Multidimensional evaluation and predicting overall speech quality

Jens Berger, Anna Llagostera

SwissQual AG, Zuchwil, Switzerland

jens.berger@rohde-schwarz.com, anna.llagostera@rohde-schwarz.com

Abstract

The quality of speech samples has been traditionally evaluated in subjective listening tests using 5-point Absolute Category Rating (ACR) scales in Listening Only Tests (LOT) as recommended in ITU-T P.800 [1]. Those tests provide the listening quality aspect of speech quality.

There are other tests are under discussion and proposed in order to assess in detail individual perceptual dimensions of speech. In this paper we investigate the relationship between the overall listening quality obtained in an ITU-T P.800 ACR subjective test and the rating of the same signals in four dimensions proposed by Wältermann [2], namely noisiness, discontinuity, coloration and loudness. The database we use is composed of conditions and speech signals extracted from an ACR LOT used in the ITU-T P.863 evaluation, processed by simulated and live telecommunication channels [3]. The signals have been re-scored using the four mentioned scales and are foreseen as contribution to the ITU-T P.AMD project.

This paper focuses on the modeling of an ACR LOT score based on individual dimensional ratings under the assumption of orthogonality of the four dimensions.

Index Terms: listening quality, subjective testing, multi-dimensional testing

10.21437/Interspeech.2015-549

1. Introduction

Subjective listening tests have been used extensively in order to assess the quality of transmitted and/or processed speech signals. Traditionally, ACR LOT experiments have been used for the speech quality evaluation in a listening situation. In such tests, subjects are asked to determine the perceived quality of a set of distorted speech samples regarding their overall impairment. The used rating scale is shown in Table 1.

However, in some situations, more detailed information about the impairments which affect a speech recording would be desired [4,5]. For this purpose, tests have been designed, which ask for individual dimensions of listening quality. They target the diagnostic quality assessment (based on attributes or characteristics) more than the overall listening quality.

In an ongoing ITU-T activity, a four-dimensional listening experiment has been established. Within such an experiment each listener is asked in sequence for perceived impairments in noisiness, discontinuity, coloration and loudness. The first three dimensions (scales) in this experiment have been extensively studied and found to be orthogonal in [2] and [6]. The loudness scale is added to complete the diagnostic quality assessment.

The goal of this paper is to share the knowledge extracted from conducting a multi-dimensional experiment on a database initially designed for a standard listening quality evaluation, and to study the relationship between the scores obtained with such a multi-dimensional test and the overall listening quality that is derived from a P.800 ACR LOT.

1.1. Single scale ACR overall quality experiments

In an ACR experiment a speech sample is presented for scoring. The scoring labels reflect absolute terms such as "good" or "bad", while the scale ranges downward from a high score to a lower score. In such a test the person must rate the quality on this absolute scale based on his or her own expectation and experience. The most common rating method for an ACR test is a five point scale, which has verbal categories in the native language of the test subjects (Table 1).

Speech quality	Excellent	Good	Fair	Poor	Bad
Score	5	4	3	2	1

Table 1. *ACR scale for Listening Only Tests*

The experiment usually consists of more than one hundred speech samples to be scored sequentially. To minimize gender and talker biases as well as training effects on known material, different talkers and sentences are used in one experiment. For further information please refer to [1].

1.2. Four-dimensional listening experiments

The subjective tests carried out in multiple dimensions are designed closely to the very well established ACR LOT according to [1]. Instead of judging on one 'overall quality' scale the subjects are asked to score the presented sample on the proposed four different scales for individual impairments [6].

The four descriptive scales are presented sequentially to the subjects, each of them with labels at both ends of the scales describing the characteristic to be judged. Aligned to ACR LOT, five-point but continual rating scales are used for this experiment. Each of the scales target an individual dimension of perception; however in telecommunications there are typical technical causes behind each dimension, which may have an impact in one or more of the four scales:

- **Noisiness**: Presence of background noises at the sending side or additive / modulated noises during transmission.
- **Discontinuity**: Interruption of the speech flow as e.g. gaps caused by unconcealed packet/frame losses in the transmission the same as perceived discontinuities by frequent and strong warping of the speech signal impairing the natural flow of the speech signal.
- **Coloration**: Bandwidth limitations, preference of spectral ranges or perceptible spectral losses.
- **Loudness**: Loudness of the signal perceived as non-optimal by the subjects for a telecommunication service.

