Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech

Friedemann Köster, Dennis Guse, Marcel Wältermann, Sebastian Möller

Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Deutschland,

Email: friedemann.koester@tu-berlin.de, dennis.guse@tu-berlin.de, mar.wael@gmx.de, sebastian.moeller@tu-berlin.de

Abstract

In this contribution the properties of quality ratings as obtained on a continuous and extended scale are investigated and a functional relation to the ITU-T P.800 Absolute Category Rating (ACR) scale is derived. Therefore, two auditory experiments were conducted in which narrowband and wideband conditions were rated on both the extended continuous and the ACR scale. It turned out that the main benefit of the extended continuous scale is the increased sensitivity, especially for conditions of lower quality. This fact renders the continuous scale attractive for wideband or beyond-wideband experiments, where only a reduced area of the rating scale is available for low-quality conditions (e.g., narrowband conditions).

Introduction

For the purpose of subjective (speech) transmission quality assessment, the usage of a discrete 5-point scale is recommended by the ITU-T in Rec. P.800 [1]. In listening-only tests, the Absolute Category Rating scale (ACR, Figure 1) is usually employed, consisting of five discrete categories, each associated with a number (given in brackets): “Excellent” (5), “good” (4), “fair” (3), “poor” (2), and “bad” (1). For each test sample, the participant is asked to choose one of these categories to judge the perceived quality. The pre-annotated numbers are usually omitted in conversational tests (there, the scale is called Category Judgment Scale).

This scale shows a number of well-known drawbacks (see Möller [2] for details):

1. Judges differ in their use of the category scale.
2. Context effects occur showing the interdependence of categorical judgments.
3. The perceived intervals between the attributes of the ACR scale are not equidistant.
4. The sensitivity of the ACR scale is relatively low.
5. Naïve participants often avoid the use of the extreme categories of the scale.

Items 1 and 2 are not necessarily only a drawback of the mentioned scale, but of absolute scaling in general. The issues related to item 3 are under ongoing discussion (see Zieliński [3] for a recent consideration). These items will not be addressed further in the present contribution.
Theoretically, however, the drawbacks related to items 4 and 5 may partly be avoided by a straight-forward redesign of the scale: The sensitivity might be increased by replacing the five discrete categories by a continuous scale, allowing the participants to also place their ratings in the space between the given attributes.

The phenomenon that participants tend to avoid extreme categories might stem from their expectation of even worse or better quality of following samples. On the other hand, a saturation effect occurs if participants rate a stimulus ’excellent’ or ’bad’, although it is in fact of better or worse quality than a preceding stimulus which has already been put in one of these categories. Although a proper training of the test participants might already reduce these effects, a further reduction might be achieved by extending the scales beyond the existing end-points.

A more sensitive scale which allows well-balanced scoring also in the edge regions can especially be beneficial in wideband and beyond-wideband tests. In such experiments, low-quality conditions (e.g., narrowband) suffer from the highly reduced sensitivity of the ACR scale.

Such a continuous and extended scale has been developed by Bodden and Jekosch ([4]; see Figure 2) and is recommended for certain items of the questionnaire used for the evaluation of telephone services based on spoken dialogue systems (ITU–T Rec. P.851 [5]).

The scale is equipped with equidistant tick marks that can be thought of as graphical substitutions for the numbers usually employed with ACR scales. Furthermore, the ticks suggest equivalent semantic distances between the attributes and thus potentially reduce the problem mentioned in item 3 in the abovementioned list (however, this assumption cannot be checked with the experiments presented here).

Quality of the speech:

excellent	good	fair	poor	bad
5	4	3	2	1

Figure 1: Listening–quality scale according to ITU–T Rec. P.800 [1]

Figure 2: Continuous rating scale according to Bodden and Jekosch [4], German version. English translations: “extremely bad”, “bad”, “poor”, “fair”, “good”, “excellent”, “ideal”.

Figure 3: Continuous–quality scale according to ITU–T Rec. P.880 [6]

Note that the scale in Figure 2 is somewhat similar to the continuous scale recommended in ITU-T Rec. P.880 ([6]; see Figure 3) and elsewhere. The P.880 scale was compared to the ACR scale in ITU-T COM 12-120 [7] for transmitted speech in different bandwidth contexts. Although the P.880 scale is continuous as well, no dedicated "overflow areas" exist. Furthermore, the attributes are assigned to ranges rather than coinciding with the tick marks on the scale. Some of the results presented in this study will qualitatively be compared to the results obtained in ITU-T COM 12-120.

It is the purpose of this contribution to provide insight into the properties of the extended and continuous scale depicted in Figure 2, and its relation to the commonly employed ACR scale. For this reason, two listening-only experiments were conducted in which two fixed sets of conditions were rated on both the ACR scale and the extended continuous scale (Experimental Setup). The results are discussed in Section 3 (Results), and concluding remarks are provided in Section 4 (Conclusion).

