Diagnosing the quality of transmitted speech with expert and naïve listeners

Friedemann Köster, Sebastian Möller
Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Deutschland,
Email: friedemann.koester@tu-berlin.de, sebastian.moeller@tu-berlin.de

Abstract

In this contribution the frequency and consistency of expert and naïve listeners in a technical causes annotation experiment are compared. For this, two experiments with experts and naïfs following the guidelines of the currently discussed ITU-Recommendation P.TCA were conducted. In these experiments, participants are asked to annotate speech files with respect to their possible degradation by choosing from a list of 47 degradations, separated in 9 impairment types. Originally intended for experts, the P.TCA procedure was expanded with exemplary listening material for naïve annotators to lift them on an expert level. The results show that experts annotate more consistent than naïfs and that the additional provided examples are not sufficient for an equal analysis. Furthermore, findings about possible improvement of the P.TCA methodology are presented.

Introduction

For the diagnostic quality assessment of transmitted speech Study Group 12 of ITU-T (International Telecommunication Union) is currently working on two different approaches. Both, called P.AMD (Perceptual Approaches for Multi-Dimensional Analysis) [1] and P.TCA (Technical Causes Analysis) [2], are intended to be able to extract diagnostic information on the basis of instrumental measurements. The first, P.AMD is proposed to estimate perceptual dimension scores of two sets. Set A describes dimensions such as “coloration”, “discontinuity”, “noisiness” or sub-optimum “loudness” [3] and Set B describes the dimensions introduced in P.MULTI [4]. The second, P.TCA is proposed to estimate technical causes of the transmission channel, which might lead to perceptual dimensions, such as sub-optimum speech level, speech spectrum, noise level, or echo (refer to Table 1).

These two approaches indicate that there are obvious links between the technical causes and the perceptual dimensions which have already been pointed out [5]. In an initial study the results of a P.TCA annotation experiment have been analyzed with respect to the reliability of the annotations, as well as with respect to the relationships between technical causes, perceptual dimensions and overall quality [6]. The results showed that there is a need for all – P.TCA, P.AMD, and overall MOS scores – as these three metrics are only partly correlated and thus contain complementary information. Furthermore, two of the major conclusions which were drawn from this study were, first, the P.TCA annotation scheme is able to capture some of the technical causes of sub-optimum quality with acceptable annotation reliability, and second, 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 which may increase the annotation reliability as well.

Taking these conclusions into account, a set of exemplary listening material was created and validated [7]. The validation of the processed exemplary listening material showed, that it is basically possible to create example files. A set of files for 8 degradations was proposed. Also, it was recommended to conduct additional experiments to further substantiate the P.TCA procedure, especially with the created exemplary listening material. It was argued that the P.TCA procedure would be easier using the created example files, which might make the procedure also accessible for non-expert annotators.

In this contribution, we present the results of an initial annotation experiment following the P.TCA guidelines conducted by naïve listeners. The results are analyzed with respect to the reliability of the annotations and are compared to the results of an earlier experiment conducted by expert listeners [6]. Finally, conclusions are drawn for the further improvement of the P.TCA methodology.

Impairment type (Level 1)	Degradation (Level 2)
Speech - level	Loud speech
	Quiet speech
	Loudness varies
	Speech level fluctuations
	Temporal speech clipping
	Choppy speech
	Self-clipping
	Speech cut-outs
Speech - spectrum	Timbre varies
	Muffled speech
	Sharp speech
	Colored speech

Table 1: Extract of the P.TCA guidelines given in [2]

Speech Data

For the experiment, the same data as in the earlier expert experiment was used, the database number 503 (SwissQual-P.OLQA-SWB-TestDatabase3) from the P.863 [8] competition. This data was used because it includes diverse types of degradations and degradation combinations for which diagnostic information is most useful. The stimuli were produced in a number of different labs according to the P.OLQA specifications; four speakers with four different sentences were used. In addition, the database was available in German, which makes it suitable for annotation in our test lab. Also, the data was used before and therefore the results can be compared.

Experimental Setup

The speech files were 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. Additionally the processed exemplary listening material from [7] was provided. The naïve annotators listened to the database in one session (one hour) in a sound-proofed booth fulfilling the listening environment requirements given in ITU–T Rec. P.800 [9]. For the diotic sound presentation through a Realtek High Definition Audio ALC 268 soundcard a AKG K601 reference headphone at a comfortable listening level was used. 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.

The task of the annotators was to identify the most prominent causes of degradations within each evaluated sample. Each sample could be listened to as many times as desired. The list of technical causes was taken from [2]. A total of 47 different impairments on Level-2, grouped into 9 categories on Level-1, were provided to the naïve listeners. An example can be seen in Table 1.

Results

Table 2 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 2, 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. This result may either be linked to the particularities of the database used (i.e. that the corresponding degradations were rare in that database, e.g. echo), or it may be linked to problems in identifying particular classes of degradations from pure listening (despite their presence in the database). It may also be linked to the 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.