2. Test design and used speech material

For the four-dimensional experiment, we have used a sub-selection of 120 speech samples, which have been taken from a data set used in the ITU-T Rec. P.863 'POLQA' competition [3]. For each sample the overall quality score was already available from the previous and origin investigation; the experiment conducted for this study has evaluated the four-dimensions only.

The 120 speech samples represent 30 different test conditions, e.g. a mobile channel, a codec or similar. In each condition a set of four different samples (double-sentences) spoken by two male and two female speakers are processed. In each condition the speech processing and applied distortion is seen as identical in the technical sense. The bandwidth is exceeding ordinary telephone band and includes so-called super-wideband speech (up to 14kHz audio band-width) and intermediate limitations as e.g. 'wideband' that ends at 7kHz.

The selected speech samples are impaired by one or more technical distortions, in a consequence a perceived impairment could be expected on one or more quality dimensions. There is a subset that has intentionally only one single dimension impaired (e.g. added noise only or only band limitation), while the majority of conditions are distorted in multiple dimensions.

The listening conditions are as defined in [1]. The listeners are located in a quiet test room and the speech samples are presented monotically via diffuse-field equalized headphones. The listeners are instructed to imagine a telephony situation where they listen to a far-end conversational partner.

Each of the speech samples has been evaluated by 24 listeners. Each subject scored the full set of 120 samples but in different order to avoid context effects. In front of the experiment, the listeners received a small set of training sequences to get familiar with the test procedure. After listening to a speech sample, the subject was requested to score one dimension after the other. The listener was able to re-play the speech sample for its confidence. The scales for noisiness, discontinuity, coloration were displayed to each subject in a random order while the loudness was always in the last position. The already given scores are not visible to listeners anymore when scoring the next dimension to avoid dependencies on the given scores. After the loudness was evaluated, the next speech sample was presented.

It is important to note, that the listening panel for the overall quality is a different one than the one invited to the four-dimensional experiment. In our opinion, the disadvantage in having a different listening panel counts less than re-inviting the same listeners again or even asking for overall quality in the same session as the individual dimensions have to be evaluated. This may lead to a bias by listening to the same conditions again under a different question.

2.1. General results and statistics

The experiment results in individual scores for each speech sample given by each listener. The individual scores are averaged for the same speech sample according to [1] and form a Mean Opinion Score (MOS) for this sample, a so-called 'MOS per-sample'. To minimize sentence, gender and talker effects of the MOS, the MOS of the samples forming a condition are averaged to a 'MOS per-condition' again.

The average scores for noisiness, discontinuity, coloration and loudness are respectively named in this paper as N-MOS, D-MOS, C-MOS and L-MOS, while the O-MOS stands for the overall listening quality according to ACR LOT [1].

The average MOS values across all samples in the test and 95% confidence intervals (Ci95%) for the four-dimensional experiment and the overall listening quality are shown in Table 2. These values just give an impression about the balance of the test regarding average quality and the confidence of the listeners while scoring.

	Multi-dimensional				***P.800***
	N	D	C	L	**O**
MOS	3.20	4.01	3.12	3.79	**2.65**
Ci95%	0.177	0.177	0.187	0.165	**0.170**

Table 2. *Avg. MOS and avg. Ci95% per condition. N, D, C, L and O denote the noisiness, discontinuity, coloration, loudness and overall quality respectively.*

Average MOS values for the four scales across all samples seem to be a bit higher than the mid-point of the five-point scale at 2.5, since the test conditions were initially designed to be balanced in the overall listening quality rather than in individual dimensions. For example, the average value for discontinuity is close to 4, since there are few conditions with medium to strong packet loss in the test. Nevertheless, the confidence intervals of the scores are in the same range.

