Analytical Assessment and Distance Modeling of Speech Transmission Quality

Marcel Wältermann, Alexander Raake, Sebastian Möller
Deutsche Telekom Laboratories, TU Berlin, Germany
marcel.waeltermann@telekom.de

Abstract
The quality of transmitted speech is based on the auditory characteristics the degraded signal provokes. In past studies, it has been shown that the main features of speech transmission can be subsumed under the orthogonal perceptual dimensions “discontinuity”, “noisiness”, and “coloration”. In order to gain more insight into the dimensional composition for arbitrary transmission conditions, an auditory method is described in this paper which allows for assessing these dimensions efficiently.

The results can be used to model the total impairment, a measure of the reduction of integral quality which is compliant with the E-model, a parametric tool for speech quality prediction. The model derived in this paper is based on a distance function and yields a correlation of $r = 0.97$ between subjective scores and model predictions for the Euclidean case.

Index Terms: Speech quality, modeling, feature decomposition

1. Introduction
The perceptual quality of speech transmission, influenced by terminals and networks, is essential for the Quality of Experience (QoE) attributed to it by the users. Any such technology might thus be subject to assessment in auditory tests, where participants rate the quality on a defined scale. Following international standards as defined in ITU-T Rec. P.800, this can be done by employing Absolute Category Rating (ACR), resulting in a one-dimensional Mean Opinion Score (MOS) ranging from 5 (“excellent” quality) to 1 (“bad” quality). A listener taking part in such a subjective test can be represented by the scheme depicted in Fig. 1 (adapted from [1][2], simplified version).

Let the (degraded) output of a terminal loudspeaker at receive side be the sound event $s$ which is multidimensional in nature and can be described by physical parameters such as its audio bandwidth and the signal-to-noise ratio (SNR). The integral quality description $b_0$ is formulated by the listener subsequent to perception and comparison to an internal reference (cf. [2][3]). Apart from overall quality, analytical information from a users’ point of view can provide further insight into the composition of the perceived quality and allow its perceptual diagnosis. This information can be obtained by asking the listener about features $\beta$ of the sound event $s$ by means of appropriately designed rating scales. For instance, a bad quality $b_0$ might be perceived as “interrupted” or “noisy” ($\beta$), caused by packet loss, or noise superimposed to the speech signal ($s$).

Several tools have been developed for the instrumental estimation of quality, i.e. $b_0$. Two prominent representatives recommended by the ITU-T are PESQ (ITU-T Rec. P.862), a signal-based model, and the E-Model (ITU-T Rec. G.107), a parameter-based model. On the other hand, there are only few and not yet standardized models for diagnostic quality prediction as proposed in [4] that are capable predicting quality features, i.e. $\hat{\beta}$. Examples include [5] and [6].

Jekosch [3] defines a quality feature as a “recognized and designated characteristic of an entity that is relevant to the entity’s quality”. Following this notion, it can be hypothesized that there exists a functional relationship $f$ between perceptual magnitudes described by the quality features $\beta$ and the overall quality $b_0$. Consequently, a new class of quality models can be conceived that a) provides diagnostic information $\hat{\beta}$, and b) is based on these features, such that $b_0 = f(\hat{\beta})$.

In this paper, a new auditory method is presented which enables the direct assessment of features $\beta$ of transmitted speech which correspond to the orthogonal dimensions of the perceptual feature space of the listener. Owing to the efficiency of the method, large numbers of conditions can be assessed. Thus, it represents a basis for the effective engineering of diagnostic quality estimators. Previous research is summarized in Sec. 2 that was done in order to reveal these features for narrowband (NB) and wideband (WB) speech transmission channels. The method itself is explained in Sec. 3, as well as its application in auditory experiments. The raw data of the experiments is investigated in Sec. 4. Based on this data, a model function $f$ mapping the dimension scores onto an E-model-compliant integral quality measure is presented in Sec. 5, where also the meaningfulness of the feature scores is discussed. Conclusions are given in Sec. 6.

2. Perceptual Dimensions
The assessment of features of transmitted speech has been the subject in a number of studies (e.g., [7][8][9][10]). Different subjective test methodologies can be applied for this purpose. One approach commonly employed is the scaling of meaningful attributes describing the perceived characteristic of speech signals, often in the fashion of a Semantic Differential (SD) [7][10]. The obtained ratings can be condensed by, e.g., principal component analysis, resulting in orthogonal components correlated with sets of the employed attribute-scales. Such components represent the latent or underlying dimensions of the perceptual feature space of the listener. A different approach for revealing such dimensions is Multidimensional Scaling (MDS) of proximity data, such as the similarity between pairs of conditions obtained from subjective tests [8][10]. Due to the relatively large number of attributes (SD) and pairwise comparisons (MDS), both methods, although necessary to explore the feature space of the listener, are relatively time-consuming.

