Qual User Exp (2017) 2:4
https://doi.org/10.1007/s41233-017-0006-5

RESEARCH ARTICLE

Towards degradation decomposition for voice communication system assessment

Friedemann Köster[1] · Falk Schiffner[1] · Sebastian Möller[1] · Ludovic Malfait[2]

Received: 12 May 2016 / Published online: 30 March 2017

Abstract This article presents the current development of degradation decomposition tools for the assessment of voice communications. Overall quality scores, represented as *Mean Opinion Scores* (MOS) produced by subjective test methodologies such as ITU-T P.800 *Absolute Category Rating* (ACR), remains the most popular quality metric in the industry. While MOS is a great indicator to evaluate quality issues, it does not provide information on the cause of issues. To address this gap, work items are currently active within ITU-T to provide the industry with means to understand the cause of lower scores by perceptual or technical degradation decompositions. The goal is to produce objective models that enable automated degradation decomposition. The first step in such a development is the construction of databases for model training and validation. For this, in sum four experiments using a potential diagnostic test method discussed within ITU-T are conducted. In addition, two optional improvements for the test method are presented and discussed. The results of the experiments show that for standardization the analyzed test method still leaves room for validation and further improvements.

✉ Friedemann Köster
friedemann.koester@tu-berlin.de

Falk Schiffner
falk.schiffner@tu-berlin.de

Sebastian Möller
sebastian.moeller@tu-berlin.de

Ludovic Malfait
Ludovic.Malfait@dolby.com

[1] Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany

[2] Dolby Laboratories, Inc., San Francisco, USA

Keywords P.TCA · speech quality · Perceptual speech quality · Technical analysis

Motivation and introduction

It is critical for telecommunication systems providers to evaluate their services for costumer satisfaction, reputation and economical reasons. In this context, the quality of transmitted speech in vocal human-to-human telephony communication as perceived by the end-users, also referred to as the *Quality of Experience* (QoE), is the major parameter that is used to validate their services. QoE is defined as "the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the users personality and current state" (Qualinet 2013). Regarding telephony services, assessing and predicting QoE is one of the main challenge of current research. For the purpose of understanding QoE, experiments involving human participants are necessary. Traditionally, auditory experiments carried out in a laboratory context are valid and reliable means. For this, naïve test participants judge the overall quality of transmitted speech signals on standard rating scales (Vary et al. 1998). The most common procedures are based on *Absolute Category Rating* tasks (ITU-T Recommandation 1996) and result in a *Mean Opinion Score* (MOS), representing the average rating for an average person for each signal or processing condition. As discussed in Köster et al. (2014), these tests unfortunately provide little insight into the reasons of sub-optimal quality. As an example, two different impaired speech files-one degraded by, say, background noise and the other one by discontinuities-can be rated with the same (low)

MOS value. Consequently, the overall MOS score does not provide diagnostic information regarding the causes of sub-optimum quality.

In order to provide more diagnostic information, *Study Group 12* (SG 12) of the *International Telecommunication Union* (ITU-T) is currently working on two work items. The aim of these two work items is to define subjective evaluation methods and instrumental prediction models able to diagnose the quality of transmitted speech. 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.

For path (1), ITU-T SG12 has developed a methodology for performing expert annotations after listening to transmitted speech files. This methodology may be proposed as a future Recommendation P.TCA (P standing for ITU-T P Series: Telephone transmission quality, telephone installations, local line networks—TCA standing for *Technical Causes Analysis*), and its goal is to identify signal characteristics such as sub-optimum speech level, speech spectrum, noise level, or echo. For path (2), a subjective evaluation method based on semantic differential attributes has been applied and is foreseen for a future Recommendation P.AMD (standing for *Assessment of Multiple Dimensions*), and aims at identifying and assessing perceptual dimensions like *coloration* or *noisiness*.

While for P.AMD validated subjective test paradigms to quantify perceptual dimensions have been defined and serve as a baseline for proposed instrumental dimension estimators, the P.TCA test methodology so far has neither been validated nor analyzed on a common database.

In this article we will present the work and studies that have been conducted to deeper understand, analyze, and improve the proposed P.TCA methodology. For this, the two approaches P.AMD and P.TCA will be presented and a deeper insight into the methodologies will be provided in "Methods for diagnosing the quality of transmitted speech".

As there are obvious links between the technical causes and the perceptual dimensions, an initial experiment is carried out to analyze if the P.TCA degradation types can be classified reliably and to determine the relationships between technical causes, perceptual dimensions, and overall quality. For this purpose, a standardized database with subjective as well as instrumental MOS and P.AMD ratings is required. The database and its characteristics are presented in "Database". In "Analyzing technical causes and perceptual dimensions" the results of the annotation experiment as well as the analyses and the conclusions are shown.

One of the main findings of the initial experiment is that the experts demanded to facilitate the P.TCA annotation method in terms of easier usage and better description. Consequently, considerable effort has been conducted to simplify the procedure. More precisely, a set of exemplary listening material and a new user interface are proposed in "Exemplary listening material".

With these improvements it is argued that the P.TCA methodology is made accessible for a wider group of annotators. Thus, a follow-up annotation experiment was conducted, where naïve participants were asked to annotate the same data as in the initial experiment, following the P.TCA guidelines and having the introduced improvements at hand. The results are presented in "Diagnosing the quality of transmitted speech with naïve annotators". The outcome shows that the performed effort is not enough to lift naïve annotators on the level of experts.

As a final summarizing experiment, an additional expert annotation study with the provided improvements was incorporated. The work is presented in "Final expert annotation experiment". It was suspected that the output of the study shows that experts achieve a higher agreement using the P.TCA method together with the improvements. However, the results show that there is still room for improvement and that the P.TCA annotation method has to be validated on more data.

Final proposals for improvement and conclusions as well as an outlook towards future work are given in "Conclusions and Outlook".

Methods for diagnosing the quality of transmitted speech

As mentioned before, for the diagnostic quality assessment of transmitted speech SG 12 of ITU-T is currently working on two different approaches, called P.AMD and P.TCA. Both are intended to be able to extract diagnostic information using subjective test paradigms and methodologies to form a basis of instrumental measurements. In this section the two approaches are introduced in detail.

Assessment of multiple dimensions (P.AMD)

The scope of the current work item P.AMD is to predict perceptual dimensions of degradations relevant to the overall speech quality in *narrowband* (NB 300-3400 Hz),

wideband (WB 50-7000 Hz) and *super-wideband* (S-WB 50-14000 Hz) telecommunication scenarios. The current state is to develop a model that aims at providing more detailed information about the individual perceptual quality dimensions in addition to overall quality estimations provided by the currently inforce standard for *Perceptual Objective Listening Quality Assessment* (POLQA) (ITU-T Recommendation 2011).

Perceptual dimensions are explained as follows:

Using a telecommunication system, the listener will be confronted with a sound event, i.e. the acoustic speech wave. This sound event causes a perceptual event inside the listener that is of multidimensional nature and might be composed of several perceptual features (Jekosch 2005). These perceptual features can for instance be described with attributes like loudness or timbre. Geometrically, the perceptual event is a point in a multidimensional perceptual space. If the coordinate system of this perceptual space is Cartesian (its basis is orthogonal) and each of the auditory features lies along one of the orthogonal axes and thus are themselves orthogonal, these features are referred to as perceptual dimensions (Wältermann 2012).

Three methodologies have been used for the identification of the perceptual dimensions: (1) scaling perceptual differences of pairwise presented stimuli, and then mapping the perceptual distance to a multidimensional space [*Multidimensional Scaling* (MDS) (Borg and Groenen 2005)]; (2) rating all stimuli independently on a set of bipolar scales [*Semantic Differential* (SD) (Osgood 1957)] and reducing the space of judgments with the help of a factor analysis [*Principal Component Analysis* (PCA)]; or (3) the *Diagnostic Acceptability Measure* [DAM (Voiers 1977)] that is a special variety of the SD method using trained listeners. The application of these three methods enabled to identify two sets of dimensions: Set A using methodologies (1) and (2), and Set B using methodology (3).

