Modeling of Integral Quality Based on Perceptual Dimensions - A Framework for a New Instrumental Speech-Quality Measure

Marcel Wältermann, Alexander Raake, Sebastian Möller

Quality and Usability Lab, Deutsche Telekom Laboratories, Berlin Institute of Technology, Germany
E-Mail: {marcel.waeltermann,alexander.raake,sebastian.moeller}@telekom.de
Web: www.qu.t-labs.tu-berlin.de

Abstract

In this contribution, the general framework for a new instrumental measure for end-to-end speech transmission quality is described. It is based on the notion that integral quality can be described by the *global* perceptual dimensions "discontinuity", "noisiness", and "coloration". The dimensions were identified through multidimensional analyses of telephone speech quality in an end-to-end context. Corresponding *dimension impairment factors* are defined which quantify the quality impairment of each dimension.

Three additional Multidimensional Scaling experiments were conducted covering effects found in each of the global dimensions. By means of the resulting *sub*-dimensions, the global dimensions can be described in more detail.

On the basis of the framework, signal-based or parametric dimension estimators are developed that predict either the quality impairment due to each dimension (i.e., the *dimension* impairment factors), or the underlying sub-dimension scores. It is shown that integral quality can be estimated by a combination of the dimension impairment factors, yielding a covered variance of $R^2 = 90\%$ of the auditory test data for a wide range of perceptually different conditions. Examples are given on how a particular dimension impairment factor can be predicted both on the basis of global dimensions and of sub-dimensions.

1 Introduction

A speech signal passes several different elements when sent through a transmission system, and its quality as perceived by the receiver might be degraded to some extent. The quality is determined by the user as the degree of deviation from the desired quality. The quality judgment is based on different *features* of the perceived speech, describing its auditory characteristics (e.g., described by attributes like *noisy*, *bright*, or *interrupted*). The perceptual features - subsumed by orthogonal dimensions - may be of different importance for the *integral* quality, and hence weighted differently in the users' process of coming to a quality judgment.

Assuming that speech quality is completely determined by a set of *known* auditory features, it is then possible to obtain a quality *estimate* by combining estimates of the features, for example using appropriate signal-based or parametric instrumental measures.

Dimension-based speech-quality measures have two important advantages over current models like PESQ [1]. Firstly, assuming that the approach is based on a more or less complete set of perceptual features of modern telephone connections, a dimension-based model is expected to reliably estimate quality even for unknown kinds of degradations. Secondly, the single dimension estimators provide perceptually adequate diagnostic information on the composition of quality. This enables system developers or network providers to identify the potential source of the quality degradation.

The remainder of the paper is structured as follows: The perceptual dimensions by which the integral quality can be described are outlined in Section 2. So-called *sub*-dimensions are introduced that provide more insight into the nature of each of the global dimensions. On that basis, the general framework for new instrumental measures for end-to-end speech transmission quality is outlined in Section 3. This way, dimension-based estimators can be developed that predict either the quality impairment due to each dimension, or the underlying sub-dimensions. In Section 4, an integral quality model is presented as a concrete realization of one part of the framework, based on the concept of *dimension impairment factors* introduced here. Furthermore, different approaches for estimating one of the dimension impairment factors are presented. Conclusions and reference to further work are given in Section 5.

2 Perceptual dimensions

2.1 Global dimensions

It is assumed that the dimension-based quality prediction approach leads to valid and reliable estimations if *all* perceptually relevant dimensions are covered by the model. In order to reveal these dimensions, extensive auditory tests (with a total of 80 participants) were carried out following the paradigms of similarity scaling of pairwise presented stimuli with subsequent Multidimensional Scaling (MDS), and attribute-scaling in the fashion of Semantic Differentials [2]. Both methods aim at representing the test stimuli in a perceptual space of low dimensionality $L \ll I$, where I is the number of conditions considered in an experiment. The space's axes reflect the underlying perceptual dimensions. Those can, in turn, be interpreted by means of the particular point configuration, where the points represent the conditions. In mathematical terms, the point configuration is represented by an $I \times L$ matrix $\mathbf{G}$ containing the point coordinates. A large distance between two points corresponds to either a high dissimilarity (Multidimensional Scaling), or to a highly differing characteristic in terms of descriptive attributes (Semantic Differential).

It is vital that the test stimuli cover virtually all kinds of potential degradations in modern networks, including electro-acoustic properties of user terminals, send-side room acoustics, narrowband and wideband codecs, signal enhancement algorithms, and variation of behavior of the underlying network (e.g., packet loss). The set of attributes for the Semantic Differential was derived from separate experiments and extended using those found in the relevant literature (see [3]). Thus, it can be assumed that they capture the whole range of semantic descriptions for the given conditions.

