Towards Training Naïve Participants for a Perceptual Annotation Task Designed for Experts

Friedemann Köster, Dennis Guse, Christian Miethaner, Sebastian Möller
Quality and Usability Lab
Technische Universität Berlin
Berlin, Germany
friedemann.koester@tu-berlin.de, dennis.guse@tu-berlin.de, chris.miethaner@gmail.com, moeller@tu-berlin.de

Abstract*—*Technical Causes Analysis **(P.TCA) is a method for identifying technical causes of sub-optimum speech transmission quality. Originally created as an expert procedure for the annotation of speech samples, its applicability to naïve listener was also studied. Due to the low agreement of naïve listener annotations, it was suggested that detailed training methods are necessary to lift naïve annotations to an agreement level of experts. The aim of this work was to develop training methods for naïve annotators. For this, two different training procedures were developed and tested in two separate annotation experiments. The results are analyzed and discussed regarding the effects of the trainings and their implications for the P.TCA annotation scheme. The outcome shows that these training methods did not meet the expectations for improving the inter-rater agreement of naïve annotators. It is concluded that trainings of 15 to 20 minutes rather confuse naïve annotators by conveying too much information in too little time, and that they are not sufficient to prepare naïve annotators. It is argued that much more extensive training is needed to raise naïve annotators to expert level, and that such a training must include both, in-depth introduction to the annotation process as well as detailed presentation and exercise regarding the P.TCA degradations.**

I. Introduction

The quality of transmitted speech in vocal human-to-human telephony communication as perceived by the service end-user — the so-called *Quality of Experience* (QoE) [1] — is one important aspect for telecommunication service providers. In this context, understanding and assessing QoE is one of the main challenges in current research. For this, subjective experiments conducted with human participants in a laboratory context are a valid and reliable means. In *Listening-Only-Test*s, naïve participants rate the perceived overall quality of a speech sample on an *Absolute Category Rating* scale [2]. The ratings are averaged to the so-called *Mean Opinion Score* (MOS) [3], representing the average rating of an "average" user. However, as described in [4], the MOS value unfortunately only provides little insight into the reason for possible sub-optimal quality. More precisely, two different speech samples — one degrade by, say, circuit noise and the other one by a signal bandwidth limitation — might be both rated with the same MOS value. Thus, the MOS value does not provide diagnostic information in terms of causes for sub-optimal quality.

To overcome this disadvantage and to provide diagnostic information, the *International Telecommunication Union* (ITU-T) is currently working on two approaches for the diagnostic quality assessment of transmitted speech. Both are intended to be able to extract diagnostic information on the basis of instrumental measurements. The first, called *Perceptual Approaches for Multi-Dimensional Analysis* (P.AMD) [5], is a subjective method with naïve participants proposed to assess perceptual dimension scores. Such dimensions are e.g. "coloration", "discontinuity", "nosiness", "loudness", or the dimensions introduced in P.MULTI [6], and the idea is that these dimensions can also be estimated by technical means. The second, called *Technical Causes Analysis* (P.TCA) [7] (see Section II for detailed information), is a subjective method with expert participants proposed to assess technical causes of the transmission channel such as sub-optimum speech level, speech spectrum, noise level, or echo.

Following the P.TCA methodology [8], experts are asked to listen to a set of impaired speech stimuli. For each stimulus, the experts are asked to select possible technical causes from a given list. This list describes 47 degradations (Level-2) grouped into 9 impairment classes (Level-1). Two initial P.TCA experiments [9], [10] reported that both expert and naïve annotators complained the high *complexity* and that the usage of the method is *difficult*. The inter-rater agreement of the experiments were compared and showed that naïve annotators did not reach a similar agreement level as expert annotators. However, it was argued that a preceding training of naïve annotators could help to raise the level of agreement and lower the level of complexity.

In this article, we present the development and evaluation of two individual training methods for the P.TCA method. The two trainings are intended to give naïve annotators a better understanding of the P.TCA procedure. This is expected to raise the level of agreement of naïve annotators to the level of expert annotations. After giving a short introduction into the P.TCA method, we present the development of the two training methods. Following, both trainings were evaluated in a P.TCA annotation experiment with naïve annotators. The results are analyzed with respect to the annotation agreement and are compared to the results of the two previously conducted experiments ([9], [10]) with expert and naïve annotators. We close the article with a conclusion and an outlook towards future work.

