4th International Workshop on Perceptual Quality of Systems (PQS 2013)
2-4 September 2013, Vienna, Austria

# Analyzing Technical Causes and Perceptual Dimensions for Diagnosing the Quality of Transmitted Speech

*Sebastian Möller* [1], *Friedemann Köster* [1], *Falk Schiffner* [2], *Janto Skowronek* [2]

[1] Quality and Usability Lab, Telekom Innovation Labs, TU Berlin, Germany
[2] Assessment of IP-based Applications, Telekom Innovation Labs, TU Berlin, Germany
sebastian.moeller@telekom.de, friedemann.koester@telekom.de, falk.schiffner@telekom.de, janto.skowronek@telekom.de

## Abstract

We present an analysis of technical causes and corresponding perceptual dimensions of the quality of transmitted speech. Four experts annotated speech files of a common database according to a methodology which is currently being discussed for the future ITU-T Recommendation P.TCA. The annotations are analyzed with respect to their frequency and consistency, and compared to overall quality values as well as perceptual dimension scores, obtained either from subjective experiments or from instrumental models. This way, links can be drawn between technical causes and perceptual dimensions in diagnostic quality assessment of transmitted speech, and the proposed procedures of several upcoming ITU-T Recommendations can be initially validated.

**Index Terms**: speech transmission quality, technical causes, perceptual quality features, evaluation methods.

## 1. Motivation and introduction

For evaluating the quality of transmitted speech, auditory experiments carried out in a laboratory context are valid and reliable means. In such experiments, naïve test participants judge the overall quality of transmitted speech signals on standard rating scales. The most common procedures are based on Absolute Category Rating tasks [1] and result in a Mean Opinion Score (MOS) for each signal or processing condition. Unfortunately, these tests provide little insight into the reasons of sub-optimal quality.

In order to provide more diagnostic information, Study Group 12 (SG 12) of the International Telecommunication Union (ITU-T) has recently opened three work items. The aim of these work items is to define subjective evaluation methods and instrumental prediction models which are able to diagnose quality. Two paths are conceivable for this purpose:

1) identification of the technical causes of sub-optimum quality, in terms of characteristics of the signal or the transmitting system which cause the lower quality judgment; or
2) identification of perceptual dimensions of the transmitted signal; these dimensions can be considered as quality features in a multidimensional space, and the overall quality judgment can be seen as a distance to an optimum point (to the perceptual reference) in this space [2].

Both paths can be followed either with subjective experiments or with instrumental quality predictors.

For path 1), ITU-T SG12 has developed a methodology for performing expert annotations after listening to transmitted speech files. This methodology is proposed for a future Recommendation P.TCA (for "Technical Cause Analysis"), and aims at identifying signal characteristics such as sub-optimum speech level, speech spectrum, noise level, echo, or alike [2]. For path 2), a subjective evaluation method based on semantic differential attributes has been solicited and is foreseen for a future Recommendation P.MULTI (for "MULTIdimensional") [3]. A proposal for the implementation of this methodology has been made by Wältermann [4], based on the four dimensions "coloration", "discontinuity", "noisiness" and sub-optimum "loudness"; another proposal is based on the Diagnostic Acceptability Measure [5] and results in 7-8 dimensions [6]. In addition, instrumental predictors for the former four dimensions have been developed by Côté [7], and proposed for an upcoming Recommendation P.AMD (for "Assessment of Multiple Dimensions"). There are obvious links between the technical causes and the perceptual dimensions, but these have never been analyzed on a common database to our knowledge.

In this paper, we present results from an annotation experiment following the preliminary P.TCA guidelines. In this experiment, a database from the ITU-T Rec. P.863 [8] competition was annotated by 4 expert listeners in Telekom Innovation Labs, TU Berlin, with respect to the assumed technical causes of the degradations. The same database had been judged in a listening experiment [9] with naïve test participants regarding the overall quality (MOS) and the four perceptual dimensions which shall be included in P.AMD. Thus, technical causes, perceptual dimensions and overall quality judgments are available for the same speech files. In addition, we calculated instrumental predictions for MOS and for the perceptual dimensions, using ITU-T Rec. P.863 (POLQA) [8] and the DIAL model [7]. The results are analyzed with respect to the annotation reliability, as well as with respect to the relationships between technical causes, perceptual dimensions and overall quality. Conclusions are drawn for improving the P.TCA methodology, as well as for the usefulness of P.TCA and P.AMD in diagnostic quality assessment.