Experimental Setup

The listening-only experiments were carried out in a sound-proofed booth fulfilling the listening environment requirements given in ITU–T Rec. P.800. A group of 20 listeners (10 f, 10 m) aged between 19 and 47 (the average age was 26.5) was recruited. None of them reported any known loss of hearing and they were paid for their participation.

Both scales, the ACR and the extended continuous scale, were applied both in a pure narrowband context (experiment 1) and in a wideband context (experiment 2). Thus, the experiments consisted of four sessions in total (2 bandwidth contexts times 2 scales). Within one context, a fixed set of conditions was asked to be judged. The scales were presented in random order, whereas the contexts were purposely not intermixed. The first two sessions always consisted of the narrowband context, whereas the last two sessions consisted of the wideband context. This procedure reflects the ecological shift of the internal quality expectation of the participants when migrating from traditional narrowband to modern wideband speech. Dedicated training samples were asked to be rated in prior to each session in order to foster the sense for the range of quality to be expected.

The source speech material consisted of recorded sentences of the EUROM text material (Gibbon [8]), spoken by one male and one female German speaker. The test samples were sampled at 8 kHz (narrowband)/16 kHz (wideband), normalized to -26 dB ASL rel. ovl., and filtered according to G.712 and IRSmod (narrowband)/P.341 (wideband) at both sending and receiving side.

The narrowband test consisted of 18 conditions including different loudness levels, noise types (babble and hoth), bandpasses, codecs, codec-tandems, MNRUs, and Packet-losses. The wideband tests consisted of 25 conditions including clean speech and different loudness levels, noise types (babble and hoth), bandpasses, wideband-codecs, codec-tandems, wideband MNRUs, and Packet-losses. The samples were randomized per test session. Each condition was rated both for the male and the female speaker, resulting in 36 samples in the narrowband test and 50 samples in the wideband test. The listening level was adjusted to 73 dB SPL (diotic presentation).

The scales in Figures 1 and 2 were incorporated in a software program that led the participants through the experiments. The ACR scale was realized with software buttons pre-annotated with the numbers 5 to 1 and the attributes according to ITU–T Rec. P.800. The extended continuous scale was depicted as a bitmap, together with a software slider. The labels were internally assigned to numbers of the interval [0,6] in such a manner that the attributes corresponding to ITU–T Rec. P.800 were exactly assigned to the numbers 1, 2, 3, 4, and 5.

Results

Characteristics of the scales

Figures 4 and 5 depict the histograms of ratings of both scales as they were used in the narrowband and wideband experiment.

Moreover, the following descriptive values were calculated: The overall mean, minimum and maximum values of the per-sample mean values, and the mean of the standard deviations, also calculated per sample (see Table 1). Since the standard deviation depends on the numerical range of each scale, the scale values of the ACR scale (S_{ACR}) and the extended continuous scale (S_{EC}) were normalized according to the following equations in prior to the calculation, resulting in $S'_{ACR} \in [0,1]$ and $S'_{EC} \in [0,1]$:

$$S'_{ACR} = \frac{S_{ACR} - 1}{4} \tag{1}$$

$$S'_{EC} = \frac{S_{EC}}{6} \tag{2}$$

As Figures 4, 5, and Table 1 reveal, no principal difference between the narrowband and wideband experiments can be observed. There is only a slight difference between the overall mean values of both scales and a consistently lower standard deviation for both scales in the wideband case.

The complete range of both scales was used for rating the test conditions, except for the upper quarter of the continuous scale in the narrowband case (see Table 1). Obviously, the lower part of the scale was sufficient for

	overall mean		min		max		std	
	ACR	EC	ACR	EC	ACR	EC	ACR	EC
narrowband	0.4274	0.4443	0.0000	0.0505	0.8500	0.7642	0.2193	0.1629
wideband	0.4400	0.4457	0.0000	0.0424	0.9000	0.8575	0.1848	0.1534

Table 1: Descriptive values for the ACR and extended continuous scale (narrowband and wideband experiment)

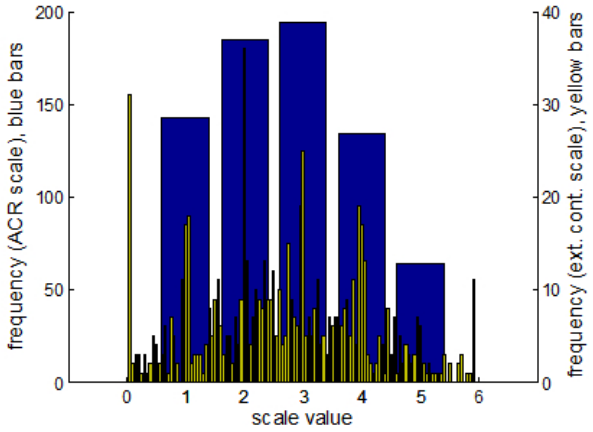

Figure 4: Histogram of the narrowband scale data.

