developed which integrate the degradations on all perceptual dimensions with an estimation of the maximum quality which can be achieved, in order to form an estimation of the integral quality of the heard speech sample. In the end, such a dimension-based model is able to provide both, an estimation of the integral quality and a diagnostic profile of dimension estimates which pinpoints to the root of sub-optimal quality.

In this paper, we follow this new dimension-based approach and provide two complementary implementations to it, one based on signals and the other based on parameters. The structure of the paper is as follows: First, candidates for dimensions and corresponding extraction methods are reviewed in Section 2. Section 3 presents a first way to estimate the dimensions and the resulting integral quality on the basis of the input and output signals of the channel. In Section 4, a second way is presented to estimate the dimensions and integral quality on the basis of parameters. Both models are evaluated in Section 5 by comparing them to standardized non-diagnostic models. Conclusions and ideas for future work are presented in Section 6.

## 2. Quality dimensions

In the reminder of the paper, we limit ourselves to the prediction of listening-only quality, because psychophysical methods for dimension extraction are available only for the listening situation, and because no reliable signal-based models for the communicative situation are available. In the listening situation, dimensions have been extracted for more than 40 years (starting with the famous paper by McDermott [5]), and focusing on different channels and processing equipment which was available at that time. An overview of the relevant literature can be found in [6] and [7].

Two approaches have mainly be followed in that literature: Either a similarity of distance scaling of pairwise-presented stimuli, and a mapping of the resulting disparities to a perceptual space with a sufficiently low dimensionality (Multidimensional Scaling, MDS [8]); or an individual scaling of all stimuli on a set of pre-selected scales which are labeled with antonyms (Semantic Differential, SD [9]) and a subsequent factor analysis. In both cases, the result is a perceptual space in which stimuli are represented as data points, and where the dimensions can be labeled either on the basis of the stimuli (MDS) or of the SD scales.

Two most recent results of such exercises need to be highlighted: Wältermann et al. [10] followed both experimental paradigms and came up with a common set of 3 orthogonal

dimensions which proved to be valid both for narrowband (300-3400 Hz) and wideband (50-7000 Hz) transmitted speech. These dimensions were termed *coloration*, *noisiness* and *discontinuity*. *Loudness* was added as a fourth dimension to account for loudness differences which are commonly balanced out in the experimental paradigm. Sen [11] followed the SD approach, in an elaborated version which is called the Diagnostic Acceptability Measure [12]. Depending on the dataset used, he separated coloration into *high-frequency absent* and *low-frequency absent*, discontinuity into *slowly-varying discontinuity* and *fast-varying discontinuity*, and noisiness into *background noise* and *signal-correlated noise*. For both dimension sets, methods have been demonstrated to directly scale each dimension on a separate scale, without the need for further time-consuming MDS or SD experiments.

For implementing our approach, we followed the three dimensions defined by Wältermann et al., because they proved to be orthogonal amongst themselves, and because they explain quality in a nearly complete way [10]. In addition, we considered the loudness dimension in the signal-based implementation.

## 3. Signal-based prediction

We implemented the approach first in terms of a signal-based model called Diagnostic Instrumental Assessment of Listening quality (DIAL). This model consists of six parts: An estimation of the maximum quality to be achieved by a *core* model (in the absence of any of the dimensions impaired), an estimation for each of the four dimensions, and a combination model integrating the five estimates into an integral quality prediction. Details of the models are given in [6].

For the core model, we used a modified version of TOSQA [13], which was a candidate in the PESQ competition and which had an adequate psychoacoustic model to serve as the basis for the quality estimation. In this model, quality is estimated by a comparison of the perceptually transformed input of the clean speech *x(t)* and degraded output *y(t)* signals. The perceptual transformation is based on Zwicker's loudness model, however compensated for non-audible degradations. The resulting similarity measure is transformed to a *MOScore* value on the MOS scale.