The confidence intervals for the scores in the four dimensions show a similar distribution to each other and the overall listening quality from the P.800 test (see Figure 1). The outlying condition in the noisiness dimension (upper left diagram) is a MNRU condition [7], where a modulated noise is applied to the speech signal. [1]

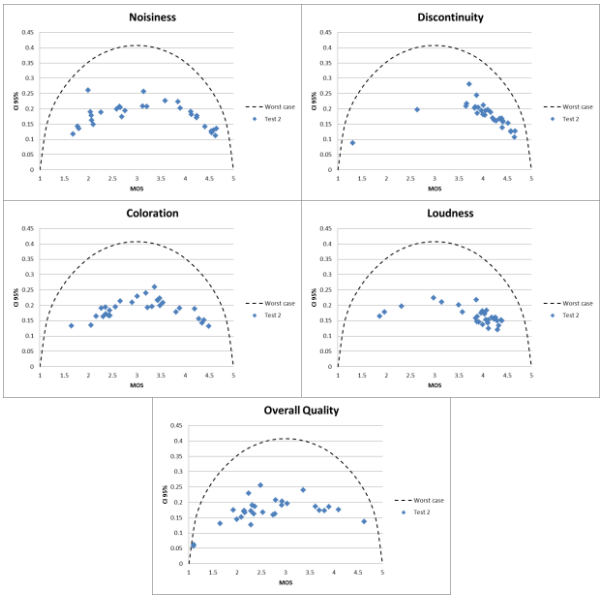

Figure 1. *95% confidence intervals vs. MOS per condition*

[1] A detailed analysis has shown that there is an individual preference, some subjects categorize MNRU as noisiness, others as discontinuity or coloration, and some of them perceive that there is an impact on several of the scales. MNRU is emulating a logarithmic PCM and has been used for decades as a reference and anchor. However, it does not reflect impairments in today's networks anymore.

2.2. Observations

Impairments clearly assignable to a noisy, interrupted or non-optimal in presentation level are scored with confidence on the corresponding scales. We have observed that distortions which cannot be clearly categorized in the three scales above as e.g. codecs are classified into the coloration scale. This is not only due to a common bandwidth limitation when applying a codec, rather the codec impairments are seen as a sort of coloration. A clear example occurs with codecs and strong amplitude clipping (see Figure 2).

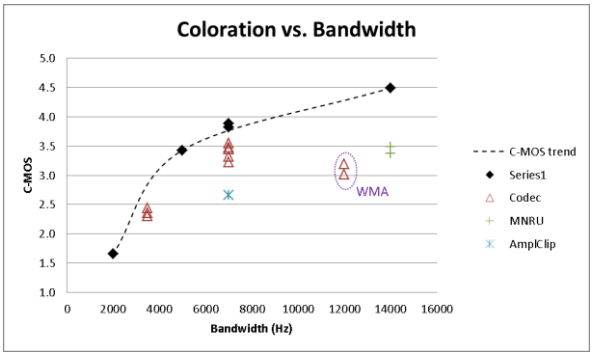

Figure 2. *Score at the coloration scale (C-MOS) vs. bandwidth on a subset of conditions with different bandwidths. The C-MOS trend is interpolated from anchor conditions with pure bandwidth limitations.*

As explained before the coloration scale reflects mainly the presence of bandwidth limitations. However, the score obtained for codecs seems to be lower than the corresponding score according to the bandwidth. Thus, other distortions caused by codecs seem also to affect significantly the score on the coloration scale. The same happens with MNRU and amplitude clipping conditions. Actually, the coloration scale seems to gather information about the 'naturalness' of speech.

3. Modelling

In this section we want to study two relationships

- How is the relation of scores on a dimensional scale vs. the overall quality if only one impairment is present?
- Is it possible to predict the overall quality acc. to P.800 [1] by the scores on the individual scales?

3.1. Relationship between individual dimensions and overall listening quality

In the first place we analyze the relationship between each of the four dimensional scales and the listening quality as obtained in an ITU-T P.800 test independently. We focus on a subset, where only a single impairment is present that affects only one dimension and the score on the remaining three is >3.5 (it means that there is no or almost no impairment in any other dimension). The results are drawn in Figure 3.