In addition, we seek bringing up the topic of training in QoE assessment to the community. Not only for speech quality assessment, but also for video, gaming, or web-browsing assessment methods, the aspect of possible training methods should be considered. We strive to provide a first start-point that could be the base for further research and discussion.

978-1-5090-0354-9/16/$31.00 ©2016 IEEE

TABLE I. EXTRACT FOR 3 OUT OF 9 LEVEL-1 DEGRADATIONS FROM THE LIST OF TELECOMMUNICATIONS IMPAIRMENTS ACCORDING TO P.TCA [7].

Impaiment type	Name	Definition
Level-1 degradation	Level-2 degradation	
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 second) 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.

II. TECHNICAL CAUSES ANALYSIS (P.TCA)

The P.TCA scheme is a method to evaluate and diagnose speech transmission quality on the basis of signal characteristics. It was designed as an annotation experiment for experts.

The purpose of the method is to link audible degradations with technical causes that can ideally be remedied by telecommunication providers. 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 [12]. The presence of technical problems is usually assumed when stimuli receive a subjective MOS < 3.0 [8]. The full list of impairments can be found in [7], and an extract for three Level-1 classes are presented in Table I. Based on this list, expert annotators are asked to identify the most prominent degradations within each evaluated stimulus. Each stimulus can be listened to as many times as desired, and no specific setup is recommended. The responses are given in two steps [8]:

(i) 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.

(ii) 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 stimulus.

In an initial P.TCA annotation experiment with expert annotators, the results were analyzed with respect to the reliability of the annotations and the inter-rater agreement [9]. The inter-rater agreement was analyzed using the *Kappa* coefficient that indicates how strongly the annotations of the different annotators agree, normalized by the per-chance agreement [11]. Since the P.TCA method is not at the final stage of development, no ground-truth data for analyzing the annotations is available. Therefore, it was decided to use the inter-rater agreement for analysis [9]. Table II shows the interpretation of the kappa values. It was discovered that with the P.TCA annotation scheme it is possible to capture some of the technical causes of sub-optimum quality with an acceptable annotation reliability and inter-rater agreement. In addition, the participating expert annotators reported that a better explanation of the named degradations, best to be provided by exemplary listening material, may increase the annotation reliability as well.

TABLE II. INTERPRETATION OF KAPPA VALUES ACCORDING TO [11].

Kappa	Inter-rater agreement
< 0	poor
$0.00 - 0.20$	slight
$0.21 - 0.40$	fair
$0.41 - 0.60$	moderate
$0.61 - 0.80$	substantial
$0.81 - 1.00$	almost perfect

Regarding this observation, a set of exemplary listening material was created and validated [12]. It was argued that with the created exemplary listening material the P.TCA procedure would be easier to use, making it also available for a wider field of users that are easier to find (non-experts or even naïve participants). Following, a second P.TCA annotation experiment with naïve annotators that provides the exemplary material to familiarize them with the degradation to be detected was conducted [10]. The results showed that the created exemplary listing material alone is not enough to lift naïve annotators on the level of experts. It was argued that a possible preceding *training* of the naïve annotators could help to introduce the P.TCA scheme and its degradations to raise the agreement level of naïve annotations. The development and evaluation of possible training methods is presented in the next sections.

III. TRAINING PROCEDURES

As mentioned before, the aim of this article is to evaluate possible trainings to lift the annotation agreement of naïve annotators to the level of experts. Since trainings for the described use-case are new in the community, it was decided to create two independent trainings. In the following, the requirements for the trainings, the two trainings, and a short summary are presented.

A. Requirements

Resulting from the observations made in earlier studies, the aim of the work presented in this article is to develop two trainings to prepare naïve annotators for the P.TCA annotation task and to raise the inter-rater agreement of naïve annotators. It is important to mention that the aim is to improve and facilitate the annotation procedure for naïve annotators – it is not the the goal to train naïve annotators to become experts. Literature review showed four important facts that should be considered when creating training tasks:

(i) The training should be limited to a duration of *15-20 minutes*. The duration is limited due to the overall time of an experiment that should not exceed 2 hours to limit the impact of fatigue.

(ii) In [13] the term "task learning" is introduced. The term describes that the *more similar a training task is to the target task, the greater the learning transfer*. Consequently, when creating a training for the P.TCA method, the task should be as similar as possible to the actual annotation task.

(iii) Regarding the "task complexity", in [13] it is argued, that the higher the complexity of a task the more learning it requires. Thus, the premise is followed, that *the "identification" of a stimulus is easier than "discrimination" of stimuli*. The demands on memory and attention are highest for an "identification" task with a greater set of stimuli. For the P.TCA training this assumption would lead to task where participants should identify specific degradation.