In [10] and [11], MDS and SD experiments were con-
directed in an extended technology context compared to [8] or [9] by considering VoIP-relevant effects like packet loss and WB speech. In [12], these tests were compared with each other and with past experiments: Three dimensions, common for both NB and WB speech, were found to cover the major part of the experimental variances. The interpretation of the configurations led to the following relevant dimensions (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 and room reverberation).

These dimensions can be conceptually related to those found by other authors. E.g., the “bubbling” dimension [9] describing the time-varying effect of codec non-linearities is covered by the dimension “discontinuity”.

The remainder of this paper builds on that own work. The three dimensions are hypothesized to completely represent all features of NB and WB speech transmission and are taken as the basic constructs for a direct and efficient scaling approach developed in the following.

3. Direct Assessment of the Dimensions

The auditory method presented here relies on the findings that the overall quality of transmitted speech can be decomposed into the three perceptual dimensions “discontinuity”, “noisiness”, and “coloration” (cf. preceding section). Since the dimensions are known, a feature decomposition of overall quality can be more efficiently achieved by directly rating these dimensions by means of three scales, each dedicated to one dimension. This way, time-consuming MDS and/or SD experiments can be avoided in this context.

As opposed to the commonly used technique of SD described above, where the ratings of a number of potentially correlated attribute-pairs are subject to a principal component analysis in order to reveal the underlying dimensions, the orthogonal dimensions themselves are subjectively rated here. With this method, feature ratings are obtained more efficiently with reduced experimental effort. This aspect is also a major difference to, e.g., the Diagnostic Acceptability Measure [7].

The scale design is depicted in Fig. 2. Each of the three dimensions is rated with a separate scale, where the letters A and B are replaced by the antonym attributes “continuous – discontinuous”, “not noisy – noisy”, and “uncolored – colored”, respectively.

Two auditory experiments were carried out. The first experiment will be described in detail in this paper and constitutes the basis for the modeling approach in Sec. 5. A total of 66 processing chains were considered in the first experiment: 8 NB (300-3400 Hz) and 7 WB (50-7000 Hz) codecs, one “clean” WB condition, 24 codec tandems, and 26 conditions including noise, uniform packet loss, and bandpass filters. The second experiment contained similar conditions with a slightly shifted focus. It will be exploited here to provide evidence on the test-retest reliability of the method, since a subset of the Experiment 1 conditions is included in Experiment 2.

Analogous to quality assessment according to ITU-T Rec. P.800, it is ensured in a prior training phase with written instructions that the scales are appropriately used by the listeners. The participants were instructed that the features or characteristics of speech samples are supposed to be judged (i.e., not the quality). Furthermore, each scale label was explained by additional synonyms in order to make sure that the listeners understand the meaning of the scales.

In addition to the written instructions, exemplary samples for each of the three scales were presented which are distorted in only the respective dimension (e.g., samples containing only packet loss, only circuit noise, and only linear distortions, respectively). The understanding of the scales was supported by presenting an undistorted sample, stating that this particular sample is “not noisy”, “continuous”, and “uncolored”.

The speech source material contained different sentences and one female and one male speaker. Two independent groups of listeners were recruited which mostly consisted of students from the local university: 20 listeners (10 f, 10 m) for Experiment 1, 24 listeners (12 f, 12 m) for Experiment 2. They were aged between 20 and 33 (average age: 27.3) in Experiment 1, and between 20 and 46 (average age: 28.7) in Experiment 2. None of them reported any known loss of hearing and they were paid for their participation. The listening-only experiments were carried out in a sound-proof booth fulfilling the listening environment requirements given in ITU-T Rec. P.800. The scales were presented separately in the test, i.e. consecutively for each stimulus. For each participant, the order of the scales was randomized. The samples were randomized per test session, sentences were assigned randomly to the conditions per session. In prior to each dimension scaling test, a separate test was conducted to collect integral quality ratings (MOS) for the given sets of conditions.

4. Raw Data Analysis

Since no experience with the described method is available so far, some key characteristics of the three dimension scales are investigated in this section. The analysis of the raw dimension scale scores $S_{dim} \in [0; 1]$, with $dim \in \{\text{dis, noi, col}\}$ in the remainder of this paper, reveals that the scales were used in an orthogonal way by the participants, indicated by correlation coefficients of $r < 0.25$ between the ratings on every two scales.