In principle the sensitivity of the scales are similar, a decreased score in one dimension shows approximately the same impact on the overall scale. For example a score of ~3.0 on the discontinuity scale will result in ~3.0 too if the same sample is scored for overall quality. The discontinuity and loudness are the scales presenting a high correlation with the overall quality, while the coloration shows a very weak relationship.

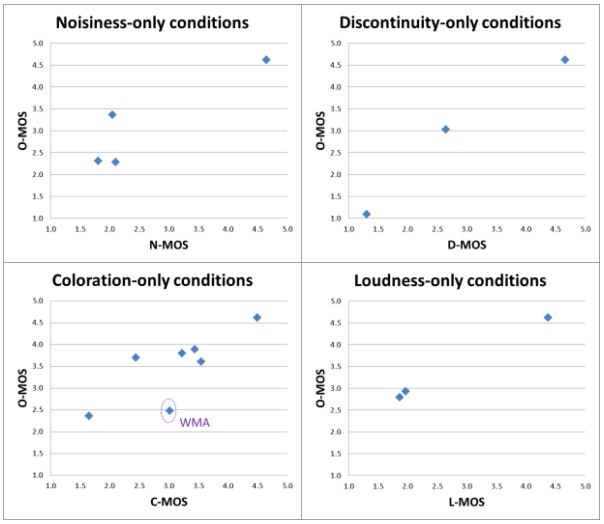

Figure 3. *Relationship between the P.800 listening quality (O-MOS) and each of the four dimensions.*

The main outlier in coloration compared to overall quality is caused by a condition coded with Windows Media Audio (WMA) at a low bitrate in which the overall listening quality is lower than expected according to the coloration score. This condition is already predicted very low at the coloration scale (see Figure 2).[1]

3.2. Modeling the overall quality

First we select the conditions that present degradations in only one of the four dimensions to estimate a model that derives the overall MOS from a linear combination of the scales. In a second step we try to see if this model fits the entire set of data and, in particular, the conditions in which several scales are impacted by the degradations.

We assume orthogonality of the four dimensions according to [6], therefore the model will not contain interaction terms between the scales. We start with a simple linear model of the form

$$O = f(N,D,C,L) = a0 + a1\,N + a2\,D + a3\,C + a4\,L \qquad (1)$$

where *N, D, C, L* correspond to the N-MOS, D-MOS, C-MOS and L-MOS and *O* corresponds to the overall speech quality.

The coefficients of the linear model are obtained using the least squares method, which minimizes the RMSE defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M}(O_i - \widehat{O_i})^2}{M-K-1}} \qquad (2)$$

Here O_i are the O-MOS values derived in the subjective experiment, $\widehat{O_i}$ the predicted O-MOS values, M the number of samples used in the regression and K the number of independent variables.

The values obtained when fitting the linear model to the 'per-condition' scores for this subset are shown in Table 3. As

[1] The reason is mostly given by the spectral representation of the higher frequency range. The WMA codec as used reproduces no spatial details in this range. Formally, it fills the bandwidth of 12 kHz but the higher spectral parts must be considered as unnatural in sound.

expected, we obtain a1, a2, a3 and a4 values that are positive, since there is a direct relationship between the overall listening quality and the noisiness, discontinuity and coloration scores (see Figure 3). As in [6], the discontinuity is the most important feature for deriving the overall listening quality.

	a0	a1	a2	a3	a4
Coeffs.	-7.947	0.641	1.046	0.467	0.672

Table 3. *Regression coefficients for formula (1)*

The four individual dimensions are found to be statistically significant for the overall speech quality estimation; in all cases the P-value<0.05. This fit leads to a correlation coefficient of 0.91 between the subjective speech quality scores per sample and its prediction based on the four dimensions.

From Table 4, it can be seen that the prediction of an overall speech quality results in lower RMSE for the samples where only one dimension is affected compared to multiple distortions. This was expected, since the conditions with only one affected dimension were used to train the predictor.

Affected dimensions	RMSE		Average error	
	Single	*Multiple*	*Single*	*Multiple*
	0.39	0.75	0.01	0.55

Table 4. *RMSE and average error ($O_i - \widehat{O}_i$) for the fit in equation (1) and the coefficients in Table 3.*

The scatterplot in Figure 4 shows the relation between the speech quality values derived in the ACR-LOT experiment and their estimations based on the four individual dimensions applying the linear model as in formula (1).