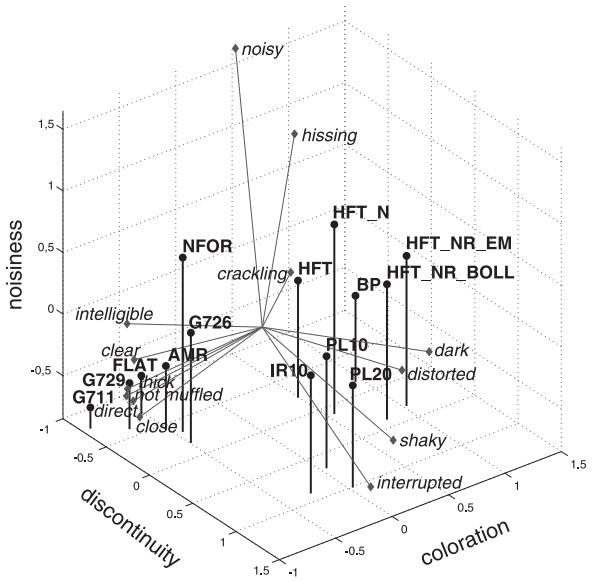

Fig. 1: Biplot for the Semantic Differential data (cf. [3]).

In addition to the experiments investigating the case of narrowband (300-3400 Hz) transmission, respective tests were conducted examining the perceptual space for future wideband transmission (50-7000 Hz). The resulting perceptual dimensions are compared in [3], showing that three *common* dimensions cover the major part of the experimental variance.

In Fig. 1, the so-called biplot of the three-dimensional narrowband configuration is depicted, representing the attributes and the dimension scores, i.e. the point coordinates stored in $\mathbf{G}$, in a single plot [3]. The configuration was derived from the Semantic Differential data by Principle Component Analysis and subsequent VARIMAX rotation. The interpretation of the underlying axes by means of their relation to the attributes of the Semantic Differential and their point coordinates leads to the following dimension labels (degradation types associated with the respective dimension are given in brackets):

- Discontinuity (packet loss, silence insertion, time-varying effect of signal-correlated noise, time-varying codec non-linearities, musical noise),
- noisiness (signal-correlated noise, additive circuit and background noise), and
- coloration (linear distortions due to bandpass filtering, electro-acoustic properties of terminal equipment, and room acoustics).

2.2 Sub-dimensions

For the development of an instrumental measure based on perceptual dimensions, more knowledge on each of the global dimensions is desirable. To this aim, a larger set of data is needed to derive adequate estimation parameters and train the measurement algorithm. Thus, three sets of stimuli were generated, each of which consists of approximately 70 degradation types associated with one of the dimensions identified in Section 2.1. The dataset associated with the "discontinuity" dimensions contained, e.g., different rates of packet loss (random and bursty) with different types of codecs and concealment techniques, front-end clipping, and different types of time-varying musical noise. A variety of spectral terminal characteristics at send side (handset, mobile phones, hands-free terminals including the effect of a realistic office room and a car interior) was included in the dataset associated with the coloration dimension. The database for the noisiness dimension contained different types of noise: signal-correlated noise, circuit and realistic background noise ("Hoth", babble, jackhammer).

For each of these data sets, i.e., in each of the quality sub-spaces "discontinuity", "noisiness", and "coloration", it is again helpful to have knowledge of their perceivable auditory *features*. Consequently, the same principle as for the global dimensions is pursued here as well. However, neither attribute-scaling nor complete similarity scaling is feasible for such a large number of stimuli (for the latter, the number of stimuli-pairs to be judged increases with $I \cdot (I-1)/2$). Thus, a more efficient scaling method, the so-called *sorting task* was deployed here as a good approximation of complete similarity scaling [4]. As an indirect measure of similarity, the frequency of occurrence of stimuli pairs in a common group is counted over all participants. The resulting similarity matrix is then fed to ordinary MDS algorithms. The extracted dimensions can thus be regarded as sub-dimensions for each global dimension "discontinuity", "noisiness", and "coloration".

The sorting experiments were repeated for four speakers. For each of the three spaces, separate experiments with different participant groups were run ("discontinuity": 20 participants, 7f, 13m, ∅26.8 years; "noisiness": 20 participants, 5f, 15m, ∅26.7 years; "coloration": 20 participants, 6f, 14m, ∅26.8 years).