## 2. Database

For our analysis, we selected database number 503 from the ITU-T Rec. P.863 competition which has been kindly provided by SwissQual AG, Solothurn. This particular database includes diverse types of degradations and degradation combinations for which diagnostic information is most useful. The stimuli were produced in a number of different labs according to the ITU-T Rec. P.863 specifications; four speakers with four different German sentences were used per condition. The database is mixed-band (narrowband 300-3400 Hz, wideband 50-7000 Hz, super-wideband 50-14000 Hz) and contains signal-correlated as


10.21437/PQS.2013-5

well as uncorrelated noise, ambient background noise of different types, temporal clipping, coding at different bitrates, temporal stretching, packet loss of different loss profiles, acoustic recordings, different frequency distortions, as well as combinations of these degradations [10].

## 3. Technical cause annotations

The speech files were annotated by four experts who have been trained for the given task through an annotation manual [2]. These experts were dealing with speech and audio processing as part of their work (either PhD students or Master students) and were particularly trained for the given task through the annotation manual. The experts listened to the database in several sessions in a quiet office room, two used Sennheiser HD 280 pro and two Beyerdynamic DT-series headphones at a comfortable listening level. Sound presentation was diotic through a Realtek High Definition Audio ALC 268 soundcard.

The experts' task was to identify the most prominent causes of degradations within each evaluated sample on two levels [11]. First, experts had to identify the most dominant types of degradations (level 1) and rate them according to whether they are highly dominant, dominant, or less dominant. Usually, there might be only one such degradation, but there should be no more than three mentioned. Second, experts should identify the detailed classes of degradations (level 2) for each of the dominant degradations. In case that they could not identify these, this was not considered as a problem. As for the level-1 degradations, each of the detailed classes of degradations should also be rated with respect to their dominance. Usually, there should be 1…2 level-2 degradation classes, but there could also be more. A total of 47 different impairments on level 2, grouped into 9 categories on level 1, were provided to the experts [2]. In accordance with **Error! Reference source not found.**, experts were asked to judge only those samples that received a subjective MOS score of 3.0 or below in the subjective scores accompanying the database; these were 33 out of the 54 conditions, listed in Table 1.

Table 1: Frequency of level-1 degradation annotations per condition.

