Effect of Questionnaire Order on Ratings of Perceived Quality and Experienced Affect

Jan-Niklas Antons, Sebastian Arndt, and Robert Schleicher

Quality and Usability Lab, Telekom Innovation Laboratories, Technische Universität Berlin, Germany

jan-niklas.antons@tu-berlin.de

Abstract

The assessment of perceived quality and experienced affect of users towards different audio and speech quality levels using subjective rating methods is a common and valid approach. In case that researchers are interested in the subjective a) quality and b) affective rating, participants have to rate the stimuli on both scales. Even though subject should just rate only one of the scales to be not influenced by the questions asked in the other questionnaire, researchers in practice frequently present various scales in a sequence to subjects. This paper examines the effect of questionnaire order on perceived quality and experienced affect by comparing the ratings of two groups of subjects which filled in questionnaires in different orders after listening to the same set of stimuli. We showed that the order of sequence had a significant effect on perceived quality and experienced affect as e.g. the quality of stimuli was rated lower if the quality rating not occurred instantly after stimulus presentation. Thus we want to create awareness for these effects and presented first guidelines to overcome them.

Index Terms: perceived quality, self-assessment manikin (SAM), experienced affect, order of questionnaires, sequence effect

1. Introduction

The assessment of perceived quality of users towards 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 codecs or technical equipment several aspects of perceived stimuli are of interest at the same time [1]. Two of the most considered aspects are the mean opinion score (MOS, using the absolute category rating [ACR] method [2]) to determine the perceived quality of the transmitted audio/speech signal, and on the other hand the affective state evoked by the stimuli. The latter one is usually measured by using the self-assessment manikin (SAM) for the scales valence, arousal and dominance [3]. Both methods are valid and approved tools measuring what they should if used singularly. In case that researchers are interested in the perceived quality and the experienced affect at the same time, subjects have to rate the stimuli on both scales. In the strict sense, too achieve appropriate results every subject should just rate one of the scales to be not influenced by the items of the other questionnaires. As this would mean much more subjects were needed, many researchers prefer to let subjects rate on all scales; in many cases with a fixed order.

Research that focuses on the effect of item-order within existing psychological inventories and scales is known. The resulting effect is addressed by randomizing the order of presented items, changing the polarity of items and mixing the items of different scales. Existing inventories can mostly be used because they were carefully developed and produce valid results. Researchers should consider the possible order effects when using two inventories simultaneously in one study.

Surprisingly there is to the knowledge of the authors no research in the quality domain performed on how the order of questionnaires impact the ratings in detail. To give an example: suppose you just rated the quality of a stimuli as low, would you still unrestrictedly state that you enjoyed listening to it, or would the previous request to explicitly consider perceived quality alter your overall experience in retrospect? Motivated by this question, we performed a study comparing the influence of the order in which questionnaires were presented to subjects on the perceived quality and experienced affect of speech stimuli in varying qualities and different affective pronunciations.

2. Method

N = 14 subjects (3 female, 11 male, average age 26.42) listened to a stimulus set of speech files (single sentence uttered by a female speaker with the length 1384 milliseconds). The stimulus set comprised five different quality levels (G.722.2 with the bit rate 6.6, 8.85, 12.65, 23.05 Kbit/s [4] and a reference stimulus in wide-band quality) and three different affective pronunciations (emotionally pronounced with the target emotions: joy, grief and rage) taken from the Kiel Affective Speech Archive [KASPAR]. The overall set of 15 stimuli was presented to every subject in randomized order. The test was performed in a laboratory room furnished following ITU-T Recommendation P.910 and using Sennheiser in-ear headphones.

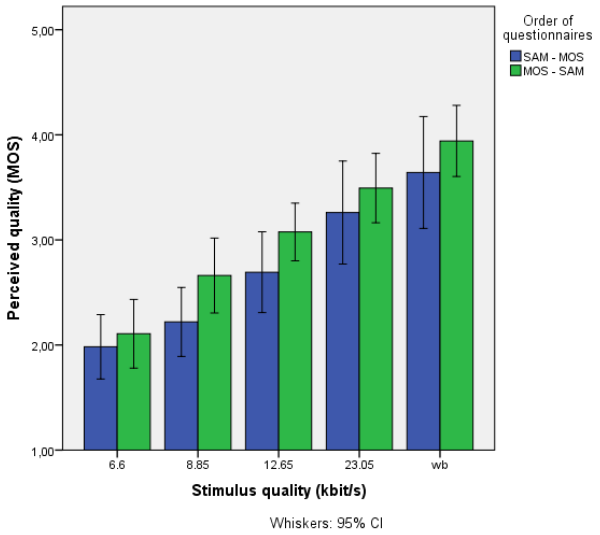

Figure 1: *MOS for the different quality levels (bit rate from 6.6 up to 23.05 Kbit/s and wide-band [wb]) and for the two questionnaire orders (SAM-MOS and MOS-SAM). Error bars indicate the 95% confidence intervals.*