Set A consists of four perceptual dimensions (Wältermann et al. 2010), namely:

- Noisiness The noisiness describes degradations such as, environmental background noise, circuit noise introduced by analogue transmission, or coding noise. It is judged on a scale labeled with the antonym pairs “not noisy” and “noisy”, which can be a proxy for “not hissing” and “hissing” for example.
- Discontinuity The discontinuity describes degradations related to isolated or non-stationary distortions. These distortions are mainly introduced by the loss of packets during a VoIP transmission and result in temporal clipping. It is judged on a scale labeled with the antonym pairs “continuous” and “discontinuous”, which can represent the terms “regular”/“steady”/“not chopped” and “irregular”/“shaky”/“chopped”, respectively.
- Coloration The coloration describes degradations resulting from frequency response distortions, e.g. bandwidth restrictions and coloration introduced by transducers. It is judged on a scale labeled with the antonym pairs “uncolored” and “colored”, which can be paraphrased with the terms “direct”/“close”/“thick” and “indirect”/“distant”/“thin”, respectively.
- Loudness The loudness plays an important role for the overall quality and the intelligibility. Extremely high or low loudness, especially in noisy environments, results in lower intelligibility. The loudness describes the impact of overall play-back level. It is judged on a scale labeled with the antonym pairs “optimum loudness level” and “non-optimum loudness level”.

For the subjective annotation, in Wältermann (2012) a procedure similar to what is currently recommended for noisy speech signals is proposed (ITU-T Recommendation 2003). Here, listeners are asked to rate the quality of the speech signal, of the background noise, and of the overall stimulus in a sequence, on three separate scales. Thus, for the subjective direct scaling each dimension is consecutively rated on a separate continuous scale. An example of the used scale for the dimension “Noisiness” can be seen in Fig. 1.

Set B was identified using methodology (3) and basically separates three of the dimensions in Set A into two sub-dimension each. Thus, set B consists of 7 perceptual dimensions grouped in three classes (ITU-T Recommendation 2014), namely:

- Quality of the speech signal
  - S-FLT (flutter) Slow-varying degradation in the speech signal, described as fluttering, babbling or discontinuous.
  - S-RUF (rough) Fast-varying degradation in the speech signal, described as rough, raspy or harsh.
  - S-LFC (low frequency coloration) Degradation of low-frequency coloration in the speech signal, described as dull or muffled.
  - S-HFC (high frequency coloration) Degradation in high-frequency coloration in the speech signal, described as small, distant or thin.

Fig. 1 Example of used scale for P.AMD Set A. “Noisiness” scale

- Quality of the background
  - B-LVL (level) Degradation due to the level of background noise, described as hissing, rushing or roaring.
  - B-VAR (variability) Degradation due to the variability of the background noise, described as bubbling, intermittent or variable.
- Loudness
  - Overall loudness in the speech signal + background noise.

Regarding the subjective annotation of the perceptual dimensions of Set B, for six of the seven dimensions, all except loudness, subjects use a magnitude estimation scale to indicate the amount of the particular perceptual quality that they judge to be present in the sample. The scales are presented simultaneously together with the overall quality scale. Figure 2 shows an example of one of the six-category rating scales used by subjects. The bottom category of the scale is labeled 0.0 (zero) to indicate that the specific perceptual quality is not detected in the sample. For loudness, a Comparison Category Rating (ITU-T Recommandation 1996) (CCR)-like scale is used, with the labeling: (1) Much quieter than preferred; (2) Quieter than preferred; (3) Preferred; (4) Louder than preferred; (5) Much louder than preferred. The subjective test methodology for Set B was validated by the ITU-T SG12 and is recommended as ITU-T Rec. P.806 (ITU-T Recommendation 2014).

Following these rating procedures, the goal of the P.AMD approach is now to develop a model that can predict the ratings of both sets to provide additional information besides the overall MOS score. For the perceptual dimensions of Set A, Scholz (2008) described initial estimators which capture the perceptual effects for the devices which could have produced them (e.g. filters for coloration). These estimators have been enhanced by Côté resulting in the *Diagnostic Instrumental Assessment of Listening quality* (DIAL) model (Côté 2011). This DIAL model is planned to cover parts of the prediction for Set A. For Set B, only algorithms for the perceptual dimensions S-FLU and S-RUF are published, yet. This model,

How would you describe amount of the quality presented in the sample?	
Overwhelming	5.0
Somewhat conspicuous	4.0
Very noticeable	3.0
Somewhat noticeable	2.0
Just detectable	1.0
Not detectable	0.0

Fig. 2 Magnitude estimation scale used for the 6 dimension ratings in Set B of P.AMD. Taken from ITU-T Recommendation (2014)

called the *Multidimensional Evaluation of Speech Quality* (MESQ) model (ITU-T Temporary Document 2009), predicts the ratings of the dimensions with the help of a non-linear cochlear model. A first version of a model combining both sets and first evaluation results are planned for end-2016.

Technical causes analysis (P.TCA)

The purpose of the P.TCA methodology is to perform a technical degradation decomposition in order to derive the technical causes, providing information for the telecommunication providers to remedy the issue. The underlying fundamental assumption is that most links between technical causes and perceptual impairments are biunique meaning that a given technical cause always leads to one specific perceptual impairment and a given perceptual impairment is always caused by one specific technical cause. However, this assumption has to be proven.

The presence of technical problems is usually assumed when either users complain or when an instrumental monitoring of quality (typically performed using P.OLQA) indicates an unexpectedly low score (for example when MOS <3.0) (ITU-T Temporary Document 2011a).

The P.TCA framework provides nine global categories of impairments (Level-1 degradations), which are further decomposed into 47 sub-classes (Level-2 degradations). The list was proposed by a group of telecommunication experts at ITU in 2011 with the aim of gathering an exhaustive list of audio degradations that can occur in widely used speech telecommunication networks. The list of impairments can be found in Table 1 and in ITU-T Temporary Document (2011a). Based on this list, expert listeners are asked to identify the most prominent degradations within each evaluated sample. Each sample can be listened to as many times as desired and no specific setup is recommended. The responses are given in two steps:

1. The experts identify the most dominant degradations based on the Level-1 categories and rate them according to whether they are highly dominant, dominant, or less dominant. In most cases, there is only one such degradation. A maximum of three Level-1 degradations can be reported.
2. The experts identify the detailed types of degradations from column "Level-2" for each of the entries of column "Level-1". It may also be reported that no suitable degradation types was found. Similar to the Level-1 ratings, experts rate each degradation with respect to its dominance. Usually, one or two Level-2 types are present in a given sample but experts may name more if they find it necessary.

A further expert rating will identify the most likely technical causes for the annotated degradations. The long-term

Table 1 List of telecommunications impairments according to the P.TCA guidelines (ITU-T Temporary Document 2011a)

Impaiment type Level-1 degradation	Name Level-2 degradation	Definition
Speech-Level	Loud speech	Speech cannot be adjusted to the preferred listening level and is too loud
	Quiet speech	Speech cannot be adjusted to the preferred listening level and is too quiet
	Loudness varies	Loudness of speech changes during call
	Speech level fluctuations	Level of speech sounds vary
	Temporal speech clipping	Words or parts of words missing
	Choppy speech	Frequent temporal speech clipping perceived as single impairment event. Sometimes sounds like person is speaking underwater
	Self clipping	*Temporal speech clipping* highly correlated with speech signal
	Speech cut-outs	Extended periods (>1 s) of missing speech
Speech-Spectrum	Timbre varies	Timbre of speech changes during call
	Muffled speech	Speech sounds unnaturally low-pitched. Also referred to as "bommy"
	Sharp speech	Speech sounds unnaturally high-pitched
	Colored speech	Timbre of speech sounds unnatural, but neither low-pitched or high-pitched
Speech-Distortion	Muddy speech	Speech always sounds unclear and spectrally smeared
	Warped speech	Short-duration (i.e., within a word) spectral and level fluctuations
	Buzzy speech	Speech has a harsh "zzz"-like sound to it
	Fuzzy speech	Speech has a "zzz"-like sound to it, but sound softer than *buzzy speech*
	Nasally speech	Speech sounds similar to someone talking while plugging their nose
	Hissy speech	Sibilant speech sounds such as "s" and "sh" more noticeable and seem exaggerated
	Rough speech	Distortion of speech signal that is described as "harsh" or "not smooth", and not covered by other impairments
Speech: Information	Poor intelligibility	Difficult to understand what is being said
	Poor speaker identification	The far end talker does not sound like himself
	Poor localization	Perceived location of talkers voice unclear or coming from undesirable location
Echo	Tunnel-sounding speech	Speech sounds reverberant, similar to someone talking inside a tunnel
	Listener Echo	Hear an echo of far end talker's voice while listening to them talk
Noise-Level	Line sounds dead	Connection is so quiet it sounds like the call has been dropped
	Loud noise	Noise is too loud
	Noise level fluctuations	Rapid changes in noise level
	Temporal noise clipping	Large drop in noise level for short periods of time (<1 s)
	Noise cut-outs	Large drop in noise level for extended periods of time (>1 s)
Noise-Steady-state	Hum	Low-pitched continuous tonal noise
	Buzz	Periodic noise with "zzz"-like sound to it
	Whine	High-pitched continuous tonal noise
	Pink noise	Low-pitched continuous random noise
	White noise	High-pitched continuous random noise
	Hiss	Very high-pitched continuous random noise that sounds similar to "s" or "sh"
	Motorboating	Low-pitched periodic "plop-plop" noise
Noise-Dynamic	Modulation noise	Random noise that is highly correlated with the speech signal
	Noise gating	Distinct change in noise level/characteristics between speech and non-speech segments of far end talker
	Musical tones	Tonal components that vary in pitch are heard intermittently in the background noise
	Distorted background noise	Background noise does not sound realistic
	Static	High-pitched intermittent random noise
	Crackling noise	White or pink noise with pops/clicks
	Wind buffeting	Constantly changing "rumbling" noise often heard with intermittent "snapping"
	GSM buzz	Intermittent *buzz* and *pops/clicks*