| Cond. | MOS | Speech Level | Speech Spectrum | Speech Distortion | Speech Inform. | Echo | Noise Level | Noise Steady-st. | Noise Dynamic | Noise Impulsive |
|---|---|---|---|---|---|---|---|---|---|---|
| C02 | 1.19 | 0 | 0 | 12 | 0 | 0 | 4 | 0 | 8 | 0 |
| C03 | 2.88 | 0 | 0 | 4 | 0 | 1 | 0 | 0 | 16 | 0 |
| C04 | 2.46 | 4 | 0 | 1 | 0 | 0 | 7 | 16 | 0 | 0 |
| C09 | 2.97 | 5 | 16 | 4 | 0 | 1 | 0 | 0 | 1 | 0 |
| C12 | 1.11 | 16 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 8 |
| C13 | 2.45 | 2 | 6 | 1 | 0 | 0 | 15 | 1 | 0 | 0 |
| C14 | 2.55 | 3 | 11 | 0 | 0 | 0 | 16 | 4 | 0 | 0 |
| C17 | 2.42 | 0 | 12 | 7 | 0 | 0 | 4 | 8 | 11 | 0 |
| C18 | 2.45 | 16 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C19 | 2.54 | 0 | 16 | 4 | 1 | 0 | 0 | 9 | 4 | 0 |
| C26 | 2.58 | 4 | 16 | 8 | 0 | 0 | 0 | 6 | 3 | 0 |
| C27 | 2.48 | 0 | 15 | 1 | 0 | 0 | 0 | 7 | 12 | 0 |
| C28 | 2.08 | 7 | 12 | 4 | 0 | 0 | 15 | 4 | 1 | 0 |
| C29 | 1.77 | 16 | 0 | 0 | 0 | 0 | 7 | 2 | 9 | 0 |
| C30 | 1.34 | 11 | 9 | 1 | 0 | 0 | 0 | 8 | 12 | 0 |
| C32 | 2.64 | 5 | 4 | 0 | 0 | 0 | 16 | 0 | 6 | 0 |
| C35 | 2.43 | 10 | 12 | 8 | 0 | 3 | 0 | 7 | 1 | 3 |
| C36 | 2.14 | 12 | 10 | 3 | 0 | 2 | 0 | 5 | 3 | 3 |
| C37 | 2.29 | 6 | 16 | 4 | 0 | 0 | 0 | 7 | 2 | 0 |
| C38 | 2.80 | 4 | 16 | 4 | 0 | 4 | 0 | 0 | 2 | 1 |
| C39 | 1.90 | 2 | 16 | 8 | 0 | 0 | 0 | 6 | 11 | 0 |
| C40 | 2.16 | 4 | 16 | 0 | 0 | 0 | 0 | 5 | 5 | 1 |
| C41 | 2.89 | 3 | 16 | 0 | 0 | 0 | 0 | 10 | 5 | 0 |
| C42 | 2.77 | 4 | 16 | 4 | 0 | 0 | 0 | 16 | 0 | 0 |
| C43 | 1.30 | 16 | 5 | 0 | 3 | 0 | 4 | 12 | 1 | 0 |
| C44 | 2.48 | 1 | 16 | 0 | 0 | 0 | 1 | 15 | 2 | 0 |
| C45 | 2.23 | 9 | 12 | 0 | 0 | 0 | 6 | 0 | 3 | 4 |
| C47 | 1.86 | 8 | 4 | 1 | 0 | 0 | 8 | 3 | 9 | 11 |
| C50 | 2.60 | 13 | 8 | 4 | 4 | 0 | 0 | 0 | 0 | 0 |
| C51 | 2.83 | 3 | 16 | 0 | 0 | 0 | 4 | 8 | 0 | 7 |
| C52 | 2.78 | 12 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| C53 | 2.80 | 1 | 16 | 0 | 0 | 2 | 8 | 9 | 1 | 2 |
| C54 | 3.00 | 15 | 14 | 0 | 0 | 0 | 0 | 6 | 0 | 0 |
| Σ | | 212 | 345 | 83 | 9 | 16 | 115 | 174 | 128 | 44 |

Table 1 shows the number of cases where experts have attributed a level-1 degradation to a speech file of the corresponding condition. As there were 4 speech files per condition and four annotators, a maximum of 16 annotations per condition and class could occur (overall 528 annotations).

The table shows that there are some level-1 classes which were annotated more frequently than others. Most labels were given to the "Speech Spectrum", "Speech Level" and "Noise Steady-state" classes. "Speech Information" and "Echo" were the classes less frequently used. This result may either be linked to the particularities of the database used (i.e. that the corresponding degradations were rare in that database, e.g. echo), or it may be linked to problems in identifying particular classes of degradations from pure listening (despite their presence in the database). It may also be linked to the task of the annotators, in the sense that a level-2 degradation like "Poor Speaker Identification", which corresponds to the level 1 degradation "Speech Information", can hardly be annotated in a listening-only task with a database of only four speakers.

Some particular conditions, e.g. condition 47, reflect a combination of degradations, for which the annotations of the experts apparently are not very congruent. The reason for this may be that in this condition it is not clearly noticeable what the "major" degradation is. On the other hand some other conditions have been annotated very homogeneously by the experts. E.g. condition 12 has a degradation of 20% time clipping which was annotated by all experts with the level-2 degradation "Temporal Speech Clipping" (corresponding to level-1 degradation "Speech Level").