The reliability of the annotation process was 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 3. It can be seen that fair to 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 agreement. The other classes only reached a poor agreement. With respect to the Level-2 degradation classes, these are obviously less frequent, as they require a corresponding Level-1 degradation to be labelled, and as there are 47 Level-2 classes instead of 9 Level-1 classes. Because of their lower frequency of occurrence and the results of [6], Level-2 degradation classes are not analyzed further, and the analysis is limited to the level-1 classes in the following paragraphs.

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 impairments 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. Also, it was demanded to have example files for all degradation, since some participants only used degradations they had examples for. Some annotators asked to have a better training before the annotation process, to get a better “feeling” for the degradations.

Experts vs. Naïve Listeners

The results of the present experiment are compared to the expert annotation experiment presented in [6] with respect to the frequency and the kappa coefficient. Table 2 shows the annotations of the expert and the naïve listeners. The maximum number of annotation per condition in the expert experiment is 16 since there were four experts rating four samples for each condition. In Table 3 the kappa coefficients and the annotation frequency for the Level-1 impairments of the naïve and the expert experiment can be seen.

Table 3 shows, that the results conducted by the experts have a higher agreement than the results conducted by the naïve annotators. The results of the calculated kappa coefficients are also shown in Figure 1. The only kappa value that is basically equal is the value for the Level-1 degradation “Speech Level”. This can also be seen in Table 2. Here multiple Conditions (C12, C18, C29, C43, C54) have a high annotation frequency 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. Further, all Level-2 degradations corresponding to “Speech Level” have example files, which could also explain an almost equal result for expert and 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 is low but in Table 2 a few Conditions (C40, C41, C42) with almost equal annotation frequency can be found. This shows that for

Cond.	MOS	Speech -level		Speech -spectrum		Speech -distortion		Speech -infor-mation		Echo		Noise -level		Noise -steady -state		Noise -dynamic		Noise -impul-sive	
		N	E	N	E	N	E	N	E	N	E	N	E	N	E	N	E	N	E
C02	1,19	4	0	1	0	6	12	2	0	0	0	5	4	5	0	26	8	2	0
C03	2,88	5	0	3	0	10	4	0	0	1	1	9	0	3	0	18	16	5	0
C04	2,46	2	4	1	0	0	1	1	0	0	0	15	7	32	16	2	0	0	0
C09	2,97	7	5	30	16	16	4	4	0	0	1	0	0	0	0	0	1	0	0
C12	1,11	40	16	1	0	0	0	4	1	0	3	0	0	1	0	0	0	2	8
C13	2,45	2	2	8	6	0	1	5	0	0	0	29	15	3	1	6	0	0	0
C14	2,55	9	3	5	11	3	0	5	0	0	0	26	16	4	4	11	0	0	0
C17	2,42	9	0	12	12	9	7	1	0	1	0	4	4	28	8	7	11	0	0
C18	2,45	37	16	11	3	6	0	3	0	0	0	1	0	0	0	0	0	0	0
C19	2,54	7	0	24	16	9	4	3	1	1	0	1	0	18	9	2	4	1	0
C26	2,58	17	4	22	16	9	8	3	0	1	0	2	0	5	6	6	3	0	0
C27	2,48	6	0	27	15	8	1	1	0	1	0	7	0	13	7	8	12	0	0
C28	2,08	6	7	16	12	5	4	3	0	0	0	23	15	6	4	13	1	0	0
C29	1,77	41	16	3	0	1	0	1	0	0	0	11	7	8	2	9	9	1	0
C30	1,34	29	11	4	9	8	1	2	0	1	0	6	0	6	8	13	12	1	0
C32	2,64	5	5	1	4	5	0	7	0	1	0	29	16	2	0	4	6	0	0
C35	2,43	20	10	25	12	4	8	2	0	1	3	5	0	8	7	4	1	1	3
C36	2,14	33	12	8	10	6	3	0	0	9	2	2	0	4	5	1	3	2	3
C37	2,29	10	6	27	16	9	4	3	0	0	0	3	0	14	7	2	2	1	0
C38	2,8	9	4	28	16	10	4	2	0	2	4	1	0	0	0	4	2	0	1
C39	1,9	3	2	25	16	10	8	3	0	0	0	3	0	2	6	25	11	1	0
C40	2,16	5	4	34	16	7	0	2	0	0	0	0	0	3	5	7	5	1	1
C41	2,89	10	3	35	16	8	0	3	0	1	0	1	0	7	10	3	5	1	0
C42	2,77	5	4	33	16	5	4	3	0	0	0	2	0	21	16	2	0	0	0
C43	1,3	38	16	8	5	3	0	8	3	0	0	4	4	8	12	0	1	0	0
C44	2,48	11	1	20	16	11	0	2	0	0	0	3	1	14	15	3	2	4	0
C45	2,23	23	9	20	12	9	0	0	0	1	0	5	6	2	0	7	3	3	4
C47	1,86	32	8	4	4	1	1	6	0	0	0	19	8	2	3	5	9	7	11
C50	2,6	40	13	20	8	5	4	0	4	0	0	1	0	0	0	1	0	2	0
C51	2,83	14	3	25	16	11	0	1	0	0	0	1	4	5	8	7	0	7	7
C52	2,78	32	12	17	16	6	0	3	0	0	0	0	0	2	0	6	0	8	4
C53	2,8	6	1	25	16	9	0	2	0	1	2	0	8	13	9	9	1	1	2
C54	3	34	15	21	14	9	0	2	0	0	0	0	0	4	6	1	0	0	0
Sum:		551	212	544	345	218	83	87	9	22	16	218	115	243	174	212	128	51	44