The model for the coloration dimension uses a perceptual representation of the frequency response of the system. Two parameters provide a simplified representation, namely the Equivalent Rectangular Bandwidth in Bark and the center frequency of the frequency response of the channel in Hz [14]. Only these two parameters are combined according to the model developed by Raake [15] providing a bandwidth impairment factor *Ibw*, which is an estimation of the quality impact of the coloration dimension. *Ibw* is further mapped to the MOS scale, in terms of a *MOSc* value, using the mapping given in [3].

The model for noisiness first estimates the additive noise in the degraded signal *y(t)* using the "silence/noisy" (i.e. without speech) frames only. The estimated parameter corresponds to a noise loudness value *Ln*. A discontinuous transmission (DTX) algorithm will avoid the transmission of the signal in silence/noisy frames; in this case, the environmental noise at the talker's side is transmitted during speech periods only, resulting in an under-estimation of *Ln*. A "Noise on Speech" (*NoS*) parameter quantifies the additive noise components during speech periods only. The final noisiness estimation *MOSn* is calculated using the maximum degradation value estimated by *Ln* and *NoS*.

The model for discontinuity follows the approach from Huo et al. [16]. It uses Weighted Spectral Slope distances and a signal-temporal loss to derive an interruption rate, an artefact rate, and a clipping rate. The dimension estimation *MOSd* is calculated as a non-linear combination of these three parameters. The model for loudness quantifies the degradation for speech heard at a non-optimum listening level. It calculates an equivalent sound pressure level *Leq* which corresponds to the mean energy of *y(t)* over all speech frames. *Leq* is then transformed into a degradation value *MOSl*.

Finally, the integration model is based on a *k*-means nearest neighbors (KNN) algorithm. It is trained on combinations of the 5 *MOSi*, $i=\{max, c, n, d, l\}$ estimates and the corresponding integral MOS value for each stimulus of the training set. The trained KNN is then used to classify unknown speech signals by mapping the predicted values of *MOSi* to the average MOS value of the *k* nearest neighbors.

## 4. Parametric prediction

The parametric model which implements our approach is called DNC (for Discontinuity, Noisiness and Coloration) and is similar to the one of the E-model, see above. The E-model bases its integral quality prediction on 19 parameters which describe the effects of different elements of the transmission channel, in terms of weighted frequency responses (so-called loudness ratings), average delay times, and average noise levels. Non-linear effects of codecs are taken into account by so-called equipment impairment factors. Time-varying degradations due to packet loss in VoIP are handled by an average loss rate, a burstiness parameter, and a robustness factor which describes the codec robustness (including any error correction schemes) against packet loss. The exact algorithm is described in [3] for narrow-band and in [18] for wideband speech.

In our dimension-based approach, we do not integrate the parameters to a one-dimensional quality estimate in a pragmatic way (as it is done in the E-model), but rather by considering the perceptual dimensions of Section 2. We use the parameters to determine 3 dimension-related impairment factors, which are subsequently subtracted from an estimation of the maximum quality (which also includes speech loudness) and finally transformed to the MOS scale. Details of the model and the exact parameter settings can be found in [7].

For the maximum quality, we use a fixed value on an internal scale which is identical to the "transmission rating scale" of the E-model. In the E-model, this value is 93.2 for an otherwise undisturbed narrowband connection with optimum loudness, and 129 for an otherwise undisturbed wideband connection with optimum loudness. For estimating the coloration, we use the same *Ibw* as in the signal-based model.

For estimating noisiness, we calculate a noisiness impairment factor as follows:

$$In = 109 + 1.5 \cdot (N_{0,n} + SLR) \quad (1)$$

where $N_{0,n}$ is the power summation of noise sources equivalent to circuit noise, send-side ambient noise and noise induced by the receiving subscriber line, and *SLR* is the loudness rating for the sending terminal. The equivalent noise sources are determined by transforming the acoustic noise levels of the background noise into electric ones, and adding a correction factor which

| No. | Transmission characteristics | DIAL | | WB-PESQ | | DNC | | WB-E-model | |
|---|---|---|---|---|---|---|---|---|---|
| | | R | ε | R | ε | R | ε | R | ε |
| A | G.722/722.2, G.711, G.723.1, G.726, G.729A, TD, BGN, BP, PL | 0.92 | 0.33 | 0.89 | 0.38 | 0.95 | 6.22 | 0.92 | 9.61 |
| B | Clean, G.722, G.722.2, G.711, G.729A, TD, MNRU, BGN, PL | 0.81 | 0.41 | 0.82 | 0.40 | 0.91 | 7.49 | 0.81 | 19.5 |