The reliability of the annotation process was analyzed with the help of the kappa coefficient, see Table 2. Kappa values show that fair to moderate agreement was obtained for all level-1 degradation classes which occur more frequently, i.e. with a minimum of 20 labels. Only the rarely-occurring classes (less frequent than 20 occurrences) show only a slight agreement.

Table 2: Kappa coefficients for level-1 degradation classes. Interpretation of kappa values: < 0: poor agreement; 0.0 – 0.20: slight agreement; 0.21 – 0.40: fair agreement; 0.41 – 0.60: moderate agreement; 0.61 – 0.80: substantial agreement; 0.81 – 1.00: almost perfect agreement [12].

| Degradation class | Frequency | Kappa |
|---|---|---|
| Speech Spectrum | 345 | 0.595 |
| Speech Level | 212 | 0.439 |
| Noise Steady-state | 174 | 0.373 |
| Noise Level | 138 | 0.592 |
| Noise Dynamic | 128 | 0.388 |
| Speech Distortion | 83 | 0.237 |
| Noise Impulsive | 44 | 0.316 |
| Echo | 16 | 0.089 |
| Speech Information | 9 | 0.118 |

With respect to the level-2 degradation classes, these are obviously less frequently annotated. Table 3 shows that there are 15 classes which have been used at least 20 times by the annotation experts. Once again, this could be due to the particular degradations contained in the database, or to the properties of the labeling procedure. Because of their lower frequency of occurrence, level-2 degradation classes are not analyzed further, and the analysis is limited to the level-1 classes in the following sections.

Table 3: Frequency of level-2 degradation annotations.

| No. | Level 2 degradation class | Frequency |
|---|---|---|
| 1 | Colored speech | 130 |
| 2 | Sharp speech | 115 |
| 3 | Pink noise | 115 |
| 4 | Loud noise | 106 |
| 5 | Muffled speech | 101 |
| 6 | Quit speech | 91 |
| 7 | Modulation noise | 70 |
| 8 | Temporal speech clipping | 66 |
| 9 | Self clipping | 40 |
| 10 | White noise | 35 |
| 11 | Noise level fluctuation | 32 |
| 12 | Clicks | 32 |
| 13 | Distorted background noise | 25 |
| 14 | Whine | 21 |
| 15 | Rough speech | 20 |
| 16 | Timbre varies | 18 |
| 17 | Hissy speech | 17 |
| 18 | Noise gating | 17 |
| 19 | Buzzy speech | 14 |
| 20 | Pops | 12 |
| 21 | Muddy speech | 11 |
| 22 | Tunnel sounding speech | 11 |
| 23 | Fuzzy speech | 9 |
| 24 | Poor intelligibility | 9 |
| 25 | Warped speech | 8 |
| 26 | Nasally speech | 7 |
| 27 | Listener echo | 5 |
| 28 | Wind buffering | 5 |
| 29 | Crackling noise | 5 |
| 30 | Choppy speech | 4 |
| 31 | Buzz | 4 |
| 32 | Speech level fluctuations | 3 |
| 33 | GSM buzz | 3 |
| 34 | Musical tones | 3 |
| 35 | Loudness varies | 2 |
| 36 | Loud speech | 2 |
| 37 | Hum | 2 |
| 38 | Hiss | 1 |
| 39 | Speech cut-outs | 0 |
| 40 | Poor speaker identification | 0 |
| 41 | Poor localization | 0 |
| 42 | Line sounds dead | 0 |
| 43 | Temporal noise clipping | 0 |
| 44 | Noise cut-outs | 0 |
| 45 | Motorboating | 0 |
| 46 | Static | 0 |
| 47 | Pre-echo | 0 |

Table 4: Spearman rank order correlation between frequencies of level-1 degradation classes and subjective judgments. Correlations with an absolute value higher than 0.35 are printed boldface.