10.21437/PQS.2013-3

Loudness was set to the individual preferred listening level at the beginning of the experiment a subsequently kept constant. Subjective quality ratings were obtained using an absolute category rating as recommended by ITU-T P.910, which varies from 1 to 5 with verbal labels Bad-Poor-Fair-Good-Excellent. The emotional self-assessment was obtained using modified versions of the Self-Assessment-Manikin (SAM) scales [5]. More specifically, listeners rated the arousal, valence and dominance dimensions using 9-point visual anchors. For one half of the subjects the order of questionnaires was MOS (perceived quality) – SAM (experienced affect) and for the other half SAM – MOS.

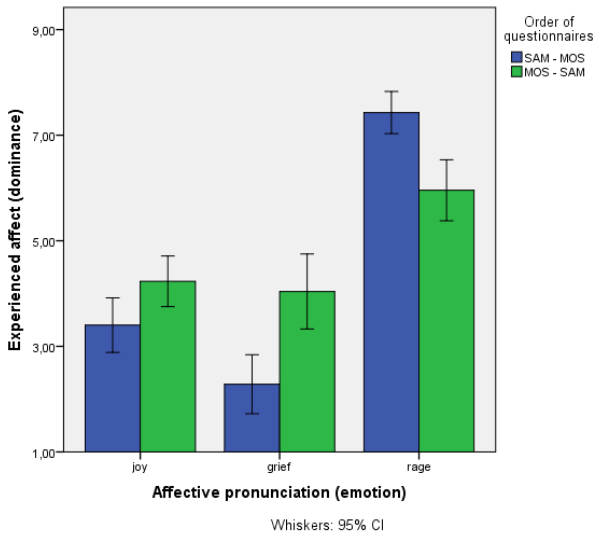

Figure 2: *Ratings on the dominance scale for the different affective pronunciations (joy, grief and rage) and for the two questionnaire orders (SAM-MOS and MOS-SAM). Error bars indicate the 95% confidence intervals.*

3. Results

All statistical analysis was done using the *linear mixed models* function of SPSS, which also automatically adjusts the degrees of freedom (df) in case the assumption of sphericity is not met. An analysis of variance (ANOVA) with *stimulus quality* and *order of questionnaires* as factors showed that as expected the *stimulus quality* influenced the subjective quality rating ($F_{4,58}$ = 30.32, p<.01). For this analysis the values for the three different affective pronunciations were averaged. The better, the stimulus quality the higher the resulting MOS value (see Figure 1).

In addition we found a significant difference caused by the factor *order of questionnaires*. MOS ratings for all quality levels were higher if first the perceived quality (MOS) and second the experienced affect (SAM) was rated ($F_{1,159}$ = 6.92, p<.01), regardless of emotional connotation. In average was the MOS 0.296 higher for the order MOS – SAM (SAM-MOS = 2.76 and MOS-SAM = 3.056).

An analysis of variance (ANOVA) with *affective pronunciations* and *order of questionnaires* as factors showed that the *affective pronunciations* influenced the subjective *dominance rating* ($F_{2,146}$ = 105.81, p<.01). For this analysis the values for the five different quality levels were averaged. We found no significant difference caused by the factor *order of questionnaires* ($F_{2,187}$ = 2.88, p=.091).