A detailed description of the test procedure, the data analysis and the derivation of the dimension labels can be found in [5] for "coloration". A preliminary and qualitative description of the respective sub-dimension of all three domains is[1]:

- "Discontinuity"
  - "Interruptedness": Perceived interruptions of transmitted speech (dependent on packet/frame loss rate and the concealment technique)
  - "Additive Artifacts": Perceived "additive" (overloud) artifacts stemming from the packet loss concealment technique and depending on the packet/frame loss rate
  - "Musical Noise": Distinction of time-varying "musical noise" due to imperfect noise reduction techniques
- "Noisiness"
  - "Speech Contamination": Perception of noise-like distortions correlated with speech or lying within the (bandlimited) transmitted speech spectrum
  - "Additive Noise Level": Perceived level of additive noise
  - "Noise Coloration": Spectral shape and spectral content of noise
- "Coloration"
  - "Directness": "Nearness"-perception due to, e.g., bandwidth restrictions or room acoustics
  - "Brightness": Similar to "sharpness", determined by the center of gravity of the magnitude spectrum of the transmission channel

[1] Since narrowband conditions were considered in these experiments, it has to be noted that it cannot be inferred that these sub-dimensions are valid also for the wideband case.

The point coordinates are stored in $I \times L$ configuration matrices $\mathbf{S}_{dis}$, $\mathbf{S}_{noi}$, and $\mathbf{S}_{col}$, respectively, with $L \in \{2,3\}$.

3 Modeling framework

The principle of the dimension-based speech quality modeling approach described so far is reflected by the block scheme in Fig. 2. The upper two layers represent the "discontinuity-noisiness-coloration" decomposition of integral listening speech quality. In turn, each of these global dimensions can further be decomposed into a set of sub-dimensions as discussed in Section 2.2.

The so-called transmission rating $R \in [0; R_{0,max}]$ is used in order to quantify the integral quality, where $R_{0,max}$ is the highest achievable quality ($R_{0,max} = 100$ in the narrowband-only context [6]). Each of the global dimensions are linked to a deviation from the maximum, or ideal quality. Thus, each of the dimensions can be transformed to the quality scale in terms of a respective dimension impairment factor $I_{dim} \in [0; R_{0,max}]$, with $dim \in \{dis, noi, col\}$. A value $I_{dim} = 0$ reflects the case of no degradation in the dimension dim, whereas the degradation is maximal if $I_{dim} = R_{0,max}$. The notions of transmission rating and impairment factors are adopted from the so-called E-Model, a computational model for predicting overall quality during telephone-network planning [6]. They can be derived from *MOS*-values usually obtained in auditory tests and have the advantage of a defined absolute zero. Instead of R, *MOS* can equally well be used for the dimension-based model, as it was done in [7].

The sub-dimensions are quantified by the magnitudes $S_{dim,k} \in \mathbb{R}$, with $k \in \{1,2,3\}$ for $dim \in \{dis, noi\}$ and $k \in \{1,2\}$ if $dim = col$. $S_{dim,k}$ corresponds to a point coordinate of a condition on the kth sub-dimension in the subspace of dim.

Following the structure shown in Fig. 2, the integral speech quality can be modeled by determining functions f_A or $f_{B,dim}$ that appropriately combine the global dimension or sub-dimension scores to the integral quality score. For describing integral speech quality by means of the three global dimensions reflected by I_{dim}, f_A needs to be known, yielding a model for R with

$$\widehat{R} = f_A(I_{dis}, I_{noi}, I_{col}). \tag{1}$$

The function f_A is determined by regressing known, i.e., measured subjective test-data for I_{dim} onto R in a least squares sense.

Similarly, sub-dimension scores can be mapped onto global dimensions by functions $f_{B,dim}$, according to

$$\widehat{I}_{dim} = f_{B,dim}(S_{dim,1}, S_{dim,2}, S_{dim,3}), \tag{2}$$

where $S_{dim,3} \equiv 0$ for $dim = col$. Since the integral quality R is estimated by the nested relation $\widehat{R} = f_A(f_{B,dis}, f_{B,noi}, f_{B,col})$ in this case, the regression coefficients may best be optimized by a repeated regression analysis in order to avoid error propagation.