(iv) In [14] the importance of an "internal representation" is discussed. It is suggested, that internal standards of naïve test participants are "unstable" and influenced by undesirable factors that are not relevant for the evaluation task. It is further assumed, that these *internal representations can be replaced by external ones through training*, for example with reference-matching tasks. Applying this concept for P.TCA training would call for a method to successfully introduce degradation speech examples to create correct internal references for test subjects.

Considering these four requirements, we developed two individual trainings for the P.TCA annotation method. The first one is called "Selection Training" and the second one is called "Association Training". Both use the exemplary listening material created in [12]. As the amount of Level-1 and Level-2 degradations can be overwhelming to untrained annotators, the two trainings primarily aim at introducing those degradations. The trainings are presented in the next sections.

B. Selection Training

The main focus of the "Selection Training" is to familiarize annotators with the Level-1 and Level-2 degradations of the P.TCA scheme. It tries to accomplish this in two ways: presentation of each degradation, its description and exemplary listening material, and subsequently asking its localization within the two-level structure of P.TCA.

There are a total of 47 trials, one per Level-2 degradation. Each trial consists of two stages: First the impairment type, name, and definition of the current degradation are shown. After reading the presented information the participant must listen to the corresponding speech example once. Stage two starts after the speech finishes. Name and definition are blanked out. Instead, a two-level list menu is enabled in which the participant has to locate the degradation presented in stage one. This menu reflects the selection process of the P.TCA annotation method and is designed accordingly. First, the Level-1 degradation is selected, then the Level-2 degradation. The annotator can at any time return to stage one, i. e., listening to the example speech again. In this case — for the duration of the example — name and definition are shown again, and the menu is disabled and reset. If the selection is correct, the next trial starts. If it is incorrect, the trial reverts to stage one.

C. Association Training

The focus of the "Association Training" is to introduce the characteristics of the Level-1 degradations. To accomplish this, annotators are presented with three (sometimes two) definitions of degradations of Level-2 degradations from within a Level-1 degradation. They have to match these with the associated speech example.

This method introduces the Level-1 degradations of the P.TCA schema. The order of Level-2 degradations is randomized within their Level-1 degradations. In a total of 19 trials, three (sometimes two) Level-2 degradations from the same Level-1 degradation are presented. The number of trials per Level-1 degradation varies according to their size. Annotators have to match the names and definitions of the Level-2 degradations with their (unmarked) speech examples. Thus, the participants are confronted with three (two) different Level-2 example speech files from one Level-1 degradation which they have to "associate" to three (two) corresponding Level-2 degradation definitions. In the "Association Training" the participants are therefore asked to identify two to three Level-2 degradations within one Level-1 degradation (e.g. "Loud speech", "Quiet speech", and "Loudness varies" inside "Speech-Level"). To prevent guessing, the matching process for each speech example may only begin after it has been listened to once. After three distinct (one degradation matched to one speech example) answers have been given, feedback about correctness is provided. Wrong matches reset the answer selection completely and force the annotator to listen to the example again.

D. Summary

We developed the "Selection" training and the "Association" training separately from each other. Both cover the aforementioned requirements of (i) being in a time limit of about 20 minutes, (ii) being similar to the actual P.TCA annotation task, (iii) identifying the degradations covered in the P.TCA guidelines, and (iv) building a new internal representation. DG For both trainings individual *Graphical User Interfaces* (GUI) were developed. The trainings were tested by two experts on P.TCA in separate test-sessions. They found that the two trainings both introduce the concept of Level-1 and Level-2 degradations as well as the individiual degradations properly. Thus, the trainings should now be evaluated in P.TCA experiments preceding to the actual annotation task. The results and outcome of the experiment are presented and discussed in the following sections.

IV. Annotation Experiment

The annotation experiment was conducted to gather data on the performance of trained annotators in contrast to untrained annotators (data taken from [10]). In the following, the used data, the test design, and the demographic data are presented.