Table 2: Frequency of Level-1 degradation. N - Naïve Listener; E - Experts

	Naïve experiment		Expert experiment	
Degradation class	Frequency	Kappa	Frequency	Kappa
Speech-spectrum	544	0.277	345	0.595
Speech-level	551	0.430	212	0.439
Noise-steady-state	243	0.226	212	0.439
Noise-level	218	0.313	138	0.592
Noise-dynamic	212	0.165	128	0.388
Speech-distortion	218	0.036	83	0.237
Noise-impulsive	51	0.056	44	0.316
Echo	22	0.069	16	0.089
Speech-information	87	0.012	9	0.118

Table 3: Kappa coefficients for Level-1 degradation classes of the naïve and the expert experiment. 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 [10].

conditions with distinctive degradations naïve listeners reach a high agreement. This, however, is not enough to yield to a high overall kappa value.

The results show, that experts have a higher agreement in the annotation experiment while some conditions can also be annotated by naïve listeners with a high frequency. This is due to the fact, that (i) the procedure was developed for experts, that (ii) the amount of information is too much for naïve listeners, that (iii) naïve listeners can only identify certain distinctive degradations and that (iv) the exemplary listening material does not cover all degradation. Therefore it can be claimed that the material is not lifting naïve listeners to the level of experts.

Conclusions and Outlook

The main outcome of the presented experiment is that the performed effort to make the P.TCA schema accessible for naïve listeners is not enough. The results show, that even with the exemplary listening material naïve listeners show lower agreements in their annotations then the experts without exemplary listening material. However, for particular Level-1 degradations ("Speech-level") the agreement and for certain explicit degradations the annotation frequency of experts and naïve listeners is almost equal. This can be explained by the understanding of the naïve listeners and the number of example files for different Level-1 degradation. Thus, there is need for more exemplary listening material and a detailed training of naïve listeners. With this additional improvements further experiments with experts and naïve listeners should be incorporated to check, if naïve listeners can achieve similar results as the experts.

Concerning the P.TCA methodology, it would be helpful to find possibilities to further investigate the annotations of listeners. One potential option would be to find a "ground truth" for the data that should be annotated. With the given database (SwissQual-P.OLQA-SWB-TestDatabase3) the results of experts and naïve listeners as well as the description of each condition can be used to determine the "correct" Level-1 degradation for each condition. Thereby more analytic possibilities are given (e.g. Precision and Recall) and next steps toward standardizing the subjective annotation schema can be made.

References

[1] 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.

[2] 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.

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

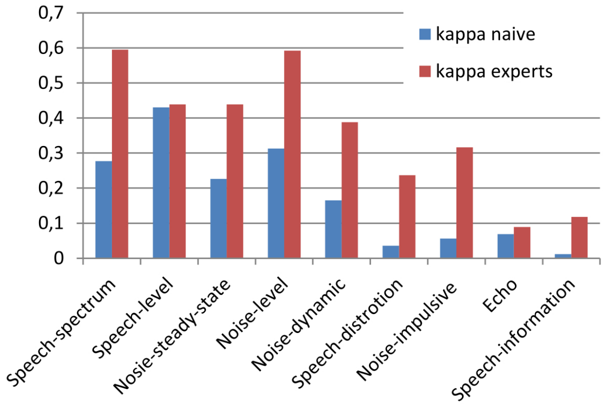

Figure 1: Comparison of the kappa coefficient for the naïve and the expert experiment.

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

[5] ITU-T Contribution COM 12-74, *Proposal for Benchmarking of the P.OLQA Degradation Decomposition*, International Telecommunication Union; Deutsche Telekom AG, TNO Information and Communication Technology (M. Wältermann, S. Möller, A. Raake, J. Beerends), 2007.

[6] 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), 3035, 2013.

[7] F. Köster and S. Möller, "Standards for diagnosing the quality of transmitted speech and their improvements," in *Globecom 2014 Workshop - Telecommunications Standards - From Research to Standards.* 2014, IEEE Globecom 2014 Proceedings.

[8] ITU-T Recommendation P.863, *Perceptual Objective Listening Quality Assessment*, International Telecommunication Union, Geneva, 2011.

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

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