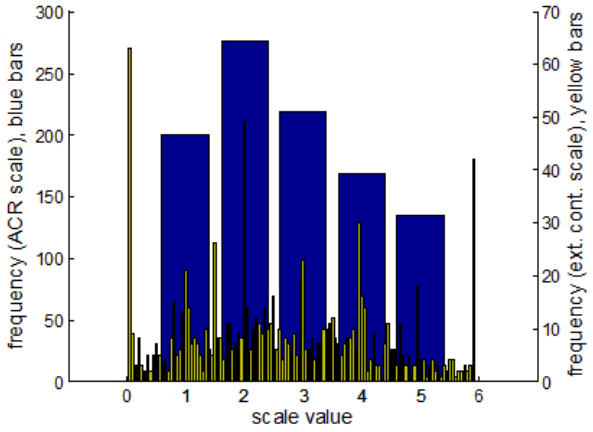

Figure 5: Histogram of the wideband scale data.

rating the narrowband stimuli, being indeed far below 'ideal' quality.

The histograms reveal the active use of the extreme categories on the extended scale. Furthermore, the subjects tended to anchor their ratings on this scale by using the attributes, leading to a quantization of the data. This is a well-known phenomenon, also observed in, e.g., Zieliński et al. and in ITU-T COM 12-120. However, there is a number of ratings placed in-between the main categories, indicating a more fine-grained differentiation by the participants. This is also reflected by the lower mean standard deviation. It can, thus, be assumed that the sensitivity of the continuous scale is potentially higher than the sensitivity of the ACR scale.

Most of the findings described in this section are consistent with the findings in ITU-T COM 12-120, where the scale depicted in Figure 3 was compared to the ACR scale. Even for the continuous scale used there, a frequent use of the scale end points could be observed. Obviously,

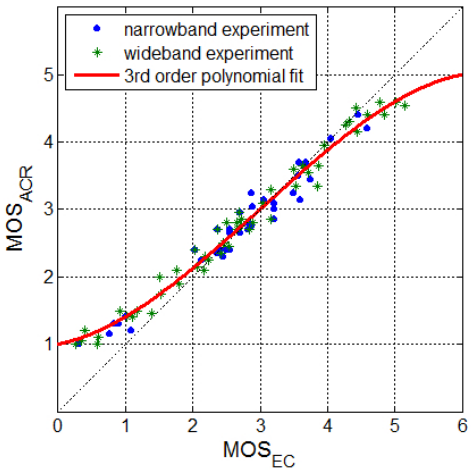

Figure 6: MOS_{ACR} vs. MOS_{EC}; fitting curve

the specific design of the continuous scale is not important in order to produce the characteristics described in this section. Principal differences between both types of continuous scales will be, however, revealed in the next section.

Relation between the scales

Figure 6 depicts the relation between the Mean Opinion Scores (MOS) as obtained on the ACR scale (MOS_{ACR} ; ordinate) and the extended continuous scale (MOS_{EC}; abscissa). The scores were calculated on a per-sample basis, i.e. separately for the male and female speaker.

As it can be seen from the figure, samples of 'fair' quality, corresponding to $MOS = 3$, were rated almost exactly in the middle area of both scales. Towards higher quality, the ratings on the extended continuous scale are generally higher, whereas they are generally lower towards lower quality. Moreover, the saturation effect of the ACR scale is apparent in Figure 6, which is to some degree compensated by the extended continuous scale. In particular, low-quality conditions (e.g., narrowband) in a wideband context are better distinguishable in the lower area of the continuous scale, reflecting the increased sensitivity. Due to the explicit scale extension, the gain in sensitivity is most probably even higher than due to the usage of a continuous scale like the one in Figure 3 (however, further experiments are needed for final conclusions).