Table 1 continued

Impaiment type Level-1 degradation	Name Level-2 degradation	Definition
Noise-Impulsive	Pops	Low-pitched impulse noise
	Clicks	High-pitched impulse noise
	Pre-echo	Noise heard at the very beginning of speech onset-especially transient sounds

goal is developing an instrumental model that predicts the annotations provided by the experts following the P.TCA methodology.

The P.TCA schema is still at an early stage and exhibits potential for improvement. The test setup to be employed can be freely chosen by the experimenter. No strict standardized methods to gather the answers has been recommended and the definitions of the degradations (column 3 in Table 1) are not considered precise and unambiguous by all experts.

However, there have been no annotation experiments conducted that follow the guidelines of the P.TCA method so far. Thus, the validity as well as the annotation reliability has never been researched. In the following sections, initial activities analyzing the P.TCA methodology are presented.

Database

For our analysis, we selected database number 503 from the ITU-T Rec. P.863 (ITU-T Recommendation 2011) competition which has been kindly provided by SwissQual AG, Solothurn, Switzerland. This particular database includes diverse types of degradations and degradation combinations (54 conditions) 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 (NB, WB, and S-WB) and contains signal-correlated as 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 (ITU-T Contribution 2013). Furthermore, the database was selected, because subjective MOS ratings as well as ratings for P.AMD Set A are available.

Analyzing technical causes and perceptual dimensions

In an initial experiment the data introduced in "Database" was annotated by experts according to the preliminary P.TCA guidelines. In this section, 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. For this, the results of the annotation experiment are compared with subjective and instrumental judgments of the MOS and the Set A perceptual dimensions of P.AMD. Parts of the work illustrated in this section is based on the data presented in a former publication (Möller et al. 2013).

Technical cause annotation

The speech files were annotated by four experts who are 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 (ITU-T Temporary Document 2011a). 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 according to the P.TCA guidelines (ITU-T Temporary Document 2011b). In accordance with ITU-T Temporary Document (2011a), 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 2.

The Table shows the number of cases where experts have attributed a Level-1 degradation to a speech file of the corresponding condition. As there were four 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

Table 2 Expert experiment: frequency of Level-1 degradation annotations per condition

Cond.	MOS	Speech Level	Speech Spectrum	Speech Distortion	Speech Information	Echo	Noise Level	Noise Steady-state	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
$\sum$		212	345	83	9	16	115	174	128	44

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 C47, 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 C12 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 (Sachs and Hedderich 2009), see Table 3. The kappa coefficient indicates how strongly the annotations of the different annotators agree,

Table 3 Expert experiment: Kappa coefficients for Level-1 degradation classes

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

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 (Sachs and Hedderich 2009)

normalized by the per-chance agreement. 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 a slight agreement.

With respect to the Level-2 degradation classes, these are obviously less frequently annotated. About 15 Level-2 degradations 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.

Annotators feedback

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 became more complicated. For instance, one expert reported difficulties in classifying 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.

Relationships between technical causes and subjective ratings

Spearman rank order correlations (Bortz 2005) between annotation frequencies of the technical causes and subjective judgments described in ITU-T Contribution COM 12-342 (2012) (both in terms of MOS and perceptual dimensions according to P.AMD Set A) 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

Table 4 Spearman rank order correlation between frequencies of Level-1 degradation classes and subjective judgments

Degradation Class	MOS	Coloration	Discontinuity	Loudness	Noisiness
Speech-Level	−0.275	0.242	−0.124	**−0.354**	**0.388**
Speech-Spectrum	**0.470**	**−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**

Correlations with an absolute value higher than 0.35 are printed boldface

"Speech-Spectrum". To a lesser extent, also "Noise-Steady-state" 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.

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 Set A. Two such models were available to us: POLQA (ITU-T Recommendation 2011) and DIAL (Côté 2011). Whereas both models provide an estimation of the overall MOS in a SWB 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.

Discussion

The annotations in this experiment have been performed by four "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. 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

Table 5 Spearman correlations between Level-1 degradation classes and instrumental quality predictions

	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

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

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 annotation analysis and the annotators feedback 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.

Overall, the results of the initial P.TCA experiment study show that (I) the P.TCA annotation scheme is able to capture some of the numerous technical causes of sub-optimum quality with acceptable annotation reliability and (II) that there is a need to assess all—P.TCA cause analysis, P.AMD perceptual dimension analysis, and overall MOS scores—to fully investigate the quality of transmitted speech, as these three metrics are only partly related and thus contain complementary types of information.

Improvements of the P.TCA methodology

As mentioned in "Motivation and introduction" and revealed in the initial annotation experiment, the P.TCA methodology still is in an early stage of development and there is room for improvement. Participating experts reported that it was hard to hold on to the top-down approach of the methodology. They also reported that the descriptions of the impairments are ambiguous. Therefore, it was decided to improve the P.TCA methodology by providing exemplary listening material and a novel user interface. In this section, the effort to make the P.TCA method easier to use is presented in detail.

Exemplary listening material

The participating experts of the initial experiment (see "Analyzing technical causes and perceptual dimensions") reported that the descriptions of the impairments covered in the P.TCA schema are in some cases hard to interpret and they can be ambiguous. According to the experts opinion, the descriptions do provide an idea of the impairments but most experts still asked themselves: **What does this really sound like?** They argued that exemplary listening material for the different impairments can reduce ambiguity and make the procedure easier and accessible for a wider field of users as well as raise the level of agreement amongst individuals. Thus, we decided to provide a MATLAB Toolbox allowing for the creation of all 47 degradations covered by the P.TCA schema. Furthermore, we validated the processed signals by means of a P.TCA-like annotation experiment. Parts of the work illustrated in this section is based on the data presented in a former publication (Köster et al. 2015).

MATLAB toolbox

Since examples for common degradations of transmitted speech signals are generally not available, the authors created a MATLAB Toolbox that allows imposing the 47 different degradations covered in the P.TCA method onto given clean input signals. The Toolbox can be obtained from P.TCA MatLab Toolbox (2016). Thirty-one MATLAB functions were created and a function for adding noise and a voice activity detection (VAD) function from external sources (Matlab-file exchange 2015; Deller et al. 2000) were included.

Additionally, the Toolbox also provides a number of audio files that contain 20 different types of foreground and background noises that can be added to the speech signals. Basically any speech file (48 kHz, Wav-format) can be used as input. The only current restriction is that it is recommended to use speech signals with a duration between 8 and 12 s. The restriction originates in the recommendation for the duration of speech signals for subjective and objective tests given by ITU-T (ITU-T Recommandation 1996; ITU-T Recommendation 2011).

Development approach

The proposed MATLAB toolbox applies signal processing in order to obtain signals that mimic the definition of each Level-2 impairment, respectively, as provided by the P.TCA guidelines (ITU-T Temporary Document 2011a). In other words, we process the files such that impairments are imposed onto a given input signal that satisfy our understanding of the impairment definitions. Additionally, we used the annotations and information from the database 503 ("Database") as orientation. The processing stages were developed in an iterative approach in which we created the corresponding speech material, evaluated it by expert listening, and modified the processing if necessary.