| Degradation Class | MOS | Coloration | Discontinuity | Loudness | Noisiness |
|---|---|---|---|---|---|
| Speech Level | -0.275 | 0.242 | -0.124 | **-0.354** | **0.388** |
| Speech Spectrum | 0.47 | **-0.889** | 0.109 | -0.024 | **0.361** |
| Speech Distortion | -0.098 | -0.053 | -0.191 | **0.402** | -0.222 |
| Speech Information | -0.195 | 0.185 | -0.036 | -0.161 | 0.331 |
| Echo | 0.138 | 0.057 | **-0.285** | **0.54** | **0.403** |
| Noise Level | -0.191 | **0.396** | 0.159 | -0.115 | **-0.609** |
| Noise Steady-state | 0.035 | **-0.365** | 0.085 | -0.05 | -0.263 |
| Noise Dynamic | -0.283 | 0.000 | **-0.488** | 0.323 | -0.365 |
| Noise Impulsive | -0.103 | -0.022 | **-0.441** | **0.385** | **0.413** |

Table 5: Spearman correlations between level-1 degradation classes and instrumental quality predictions. MOS-C: Coloration estimation; MOS-D: Discontinuity estimation; MOS-L: Loudness estimation; MOS-N: Noisiness estimation. All estimations are provided on the MOS scale, with 1 being the worst and 5 being the optimum score. Correlations with an absolute value higher than 0.35 are printed in boldface.

| | DIAL MOS-C | DIAL MOS-D | DIAL MOS-L | DIAL MOS-N | DIAL MOS | POLQA MOS |
|---|---|---|---|---|---|---|
| Speech Level | -0.006 | 0.046 | **-0.382** | -0.146 | -0.191 | -0.284 |
| Speech Spectrum | **-0.758** | -0.295 | 0.015 | **0.484** | **0.401** | **0.524** |
| Speech Distortion | 0.215 | -0.226 | 0.334 | -0.087 | 0.128 | -0.012 |
| Speech Information | -0.019 | 0.009 | **-0.384** | 0.146 | 0.02 | -0.134 |
| Echo | 0.224 | -0.294 | 0.118 | 0.318 | 0.073 | 0.193 |
| Noise Level | 0.295 | **0.445** | 0.246 | **-0.611** | **-0.469** | -0.263 |
| Noise Steady-state | **-0.467** | -0.057 | 0.071 | 0.104 | 0.236 | 0.226 |
| Noise Dynamic | 0.198 | **-0.423** | 0.169 | -0.057 | -0.062 | -0.201 |
| Noise Impulsive | -0.002 | **-0.415** | -0.079 | 0.196 | -0.207 | 0.047 |

After the annotation process, the experts reported different difficulties in annotating the level-1 and level-2 degradation classes. Assigning a level-1 degradation class to a certain condition was at first glance easy, as those class names allowed to subsume a rather broad range of distortions, for instance "Speech Distortion". When the experts attempted to assign also the level-2 degradations, the task got more complicated[1]. For instance, one expert reported his difficulties in case of short speech segments with a metallic voice character, e.g. artefacts that can be observed for packet loss concealment (e.g. C38) or noise reduction (e.g. C30). It would have been easy to assign such distortions to "Speech Distortion". However, the closest available level-2 degradation for the expert was "Timbre Changes", as the voice character is changing for a short moment, but this was defined to be subsumed under level-1 class "Speech Spectrum".

A second related aspect that the experts reported was that the broad meaning of the level-1 class names without a further description of that name triggered the experts to use the level-2 degradations with corresponding descriptions as "definitions" for the aspects of that level-1 class. This essentially results in some kind of bottom-up approach, while the instructions are intended to be used in a top-down approach. Overall the experts reported that the annotation would have been much easier if there were example files for the level-2 degradations.

[1] Note that we do not refer here to the "trivial" fact that one had to listen more carefully to identify the details of the distortion.

## 4. Relationships between technical causes and subjective ratings

Spearman rank order correlations between annotation frequencies of the technical causes and subjective judgments described in [9] (both in terms of MOS and perceptual dimensions) were analyzed. The results are shown in Table 4.