Obviously, the extension of the regular ACR scale towards 'ideal' (MOS_{EC}= 6) and 'extremely bad' (MOS_{EC}= 0), both in terms of graphical design and semantic labelling, leads to the (non-linearly) expanded scores. One possible explanation of this expansion is that the participants did not use the attributes/numbers as absolute anchors for their ratings, but rather employed

the whole range of each scale for rating the quality, ignoring the attributes. Zieliński et al. compared the scale depicted in Figure 3 to an attribute-free scale. The authors found that the scores obtained on both scales are equal, meaning that the attributes on the labelled scale are essentially ignored by the participants. However, the histograms in Figures 4 and 5 provide evidence that either the attributes, the tick marks, or both were in fact used as anchors. Thus, an alternative plausible explanation would be that the meaning of the attributes might be shifted in dependence of the presented scale. Due to the additional attributes 'extremely bad' and 'ideal', the relation to the less extreme neighbour-attributes 'bad' and 'excellent', respectively, might be shifted as compared to the discrete ACR scale. This distortion decreases for the even more distant attributes 'poor' and 'good', and vanishes for the mid-scale attribute 'fair'. The distortion is more pronounced towards the lower end of the scale. This observation implies that the attributes are, however, not used in an absolute way, but relative to the particular scale presented to the subjects.

In ITU-T COM 12-120, it was found that the absolute values of the ratings were equal for subjective ratings obtained on the regular ACR scale and the continuous scale depicted in Figure 3. Thus, the attributes can be considered as absolute anchors in that experiment. That is, only the explicit extension of the regular ACR scale as it was done here (Figure 2), i.e. both in terms of graphically distinctive design and semantic labelling of the scale end points, obviously causes the non-linear relation between the scale values.

The relation between the MOS values as obtained on both scales allows a derivation of a function $f : [0, 6] \rightarrow [1, 5]$, mapping the scores MOS_{EC} obtained on the extended continuous scale onto the ACR scale, such that estimates $\widehat{MOS_{ACR}}$ can be calculated. The S-shaped point scatter in Figure 6 can, for example, be approximated with a third-order polynomial function. By means of non-linear curve fitting in a least square sense, with the constraints $f(0) = 1$ and $f(6) = 5$, the following mapping function has been obtained:

$$\begin{aligned} \widehat{MOS_{ACR}} =& f(MOS_{EC}) \\ =& -0.0262 * MOS_{EC}^3 \\ &+ 0.2368 * MOS_{EC}^2 \\ &+ 0.1907 * MOS_{EC} + 1 \end{aligned} \tag{3}$$

The correlation between MOS_{ACR} and $\widehat{MOS_{ACR}}$ amounts to $r = 0.987$ (root mean square error $rmse = 0.162$). Due to the strong relation to the standard ACR scale and the improved scale properties, it can be assumed that both the reliability and the validity of the extended and continuous scale are at least as high as the standard ACR scale.

Conclusions

In this contribution, a comparison between ratings on the ITU-T P.800 ACR scale and a particular extended and continuous quality scale was made. It has been shown that there is a strong relation between the scores as obtained on the ACR and the extended continuous scale, which is of non-linear nature in the edge regions. The empirically derived S-shaped transformation curve between the two scales reveals an implicit compensation of the compression effect known from ACR scales. Moreover, the scores are more spread along lower part of the scale. Together with the lower mean standard deviation, this fact provides some evidence of improved sensitivity.

The higher sensitivity, which is most pronounced in the lower part of the scale, is advantageous for subjective experiments with very good and very bad quality conditions, such as beyond-narrowband tests including narrowband conditions. Since such conditions are typically located in the lower scale area, they are judged with an increased sensitivity on the extended continuous scale.

Due to the strong relation to the standard ACR scale and the improved scale properties, it is also assumed that both the reliability and the validity of the extended and continuous scale are at least as high as the standard ACR scale. Thus, the usage of this scale for subjective assessment of transmission quality seems to be generally advantageous.

References

[1] ITU-T Recommandation P.800, *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, Geneva, 1996.

[2] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*, Kluwer, Boston, 2000.

[3] F. Rumsey S. Zieliński, P. Brooks, "On the use of graphic scales in modern listening tests," in *Proc. 123th AES Convention*, US-New York NY, 2007.

[4] U. Jekosch M. Bodden, "Entwicklung und durchführung von tests mit versuchspersonen zur verifizierung von modellen zur berechnung der sprachübertragungsqualität," Final report on project with Deutsche Telekom AG (unpublished), Institute of Communication Acoustics, Ruhr-University Bochum, 1996.

[5] ITU-T Recommendation P.581, *Subjective quality evaluation of telephone services based on spoken dialogue systems*, International Telecommunication, Geneva, 2003.

[6] ITU-T Recommendation P.880, *Continuous evaluation of time-varying speech quality*, International Telecommunication Union, Geneva, 2004.

[7] ITU-T Contribution COM 12-120, *Investigating the proposed P.OLQA subjective test method*, International Telecommunication Union, Geneva, 2007.

[8] D.Gibbon, "Eurom.1 german speech database," ESPRIT Project 2589 Report (SAM, Multi-Lingual Speech Input/Output Assessment, Methodology and Standardization), University of Bielefeld, DE-Bielefeld, 1992.