We succeeded to propose processing for all of the 47 target degradations. To give an idea of the processing, Table 6 shows the Level-2 degradations, their definitions given in the P.TCA document, and the corresponding methods to process the exemplary listing material for the Level-1 category "Speech-level". A detailed description of the processing steps can be found in Köster et al. (2015) or P.TCA MatLab Toolbox (2016).

Validation experiment

To validate the processed exemplary listening material, a P.TCA-like experiment was conducted. Expert listeners annotated the processed files according to the P.TCA guidelines. None of the authors of this article served as expert.

An example database was created with the MATLAB Toolbox presented above. For this purpose, the female reference stimulus from the database number 503 (see "Database") was used. This reference signal was processed with the *ptcaexamples.m* MATLAB script, resulting in 47 example stimuli. The example signals were annotated by five experts with terms from the list of the nine Level-1 and 47 Level-2 degradations. Again, the experts were all working with speech and audio processing on a daily basis and have been particularly trained for the given task through the annotation manual (refer to ITU-T Temporary Document 2011a, b). The experts listened to the database in several sessions in a quiet office room, each using their own setup (computer, sound card, and headphones were free to choose). It was possible to replay each sample as many times as desired. In line with the P.TCA method (ITU-T Temporary Document 2011b), the annotations were performed without knowledge on the annotations created by the other experts.

Note that the annotation task was performed slightly differently compared to the actual P.TCA annotation procedure given in ITU-T Temporary Document (2011b) and described in "Methods for diagnosing the quality of transmitted speech". In the present experiment, the focus was on the identification of the Level-2 degradations as the processing of the stimuli was developed according to the definitions of the different Level-2 impairments. Therefore, the first step of the actual P.TCA annotation procedure – the identification of the Level-1 category—was skipped. The experts were asked to annotate solely based on a list of

Table 6 The Level-1 category "Speech-level" with its corresponding impairments (Level-2) according to ITU-T Temporary Document (2011a) as well as their processing documentation (VAD: Voice Activity Detection)

Level-1	Level-2	Definition	Processing documentation
Speech-Level	Loud speech	Speech cannot be adjusted to the preferred listening level and is too loud	Signal multiplied by 4.5
	Quiet speech	Speech cannot be adjusted to the preferred listening level and is too quiet	Signal multiplied by 0.05
	Loudness varies	Loudness of speech changes during call.	Fade out of the signal to 0.05 of the original amplitude
	Speech level fluctuations	Level of speech sounds varies.	Signal multiplications: changes from 2.5 to 0.5 and back
	Temporal speech clipping	Words or parts of words missing.	VAD, cut random active segments
	Choppy speech	Frequent temporal speech clipping perceived as single impairment event	VAD, cut one seperate active segment
	Self clipping	Temporal speech clipping highly correlated with speech signal	VAD, cut every active segment
	Speech cut-outs	Extended periods (>1 s)	VAD, active parts of missing speech >1 s get 1 s cut

all Level-2 degradations. It was possible to pick more than one Level-2 degradation for a given stimulus.

Results

The results of the annotation experiment are presented in Table 7. As can be seen, some example files were annotated more reliably than others. Perfect results were obtained, for example, for the annotation of the degradation "Listener echo". Five out of five experts identified the example signal correctly. In contrast, no expert annotated the degradation "Buzz" correctly. However, four out of five experts responded correctly with respect to the Level-1 category, meaning that they identified correctly that they were dealing with noise of some sort. To clarify, the annotators were not asked for a Level-1 identification. However, their Level-2 identification (and the Level-1 degradation the annotations are connected to) allows to categorize whether the annotators are right in a Level-1 sense. Potential causes for incorrect annotations can be:

1. Inappropriate processing (the result does not sound as intended),
2. the degradation is not prominent enough, or
3. the description of the degradation is ambiguous.

Remarkably, the results for some stimuli were inverted (e.g. "Loudness varies" and "Speech-level fluctuation"). On average, the annotators identified 55% on Level-2 and 72% on Level-1 correctly.

Further insight is gained by analyzing the results using the *precision* and *recall* values that were proposed in Powers (2011). There are 47 stimuli presented to 5 subjects resulting in a total of $5 * 47 = 235$ data points. For each Level-2 degradation, 5 stimuli can be annotated correctly while 230 annotation options are false. In the case of e.g. "Temporal speech clipping", a total of 7 annotations were given of which 3 were correct (True Positive *tp*), 4 were not correct (False Positive *fp*), 2 true stimuli were annotated incorrectly (False Negative *fn*), and 226 were correctly not respected (True Negative *tn*). This results in a confusion matrix for "Temporal speech clipping" as shown in Table 8.

Now, precision and recall are calculated according to Powers (2011):

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

and

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

Precision (also termed *confidence*) represents the probability that a given annotation is a correct annotation. A precision of 100% is therefore the best possible result. Recall (also termed *sensitivity*) represents the probability that a correct annotation belongs to the bulk of possible correct annotations. A recall value that tends to 100% therefore represents a situation in which all correct responses are given. A recall value that tends to 0 represents a situation in which only very few options in a large pool of correct responses are picked. The corresponding values for each example file are also indicated in Table 7.

For the Level-2 degradation "Temporal speech clipping", the calculations result in $Precision = \frac{3}{3+4} = 0.43$ and $Recall = \frac{3}{3+2} = 0.6$. The values can generally range between 0 and 1 and the figures stated above are in a medium area for this Level-2 degradation. We explain the fact that the values do not represent perfect performance with the occurrence of confusion between the Level-2 degradation "Temporal speech clipping", "Choppy speech", and "Self clipping". All these degradations describe temporal clipping artifacts whereby differences are rather small. Further confusion is obtained by the definition of the degradation "Self clipping", that directly refers to "Temporal speech clipping" (cf. Table 1). As can be seen in Table 7, the three degradations have a low correctness rate on Level-2 but a perfect rate on Level-1. This represents the fact that the experts inverted their annotations.

A possible solution is to process the impairments with more prominent characteristics in future versions of the example material. Or, it might be proposed to adopt the P.TCA methodology in terms of merging the three Level-2 degradation to one degradation avoiding confusion and making the method easier.

Overall, it can be seen (Table 7) that for both precision and recall, 28 stimuli exhibit values above or equal to 0.5. This means that for all 47 stimuli, approximately 60% were annotated correctly with a probability of 50% or higher. There are also a few examples ("Rough speech", "Poor speaker identification", "Buzz" noise, and "Musical tones") that were not annotated correctly at all (precision and recall vanish). It has to be investigated if this is due to the definition of the degradations or whether our interpretation of the definition is not appropriate.

In sum, the results of the validation experiment show that the processed signals exhibit the intended degradations in most cases. More precisely, the processed Level-2 degradations were mostly recognized correctly by the annotators. The presented improvement of the P.TCA method provides a validated set of algorithms for imposing the impairments described in the P.TCA guidelines. The work on the toolbox will be continued to further improve the validity of the generated stimuli.

Table 7 Results of the exemplary listening material annotation experiment; CL1 and CL2 are the number of correct annotations on Level-1 and Level-2, respectively