Table 1. *Model performance for integral quality (MOS) estimations. G.NNN: codecs according to ITU-T Rec. NNN; TD: codec tandems; BGN: background noise; BP: bandpass filtering; PL: packet loss; MNRU: signal-correlated noise.*

accounts for the perceptual effect of the particular noise spectrum and temporal structure compared to white noise.

For estimating discontinuity, we calculate an impairment factor

$$Idis = Idis,o + (72 - Idis,o) \cdot \frac{Ppl}{Ppl+Bpl+Mn} \qquad (2)$$

where *Idis,o* is a codec-intrinsic discontinuity impairment, *Ppl* is the packet-loss rate, *Bpl* is the robustness factor against packet loss, and *Mn* is a masking factor which accounts for the masking of discontinuities by noise.

Finally, the integration model for the parametric approach follows the assumption of a Euclidean vector space. In this space,

$$Itot = \left(\sum_{dim} I_{dim}^2\right)^{1/2} \qquad (3)$$

is the total impairment value resulting from the 3 dimension estimates which is subtracted from the maximum quality value, and then transformed onto the MOS scale, using the relationship between the impairment factor scale and the MOS scale given with the E-model.

## 5. Evaluation

The two models are evaluated two ways. First, the integral quality estimations are compared to the ones of standardized models. In the case of the signal-based implementation, we used the wideband version of the former standard PESQ (WB-PESQ, [17]) for this purpose, as the current standard POLQA was not available to us. For the parametric implementation, we compared to the wideband extension of the E-model (WB-E-model, [18]). Second, we analyze the diagnostic performance of both models by comparing them to perceptual dimension judgments which have been obtained using the direct dimension scaling method described in [7].

Two mixed narrowband and wideband databases (A and B) were used for this comparison. These were the only databases for which we have both subjective integral quality judgments (MOS) as well as dimension judgments available. Further dimension-based databases are currently under preparation in Study Group 12 of the International Telecommunication Union (ITU-T), but this work is not yet finished. An overview of the conditions contained in the databases is given in Table 1, and more details can be found in [6] and [7].

### 5.1. Performance for integral quality prediction

Table 1 presents the performance of the four models in terms of the Pearson correlation coefficient *R* and the root mean squared error ($\varepsilon$). For the signal-based models (DIAL and WB-PESQ), the values are calculated on the MOS scale [1; 5] which is the primary output scale of these models. The comparison of the parametric models (DNC and the WB-E-model) is carried out on the impairment factor scale [0; 129], which is the primary output scale of the E-model. Thus, the absolute values of $\varepsilon$ cannot be directly compared between the signal-based and the parametric models, but only within each group of models.

The results show that both dimension-based models reach a high correlation to the subjective integral quality judgments, and a relatively low prediction error. Only in one case (database B and DIAL) the correlation falls below 0.9. Comparing the dimension-based approach to the non-dimension-based one, DIAL outperforms WB-PESQ on database A, in that it provides higher correlation coefficients and lower prediction errors than WB-PESQ, and is close in performance to WB-PESQ for database B. On the parametric side, DNC performs considerably better than the WB-E-Model. However, it has to be noted that these two databases were also used for the DNC model derivation; thus, this cannot be seen as a proof for a better performance than the standard.

