Speech Quality Assessment in Crowdsourcing: Influence of Environmental Noise

Babak Naderi, Sebastian Möller, , Gabriel Mittag
Quality and Usability Lab, Technische Universität Berlin, Germany
E-mail:{babak.naderi, sebastian.moeller, gabriel.mittag}@tu-berlin.de

Abstract

Micro-task crowdsourcing opens up new possibilities for investigating the influence of a variety of realistic environmental factors on the quality of transmitted speech as perceived by the user. This paper reports the influence of environmental noise on speech quality assessment ratings using crowdsourcing approach. In a two-phase experiment, subjects assessed the quality of speech stimuli from a standard dataset (SwissQual 501 speech database from the ITU-T Rec. P.863 competition) in different environments. Phase A was conducted in the laboratory, in either silent or simulated environments with background noise. In phase B, the same group of participants completed the same task in different crowdsourcing environments. The Mean Opinion Score (MOS) values, representing perceived overall quality, were calculated for each degradation condition and compared to the scores reported from the standard laboratory test. The highest correlation with standard laboratory test was achieved in the silent-laboratory environment ($r_s = .97$). In the noisy (simulated) environments higher correlation was achieved when subjects were wearing in-ear headphones, and in crowdsourcing condition when they were performing their task in their living-room. It was also discovered that perceived loudness of the stimuli negatively correlates with the difference between MOS values obtained in test environmental conditions and the MOS values reported in the standard laboratory.

Introduction

The quality of transmitted speech as perceived by the user, the so-called Quality of Experience (QoE) [1], is a key performance indicator for the telecommunication system providers. ITU Telecommunication Standardization Sector (ITU-T) provides a Recommendation on methods for subjective determination of transmission quality including both active (conversation-opinion tests) and passive (Listening-opinion tests -LOT) tests in controlled laboratory environment [2]. Absolute Category Rating (ACR) is the recommended LOT method. Within ITU-T P.800 Recommendation properties of recommended listening environment and system are listed including the environmental noise which "should be set to the appropriate level" and measured at least twice (at the beginning and end of the experiment) [2]. However, LOTs conducted in a laboratory setting exhibit some limitations, as they are time intensive, expensive and has limited external validity [3].

Meanwhile, micro-task crowdsourcing provides a remarkable opportunity for academic and industry sectors by offering a high scale, on demand and a low-cost pool of geographically distributed crowdworkers that can participate in online QoE experiments in different environment [4]. Therefore, in contrast to the laboratory test, there is no control over surrounding environment and the devices used by participants in crowdtesting. The surrounding environment of worker (being noise or quite) and used system may influence the result of speech quality assessment in crowdsourcing. In this paper, we investigate the influence of environmental noise on a standard ACR task performed in crowdsourcing approach.

This paper organized as following. In the next section, the study design is briefly explained following by the employed dataset, different study conditions, and how environments were simulated in the laboratory. Next, data collection and screening procedures are shortly described. Results are presented in two sections. First, a comparison between different study condition and the standard laboratory ratings are given, later a model for predicting the difference between MOS values collected in different environments and quite laboratory condition is presented. The last section contains the conclusion and a short discussion about the results.

Method

A two-phase experiment has been conducted to evaluate influence of background noise in the environment on the speech quality rating. Phase A was performed in a controlled laboratory environment and phase B in crowdsourcing. Each phase includes two different sessions. 56 participants were randomly assigned each to one laboratory and one crowdsourcing session. Within each they performed one or more ACR task using a standard database in different environmental conditions and using different devices. Correlations between MOS collected in each condition with laboratory ratings and the Root Mean Square Deviation (RMSD), are used to evaluate the influence of listening environment and device on the given ratings. In the following, each part is explained in detail.