Impairment type (Level-1)	Detailed description (Level-2)	ID of example	ID of example as annotated by subjects					CL2	CL1	Precision	Recall
			P1	P2	P3	P4	P5				
Speech-Level	Loud speech	37	37	37		37	37	4	4	1	0.8
	Quiet speech	29	29	29	29	29	29	5	5	1	1
	Loudness varies	17	46	46	17, 46	17	17	3	5	0.5	0.6
	Speech level fluctuations	46	17	17	28	46	46	2	5	0.4	0.4
	Temporal speech clipping	4	9	9	4, 8, 9	4	4	3	5	0.43	0.6
	Choppy speech	9	8	4	10		9, 13	1	5	0.2	0.2
	Self clipping	8	4	8		8,9	8	3	5	0.6	0.6
	Speech cut-outs	32	32	32	32	32	32	5	5	1	1
Speech-Spectrum	Timbre varies	2	2	47, 28		28	2, 14, 28, 47	2	2	0.25	0.4
	Muffled speech	22	22	22	22	22	22	5	5	1	1
	Sharp speech	11	11	14	11, 30	11	11, 30	4	4	0.57	0.8
	Colored speech	30	10	44	47	30, 35	10	1	3	0.17	0.2
Speech-Distortion	Muddy speech	10	47			44	44	0	2	0	0
	Warped speech	28	28	2	2, 14, 35	2	35	1	1	0.5	0.2
	Buzzy speech	7		11		7	38	1	1	0.33	0.2
	Fuzzy speech	40	30	10	38	40	5	1	1	0.2	0.2
	Nasal speech	44	44	30	44	14	44	3	4	0.6	0.6
	Hissy speech	38	38			38	36	2	4	0.66	0.4
	Rough speech	35	14		37	10	38	0	2	0	0
Speech-Information	Poor intelligibility	47	39	32		47	4	1	1	0.25	0.2
	Poor speaker identification	14	35					0	0	0	0
	Poor localization	16	16	16	16	16	16	5	5	1	1
Echo	Tunnel-like sounding speech	24	24		24	24	24	4	4	1	0.8
	Listener Echo	6	6	6	6	6	6	5	5	1	1
Noise-Level	Line sounds dead	36				36	36	2	2	1	0.4
	Loud noise	39	27	39	1, 39	39	39	4	4	0.66	0.8
	Noise level fluctuations	3	3	3	34	3, 34	3	4	5	0.8	0.8
	Temporal noise clipping	31	31	34	3	31	31	3	3	0.6	0.6
	Noise cut-outs	19	34	19	19		19	3	3	0.75	0.6
Noise-Steady-state	Hum	43	5, 43	43	5, 43	5, 43	43	5	5	0.625	1
	Buzz	18	21	21		33	40, 34	0	4	0	0
	Whine	20	20	20	21	20	18	3	3	0.6	0.6
	Pink noise	41	41	41	41	18	41	4	4	0.8	0.8
	White noise	27	1	1, 27	1, 11, 26, 27	27, 41	27	4	4	0.4	0.8
	Hiss	1	18	18	18	1	1	2	5	0.4	0.4
	Motorboating	5		5			12	1	4	0.4	0.2

Table 7 continued

Impairment type (Level-1)	Detailed description (Level-2)	ID of example	ID of example as annotated by subjects					CL2	CL1	Precision	Recall
			P1	P2	P3	P4	P5				
Noise-Dynamic	Modulation noise	23	23	35	23, 36	23	26, 23	4	4	0.57	0.8
	Noise gating	26	26	36	25, 26	26	33	3	4	0.5	0.6
	Musical tones	33	7	40, 7	7, 40		7	0	4	0	0
	Distorted background noise	21	33	33	33	21	20, 21	2	2	0.33	0.4
	Static	34	19		31	19	19	0	0	0	0
	Crackling noise	42	42	42	42	42	42	5	5	1	1
	Wind buffeting	13	13	13	13	13		4	4	1	0.8
	GSM buzz	45	45	12, 15, 45	45	45	45	5	5	0.71	1
Noise-Impulsive	Pops	12	12	25	12	12	12, 13	4	4	0.66	0.8
	Clicks	15	15	38, 31	15	15	15	4	4	0.8	0.8
	Pre-echo	25	25			25	25	3	4	1	0.6

Table 8 Confusion matrix for "Temporal speech clipping"

	Correct	Not correct
Annotated	$tp = 3$	$fp = 4$
Not annotated	$fn = 2$	$tn = 226$

Discussion

In the (detailed) Level-2 context, the experts were only able to identify slightly more than half of the material reliably. It has to be investigated if these results are due to inappropriate processing or due to a general problem in the annotation procedure. An argument supporting the former is the observation that a significant number of inverted annotations occurred which suggests that the imposed degradations were not prominent enough. An argument supporting the latter explanation is the fact that results that are similar to the presented ones were obtained in the inital study in "Analyzing technical causes and perceptual dimensions". A reduction of the number of degradation definitions, for example by merging degradations with similar or ambiguous definitions, should be considered. Nevertheless, the conducted validation experiment shows that the creation of example signals with a certain annotation robustness is possible. Keeping the current set of degradations allows for obtaining more detailed information on a given signal under test than with a reduced set which should support participating experts in future P.TCA annotation experiments. However, in future versions of the exemplary listening material an identification rate close to 100% should be archived.

User interface

Currently, in annotation experiments following the P.TCA guidelines, experts are asked to use an Excel-Sheet template for the annotation process. After listening to a speech sample, the annotators have to switch to the Excel-file, and then identify in two columns the corresponding degradations; they actually have to type in the names of the degradations. As mentioned before ("Analyzing technical causes and perceptual dimensions"), it was reported, that this procedure is complicated, even for expert annotators. Additionally, this procedure violates the top-down approach of the methodology since experts tend to use the Level-2 degradations to identify the Level-1 impairments. To make the annotation procedure more usable and easier, for experts and and possibly also for naïve listeners, a Graphical-User-Interface (GUI) was developed that sums all steps of the annotation process following the guidelines of ITU-T Temporary Document (2011b). A screenshot of the GUI is given in Fig. 3.

It can be seen in the screenshot presented in Fig. 3 that the created GUI is divided into three sections:

– Section 1 (upper left): **Audio evaluation** Here, the participants can listen to the samples which they are asked to annotate. For each condition four (or how much are desired) samples and four corresponding reference files can be listened to (male and female). The green (already listened to) and red (still needs to be listened to) marks indicate whether an expert already listened to a file or not. Note, experts have to listen to all files before they can continue with the actual

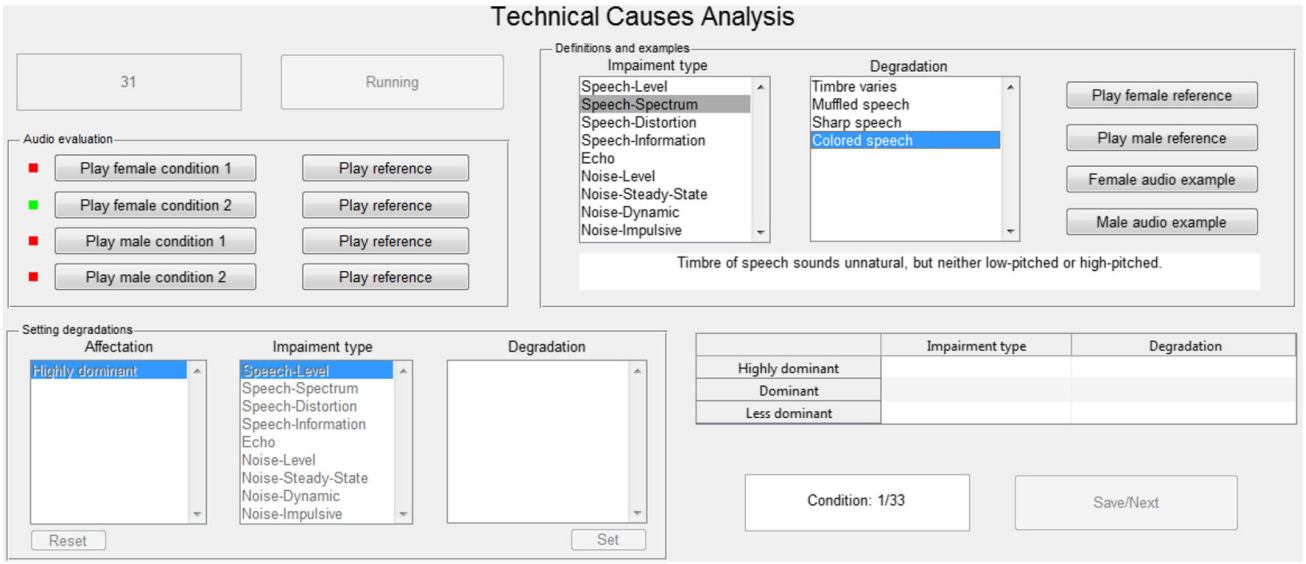

Fig. 3 Screenshot of the graphical-user-interface (GUI)

annotation. However, each file can be listened to as often as desired as long as all files are listened to.

– Section 2 (upper right): **Definitions and Examples** Here, the participants can read the definition of each degradation and can listen to the corresponding created exemplary listening material presented in "Technical cause annotation". On the left the participants can choose one of the nine Level-1 impairment types. In the middle block the corresponding Level-2 degradation appear according to the selection. Depended on the selected Level-2 degradation the definition from Table 1 appears in the window below and on the right the participants can listen to one relevant male and one female example file (and the clean reference speech files). The participant can listen to each example file as often as desired. This section is supposed give the participants on overview (in terms of definitions and examples) of the degradation that can be selected and might act as an replacement for Table 1 ("Technical causes analysis (P.TCA)") in future experiments.
– Section 3 (bottom): **Technical Causes Analysis** In the part "Setting degradation", the participants are asked to conduct the actual annotation procedure. First, the experts are asked to identify the most prominent Level-1 degradation ("highly - dominant"). After selecting a Level-1 degradation ("Impairment type"), the corresponding Level-2 degradations appear on the right, and have to be selected by the participant ("Degradation"). By clicking the "Set" button the selection is locked and appears in the table in the lower right. This procedure can then be repeated with the "dominant" and "less-dominant" impairments that appear after a "highly-dominant" degradation is selected. The Top-Down approach is reflected with this procedure. Clicking the "Save" button saves the annotations and the participant can continue with the next condition.