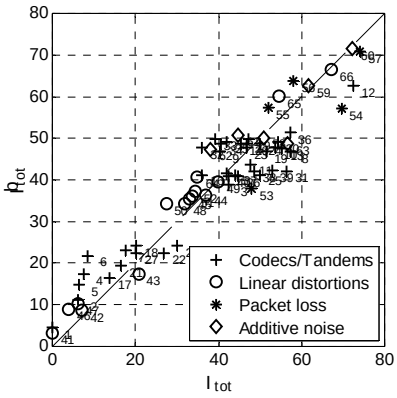


Figure 1. *Estimated integral quality values Îtot of the DNC model and subjective ratings Itot for database A.*

The models perform equally well in predicting overall quality for different types of degradations. As an example, Fig. 1 shows the scatter plot of the DNC model for database A, separating between codecs, linear distortions, packet loss and noise. Linear distortions are expected to mostly influence the coloration dimension, packet loss the discontinuity dimension, noise the noisiness dimension, whereas codecs usually provoke a mixture of these three perceptual dimensions. The scatter plot shows that the integration model of DNC is able to weight these degradations correctly with respect to their impact on integral quality, so that the overall impairment estimations are quite accurate.

### 5.2. Analysis for diagnostic power

The prediction accuracy for individual perceptual dimensions can be analyzed by calculating the correlations between the subjective dimension scores to the dimension estimates provided by DIAL and DNC. For DIAL, this has been done by calculating the correlation between *MOSc*, *MOSn* and *MOSd* to the subjective dimension judgments. For DNC, the correlations are calculated again on the impairment factor scale. The values are given in Table 2.

| No. | DIAL | | | | | | DNC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Discontinuity | | Noisiness | | Coloration | | Discontinuity | | Noisiness | | Coloration | |
| | R | ε | R | ε | R | ε | R | ε | R | ε | R | ε |
| A | 0.826 | 0.429 | 0.780 | 0.499 | 0.969 | 0.190 | 0.933 | 6.105 | 0.919 | 6.194 | 0.964 | 4.779 |
| B | 0.811 | 0.496 | 0.675 | 0.561 | 0.975 | 0.234 | 0.977 | 4.247 | 0.953 | 6.953 | 0.971 | 3.838 |

Table 2. *Model performance for perceptual dimension estimations on Databases A and B.*

The DIAL model reaches very high correlations (>0.95) and a very low prediction error for the coloration dimension, whereas the estimations for the discontinuity dimension are slightly lower, and for the noisiness dimension are only moderate-to-high. Apparently, the signal-based implementation still has some problems in estimating noisiness in an appropriate way. The DNC model reaches high correlations (> 0.90) and a low prediction error on all three scales. However, it has to be noted that this model has also used the two databases for training. In summary, the high correlation >0.8 for all but one scale shows that both model implementations are well able to provide diagnostic information with respect to the perceptual dimensions causing sub-optimal quality.

## 6. Conclusions and future work

In this paper, we presented a framework for a diagnostic as well as holistic estimation of transmitted speech quality. By combining dimension-based quality estimators to an estimation of the maximum quality achievable in a certain situation we aim at providing reliable as well as diagnostic information on the speech quality linked to a specific transmission channel. We based our framework on perceptual dimensions which are orthogonal, and which seem to be valid for transmission channels of different bandwidths.

For this framework, we provided two complementary implementations. One is based on the input and output signals of the transmission channel, and can be used for online, per sample measurement and diagnosis as well as monitoring of the system. The other is based on parameters which are known during the network planning process, and thus can be used in a prospective manner for characterizing entire channels, even before a channel has been set up. We evaluated both approaches with databases for which not only the integral quality judgments, but also the dimension judgments were available. Although the evaluation could not be carried out on unknown test databases in the parametric case, the results showed that the dimension-based approach is not worse than a standard (non-diagnostic) approach for estimating integral quality, and on most of our databases even outperformed a direct MOS estimator in terms of WB-PESQ and the WB-E-model. The big advantage of the dimension-based approach, however, is that it provides diagnostic value, in the sense of information on which perceptual dimension is impacted by a certain channel. The generally high correlations to the subjective dimension judgments show that these diagnostic predictions are reliable as well.