Apart from visualizing the modeling procedures of auditory data, Fig. 2 also represents a framework for different approaches of instrumental measures. The measurement approaches predict quality on the basis of a set of parameters that are extracted, e.g., from the transmitted signal (with or without comparison of the transmitted signal with the undistorted reference). Conventional measures like PESQ [1] or the E-Model [6] follow the type I estimation approach, i.e., no perceptual dimensions are explicitly taken into account. With the type II approach, global dimensions are predicted, each on the basis of a particular combination of a number of N_{dim} parameters, here referred to as impairment parameters $p_{dim,i}$ of a certain unit, with $dim \in \{dis, noi, col\}$ and $i = 1, ..., N_{dim}$. In this way, estimates of I_{dim} are found by respective impairment models g_{dim}:

$$\widehat{I}_{dim} = g_{dim}(p_{dim,1}, p_{dim,2}, ..., p_{dim,N_{dim}}). \tag{3}$$

Quality estimates are then obtained using the aforementioned function f_A. Error propagation might be avoided by re-running a regression analysis and optimizing the coefficients (cf. [7]), rendering $\widehat{R} = f_A(g_{dis}, g_{noi}, g_{col})$.

Finally, dimension-based measures might estimate sub-dimension scores (type III, e.g., [9]). A single sub-dimension $S_{dim,k}$ is estimated following the relation

$$\widehat{S}_{dim,k} = g_{dim,k}(p_{dim,k,1}, p_{dim,k,2}, ..., p_{dim,k,N_{dim,k}}), \tag{4}$$

with $k \in \{1,2,3\}$ if $dim \in \{dis, noi\}$ and $k \in \{1,2\}$ if $dim = col$. $N_{dim,k}$ denotes the number of sub-dimension parameters $p_{dim,k,i}$ for each sub-dimension k with $i = 1, ..., N_{dim,k}$. Note that the parameters of the type III approach might be different from those following the type II approach.

The two-step estimation of integral quality follows the model $\widehat{R} = f_A[f_{B,dis}(g_{dis,1}, g_{dis,2}, g_{dis,3}), f_{B,noi}(g_{noi,1}, g_{noi,2}, g_{noi,3}), f_{B,col}(g_{col,1}, g_{col,2})]$. The regression coefficients can be optimized here as well, analogous to the procedure described above.

4 Model realizations

4.1 Integral model

To derive a model for integral quality, a total of four databases containing different conditions and corresponding integral quality scores were available. Three of these databases were obtained in the context of the sub-dimension tests: For a subset of the conditions described in Section 2.2, quality ratings were collected prior to the Multidimensional Scaling experiment. Thus, for each of the dimensions "discontinuity", "noisiness", and "coloration", dimension impairment factors I_{dis}, I_{noi}, and I_{col} were available. A fourth test was run, with 50 conditions rated by 20 participants (11 f, 9 m, ∅32.2 years), and covering the *whole* range of relevant dimensions. In addition, 19 combinations of the dimensions were produced in order to test the influence of degradation-interactions on overall quality. The conditions from this fourth test constitute the target variable for a multiple non-linear regression.

The following model, as a realization of Eq. (1), can be formulated[2] (for details, see [8]):

$$\begin{aligned}\widehat{R} = f_A(I_{dis}, I_{noi}, I_{col}) &= 100 - I_{dis} - I_{noi} - I_{col} \\ &+ (9.319 \cdot I_{dis} \cdot I_{noi} + 6.047 \cdot I_{dis} \cdot I_{col} \\ &+ 8.274 \cdot I_{noi} \cdot I_{col}) \cdot 10^{-3}\,.\end{aligned} \tag{5}$$

The model covers a variance of $R^2_{adj} = 90\%$ of the data. A diagnostic quality information as it is pursued in the present study can easily be obtained and intuitively be interpreted based on Eq. (5), since it assumes a simple addition of the dimensions and their one-way interactions; if no degradation is existent in one dimension, a value

[2] A modeling example based on *MOS* prediction (including type II estimation) can be found in [7].

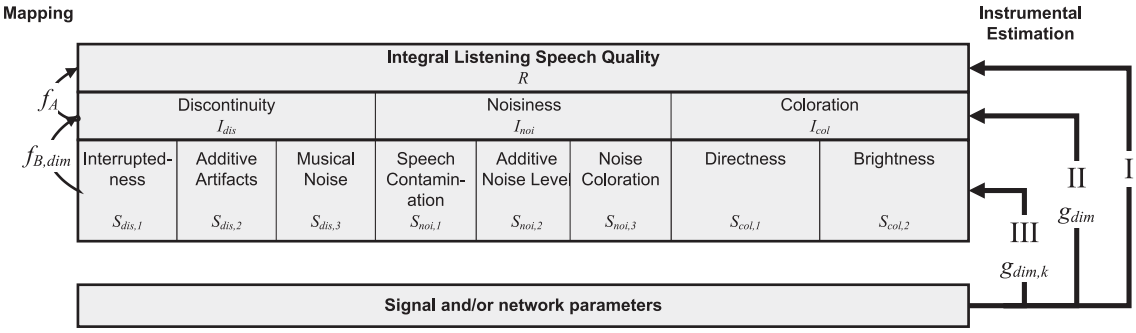

Fig. 2: Dimension-based modelling framework.

of zero is defined for the corresponding impairment factor (cf. Sec. 3). The interaction terms, subtracted from the one-dimensional impairments, reflect the discrepancy which arises from simply adding the single impairments. Thus, it is a measure for the mutual "masking" of the respective impairments.