The usage of the developed GUI is in line with the guidelines of ITU-T Temporary Document (2011b). However, there are three appreciable particularities: (I) Following the guidelines, the annotators are first asked to identify all Level-1 degradations; in the GUI this step is done consecutively. (II) Using the presented GUI, the listeners annotate on a condition basis. In the first experiment ("Analyzing technical causes and perceptual dimensions") the annotation were done on a per-sample basis. (III) As in the guidelines, the annotators do not have to identify a corresponding Level-2 degradation. It is enough to just select a Level-1 impairment type.

Conclusion

In this section we presented two profound improvements for the P.TCA methodology. The improvements were developed with respect to the feedback given by the experts participated in the initial P.TCA annotation experiment presented in "Analyzing technical causes and perceptual dimensions". Using the exemplary listening material in combination with the GUI, the authors argue that the P.TCA annotation process enhances in terms of usage and understanding, making the method accessible for a wide field of users. However, this still has to be proven.

Diagnosing the quality of transmitted speech with naïve annotators

Taking the conclusion of the initial P.TCA annotation experiment ("Analyzing technical causes and perceptual dimensions") into account, a set of exemplary listening material and a GUI was introduced in "Improvements of the P.TCA methodology". It is argued that the improvements would help the P.TCA method to be easier to use and understand. It was decided to conduct an annotation experiment following the P.TCA guidelines and using the introduced improvements with naïve listeners to prove this assumption. The underlying idea of this approach is the following one: if the introduced improvements show to provide a higher annotation reliability with naïve listeners, it would be obvious that the improvements also raise the annotation reliability of experts. In addition, the approach shows whether or whether not the P.TCA method could be used with naïve listeners. To compare the results of the experiment with the expert annotations the same data was used ("Database"). In this section, the experiment, its results, and the comparison with the expert annotations are presented in detail.

Test design

The data was annotated by 41 (15 f, 26 m) naïve listeners, aged between 17 and 50 (Mean: 25.1; SD: 5.16), in Telekom Innovation Labs, TU Berlin. These naïve annotators have not been particularly trained for the given task. As an introduction they were asked to read the annotation manual as it is presented in the P.TCA guidelines. The naïve annotators listened to the database in one session (one hour) in a sound-proofed booth respecting the listening environment requirements given in ITU-T Recommandation (1996). Due to the limited time the listeners were asked to listen only to two of the four files per condition. Also, according to the presented GUI, the naïve listeners annotated per-condition and not per-sample as the experts did ("Analyzing technical causes and perceptual dimensions"). A *Realtek High Definition Audio ALC 268* soundcard and a *AKG K601 reference headphone* at a comfortable listening level was used for the diotic sound presentation. The naïve listeners annotated the speech material independent from each other and without knowing the annotations of the other participants. Before the annotation process, it was recommended but not required, to listen to all the exemplary listening material and read all the descriptions of the degradations. After a introduction into the GUI and the process of the P.TCA methodology the naïve listeners followed the guidelines given in ITU-T Temporary Document (2011a) and annotated the speech data.

Results

Table 9 shows the numbers of cases where naïve annotators have attributed a Level-1 degradation to a speech file of the corresponding condition. As there were 41 annotators, a maximum of 41 annotations per condition and class could occur. The table shows, the number of the condition, the subjective MOS value and the annotations for the nine Level-1 degradations.

As can be seen from Table 9, there are some Level-1 classes which were annotated more frequently than others. Most labels were given to the "Speech Spectrum" and "Speech Level" classes. "Speech Information", "Echo", and "Noise impulsive" were the classes less frequently used. Again, this result may either be linked to the particularities of the database used (i.e. that the corresponding degradations were rare in that database), or it may be linked to problems in identifying particular classes of degradations from pure listening ("Speech-Information"). It may also be linked to the fact, that naïve listeners could use some classes better than others since they don't really understand what these classes describe in particular. That results in using rather the classes they understand (e.g. "Speech-Level").

The reliability of the annotation process was again analyzed with the help of the kappa coefficient, which indicates how strongly the annotations of the different naïve annotators agree, normalized by the per-chance agreement, see Table 10. It can be seen that moderate agreement was obtained for only one Level-1 degradation class, namely "Speech level". "Speech Spectrum", "Noise-steady-state" and "Noise-level" reached a slight to fair agreement. The other classes only reached a poor agreement.

Naïve feedback

In sum, almost all naïve annotators reported that the amount of new information is too much. It is very hard to learn the 47 Level-2 degradations grouped into 9 Level-1 degradations in a short period of time. Even with the help of the exemplary listening material and the instructions the annotators reported that they only had a limited overview of the degradations they could use. Some annotators asked to have a better training before the annotation process, to get a better "feeling" for the degradations.

Table 9 Naïve listener experiment: frequency of Level-1 degradation annotations per condition

Cond.	MOS	Speech Level	Speech Spectrum	Speech Distortion	Speech Information	Echo	Noise Level	Noise Steady-state	Noise Dynamic	Noise Impulsive
C02	1.19	4	1	6	2	0	5	5	26	2
C03	2.88	5	3	10	0	1	9	3	18	5
C04	2.46	2	1	0	1	0	15	32	2	0
C09	2.97	7	30	16	4	0	0	0	0	0
C12	1.11	40	1	0	4	0	0	1	0	2
C13	2.45	2	8	0	5	0	29	3	6	0
C14	2.55	9	5	3	5	0	26	4	11	0
C17	2.42	9	12	9	1	1	4	28	7	0
C18	2.45	37	11	6	3	0	1	0	0	0
C19	2.54	7	24	9	3	1	1	18	2	1
C26	2.58	17	22	9	3	1	2	5	6	0
C27	2.48	6	27	8	1	1	7	13	8	0
C28	2.08	6	16	5	3	0	23	6	13	0
C29	1.77	41	3	1	1	0	11	8	9	1
C30	1.34	29	4	8	2	1	6	6	13	1
C32	2.64	5	1	5	7	1	29	2	4	0
C35	2.43	20	25	4	2	1	5	8	4	1
C36	2.14	33	8	6	0	9	2	4	1	2
C37	2.29	10	27	9	3	0	3	14	2	1
C38	2.80	9	28	10	2	2	1	0	4	0
C39	1.90	3	25	10	3	0	3	2	25	1
C40	2.16	5	34	7	2	0	0	3	7	1
C41	2.89	10	35	8	3	1	1	7	3	1
C42	2.77	5	33	5	3	0	2	21	2	0
C43	1.30	38	8	3	8	0	4	8	0	0
C44	2.48	11	20	11	2	0	3	14	3	4
C45	2.23	23	20	9	0	1	5	2	7	3
C47	1.86	32	4	1	6	0	19	2	5	7
C50	2.60	40	20	5	0	0	1	0	1	2
C51	2.83	14	25	11	1	0	1	5	7	7
C52	2.78	32	17	6	3	0	0	2	6	8
C53	2.80	6	25	9	2	1	0	13	9	1
C54	3.00	34	21	9	2	0	0	4	1	0
$\sum$		551	544	218	87	22	218	243	212	51

Experts vs. Naïve listeners

The results of the present experiment are compared to the expert annotation experiment presented in "Analyzing technical causes and perceptual dimensions" with respect to the frequency and the kappa coefficient. In Table 10 the kappa coefficients and the annotation frequency for the Level-1 impairments of the naïve experiment can be seen. In contrast, the results of the calculated kappa coefficients for the expert and the naïve experiment are illustrated in Fig. 4 (Tables 3, 10).

The tables and the figure show, that the results obtained by the experts have a higher agreement than the results conducted by the naïve annotators. The only kappa value that is basically equal is the value for the Level-1 degradation "Speech Level". This can also be seen by looking at the annotation frequency of both experiments in Tables 2 and 9. Here multiple Conditions (C12, C18, C29, C43,

Table 10 Kappa coefficients for Level-1 degradation classes (Naïve listener P.TCA experiment)

Degradation class	Frequency	Kappa
Speech-Spectrum	544	0.277
Speech-Level	551	0.430
Noise-Steady-state	243	0.226
Noise-Level	218	0.313
Noise-Dynamic	212	0.165
Speech-Distortion	218	0.036
Noise-Impulsive	51	0.056
Echo	22	0.069
Speech-Information	87	0.012

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 (Sachs and Hedderich 2009)

C54) have a high annotation frequency (for "Speech Level") on the expert and the naïve listener side. Again, this can be explained by the fact that degradations related to "Speech Level" are probably easy to understand, also for naïve listeners.