In the future, we would like to complement the framework by a non-intrusive signal-based implementation, i.e. an implementation which relies on the degraded output signal *y(t)* alone, combining ideas from DIAL and DNC. In addition, we will validate the approach with independent databases, and compare it to the new signal-based intrusive standard POLQA. A study item for a new dimension-based Recommendation P.AMD has been opened in Q.9 of ITU-T Study Group 12, and the DIAL model, after some improvements, could form a candidate for such a standard. We are confident that the dimension-based approach will also be applicable to other types of quality estimation problems, e.g. quality estimation for synthesized speech or for audio-visual transmission channels.

## 7. References


[1] ITU-T Rec. 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. Int. Telecomm. Union, Geneva, 2001.

[2] ITU-T Rec. P.863, Perceptual Objective Listening Quality Assessment (POLQA). Int. Telecomm. Union, Geneva, 2011.

[3] ITU-T Rec. G.107, The E-Model, a Computational Model for Use in Transmission Planning, Int. Telecomm. Union, Geneva, 2011.

[4] Heute, U., Möller, S., Raake, A., Scholz, K., Wältermann, M., "Integral and Diagnostic Speech-Quality Measurement: State of the Art, Problems, and New Approaches", in: Proc. Forum Acusticum 2005, Budapest, 1695-1700.

[5] McDermott, B., "Multidimensional Analyses of Circuit Quality Judgments". J. Acoust. Soc. Am., 45(3), 774–781, 1969.

[6] Côté, N., Integral and Diagnostic Intrusive Prediction of Speech Quality, Springer, Berlin, 2011.

[7] Wältermann, M., Dimension-based Quality Modeling of Transmitted Speech, Doctoral Dissertation, TU Berlin, 2012.

[8] Carroll, J., "Individual Differences and Multidimensional Scaling. Multidimensional Scaling", in: Theory and Applications in the Behavioral Sciences Volume I — Theory (R.N. Shepard, A.K. Romney, S.B. Nerlove, eds.), 105–155, 1972.

[9] Osgood, C., Suci, G., Tannenbaum, P., The Measurement of Meaning, University of Illinois Press, Urbana IL, 1957.

[10] Wältermann, M., Raake, A., Möller, S., "Quality Dimensions of Narrowband and Wideband Speech Transmission", Acta Acustica united with Acustica, 96(6), 1090–1103, 2010.

[11] Sen, D., "Determining the Dimensions of Speech Quality from PCA and MDS Analysis of the Diagnostic Acceptability Measure". In: Proc. MESAQUIN 2001, Prague, 2001.

[12] Voiers, W.D., "Diagnostic Acceptability Measure for Speech Communication Systems". In: Proc. ICASSP'77, 204-207, Hartfort CT, 1977.

[13] ITU-T Contrib. COM12-34, TOSQA – Telecommunication Objective Speech Quality Assessment, Int. Telecomm. Union, Geneva, 1997.

[14] Scholz, K., Wältermann, M., Huo, L., Raake, A., Möller, S., Heute, U., „Estimation of the Quality Dimension 'Directness/Frequency Content' for the Instrumental Assessment of Speech Quality", in: Proc. of Interspeech, 1523-1526, Pittsburgh PA, 2006.

[15] Raake, A., Speech Quality of VoIP — Assessment and Prediction, John Wiley & Sons, Chichester, West Sussex, 2006.

[16] Huo, L., Wältermann, M., Heute, U., Möller, S., "Estimation Model for the Speech-Quality Dimension 'Continuity'". In: Proc. 8. ITG-Fachtagung Sprachkommunikation, ITG Fachbericht Band 211, VDE Verlag, Berlin, 2008.

[17] ITU-T Rec. P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs, Int. Telecomm. Union, Geneva, 2007.

[18] ITU-T Rec. G.107.1, Wideband E-Model, Int. Telecomm. Union, Geneva, 2011.