Feature Decomposition and Modeling of Speech Quality for Various Wideband Conditions

Marcel Wältermann, Alexander Raake, Sebastian Möller
Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Email: marcel.waeltermann@telekom.de

Introduction

The integral quality of transmitted speech can be seen as a combination of features which are recognized by the listener in the auditory domain. It has been shown recently that traditional narrowband as well as wideband speech quality can sufficiently well be quantified by three orthogonal dimensions, as revealed by the techniques of Semantic Differential and Multidimensional Analysis [1]: "Discontinuity", "noisiness", and "coloration". These dimensions can not only be used to model the integral quality, they also provide perceptually adequate diagnostic quality information, which is the focus of this paper.

In wideband speech transmission (50-7000 Hz), the perceptual effects of degradations due to transmission impairments are even widerly spread across the perceptual space than in the narrowband case. In order to investigate common wideband transmission scenarios with respect to the mentioned three perceptual dimensions, a series of real-world conditions are quantified in the study presented here. This quantification was done by directly scaling the dimensions "discontinuity", "noisiness", and "coloration" following a new test method. Furthermore, the integral quality was assessed. The obtained results allow to characterize the test conditions in terms of the aforementioned dimensions.

Method for Direct Assessment of Quality Dimensions

A test protocol has been developed for assessing the relevant quality dimensions "discontinuity", "noisiness", and "coloration" directly by means of three pre-defined scales, each reflecting a single orthogonal dimension.

The scale design is depicted in Figure 1. 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.

Figure 1: Scale design.

Since the labels on the left ends of the scales describe zero impairment in the respective dimension, whereas the labels on the right ends describe maximum impairment, the scales can be considered as unipolar scales.

In order to ensure that the scales are understood and appropriately used by the listeners in a subjective experiment, a prior training phase is essential. The participants should be instructed that the *features* or *characteristics* of speech samples are supposed to be judged (i.e., not the quality), that the assessment is done by means of three scales, and that each scale is labeled with an attribute at each end that describes the characteristic to be judged upon.

Furthermore, each scale should separately be described by the specific dimension label, and each scale label should be explained by describing synonyms in order to make sure that the listeners understand the meaning of the scales. The synonyms chosen here correspond to those attributes which are very highly correlated with the principal components reflecting the perceptual dimensions (see, e.g., [2]). In detail, the participants should be instructed that

- with the scale labeled with "continuous - discontinuous", the "discontinuity" of the sample is supposed to be judged; the labels "continuous" and "discontinuous" can be paraphrased with the terms "regular" / "steady" / "not chopped" / "not bubbling" / "not ragged" and "irregular" / "shaky" / "chopped" / "bubbling" / "ragged", respectively
- with the scale labeled with "not noisy - noisy", the "noisiness" of the sample is supposed to be judged; the labels "not noisy" and "noisy" can be paraphrased with the terms "not hissing" and "hissing", respectively
- with the scale labeled with "uncolored - colored", the "coloration" of the sample is supposed to be judged; the label "uncolored" and "colored" can be paraphrased with the terms "direct" / "close" / "thick" / "not nasal" and "indirect" / "distant" / "thin" / "nasal", respectively

In addition to the written instructions, exemplary samples for each of the three scales should be 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 can be supported by presenting an undistorted sample, stating that this particular sample is completely "not noisy", "continuous", and "uncolored".

Auditory Experiments

An auditory experiment was carried out according to the protocol described in the preceding section. A total of 66 processing chains were considered and applied to speech source material (different sentences, one female and one male speaker): 8 narrowband (300-3400 Hz) and 8 wideband codecs, 24 codec tandems, 12 estimated "transfer functions" of some of the codecs and tandems,

and 14 conditions of which it is known which perceptual dimension(s) is/are mainly affected [1], including noise, packet loss, and filters.

A group of 20 listeners (10 f, 10 m) was recruited. They were aged between 20 and 33 (the average age was 27.3). None of them reported any known loss of hearing and they were paid for their participation.

It turned out that the duration of the training phase took not longer than 15-20 minutes until the listeners confirmed that they understood the meaning of the scales. The scales were presented separately in the test, i.e. consecutively for each stimulus. For each participant, the order of the scales was randomized.

In prior to the dimension scaling test, integral quality ratings were collected for the set of conditions, resulting in Mean Opinion Scores (MOS) ranging from 5 ("excellent") to 1 ("bad") [3].

