prof. Ing. Peter Počta, PhD.

University of Žilina

Faculty of Electrical Engineering and Information Technology

Department of Multimedia and Information-Communication Technology

8215/1 Univerzitná

Žilina, Slovakia

Žilina, March 16th 2021

A reviewer report for a Ph.D. thesis of Hakob Avetisyan entitled Parallel Task in Subjective Audio Quality and Speech Intelligibility Assessments

The Ph.D. thesis deals with a comparison of different laboratory test results of subjective speech quality measurement in terms of repeatability, call duration dependency on the quality of VoIP calls, parallel task in subjective speech quality and intelligibility measurement and testing the speech intelligibility for Czech language. In my view, the research topics covered by the thesis are actual and of interest of the quality community, to be more precise both parts of the community, i.e. academia and industry.

The first part of the thesis clearly describes state-of-the-art of the topics to be researched in the thesis and objective of the thesis and research to be done. Moreover, it also provides a technical background for the topics to be researched in the thesis. So, I have found this part useful for the interested reader, which does not have the particular background.

The second part of the thesis describes a plenty of the subjective (mostly) experiments done by the Ph.D. student and their results. First of them focuses on the comparison of different laboratory test results of subjective speech quality measurement in terms of repeatability. The results of the tests show the level of inter-lab and intra-lab repeatability when the identical test speech samples were deployed and the requirements of the ITU-T Rec. P.800 and ITU-T Rec. P.835 were strictly followed. This confirms that the performed tests were highly repeatable. The results of this experiment were published in the renowned International Journal of Speech Technology. It is worth noting here that this experiment was done in a cooperation with the Boston University. Second one has analysed the call duration dependency on the quality of VoIP calls. It turns out that alongside the assumption that better user experience brings longer call durations is valid only for moderate to high quality, and lower quality also yields longer call durations, which contradicts common expectations. This analysis was published in the well-renowned IEEE Wireless Communications Letters. It is worth noting here that this analysis was done in a cooperation with the German company called Voipfuture. Third and fourth ones deal with the parallel task in subjective speech intelligibility and quality measurement, respectively. In both cases, a novel subjective testing methodology has been designed and demonstrated. The purpose of the parallel task during subjective testing was to bring the test results closer to realistic conditions. When it comes to the speech quality measurement, the test results were highly correlated, certain conditions indicate different pair rankings after the parallel task was introduced. The resulting analysis indicated voting mistakes because of loss of subjects' concentration due to parallel task introduction. Therefore, it was concluded that the ITU-T Rec. P.835 methodology is too complicated to be combined successfully with a complex parallel task. Regarding the speech intelligibility measurement, there were certain samples where intelligibility values were counterintuitive, which proves that tests done in laboratory conditions cannot be considered as an etalon of speech intelligibility testing, and parallel-task techniques are highly recommended for subjective speech intelligibility evaluation. These two studies were published in the renowned Acta Acustica united with Acustica and PLOS ONE journals. Fifth experiment deals with a design and performance evaluation of the speech intelligibility tests in the Czech language. As before, these results also had counterintuitive values. However, all of them were statistically insignificant since they were in ranges of the Standard Deviation of Arithmetical Mean. This study was published in the renowned Journal of Audio Engineering Society. To sum up, all the experiments provide very interesting results, which are of great value for the corresponding research community. I strongly believe that the ITU-T SG12 and ETSI TC STQ (as far as I know some of the results were already presented to the ETSI TC STQ and the ETSI TC STQ has found them very interesting) would be highly interested in these results.

Despite some minor issues, see my comments below for more detail, I have found all the presented results useful for the research community. When it comes to the objective of this thesis, more precisely to its fulfilment, the objective detailed in chapter 1.4 of the thesis was, in my view, fulfilled. Regarding the deployed approaches, I have found them all up-to-date and legitimate.

I have a couple of comments to the thesis and questions for the Ph.D. student. Please see them below.

Comments:

English is sometimes hard to follow, e.g. Page 3, Chapter 1.5, first paragraph, first sentence: .....“written in the format of a thesis by publication approved by the Dean of Faculty of Electrical Engineering”..., Page 11, Chapter 3.1, first paragraph, fifth sentence: ...”with the deployment of identical test speech samples, and in strictly followed rules and requirements of ITU-T P.800 and ITU-T P.835.”, Page 57, Chapter 8.1.1, first sentence: “The main contributions in the field mentioned in this thesis are”...

Page 56, Chapter 8.1, fourth paragraph, third sentence: I would be quite careful with the wording here. Maybe, there is another reason for that inconsistency. According to the analysis, which is presented in the paper, the corresponding samples are rather specific. So, it can be simply a reason for this discrepancy.

Questions:

Page 18, Chapter 4.1, first paragraph, seventh sentence: I would be quite careful with the wording here. I think that it is more legitimate to say "better quality perceived by the end user" instead of "better user experience" as not all aspects of QoE are considered in the study. Unfortunately, it is not fully clear from the paper, how the MOS scores were collected. According the provided description, I assume that the MOS scores represent estimates of the E-model. Is this assumption correct?

Page 31, Chapter 6.1, first paragraph, ninth sentence: When it comes to the reported voting mistakes, I am just wondering, why have you selected the P.835 methodology for this study? In my experience, it is the most demanding methodology when it comes to the speech quality testing and some subjects find it rather difficult even without the parallel-task. To sum up, I would go for a less complicated/demanding methodology, e.g. the ACR test defined in the ITU-T Rec. P.800.

Page 57, Chapter 8.1.1, fifth bullet point: Was that really a recommendation, what was sent to the ETSI for a standardization process??? Anyhow, I very appreciate the fact that some of the results presented in thesis had made a basis for the ETSI Technical Reference.

To sum up, as the Ph.D. thesis contains an original and published research results of the author of the thesis and also fulfils the conditions of independent research work, I recommend this thesis for a final defence.

prof. Ing. Peter Počta, PhD.