Dataset

For this experiment, we used subset of the SwissQual 501 database from the ITU-T Rec. P.863 competition, which has been kindly provided by SwissQual AG, Solothurn. This database includes variable types of degradations and degradation combinations and prepared base of the ITU-T Rec. P.800 specification. Within the dataset, 200 stimuli are arranged to carry 50 conditions. Each condition describes one degradation or a combination of degradati-

ons and each is composed of four stimuli (with the same degradation) recorded by four speakers with four different German sentences. The database contains 24 quality ratings from German natives per stimulus, which were obtained in a lab environment in accordance with ITU-T Rec. P.800. The resulting MOS per stimulus and test condition serve as a reference.

For this experiment 20 conditions carried by 80 stimuli were used. The selected conditions include eight anchoring conditions and twelve other conditions that their MOS obtained in crowdsourcing studies were previously reported to be significantly different from their MOS obtained in the laboratory test [3].

Environmental conditions under study

Each participant was randomly assigned to one laboratory and one crowdsourcing session which performed in a random order. Participants were asked to use their own device (Android based smartphone and headphone/earphone) or a provided device (Samsung Galaxy S4 and a professional high-impedance studio headphone[1]). In the following each session is explained in details. Table 1 summarize the experimental design.

Lab1- Laboratory Quiet: Participants assigned to this session assessed the quality of all stimuli in the dataset once with their own listening device (E1) and once with the provided device (E2) in a quiet laboratory room meeting P.800 conditions. The whole procedure took about 60 minutes.

Lab2- Laboratory Simulated: Same as Lab1, except environmental noise is played to simulate being in a crowdsourcing ambience. While handling a shorter testing process, environmental noise was changed twice (E3-4). The order of the two noises was picked randomly. After first two sequences were handled with own equipment, environmental noise was randomly changed to one of the first two noises and listening equipment was changed to laboratory headphones (E5-6). All 80 audio files were assessed at least once and the whole procedure took about 60 minutes

CS1- Crowdsourcing Specified environment: Participants were asked to perform an ACR test in an instructed environment not inside the lab. Two different environments were selected "living room" (E7) or "café/cafeteria" (E8). They used their own Android smartphone and headphones and all 80 audio files were assessed once.

CS2- Crowdsourcing not specified environment: Participants were not instructed to perform the ACR test in a specific environment. During this procedure, they used own Android smartphone and headphones and assessed all 80 audio files.

Environment simulation

Within session Lab2, two-different surrounding noise were simulated in the controlled laboratory environment as our aim was to investigate the influence of disturbance caused by environmental noise on the quality ratings. Following background noises, provided in ETSI TS 103 224 [5] were used to simulate two environments:

- Outside Traffic Street Noise: "Crossroad" (64.7 dBA),
- Public Places Noise: "Cafeteria" (62.7 dBA).

To implement the correct playback of the background noises, a simple setup of a pair of active stereo loudspeakers[2] was positioned in an angle of 45 degrees to the subject's position in the middle of the room. The loudspeakers were mounted on 1.4 meter stands, each in 2.0 meters distance, providing the creation of ideal stereo audio perception. The room setup is illustrated in Figure 1. Since a simple stereo setup was preferred, only two channels of the noise audio files were used. The sound pressure levels were checked regularly during the study to guarantee consistent testing conditions

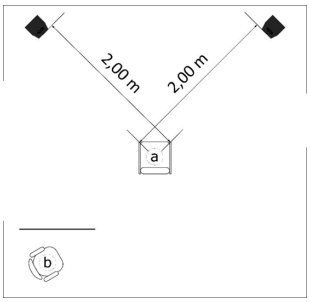

Abbildung 1: Room setup. a) participant's chair, b)supervisor's chair during experiment

Data collection and results

Overall, 56 participants (32 f, 24 m, M_{age}=31.2 y., German native, no hearing impairment) were randomly assigned to eight study groups. Participants in each study group perform one laboratory session and one crowdsourcing session in a specified order. For both laboratory and crowdsourcing studies the infrastructure of the Crowdee mobile crowdsourcing platform[3] were used. Corresponding qualification, training and ACR jobs were created following principles specified in [3]. In each ACR task, participants assessed the speech quality of nine stimuli (including one trapping question, see [3] for details).

