Challenging ITU-T P.835: Searching for the optimal order of scales for assessing the quality of complex speech signals

Sebastian Arndt, Sebastian Möller, Jan-Niklas Antons

Quality and Usability Lab, TU Berlin, E-Mail: sebastian.arndt@telekom.de

Abstract

When evaluating the quality of speech samples which contain background noise, the International Telecommunication Unit (ITU) has released the ITU-T P.835 standard for assessing the quality on several scales. The scales contain ratings on signal quality, background noise annoyance, and the overall quality. Hereby, the order is fixed in the recommendation, such that the overall quality should be always assessed last, and the signal quality and background noise randomly either first or second. However, no clear argumentation is given why only the first two items are supposed to be randomized but the overall quality should be assessed last. Furthermore, research has shown that the order of scales has an effect on the judgment made. In this paper, we will show that the order of scales has an effect on the rating scores in this scenario. Therefore, we conducted an experiment conform to the recommendation, and altered the order of scales. We used a complex speech sample which contained different types and intensities of background noise; these were also suppressed using noise suppression algorithms. These may have an influence on the speech signal itself. We assessed the proposed scales and differed the order of those in a within subject study design.

Motivation

The assessment of perceived quality evoked by different audio and speech quality levels using subjective rating methods is an established procedure in Quality of Experience (QoE) research. For developers of codecs and products as e.g. telecommunication products or technical equipment several aspects of perceived stimuli are of interest at the same time [3]. Two of the most considered aspects are the mean opinion score (MOS, using the absolute category rating [ACR] method [2]) to determine the perceived overall quality of the transmitted audio/speech signal, and subscales to assess the speech signal alone and the background noise alone. All scales are valid and approved tools measuring what they should, if used as a single scale. In case, researchers are interested in the perceived overall quality and the subscales at the same time, subjects have to rate the stimuli on all scales. In the strict sense, to achieve appropriate results every subject should just rate one of the scales to be not influenced by the items of the others. As this would mean much more subjects would be needed, many researchers prefer to let subjects rate on all scales; in many cases with a fixed order. It could be shown that e.g. the order of multiple questionnaires in a sequence assessing the overall quality and the emotional state has a significant influence on the given ratings [4].

In the ITU-T Recommendation P.835 the order of presented rating scales is pseudo-randomized so that the last rated scale is always the overall quality. The two subscales (speech signal alone, the background noise alone) are in randomized order on position one or two.

As presented in ITU-T Contribution 12-113 [1], the order of items presented in blocks of three after the stimulus presentation had a significant impact on the rating of the speech signal and background noise. The comparison of ratings on single scales versus ratings of multiple scales was not performed so far.

Motivated by this, we performed a within subject designed study, comparing quality ratings on single scales (speech signal alone, the background noise alone and overall quality) versus ratings of all three scales in one block as proposed in ITU-T Recommendation P.835.

Experiment

Participants: 33 students (21 males, average age 24.38 years) participated in the study. 28 of those were German native speakers. All participants reported normal auditory acuity and no medical problems. Participants gave informed consent

Stimuli: Three different sentences spoken by a German male and female speaker were used for the experiment, each sentence had an approximate length of 8 s. For the background noise we chose babble and cafeteria noise, both with a SNR of 10 dB and 30 dB. The method of spectral subtraction (SS) was applied to reduce the noise for half of the stimuli. Resulting to a total stimulus corpus of 72 stimuli: 3 sentences * 2 speakers * 3 SNR levels * 2 noise types * 2 SS/none.

Experimental Design and Procedure:

The experiment was divided into 5 blocks. In each block one scale setup (3 single, 2 multiple) was tested on all 72 stimuli, the sequence of stimuli was randomized. Thus, test participants were listening to 360 speech samples (within subject study design) which resulted into average test duration of 90 min, including a pre-questionnaire and breaks between the blocks.

The tested conditions – scales that were used – between the blocks were:

a) Only Background
b) Only Speech Signal
c) Only Overall Rating
d) Background ↔ Signal; Overall
e) Overall; Background ↔ Signal

Whereat, ↔ indicates random order between trials within a participant.

The order of blocks was randomized between participants. In blocks a), b), and c) only one scale was asked for. In block d) randomly first either background noise or signal rating was asked, and last the overall rating, as the P.835 suggests originally. In block e) first the overall quality, and subsequently background noise and signal rating was asked in randomized order, as already proposed in ITU Contribution 12-113.

For the rating scales proposed in ITU-T P.835 the German labels were used.

Results

Obviously, the SNR level had a statistically significant effect on the rating of the speech file (F(2,64)=152.69, p<0.01, η²=0.83). As can be seen in Fig. 1-3, the ratings for the conditions with lower SNR result into lower quality ratings on all scales and all scale setups.

The scale setup had a significant effect on the rating of the participants (F(2,64)=8.64, p<0.01, η²=0.21). Bonferoni pairwise comparisons show significant difference between the single scale and both other scenarios. As can be seen in Fig. 1 for the speech signal quality, ratings were lowest for asking if just the single scale was used. For ratings on the background noise, no influence of the scale setup was found (see Fig. 2). The overall quality ratings were higher if the ratings were given in blocks compared to ratings on the single scale (see Fig. 3). Scenario*Scale had a significant effect: F(4,128)=3.69, p=0.017, η²=0.103).