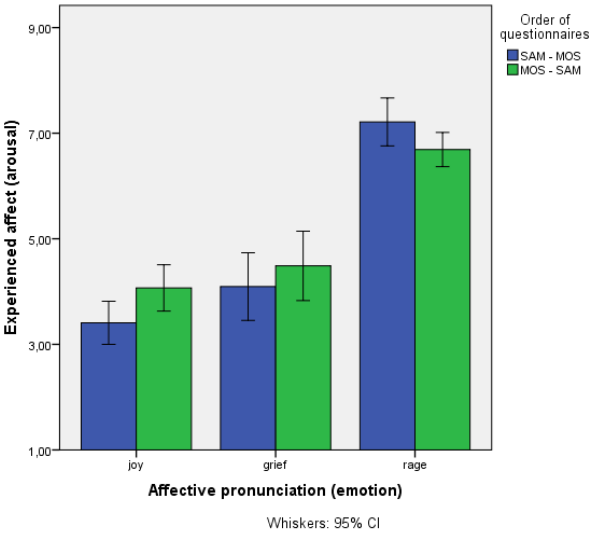

Figure 3: *Ratings on the arousal scale for the different affective pronunciations (joy, grief and rage) and for the two questionnaire orders (SAM-MOS and MOS-SAM). Error bars indicate the 95% confidence intervals.*

However as can be seen in Figure 2 the interaction of *order of questionnaires*affective pronunciations* was significant ($F_{2,187}$ = 19.31, p<.01). When a speech stimulus with the affective pronunciations *rage* was presented, values for dominance were on average 1.5 lower for the order of questionnaires *MOS-SAM*. To the contrary, for affective pronunciations *joy* and *grief* the average values were higher for the order MOS-SAM, 0.9 and 1.8 respectively. A similar effect was found for the arousal ratings (see Figure 3).

The interaction of *order of questionnaires*affective pronunciations* was significant ($F_{2,149}$ = 4.79, p<.01). The values for the affective pronunciations *joy* and *grief* were higher for the order MOS-SAM, 0.4 each. The arousal rating for the affective pronunciation *rage* was lower for the order MOS-SAM (0.6). For the valence rating (see Figure 4), only a non-significant tendency was found for the interaction *order of questionnaires*affective pronunciations* ($F_{2,135}$ = 3.46, p=.034)

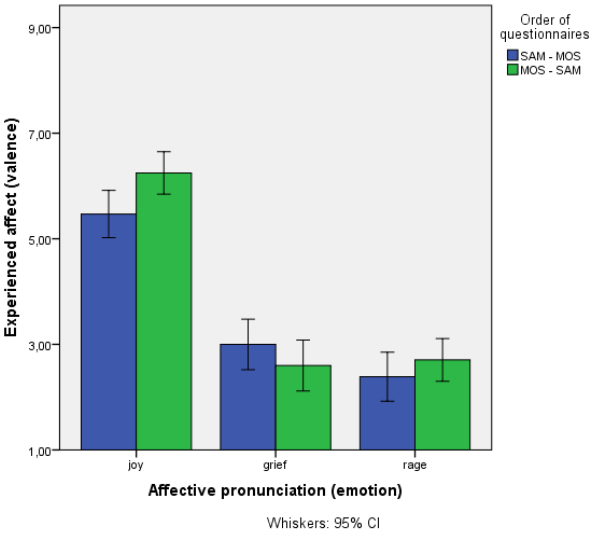

Figure 4: *Ratings on the valence scale for the different affective pronunciations (joy, grief and rage) and for the two questionnaire orders (SAM-MOS and MOS-SAM). Error bars indicate the 95% confidence intervals.*

To encounter effects due to the between-subject design of the study we analyzed differences between the two groups of subjects: order group SAM-MOS vs. order group MOS-SAM. For that analysis we divided the data in two new sub-groups. Stimuli with the bit rate 6.6, 8.85, and 12.65 Kbit/s were ascribed to the sub-group "low-quality" and stimuli with the bit rate 23.05 Kbit/s and the reference stimulus in wide-band quality were ascribed to the second sub-group "high-quality". The analysis showed no significant differences between the two stimulus classes (low and high quality), and the two groups of subjects SAM-MOS and MOS-SAM (see Figure 5). The interaction *order of questionnaires*stimulus sub-group* was not significant ($F_{1,198} = 0.043$, p=.83).