A. Speech data and exemplary listening material

To ensure the comparability, in the experiment the same speech data was used as in the previous experiments [9], [10]. This includes mixed-band (narrowband, wideband, super-wideband), noise (signal-correlated, uncorrelated), ambient background noise of different types, temporal clipping, coding, temporal stretching, packet loss, acoustic recordings, and different frequency distortions. Also combinations of these degradations are included. The data provides 33 conditions; for details see [9]. For each condition, two speech stimuli of different sentences by one male speaker and one female speaker were annotated. For each condition, reference speech was available. The language of the database is German. For the exemplary listening material the reference stimuli from the database were used. A complete set of speech examples for all 47 P.TCA degradations was available - one male and one female example per degradation [12].

B. Test design

The experiment was divided into two phases — the training phase and the annotation phase — to be completed within one session. Two groups of participants were either trained through the "Selection" or the "Association" training in the first phase. In the second phase, both groups performed the same P.TCA annotation task (with the same conditions). For this phase, the participants used the same setup as the untrained naïve annotators in [10]. Instructions for the training and annotation task were given in written form. Here, the identical instructions for the P.TCA method as used by were presented to the annotators. In addition to the method, these describe the GUI. The experiment was conducted in a room suited for auditory experiments according to [3]. Speech was presented using an *Edirol UA-25EX audio interfaces* and a pair of *Sennheiser HD 280 Pro* headphones.

C. Participant Data

30 participants (17 female, 13 male) aged between 19 and 58 years ($\mu = 35$, $\sigma = 10$) took part in the experiment. All participants were native German speakers, had no hearing difficulties, and had no prior experience related to P.TCA. They were mainly students from TU Berlin. The "Selection" group consisted of 6 female and 9 male participants ($\mu = 31$, σ: 11). The "Association" group consisted of 11 female and 3 male participants ($\mu = 32$, σ: 9).

TABLE III. Kappa coefficient for Level-1 degradation classes for expert, trained, and untrained naïve annotators. See visualized in Figure 1

Degradation class	Experts	Naïve annotators		
		No Training	Selection Training	Association Training
Speech - Level	0.44	0.43	0.34	0.41
Speech - Spectrum	0.59	0.28	0.23	0.21
Speech - Distortion	0.24	0.07	0.09	0.12
Speech - Information	0.12	0.12	0.01	0.11
Echo	0.09	0.01	0.01	0.04
Noise - Level	0.59	0.31	0.08	0.01
Noise - Steady-state	0.37	0.23	0.23	0.19
Noise - Dynamic	0.39	0.17	0.17	0.06
Noise - Impulsive	0.32	0.06	0.06	0.04

V. Results

To analyze the results, the inter-rater agreement regarding the Level-1 annotations of the respective groups are analyzed using the kappa coefficient. These will be compared to the kappa values of previous experiments (expert and untrained naïve annotators) [9], [10]. Kappa values for the trained groups were calculated in the same way as in the previous experiments (according to [11]). In addition, participant feedback will be analyzed.

A. Selection Training

In average, participants needed 19 minutes to complete the "Selection" training. Table III shows the kappa values for expert annotators, naïve annotator that received no training, and naïve annotators that received "Selection" or "Association" training ahead of the annotation task. The fourth row shows the kappa values for participants that received "Selection" training. The results show, that the "Selection" training did not improve the inter-rater agreement. In comparison to the untrained naïve annotators (cf. the third column in Table III) they performed substantially worse in "Speech-Level" and "Noise-Level" agreement and slightly worse regarding "Speech-Spectrum". For the remaining Level-1 degradations about the same inter-agreement was reached, however, receiving a training should have raised the agreement level. With regard to the interpretation scale, whereas a "fair" agreement in at least one class was reached by untrained annotators, "Selection"-trained annotator reached "slight" agreement at best.

B. Association Training

On average, participants needed 14 minutes to complete the "Association" training. The fifth column in Table III shows the kappa values reached by "Association"-trained annotators. Compared to untrained naïve annotators, "Association"-trained annotators reached a better agreement regarding "Speech-Distortion" and "Speech-Information", but a worse agreement regarding "Speech-Spectrum", and "Noise-Dynamic" and especially for "Noise-Level". Again, for the remaining Level-1 degradations about the same inter-agreement was reached. Whereas untrained annotators achieved at least "slight" agreement in 4 classes, "Association"-trained annotators did so only for two Level-1 classes. It is worth noting, that the "Association"-trained annotators reached a "fair" agreement for speech level, about the same (but not a higher) as the untrained naïve annotators (0.43).