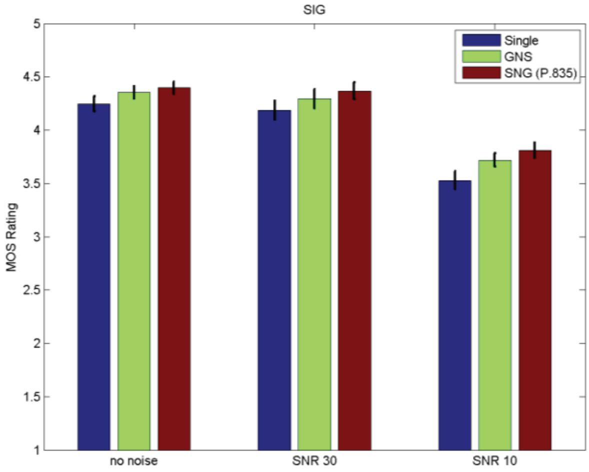

Fig. 1 Average ratings of the speech signal (SIG) on the single scale and the two scale orders (first OVRL → second SIG, BAK and first SIG, BAK → second OVRL) as a function on SNR. Whiskers denote 95% confidence intervals.

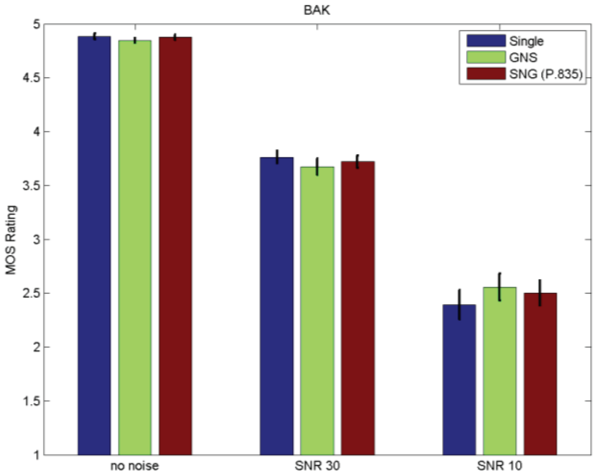

Fig. 2 Average ratings of the background noise (BAK), on the single scale and the two scale orders (first OVRL → second SIG, BAK and first SIG, BAK → second OVRL) as a function on SNR. Whiskers denote 95% confidence intervals.

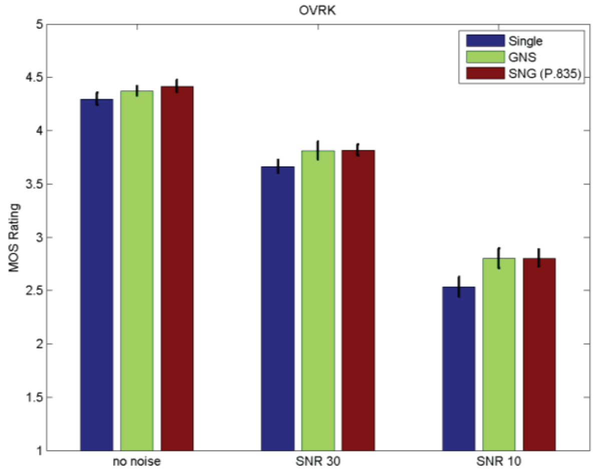

Fig. 3 Average ratings of the overall quality (OVRL), on the single scale and the two scale orders (first OVRL → second SIG, BAK and first SIG, BAK → second OVRL) as a function on SNR. Whiskers denote 95% confidence intervals.

The results show the averaged values over both speakers (male and female), noise types and noise reduction algorithm versus no noise reduction.

Discussion and future work

The significant main effect of the SNR-level showed that, as expected lower SNR conditions resulted into a lower quality rating. This was considered as a validation of the used test method and showed that, as intended quality variations could be measured.

Also for the scale setup a significant effect on the quality ratings was found. If participants rated the quality of the speech alone or the overall quality, the ratings were lower if a single scale was used. This effect could be assigned to the tendency that subjects were probably more precise when rating on a single scale, resulting in a lower rating. The possibility of adapting a later rating according to compensate for the not precise first rating, probably let to an overestimated quality. Based on these initial results, influences on the global average ratings were shown. In a deeper analysis, the focus should be on possible interactions between other influencing parameter such as the strength of degradation i.e. level of applied noise or the type of noise reduction method. Furthermore, it would be interesting to analyze if the reaction time of subjects corresponds to the precision of rating. Thus, whether subjects give an overestimated fast rating when three scales are presented, will the response be faster.

Summary

This contribution shows that there is a significant effect introduced by the scale setup. If listeners have to rate stimuli on multiple scales, the resulting ratings seem to overestimate the quality of the presented speech material. Even though it is still not completely understood what leads to this effect and that a deeper analysis of the recorded data is necessary, it is important to deepen the knowledge of these influences.

Literature

[1] ITU-T Contribution COM 12-113 (2013). Effects of rating scale order on subjective quality ratings of the speech signal alone, the background noise alone, and overall quality, Deutsche Telekom AG (Authors: S. Möller, S. Arndt), ITU-T SG12 Meeting, 03 – 12 Dec. 2013, CH-Geneva.

[2] ”Methods for Subjective Determination of Transmission Quality”, ITU-T Recommendation P.800, International Telecommunication Union, Geneva, 1996.

[3] “Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm”, ITU-T Recommendation P.835, International Telecommunication Union, Geneva, 2003.

[4] J.-N. Antons, S. Arndt and R. Schleicher, ”Effect of Questionnaire Order on Ratings of Perceived Quality and Experienced Affect” In: Proc. Perceptual Quality of Systems (PQS), 2013.