The means of the standard deviations $\sigma_{std_{dim}}$ were calculated on a per-file basis. They amount to $\sigma_{std_{dis}} = 0.195$, $\sigma_{std_{noi}} = 0.177$, and $\sigma_{std_{col}} = 0.202$. These values lie well within the range of standard deviations obtained on ACR scales of standard quality tests (see, e.g., [13], p.151, Table 6.2).

Single univariate mixed-model ANOVAs were separately applied to the data obtained from the “discontinuity” scale, the “noisiness” scale, and the “coloration” scale, respectively. The factor subject was included as a random variable, whereas the remaining experimental factors speaker, codec, packet loss rate, noise, and filter were included as fixed variables. The main effects and 2-way interactions (where possible) were tested at the 1%-level ($p \leq 0.01$). Although we refrain from complete ANOVA tables here due to space constraints, in summary it can be stated that:

- “discontinuity” is influenced by codec ($F = 16.1, df = 39$) and packet loss rate ($F = 112.0, df = 4$),
- “noisiness” is influenced by codec ($F = 30.1, df = 39$), noise ($F = 159.8, df = 4$), and their interaction ($F = 44.8, df = 2$), and

Figure 2: Scale design.
“coloration” is influenced by codec \((F = 32.8, df = 39)\) and filter \((F = 36.0, df = 14)\).

Occasionally, the factors subject and speaker turned out to be significant with relatively low \(F\)-values which thus causes “weaker” effects than the fixed experimental factors actually of interest. Apart from that, certain subject- and speaker-dependencies seem to be usual also for MOS data and are thus not considered here in detail.

Apparently, the perceptual scales reflect the physical effects intuitively associated with them (cf. Sec. 2): “Discontinuity” perception depends on the packet loss rate, the “noisiness” scale captures the impact of noise (noise is perceived differently depending on the codec), and linear filtering affects “coloration” perception. Codes are inherently of multidimensional nature and thus have an influence on all three scales (see Sec. 5 for further details).

The common conditions included in Experiments 1 and 2 allow to provide some insight into the test-retest reliability of the subjective method presented here. These 12 reference conditions include the clean WB PCM, different codecs at different bitrates, bandpass filters, noise of different levels and different rates of packet loss. They were selected in such a manner that they are as evenly spread over each of the three dimension scales as possible. The correlation coefficients \(r\) and root mean square errors \((RMSE)\) between the raw scale ratings of Experiments 1 and 2 amount to \(r_{\text{dis1.dis2}} = 0.96\), \((RMSE)_{\text{dis1.dis2}} = 0.10\), \(r_{\text{no1.no2}} = 0.98\), \((RMSE)_{\text{no1.no2}} = 0.08\), and \(r_{\text{col1.col2}} = 0.98\), \((RMSE)_{\text{col1.col2}} = 0.09\). The close-to-linear agreement between the absolute scale values of the two tests provide evidence that the method is reliable to a high degree.

5. Quality Modeling

In this section, it is assumed that the perceptual dimensions described in Sec. 2 are appropriate to describe the overall quality in a complete way, i.e., the perceptual dimensions are in fact quality dimensions. In order to show that, “dimension impairment factors” \((\text{DIFs})\) \(I_{\text{dim}}, I_{\text{noi}}, \text{ and } I_{\text{col}}\) are introduced, each quantifying the decrease in overall quality due to the dimensions discontinuity, noisiness, and coloration, respectively. According to the considerations in Sec. 1, a model function \(f\) is sought that maps the DIFs (corresponding to \(\beta\)) onto quality scores \(b_0\).

In the remainder of this paper, this integral quality measure \(b_0\) is represented by the total impairment \(I_{\text{tot}}\), combining the dimension impairment factors according to a model \(f\). \(I_{\text{tot}}\) is defined to be compliant with the sum of the impairment factors of the E-model (ITU-T Rec. G.113) as close as possible. In particular, the clean condition (WB PCM) is set to \(I_{\text{tot}}(\text{clean}) = 0\).

The coefficients \(a_{\text{dim}}, G_{\text{dim}}\), and \(p\) are determined through multiple non-linear regressions in a least-squares sense.