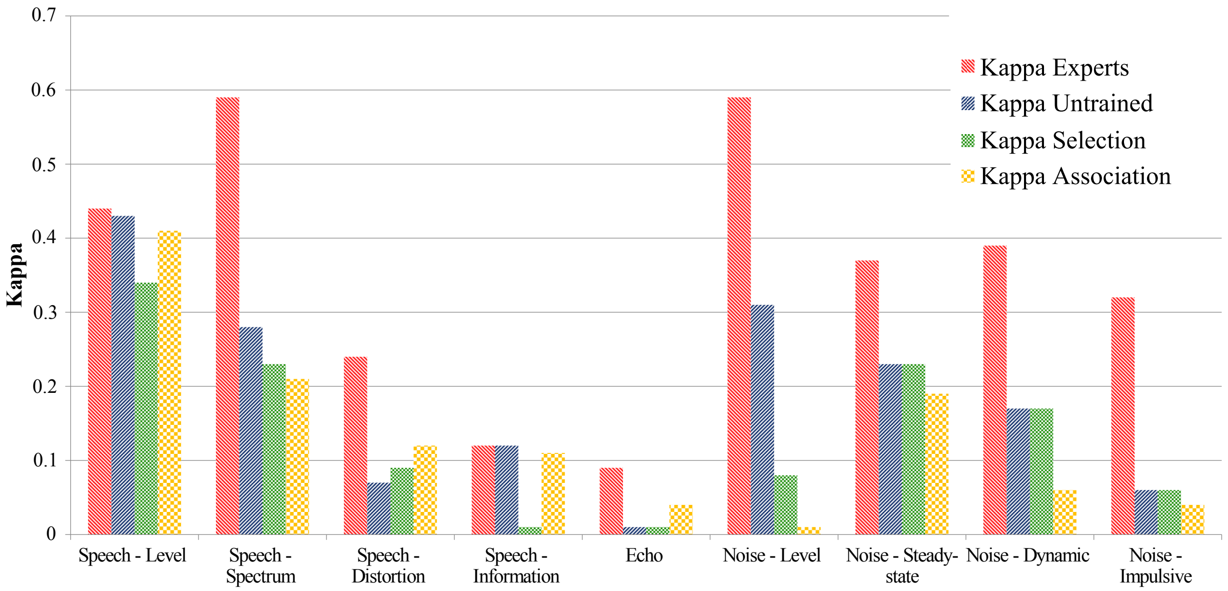

Fig. 1. Kappa values for all four annotator groups. Based on values presented in Table III.

C. *Comparing both Trainings*

"Association"-trained annotators reached a better agreement regarding the classes "Speech-Level" and "Speech-Information". "Selection"-trained annotators reached better agreement regarding "Noise-Level", "Noise-Steady-state" and "Noise-Dynamic". In an absolute context, "Selection"-trained annotators reached at least "slight" agreement in 3 classes whereas "Association"-trained annotators only reached "slight" agreement in "Speech-Level". However, neither training method improved the inter-rater agreement compared to naïve annotators without training. Quite the contrary, they both failed to reach a similar agreement level to naïve annotators without training.

D. *Overall Comparison*

Figure 1 shows a graphical comparison of all annotator groups. It can be concluded that neither untrained nor trained annotators reached kappa values close to the expert annotators with the exception of untrained and "Association"-trained groups regarding "Speech - Level". Furthermore, with the exception of "Speech-Distortion", untrained annotators shared greater agreement than both trained groups.

E. *Participants Feedback*

Participants were asked to give feedback on their respective training and the annotation task. This was done in an qualitative way. Some participants' comments revealed that they had not fully understood their tasks. One did not realize he could listen to the example speech during the annotation task. Two assumed they had to always find three degradations. Also, many "Selection"-trained participants reported, they did not understand the "look-for-degradation" concept at first, expecting a more difficult training task. Participants from both groups reported not being able to recall many degradation names and almost any speech examples after training. Three felt that only certain "distinct‘" examples were easy to learn and remember. Many stated that the annotation task was too complex. This was also reported in [10].

VI. Discussion

Compared to the kappa values of untrained annotators (refer to [9], [10]), the "Selection" training resulted in drops of kappa values for more frequently used classes: "Speech-Level", "Speech-Spectrum", and "Noise-Level". This leads to the assessment that the "Selection" training merely impaired results, as no new effect was observed. A possible explanation could be that introducing participants to all possible degradations built an expectation for the whole variety of degradations to appear in the annotation task. Participants may have felt the need to alternate between degradations, trying to apply their new knowledge. Such an effect could even turn out positive in creating an incentive to consider the full variety of classes and degradations. But therefore, the training would have to be much more extensive to ensure all necessary information can be internalized.