The table shows that there is no simple relationship between the occurrence of individual degradations and MOS. The picture becomes clearer when the perceptual dimension judgments are considered. Coloration is significantly negatively correlated with degradations in the speech spectrum. To a lesser extent, also steady-state noise degradations (hum, buzz, etc.) may impact coloration. The noise level shows a moderate positive correlation with coloration, indicating that noise may mask colorations of the speech signal to a certain degree. Discontinuity is negatively correlated with dynamic and impulsive noise components. Furthermore, there are echo degradations (tunnel-sounding speech, listener echo) which also contribute to the impression of discontinuity. Loudness moderately correlates with the speech level (e.g. loud speech, quiet speech). The perceptual effect of noisiness correlates most notably with the noise level, and to a lesser extent also with the presence of impulsive noise and echo components.

## 5. Relationship between technical causes and instrumental predictions

In addition to the subjective ratings, instrumental models were used for predicting the overall quality of the processed speech files, and also for predicting the perceptual dimension scores in the way which is foreseen by P.AMD. Two such models were available to us: POLQA (ITU-T Rec. P.863, [8]) and DIAL [7]. Whereas both models provide an estimation of the overall MOS in a super-wideband context, only DIAL is able to predict the perceptual dimension scores, as it is targeted for P.AMD. The corresponding Spearman rank order correlation coefficients are given in Table 5.

There are three moderate positive and negative correlations between predicted MOS values and the frequency of level-1 degradation classes. Positive correlations are found for the "Speech Spectrum" class with both POLQA and DIAL, and a negative correlation for the "Noise Level" class and DIAL. Furthermore, there is a slight negative correlation with the "Speech Level" class, and a slight positive correlation with the "Noise Steady-state" class.

When predicting perceptual dimensions with DIAL, the coloration estimate correlates strongly negatively with the "Speech Spectrum" degradation class. Further contributions to this dimension estimate come from the "Noise Steady-state" class. The "Noise Level" class has a slight positive correlation to this dimension estimate. The discontinuity estimate correlates most strongly with the presence of "Noise Dynamic" and "Noise Impulsive" classes of degradations. It may be masked by high noise levels, as the positive correlation with "Noise Level" indicates. Loudness dimension estimates are most strongly (negatively) correlated with the "Speech Level" and "Speech Information" classes. For noisiness, the highest correlation is observed with the "Noise Level" class, followed by the "Speech Spectrum" class of degradations. This indicates that also degradations on the speech signal itself can contribute to the impression of noisiness.

## 6. Discussion

The annotations described in our paper have been performed by 4 "experts" who have a reasonable experience in speech quality assessment, and who have been familiarized with the annotation task via the P.TCA annotation manual [11]. Whereas it would of course have been desirable to have more experts at hand to more precisely determine inter-rater agreement, we think that the results are nevertheless useful for the given purpose, and they might be realistic for a real-life situation in which hardly more annotators can be found.

The results of our analysis show that many of the P.TCA degradation classes can be annotated with a fair or moderate level of reliability. Particularly the degradation classes which occur most frequently, such as "Speech Spectrum", "Speech Level", "Noise Steady-state", "Noise Level" and "Noise Dynamic" were annotated quite consistently by our four annotators, with kappa coefficients larger than 0.35. The degradation classes detected less frequently are also less reliable in their annotation; these include the classes "Speech Distortion", "Noise Impulsive", "Echo" and "Speech Information". From the limited experimental data available to us, it is difficult to decide whether the lower annotation reliability for these classes stems from the particularities of the database (in our case the SwissQual 503 database), or whether there is a general problem in identifying the related degradation causes from pure listening.

Overall, the results show that experts need a better explanation of the named degradations, best to be provided by exemplary listening material given to expert listeners together with the instructions. This may increase the annotation reliability as well, and should be considered in the future set-up of ITU-T Rec. P.TCA.