Table 1 summarizes the coefficients \(a_{\text{dim}}, G_{\text{dim}}\), and \(p\), together with goodness-of-fit parameters \(r\) and \((RMSE)\) between subjective and estimated scores. The first column shows the results for \(p\) being a free parameter. The relatively high correlation coefficient and the low \((RMSE)\) provide evidence of the appropriateness of the model Eq. (2). The value \(p = 1.61\) lies in-between the City-Block-Metric \((p = 1)\) and the Euclidean Metric \((p = 2)\), but closer to the latter. In order to obtain a more intuitive model, the parameter \(p\) is thus set to \(p = 2\), while Eq. (2) is subject to refitting. As can be seen from the resulting goodness-of-fit parameters \(r\) and \((RMSE)\) in Table 1 (second column), a comparatively accurate model is obtained. Setting \(p \geq 1\) leads to a slightly worse fit (third column in Table 1). Thus, the relation between the DIFs and the total impairment seems to best follow an Euclidean distance model, suggesting that the perceptual space of the listener is Euclidean as well. This thought has already been arisen in \([14]\) and differs from linear approaches as discussed in, e.g., \([8]\) and \([9]\).

Also in the E-Model, a simple summation of impairment factors is assumed. The impairment factors defined in the E-model, however, are not based on perceptual dimensions but rather on physically distinguishable features (e.g., signal-correlated distortions, delayed distortions). The relatively high magnitudes of the weights \(a_{\text{noi}}\) suggest that “discontinuity” is of highest importance for overall quality.

\[ I_{\text{dim}} = a_{\text{dim}} \cdot \left( \frac{\overline{\text{dim}}}{\overline{G_{\text{dim}}}} \right)^{\frac{1}{2}} \]

We also experimented with other types of models, e.g. models extending Eq. (2) by (weighted) interaction terms. However, no significantly better fits could be achieved, even at the expense of additional free parameters.

| Table 1: Model coefficients and goodness-of-fit parameters. |
|---|---|---|
|   | p = 1.01 | p = 2 | p = 1 |
| r | 0.97 | 0.96 |
| RMSE | 4.5 | 4.6 | 5.3 |
| $a_{\text{dis}}$ | 52.3 | 54.3 | 43.6 |
| $a_{\text{nai}}$ | 38.5 | 43.3 | 22.4 |
| $a_{\text{col}}$ | 16.9 | 18.8 | 14.6 |
| $G_{\text{dis}}$ | 2.08 | 2.13 | 2.00 |
| $G_{\text{nai}}$ | 1.39 | 1.67 | 0.76 |
| $G_{\text{col}}$ | 1.12 | 1.19 | 1.20 |

The model variants are intuitively meaningful, and the fits are quite comparable. Thus, only future experiments in this direction will show which model is the most general. For the time being, $p = 2$ is considered in the following.

In Fig. 3, DIFs are depicted for a variety of conditions as grey-scale-coded bars. As intended, distortions assumed to be of perceptually unidimensional nature (e.g., packet loss, noise, linear distortions) provoke high values $I_{\text{dim}}$ for a single dimension only (cf. the ANOVA in Sec. 4) and reflect the degree of the distortion (e.g., the packet loss rate) in a plausible way.

The $I_{\text{dim}}$ values for two-dimensional distortions roughly correspond to those of the respective two single-dimensional conditions. Although slightly varying, $I_{\text{dis}}$ lies around the value of 33 for NB conditions, reflecting the almost constant “coloration” perception of NB conditions; $I_{\text{nai}}$ has a roughly constant value for both the NB and WB background noise condition etc. In general, both $I_{\text{dis}}$ and $I_{\text{nai}}$ monotonically increase with decreasing codec bitrate per codec scheme, reflecting both the increasing “noisiness” and “discontinuity” (“bubbling” in [9]).

6. Conclusions

A new efficient and reliable test method was presented that allows speech quality to be decomposed into its orthogonal features by directly scaling the relevant speech quality dimensions “discontinuity”, “noisiness”, and “coloration” for sets containing larger numbers of conditions. As it has been shown, the judgments obtained from an auditory test are orthogonal in nature and provide meaningful diagnostic information.

The dimension components were quantified as dimension impairment factors (DIFs) on a psychological continuum, the $R$-scale of the E-Model, and can intuitively be mapped onto total impairment values by applying a distance function. It turned out that the Euclidean distance between zero impairment and “dimensional impairment” best fits the available experimental data, suggesting the multidimensional psychological space being of Euclidean nature.

In future work, the reliability of the scaling method needs to be confirmed by other laboratories. Apart from that, further investigations are needed to relate physical parameters to the three DIFs, and eventually develop and improve parametric models as the E-Model and diagnostic signal-based speech quality models (first developments on the basis of limited auditory data were demonstrated in, e.g., [6]).

7. References