The kappa coefficients for the other Level-1 degradations show higher values for the expert listeners than the naïve listener. However, for the Level-1 degradation "Speech-spectrum" the kappa coefficient for naïve annotators is low, but looking at the annotation frequency (Tables 2, 9) a few conditions (C40, C41, C42) with almost equal annotation frequency can be found. This shows that for conditions with distinctive degradations also naïve listeners reach a high agreement. This, however, is not enough to yield to a high overall kappa value.

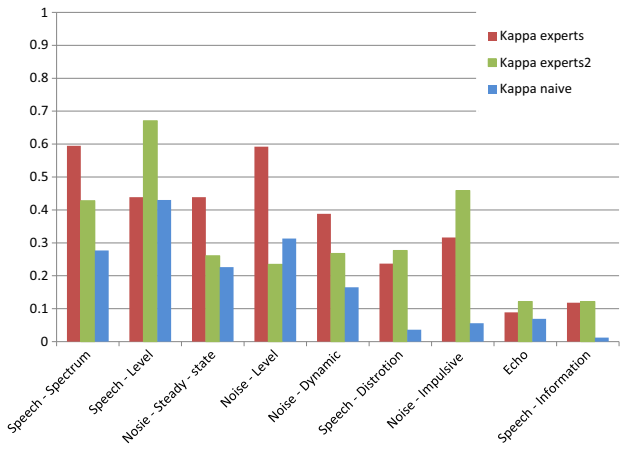

Fig. 4 Comparison of the kappa coefficient for the naïve and the two expert experiments. experts—initial experiment; experts2—final experiment

Conclusion

The outcome of the naïve experiment shows, that experts achieve a higher annotation agreement while some conditions can also be annotated by naïve listeners with a high frequency. For example, for some Level-1 degradations ("Speech-level") the agreement and for certain explicit degradations the annotation frequency of experts and naïve listeners is almost equal. The different results may be due to the fact, that (1) the procedure was developed for experts, that (2) the amount of information is too much for naïve listeners, and that (3) naïve listeners can only identify certain distinctive degradations. Therefore it can be claimed that the proposed improvements are not lifting naïve listeners to the level of experts.

Final expert annotation experiment

The results of the P.TCA annotation experiment conducted with naïve listeners ("Diagnosing the quality of transmitted speech with naïve annotators") showed that the performed effort to simplify the P.TCA methodology was not enough to lift naïve listeners to the level of experts. However, the authors still argue that the introduced improvements help annotators (experts and naïve listeners) to better understand the P.TCA scheme and that the annotation process is now more practical. Finally, the assumption that the improvements also help experts to achieve a higher agreement in their annotations has to be proven. Thus, it was decided to conduct a further summarizing subsequent annotation experiment with the proposed improvements and experts. Again, the same data ("Database") as in the two previous experiments was used to compare the results. In this section, the experiment, its results, and the comparison between the three conducted experiments are presented in detail.

Test design

The data was annotated by six experts in Telekom Innovation Labs, TU Berlin (these experts were different from the four expert of the experiment presented in "Analyzing technical causes and perceptual dimensions"). The experts are dealing with speech and audio processing as part of their work (either PhD students or Master students). Basically, the experts conducted the same experiment as the naïve listeners. Thus, the experts used the introduced GUI, they listened to the same data in one session (again one hour) in the same sound-proof booth using the same hardware ("Diagnosing the quality of transmitted speech with naïve annotators"). Similar to the naïve listeners experiment, the experts annotated per-condition. The

experts annotated the speech material according to ITU-T Temporary Document (2011a) independent from each other and without knowing the annotations of the other experts.

Results

Table 11 shows the numbers of cases where experts have attributed a Level-1 degradation to a speech file of the corresponding condition. As there were 6 annotators, a maximum of 6 annotations per condition and class could occur. The table shows the condition number, the subjective MOS value and the expert annotations for the nine Level-1 degradations.

Table 11 shows that there are some Level-1 classes which were annotated more frequently than others. Just like in the other two experiments, most labels were given to the "Speech Spectrum" and "Speech Level" classes while "Speech Information", "Echo", and "Noise impulsive" were less frequently used. Still, possible explanations for this recurring result may be the particularities of the database or general problems in the pure listening procedure ("Diagnosing the quality of transmitted speech with naïve annotators").

Table 11 Expert listener experiment 2: frequency of Level-1 degradation annotations per condition

Cond.	MOS	Speech level	Speech spectrum	Speech distortion	Speech information	Echo	Noise level	Noise steady-state	Noise dynamic	Noise impulsive
C02	1.19	0	0	3	0	0	0	0	3	0
C03	2.88	0	0	2	0	0	1	0	3	0
C04	2.46	1	0	0	0	0	1	3	2	0
C09	2.97	1	4	3	0	0	0	0	0	0
C12	1.11	6	1	0	0	0	0	0	0	0
C13	2.45	1	1	1	0	0	4	0	1	0
C14	2.55	2	2	0	1	0	4	1	0	0
C17	2.42	0	3	4	0	0	1	3	1	0
C18	2.45	6	2	0	0	0	0	0	0	0
C19	2.54	0	5	1	0	0	0	3	0	0
C26	2.58	6	4	2	0	0	1	0	0	0
C27	2.48	0	4	4	0	0	1	0	3	0
C28	2.08	0	3	1	0	0	4	1	1	0
C29	1.77	6	0	0	0	0	3	0	0	0
C30	1.34	4	2	1	0	0	1	1	3	2
C32	2.64	5	0	1	0	0	3	0	0	0
C35	2.43	5	5	0	0	0	3	2	0	0
C36	2.14	6	1	2	0	1	0	2	0	0
C37	2.29	1	5	3	0	0	0	2	0	0
C38	2.80	1	5	2	0	1	0	0	0	0
C39	1.90	0	2	6	0	0	1	1	1	0
C40	2.16	1	5	2	0	0	1	0	1	0
C41	2.89	0	5	1	0	0	1	1	1	0
C42	2.77	0	6	0	0	0	1	3	0	1
C43	1.30	6	0	1	0	0	1	2	0	0
C44	2.48	0	5	1	0	0	1	4	0	0
C45	2.23	3	4	2	0	0	2	0	0	0
C47	1.86	5	0	1	0	0	2	0	1	4
C50	2.60	6	1	0	0	0	0	0	0	0
C51	2.83	2	4	1	0	0	1	2	0	0
C52	2.78	5	4	0	1	0	0	0	0	0
C53	2.80	0	6	0	0	0	0	0	0	0
C54	3.00	4	2	2	0	0	1	1	0	0
$\sum$		83	91	47	2	2	39	32	21	7

Table 12 Kappa coefficients for Level-1 degradation classes (Expert listener P.TCA experiment 2)

Degradation class	Frequency	Kappa
Speech-Spectrum	91	0.428
Speech-Level	83	0.671
Noise-Steady-state	32	0.261
Noise-Level	39	0.235
Noise-Dynamic	21	0.268
Speech-Distortion	47	0.277
Noise-Impulsive	7	0.459
Echo	2	0.122
Speech-Information	2	0.122

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 (Sachs and Hedderich 2009)

As in the two preceding experiments, the reliability of the annotation process was analyzed with the help of the kappa coefficient ("Analyzing technical causes and perceptual dimensions"), see Table 12. Besides the two Level-1 degradation "Echo" and "Speech-Information" that obtained slight agreement, at least fair to moderate agreement was obtained for all 7 remaining Level-1 degradations. Substantial agreement was obtained for "Speech-Level".

Experts vs. experts vs. Naïve listeners

The results of the present expert experiment are compared to the two preceding experiments with naïve listeners ("Diagnosing the quality of transmitted speech with naïve annotators") and experts ("Analyzing technical causes and perceptual dimensions") with respect to the kappa coefficient. The kappa results of the three experiments are summarized in Fig. 4 (Tables 3, 10, 12). The figure shows that the annotation agreement of the experts (*experts* for the initial experiment and *experts2* for the final experiment) outperforms the annotation agreement of the naïve listeners. This was already observed in "Diagnosing the quality of transmitted speech with naïve annotators".