The relationship between the frequency of occurrence of particular classes of degradations (P.TCA) and corresponding MOS values is not a simple one. Only one degradation class ("Speech Spectrum") shoes a correlation higher than 0.30 with the subjective MOS scores. This indicates that technical causes, as annotated according to the P.TCA scheme, are not enough in judging whether a particular speech sample is of good or bad quality. However, the relationship becomes better explainable when perceptual dimensions, as they are chosen for P.AMD, are taken into account. There are commonly several P.TCA degradation classes which correlate with perceptual P.AMD dimensions. The results are mostly congruent for the subjective dimension scores and their instrumental counterparts, as they have been estimated with the DIAL model.

## 7. Conclusions and future work

The results of our comparative study show that the P.TCA annotation scheme is able to capture some of the technical causes of sub-optimum quality with acceptable annotation reliability. For some other causes, this evidence is still missing, either because of the limited test material which was available to us, or because there are principled problems in the annotation of those causes. This cannot be decided on the basis of the available data, and thus additional experiments are required to further substantiate the P.TCA annotation procedure.

The results also show that there is a need for all – P.TCA cause analysis, P.AMD perceptual dimension analysis, and overall MOS scores – as these three metrics are only partly related and thus contain complementary types of information. This statement refers to the "subjective" procedures (manual annotation for P.TCA, subjective dimension judgment as it has been proposed by Wältermann [2] for the P.AMD specification or is expected in the P.MULTI recommendation, and subjective MOS testing) as well as to their instrumental counterparts (instrumental identification of technical causes in P.TCA, estimation of perceptual dimensions in P.AMD, and MOS estimation as in P.OLQA).

## 8. References


[1] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality", Int. Telecomm. Union, Geneva, 1996.

[2] ITU-T Temporary Document TD 650rev1 (GEN/12), "Requirement Specifications for P.TCA (Technical Cause Analysis)". Source: Rapporteur Q.16/12 (L. Malfait), International Telecommunication Union, CH-Geneva, 2011.

[3] ITU-T Temporary Document TD 114rev2 (GEN/12), "Revised Status Report of Question 7/12". Source: Acting Rapporteurs of Question 7/12 (A. Sharpley, P. Usai), International telecommunication Union, CH-Geneva, March 2013.

[4] Wältermann, M., "Dimension-based Quality Modeling of Transmitted Speech", Springer, Berlin, 2013.

[5] Voiers, W.D., "Diagnostic Acceptability Measure for Speech Communication Systems", in: *Proc. ICASSP'77*, 204-207, Hartfort CT, 1977.

[6] Sen, D., “Determining the Dimensions of Speech Quality from PCA and MDS Analysis of the Diagnostic Acceptability Measure”. In: Proc. MESAQUIN 2001, CZ-Prague, 2001.
[7] Côté, N., “Integral and Diagnostic Intrusive Prediction of Speech Quality”, Springer, Berlin, 2011.
[8] ITU-T Rec. P.863, “Perceptual Objective Listening Quality Assessment (POLQA) — An Advanced Objective Perceptual Method for End-to-end Speech Quality Evaluation of Fixed, Mobile, and IP-based Networks and Speech Codecs Covering Narrowband, Wideband, and Super-wideband Signals”. International Telecommunication Union, CH-Geneva, 2011.
[9] ITU-T Contribution COM 12-342, “Results from a Multidimensional Rescaling Experiment of a P.OLQA SWB Test Database”. Source: Deutsche Telekom AG (M. Wältermann, S. Möller), International Telecommunication Union, CH-Geneva, 2012.
[10] ITU-T Contribution COM 12-40, “Validation of the P.TCA Annotation Methodology and Comparison to Perceptual Dimensions from P.AMD”. Source: Deutsche Telekom AG (F. Köster, S. Möller, F. Schiffner, J. Skowronek), International telecommunication Union, ITU-T SG12 Meeting, 19 – 28 Mar. 2013, CH-Geneva.
[11] ITU-T Temporary Document TD 686 (GEN/12), “Expert Listening for P.TCA”. Source: Rapporteur Q.16/12 (L. Malfait), International Telecommunication Union, CH-Geneva, 2011.
[12] Sachs, L. , Hedderich, J., „Angewandte Statistik“, Springer, Berlin, 2009.