4.2 Example: Dimension model for I_{col}

In [5], a realization of Eq. (2) is given that predicts the "coloration" impairment factor, i.e. $\widehat{I}_{col} = f_{B,col}(S_{col,1}, S_{col,2})$ with a covered variance of 77.1%. With this result and the mapping function f_A defined in the previous Section, a type III quality prediction approach can be realized.

Therefore, two sub-dimension models according to Eq. (4), $g_{col,1}$ and $g_{col,2}$, could be found that approximate the point coordinates stored in the configuration matrix $\mathbf{S}_{col}$ of the "coloration" sub-space. The sub-dimension scores are estimated by means of the parameters *ERB* (equivalent rectangular bandwidth in Bark) and f_c (center frequency of the transmission channel's magnitude of the transfer function in Hz), i.e., $\widehat{S}_{col,1} = g_{col,1}(ERB)$ and $\widehat{S}_{col,2} = g_{col,2}(f_c)$. Further details can be found in [5]. It can also be shown that the parameters *ERB* and f_c can directly be employed in order to predict I_{col}, namely in terms of the bandwidth impairment factor I_{bw} derived in [10] as a realization of Eq. (3) (it holds $I_{col} \equiv I_{bw}$). In this case, the covered variance amounts to 93.0%. Thus, a type II estimation approach can be employed for I_{col}, too.

5 Conclusions and further work

In this contribution, we introduced a general framework for speech quality estimation methods that are based on perceptual dimensions. The prediction can be done either by means of combining scores of the global dimensions "discontinuity", "noisiness", and "coloration", reflected by so-called dimension impairment factors, or on the basis of sub-dimensions. As an example, a concrete realization of an integral speech quality model based on dimension impairment factors is provided. Furthermore, an example is given for estimation approaches for the "coloration" impairment factor I_{col}; it can be predicted both by means of estimating sub-dimension scores (type III approach) and of directly estimating global dimensions scores (type II approach). In [7] and [8], a complete set of dimension estimators can be found. The development of further signal-based and parametric measures - pertaining to the proposed framework - have already been published, e.g., in [9], or are currently under way.

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*. International Telecommunication Union, CH–Geneva, 2005.

[2] M. Wältermann, K. Scholz, A. Raake, U. Heute, and S. Möller. Underlying quality dimensions of modern telephone connections. In *Proc. 9th Int. Conf. on Spoken Language Processing (ICSLP 2006)*, USA–Pittsburgh PA, 2006, 2170–2173.

[3] M. Wältermann, A. Raake, and S. Möller. Quality dimensions of narrowband and wideband speech transmission. *Submitted to acta acustica*.

[4] L. Tsogo, M. H. Masson, and A. Bardot. Multidimensional scaling methods for many-object sets: A review. *Multivariate Behavioral Research*, 2000, 35(3):307–319.

[5] M. Wältermann, A. Raake, and S. Möller. The sound character space of spectrally distorted telephone speech and its impact on quality. In *Proc. 124th AES Convention*, NL-Amsterdam, 2008, Paper No. 7464.

[6] ITU–T Rec. G.107. *The E-Model, a Computational Model for Use in Transmission Planning*. International Telecommunication Union, CH–Geneva, 2005.

[7] K. Scholz and U. Heute. Dimension-based speech quality assessment: Instrumental measure for the overall quality of telephone-band speech. In *Proc. 8. ITG-Fachtagung Sprachkommunikation*, D–Aachen, 2008.

[8] M. Wältermann, K. Scholz, S. Möller, L. Huo, A. Raake, and U. Heute. An instrumental measure for end-to-end speech transmission quality based on perceptual dimensions: Framework and realization. In *Proc. 11th Int. Conf. on Spoken Language Processing (ICSLP 2008)*, AU–Brisbane, 2008.

[9] L. Huo, M. Wältermann, U. Heute, and S. Möller. Estimation model for the speech-quality dimension "continuity". In *Proc. 8. ITG-Fachtagung Sprachkommunikation*, D–Aachen, 2008.

[10] Alexander Raake. *Speech Quality of VoIP – Assessment and Prediction*. John Wiley & Sons, UK-Chichester, West Sussex, 2006.