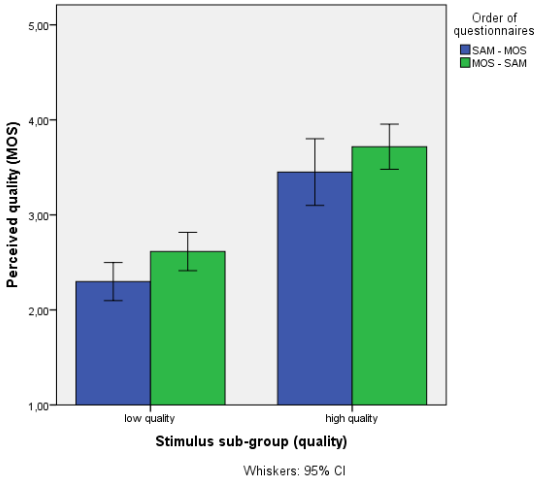

Figure 5: *MOS for the two stimulus sub-groups (low and high quality) and for the two questionnaire orders (SAM-MOS and MOS-SAM). Error bars indicate the 95% confidence intervals.*

4. Discussion and conclusions

As expected, we could show that the *stimulus quality* had an effect on the perceived quality and *affective pronunciations* had an effect on the experienced affect. These analyses were considered as a sanity check. In addition we could show that if the quality was judged directly after stimulus presentation, ratings were elevated. We ascribe this effect to the fact that subjects if first rated the experienced affect were more conservative rating the quality afterwards i.e. rated a lower quality.

For the experienced affect we observed a variety of effects: ratings on the dominance scale were elevated for direct assessment of experienced affect for rage, and in contrast lower for joy and grief. This is due to the fact that rage is associated with high arousal and increased dominance, but both feelings apparently are not very long-lasting. The contrary is true for the affective pronunciations *joy* and *grief*; with more time between the presentation and rating, subjects felt less non-dominant during presentation in hindsight. Even though there was no clear trend for the arousal and valence ratings recognizable, there was an influence of the order of questionnaires present. For the scale arousal we showed a increase for the affective pronunciation *rage* if the experienced affect was assessed directly. Again, the impact of the *affective pronunciation* was bigger if directly assessed, but if time passed due to prior assessment of perceived quality, the arousal level of subjects got lower.

Despite that there was no difference between the two groups of subjects (SAM-MOS and MOS-SAM) for the two stimulus sub-groups (low and high quality), one important issue needs to be addressed: due to the fact that the order of questionnaires was assigned to two different groups of subjects and the fact that the sample size was rather small, no generalization of the mentioned effects is possible without consideration. The main intend of this paper is not to show a general effect for the population but rather show that the effects of questionnaire order should be taken into account while planning studies in the quality domain.

We summarize the gathered results in the following guidelines: A) the order of questionnaires has a significant influence on the perceived quality and experienced affect, which has to be taken into account during the planning of experiments: both, perceived quality as well as experienced affect are obviously evanescent phenomena, and thus the aspect that is of higher importance in the study should be rated first. B) the influence was considerable large: on average 0.296 (7.4 % of scale range) on a MOS-scale ranging from 1 to 5 and up to 1.8 (22.5 % of scale range) for the SAM-scales ranging from 1 to 9. C) the effect on the experienced affect will be moderated by the emotional content of used stimuli. One possible solution to overcome this undesired effect of questionnaire sequence could be to randomize the order of questionnaires for stimuli of the same class (emotion or quality) within subjects. Order effects will not be prevented but distributed by counterbalancing across the sample or within one subject over conditions.

5. Acknowledgments

We are thankful to Ahmad Abbas for his support during data acquisition and to Sebastian Möller for technical advices. This work was co-funded by the Bernstein Focus: Neurotechnology Berlin (BFNT-B) which was supported by the Federal Ministry of Education and Research (BMBF) FKZ 01GQ0850 and the German Research Foundation (DFG - 1013 Prospective Design of Human-Technology Interaction).

6. References

[1] P. Le Callet, S. Möller, and A. Perkis, "Qualinet White Paper on Definitions of Quality of Experience (QoE) and Related Concepts." Dagstuhl, Germany, 2012.

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

[3] M. M. Bradley and P. J. Lang, "Measuring emotion: the Self-Assessment Manikin and the Semantic Differential," Journal of Behavior Therapy and Experimental Psychiatry, vol. 25, no. 1, pp. 49–59, 1994.

[4] "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)"', ITU-T Recommendation G.722.2, International Telecommunication Union, Geneva, 2002.

[5] P. J. Lang, "Behavioral treatment and bio-behavioral assessment: computer applications", in J. Sidowski, J. Johnson, and T.Williams (Eds.), "Technology in mental health care delivery systems", pp. 119-l37, NJ, 1980.