However, looking at the two expert experiments the comparison becomes more complicated. Initially, it was assumed that the GUI and the example files would help the experts to archive a constant high agreement level. In contrast to the first expert experiment without improvements, higher agreement is only obtained for two Level-1 degradations, namely "Speech-Level" and "Noise-Impulsive". For "Speech-Level" degradations thus, the GUI and the example files seem to actually facilitate the annotation procedure as naïve listeners also obtained a high agreement (see the high kappa values in Fig. 4). The example files give the annotators an overview about expected degradations and the GUI facilitates the annotations procedure in terms of complexity.

Looking at the "Speech-Spectrum" degradation the results show that in the second expert experiment the agreement is lower than in the first expert experiment but still levels at a moderate agreement level. Thus, this difference might be explained with natural fluctuation in the annotation agreement.

However, the results for the three noise degradations "Noise-Level", "Noise-Steady-state", and "Noise-Dynamic" show an explicit lower agreement level in the second expert experiment than in the first one. While for "Noise-Dynamic" it might also be explained with natural fluctuations for "Noise-Level" and "Noise-Steady-state" it has to be investigated whether the difference is due to the proposed improvements or due to general problems of the P.TCA annotation method. An argument for the first explanation are the high number of low precision values for the example files of the two degradation classes (Table 7). This might have led to confusion that was not present in the first expert experiment. An argument for the latter explanation are the observations made in the two first experiments showing that stable results are generally hard to archive with the P.TCA method.

Conclusion

The outcome of the second expert experiment shows that the introduced improvements for the P.TCA method only partially help experts to achieve a higher annotation agreement. Especially for the three Level-1 degradation describing noise ("Noise-Level", "Noise-Steady-state", and "Noise-Dynamic") the results of the second experiment show a lower agreement than the results of the first experiment without any improvements. Thus, it could be argued that the exemplary listening material seems to confuse expert listeners. While having an idea of a specific degradation in mind, the examples might be different and thus lead to an undesired annotation.

However, for the remaining Level-1 degradations the improvements used in the second expert experiment lead to similar or higher agreement levels than in the first expert experiment. For the first time, the kappa values showed a substantial agreement for Level-1 degradation ("Speech-Level"). Hence, it can be claimed that the improvements still provide helpful information and facilitate the P.TCA method. Nevertheless the unexpected results with the noise Level-1 degradations should be investigated to determine if it is due to the P.TCA method, for example there might be

too many possible annotation noise types, or due to the confusion introduced by the example material.

Conclusions and outlook

The objective of the work presented in this article is to analyze and give insights in the newly proposed diagnostic expert annotation method for the ITU-T P.TCA (Technical Cause Analysis) activity. A first expert annotation study revealed that the P.TCA annotation method is able to capture some of the technical causes of sub-optimum quality with acceptable annotation reliability. Nevertheless, participating experts reported difficulties in the annotation method due to complicated usage and ambiguous description. Therefore the authors developed exemplary listening material to provide better description as well as a user-interface to facilitate usage and to increase the usability. The improvements were tested in two subsequent P.TCA experiments with naive and expert annotators. The results showed that the improvements do not lift naive annotators to the same annotation agreement level reached by experts. As a result it is apparent that the P.TCA method should only be used with expert annotators for the time being. The expert annotators reached similar or higher agreement levels on most Level-1 degradation using the proposed improvements with the exception of three materials featuring background noise where lower agreement was achieved. The authors argue that this result might be due to too many possible annotation choices for noise types.

In fact, in all conducted experiments the participating expert or naive annotators reported that the amount of possible annotation types is overwhelming. At the time of writing the list of annotation types consist of 47 Level-2 degradations grouped in 9 Level-1 degradations. The authors argue that it might be helpful to discard or merge some of the Level-2 degradations. For example the three Level-2 degradations "Temporal speech clipping", "Choppy speech", and "Self clipping" all describe missing parts of words so these categories could be merged to a single Level-2 degradation type. This might reduce the confusion of the annotators, facilitate the annotation process and improve annotation agreement.

The annotation method should be validated with additional data and in different laboratories to confirm the reported observations and to possibly reveal new findings. Third party evaluation is a necessary step in a standardization process to verify that the method provides reliable annotations on independent data. Once the annotation methodology is settled, two main developments remain: the objective prediction of degradation classes for automated classification and the mapping of each of the degradation classes to the potential technical causes.

Acknowledgements We would like to thank Maxim Szepansky, Lucas Almeida, and Question 16 of SG12 at ITU-T for their help and comments. On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Borg I, Groenen P (2005) Modern multidimensional scaling—theory and applications, 2nd edn. Springer Series in Statistics, New York

Bortz J (2005) Statistik. Springer, Berlin

Côté N (2011) Integral and diagnostic intrusive prediction of speech quality. Springer, Berlin

Deller J, Proakis J, Hansen F (2000) Discrete-time processing of speech signals. IEEE Press, p 246

ITU-T Contribution COM 12-342 (2012) Results from a multidimensional rescaling experiment of P.OLQA SWB test database. International Telecommunication Union, Deutsche Telekom AG (M. Wältermann, S. Möller)

ITU-T Contribution COM 12-40 (2013) Validation of the P.TCA annotation methodology and comparision to perceptual dimensions from P.AMD. International Telecommunication Union SG12 Question 16, Deutsche Telekom AG (F. Köster, S. Möller, F. Schiffner, J. Skoworonek)

ITU-T Recommendation P.800 (1996) Methods for subjective determination of transmission quality. International Telecommunication Union, Geneva

ITU-T Recommendation P.806 (2014) A subjective quality test methodology using multiple rating scales. International Telecommunication Union, Geneva

ITU-T Recommendation P.835 (2003) Subjective test methodology for evaluating speech communication systems that include noise supression algorithm. International Telecommunication Union, Geneva

ITU-T Recommendation P.863 (2011) Perceptual objective listening quality assessment. International Telecommunication Union, Geneva

ITU-T Temporary Document TD 136 (GEN/12) (2009) Proposal for multidimensional evaluation of speech quality (MESQ). International Telecommunication Union, Geneva

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

ITU-T Temporary Document TD 686 (GEN/12) (2011) Expert Listening for P.TCA. International Telecommunication Union, Rapporteur Q.16/12 (L. Malfait)

Jekosch U (2005) Voice and speech quality perception: assessment and evaluation. Springer Science and Business Media, Berlin

Köster F, Möller S, Antons J-N, Arndt S, Guse D, Weiss B (2014) Methods for assessing the quality of transmitted speech and of speech communication services. Acous Austr 42(3):179–184

Köster F, Schiffner F, Guse D, Ahrens J, Skowronek J, Möller S (2015) Towards a MATLAB toolbox for imposing speech signal impairments following the P.TCA schema. In: Audio engineering society convention 139

Matlab-file exchange: Add noise. http://www.mathworks.com/matlabcentral/fileexchange/32136-add-noise/content/addnoise/addnoise.m. Accessed 2015-05-15

Möller S, Köster F, Skowronek J, Schiffner F (2013) Analyzing technical causes and perceptual dimensions for diagnosing the quality of transmitted speech. Proc. 4th International Workshop on Perceptual Quality of Systems (PQS 2013), p 3035

Osgood C (1957) The measurement of meaning. University of Illinois Press, Urbana

Powers DM (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol 2(1):37–63

Qualinet (2013) Qualinet white paper on definitions of quality of experience, 2013, (Version 1.2, eds. P. Le Callet, S. Möller, A. Perkins). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland

Sachs L, Hedderich J (2009) Angewandte Statistik. Springer, Berlin

Scholz K (2008) Instrumentelle Qualitätsbeurteilung von Telefonbandsprache beruhend auf Qualitätsattributen. Shaker, Kiel

P.TCA MatLab Toolbox. https://github.com/QULab/ptca-matlab-toolbox/. Accessed 2016-05-01

Vary P, Heute U, Hess W (1998) Digitale Sprachsignalverarbeitung. Teubner, Stuttgart

Voiers W (1977) Diagnostic acceptability measure for speech communication systems. International Confeence on Acoustics, Speech and Signal Processing (ICASSP), Hartford

Wältermann M (2012) Dimension-based quality modeling of transmitted speech. Springer, Berlin

Wältermann M, Raake A, Möller S (2010) Quality dimensions of narrowband and wideband speech transmission. Acta Acustica united Ac 1090–1103