Looking at the results of the "Association" training, the most notable difference to untrained annotators is a high drop in the kappa value of "Noise-Level". In addition, the training was completed faster (14 minutes) than the "Selection" training (19 minutes), but it is questionable if this difference had any negative effects. It is hard to reach specific conclusions for why the "Association" training failed to generate better results. It is suggested, that the attempt to train participants perceptually or create (replace) internal standards was too ambitious for the ultimately only 15-20 minute long training.

Summarizing, the developed training methods did not meet the expectations for improving the results of naïve annotators. Trained annotators reached worse inter-rater agreement and matched the expert's results less often than untrained annotators. It is concluded that 20-minute trainings rather confuse naïve annotators by conveying too much information in too little time. They are not sufficient to prepare untrained annotators for the requirements of the P.TCA annotation task. It appears much more extensive training is needed to raise naïve annotators to expert level. It must include both in-depth introduction to the annotation process as well as detailed presentations and exercise regarding P.TCA degradations. If shorter trainings are still to be developed, it is suggested to reduce the complexity of the actual experimental task to make it suitable to naïve annotators.

VII. Conclusion

We presented the development and evaluation of two individual training approaches with the aim to lift naïve participants to the level of expert participants. After introducing the P.TCA annotation method, we presented the two different trainings. Both trainings were evaluated in a P.TCA experiment and the results showed that the trainings were not sufficient to facilitate an expert annotation method for naïve participants. In fact, it can be observed that the trainings rather seem to confuse participants than helping them, as the inter-rater agreement decreased. This might be due to the presentation of extensive training material in rather short time frame of up to 20 minutes.

It can be concluded that future training methods demand to be more extensive in terms of duration and introduction to the experimental task for applying the P.TCA method. These findings should also be considered when creating possible training procedure for other QoE assessment disciplines like, video, gaming, or web-browsing assessment. Here, only little practical knowledge is available on how to train (naïve) participants and under which circumstances training is overwhelming for participants.

References

[1] Qualinet, "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. [Online]. Available: http://www.qualinet.eu/images/stories/QoE_whitepaper_v1.2.pdf

[2] P. Vary, U. Heute, and W. Hess, *Digitale Sprachsignalverarbeitung*. Teubner Verlag, 1998.

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

[4] F. Köster, S. Möller, J.-N. Antons, S. Arndt, D. Guse, and B. Weiss, "Methods for Assessing the Quality of Transmitted Speech and of Speech Communication Services," *Acoustics Australia*, vol. 42, no. 3, pp. 179 – 184, December 2014.

[5] ITU-T Temporary Document TD 438rev1 (GEN/12), *Requirement Specifications for P.AMD (Perceptual Approaches for Multi-Dimensional Analysis)*. International Telecommunication Union; Rapporteur Q.9/12 (J. Berger), 2014.

[6] ITU-T Recommendation P.806, *A Subjective Quality Test Methodology using Multiple Rating Scales*. International Telecommunication Union, 2014.

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

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

[9] S. Möller, F. Köster, J. Skowronek, and F. Schiffner, *Analyzing Technical Causes and Perceptual Dimensions for Diagnosing the Quality of Transmitted Speech.* Proc. 4th International Workshop on Perceptual Quality of Systems (PQS 2013), 30-35, 2013.

[10] F. Köster and S. Möller, "Diagnosing the quality of transmitted speech with expert and naive listeners," in *Fortschritte der Akustik – DAGA 2015: Plenarvortr. u. Fachbeitr. d. 41. Dtsch. Jahrestg. f. Akust.* Berlin: DEGA, 2015, pp. 143 – 146.

[11] L. Sachs and J. Hedderich, *Angewandte Statistik*. Springer, 2009.

[12] F. Köster, F. Schiffner, D. Guse, J. Ahrens, J. Skowronek, and S. Möller, "Towards a MATLAB Toolbox for Imposing Speech Signal Impairments Following the P.TCA Schema," in *Audio Engineering Society Convention 139*. New York, NY: AES, 2015, pp. 1–8.

[13] K. Robinson and Q. A. Summerfield, "Adult auditory learning and training," *Ear & Hearing*, vol. 17, no. 3, pp. 51S – 65S, 1996.

[14] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and B. G. S., "Perceptual evaluation of voice quality - review, tutorial, and a framework for future research," *Journal of Speech, Language, and Hearing Research*, vol. 36, pp. 21 – 40, 1993.