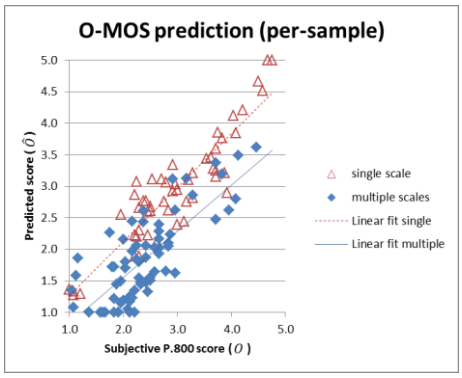

Figure 4. *Scatterplot O-MOS vs. predicted O-MOS*

The unfilled triangular symbols in Figure 4 represent the speech samples which are only impaired in one dimension and have been used for the determination of the model coefficients. The filled diamonds are representing all samples, where the impairment affects more than one dimension (e.g. a bandwidth limitation and additive noise in the same speech sample).

As expected the prediction for conditions which have not been used for determining the optimal fitting coefficients (diamonds) scatter more than the predictions for conditions used for training the predictor.

The more interesting observation is that there is a bias between predictions of conditions where only one dimension is affected to conditions where multiple dimensions are impaired. In case a speech sample is affected by more than one distortion and is penalized on multiple dimensions, the additive prediction of the overall speech quality is too pessimistic compared to an integrative score as in an ACR-LOT experiment, that leads to an under-prediction. This bias is also visible as average prediction error in Table 4 and it is in a range of about 0.5 MOS.

In this study only one experiment and 120 speech samples have been used. To avoid over-training effects a simple linear additive model with four terms was chosen. However, even if a non-linear approach was used the observed bias would remain if the individual terms just represent one individual dimension. This bias can only be avoided, if terms combine scores from multiple dimensions to compensate dependencies of multiple individual dimensions to the overall score as e.g.

O = a0 + a1 N + a2 D + a3 C + a4 L + b1 ND + b2 NC + …

This observation is not in contradiction to the reported orthogonality of the used four dimensions in [2,6], it is pointing to the hypothesis that integration of multiple impairments to an overall quality in the human brain is different to sub-dividing impairments on individual scales. The influence to an overall speech quality when multiple distortions are present is strongly influenced by inter-dimensional masking effects which are not considered in the model as targeted in this study, in formula (1).

4. Conclusions

The discussed test methodology provides reliable results in assessing four dimensions of speech quality in one single experiment. The confidence intervals are similar to those obtained in common-practice ACR-LOT acc. P.800 tests.

The coloration scores seem to be strongly linked to the bandwidth of the analyzed condition. In addition, the subjects reflect in the coloration scale distortions that are not clearly to classify to any of the other three dimensions. The dimension 'Coloration' acts as a kind of 'Naturalness'.

It is not suitable to predict an overall speech quality score in case of multiple distortions based on a model where the individual terms are representing only one dimension. Such a model does not consider inter-dimensional masking effects and tends to under-predict speech samples with multiple distortions.

5. References

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

[2] M. Wältermann, "*Dimension-based Quality Modeling of Transmitted Speech,*" Berlin: Springer, 2012.

[3] ITU-T Rec. P.863, "*Perceptual Objective Listening Quality Assessment (POLQA)*". Geneva: International Telecommunication Union, 2011.

[4] N. Côté, V. Koehl, S. Möller, A. Raake, M. Wältermann, and V. Gautier-Turbin, "Diagnostic Instrumental Speech Quality Assessment in a Super-Wideband Context*," Journal of the Audio Engineering Society*, 2012, 60 (3), pp.156-164.

[5] Côté, N., "*Integral and Diagnostic Intrusive Prediction of Speech Quality*," Berlin: Springer, 2011.

[6] 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 Proc.(ICSLP'06), Pittsburgh, PA*, Sept. 17–21 2006, pp. 2170–2173.

[7] ITU-T Rec. P.48, "*Specification for an intermediate reference system*," Geneva: International Telecommunication Union, 1988.