Results

The analysis of the raw dimension scale scores S_{dim}, with $dim \in \{\text{dis, noi, col}\}$ and $S_{dim} \in [0;1]$, reveals that the scales were used in an *orthogonal* way by the participants, indicated by a correlation coefficient of $r < 0.25$ between every two scales. The average standard deviation of $\varnothing std < 0.2$ suggests a high inter-subject agreement on the usage of the scales.

In the following, a integral quality model is derived. Therefore, the raw scale data $\overline{S}_{dim}$ (mean over speakers and participants) is linearly transformed to "dimensional MOS values" MOS_{dim}:

$$MOS_{dim} = \frac{\overline{S}_{dim} - \overline{S}_{dim,min}}{\overline{S}_{dim,max} - \overline{S}_{dim,min}} \cdot (MOS_{max} - 1) + 1 \ , \tag{1}$$

where MOS_{max} corresponds to the maximum MOS value obtained from the integral quality test.

The integral quality judgments can be described by the dimension data, according to a linear relation:

$$\widehat{MOS} = MOS_{max} - \sum_{dim} \underbrace{a_{dim} \cdot (MOS_{max} - MOS_{dim})}_{=\Delta MOS_{dim}} \tag{2}$$

The coefficients a_{dim} were found through curve fitting: $a_{dis} = 0.51$, $a_{noi} = 0.45$, and $a_{col} = 0.52$. The model covers a variance of 90.6 %.

In Figure 2, the deviation from the maximum integral quality, corresponding to the second term of Eq. (2) and reflected by the heights of the bars, is depicted for a variety of conditions. The proportion of the influence of each of the dimensions, ΔMOS_{dim}, is color-coded.

As intended, more or less unidimensional distortions (e.g., packet loss, noise, linear distortions) provoke high values ΔMOS_{dim} for a single dimension only. Two-dimensional degradations (e.g., narrowband and noise) lead to high values ΔMOS_{dim} for both corresponding dimensions. The ΔMOS_{dim} values of two-dimensional distortions roughly correspond to those of the respective two single-dimensional conditions, e.g. ΔMOS_{col} is approximately constant for narrowband conditions; ΔMOS_{noi} is approximately constant for background noise etc. Both the discontinuity and noisiness components increase with decreasing codec bitrate. Furthermore, the coloration component varies between different wideband codecs and bitrates, whereas it is constant for narrowband codecs.

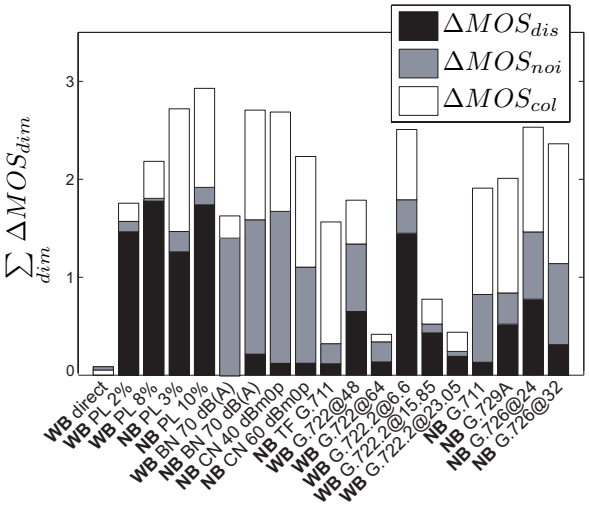

Figure 2: Deviation from max. integral quality for selected conditions (NB: narrowband, WB: wideband, PL: packet loss, BN: background noise, CN: circuit noise, TF: "transfer function"; for some codecs, the bitrate is given in kbit/s).

Altogether, it is evident that the scales were used in a meaningful way.

Conclusions

An efficient test method was presented that allows to decompose speech quality into its orthogonal features by directly scaling relevant speech quality dimensions. The judgments obtained from auditory tests provide meaningful diagnostic information, as it has been shown for various conditions in a wideband context. A model has been derived on the basis of MOS values. In future work, the reliability of the method needs to be confirmed by further experiments. Moreover, refined models will be developed.

References

[1] Wältermann, M., Scholz, K., Möller, S., Huo, L., Raake, A., Heute, U. (2008). "An Instrumental Measure for End-to-end Speech Transmission Quality Based on Perceptual Dimensions: Framework and Realization". In: Proc. 11th Int. Conf. Spoken Language Processing, AU-Brisbane, 61-64.

[2] Wältermann, M., Raake, A., Möller, S. (2006). "Perceptual Dimensions of Wideband-transmitted Speech." In: 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, DE-Berlin, 103-108.

[3] ITU-T Rec. P.800 (1996). "Methods for Subjective Determination of Transmission Quality". International Telecommunication Union, CH-Geneva.