Overall six participants did not complete their tasks (i.e. not participating in the second phase of study), and responses from one participant were removed based on the answers provided to trapping questions. For all listening environments and system conditions, MOS values were calculated for each of the 80 stimuli, and 20 degradation conditions.

Comparison between conditions

Spearman's rank-order correlations and RMSD between the MOS ratings obtained from each study condition

[1] AKG K702 (open-back) was used.

[2] Genelec 8040A Studio Monitor

[3] https://crowdee.de

Tabelle 1: Overview of conditions under study

Session	Condition	Location	Environment	Listening system	Participants
Lab1	E1 E2	Laboratory	Quiet	Own PRO.	28
Lab2	E3 E4 E5-6	Laboratory	Simulated Cafeteria Simulated Crossroad, Simulated (either)	Own Own PRO.	28
CS1	E7 E8	Crowdsourcing	Living room Cafeteria	Own Own	28
CS2	E9	Crowdsourcing	Not specified	Own	28

and the MOS ratings reported in the original dataset (in the SwissQual lab) were computed (Table 2). Results show that ratings collected in the quiet laboratory condition strongly correlate with the ratings reported from standard laboratory despite the listening device used ($r_s = .972$ and $r_s = .964$ for own and professional headset respectively). Following by results from crowdsourcing study either when workers are instructed to perform task in their living room at home, or not instructed ($r_s = .9$). Worse results were achieved in the laboratory environment with simulated background noise. When participants used their own headset their ratings correlate higher with the standard laboratory ratings. It was expected as most of participants brought their in-ear headset whereas the provided headset was open-backed.

Predicting MOS differences

To examine which factors have the largest impact on the deviations between MOS values obtained from the crowd and the ones from the lab, we build a model that predicts the difference between the MOS values: $\Delta MOS = MOS_{\text{Crowd}} - MOS_{\text{Lab}}$. Different input features were available for this purpose:

- Subjective ratings of the environment: Three questions about the loudness, noisiness and disturbance of the environment that was present during the crowd task, rated on a scale from 1 to 5.
- Subjective ratings of the stimuli: Overall quality MOS ratings from a previous lab experiment, and also the MOS ratings of their individual quality dimensions Noisiness, Coloration, Discontinuity, and Loudness.

Based on these features we found that the overall quality and the quality in terms of loudness has the largest influence on the difference, yielding following linear regression model:

$$\Delta\widehat{MOS} = 0.439 - 0.215MOS_{\text{loud}} + 0.135MOS, \quad (1)$$

with a resulting prediction error of RMSE = 0.91. As can be seen from the regression coefficients, the quality dimension loudness has the largest influence on the difference prediction. Figure 2 shows how participants from the crowd tend to overate conditions with a low loudness

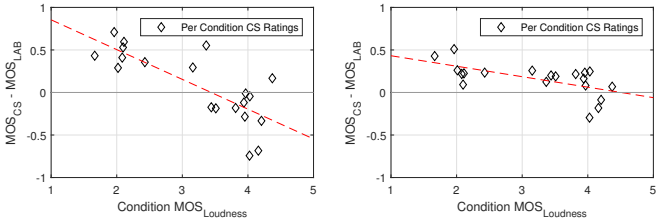

Abbildung 2: Difference vs Loudness MOS. *LEFT:* Crowd in cafe environment. *RIGHT:* Crowd in living room environment

score. A low loudness rating in the database under study is usually caused by a condition that was perceived as too quiet. Thus, we can conclude that, compared to the lab, disturbances in the conditions with a low loudness level are not perceived as strongly by the participants in the crowd. We assume that the environmental noises, which the crowdworkers are exposed to, conceal the disturbances in low loudness conditions. This can also be seen by comparing the two graphs in Figure 2, on the left side the scores from crowdworkers in a cafe are seen. Here the influence of MOS_{loud} is much stronger than in the quieter living room environment on the right hand-side. Due to these effects we calculated separate models depending on the environment of the crowdworker, based on their loudness rating of the environment (resulting in five different model fit). Applied on all files (café, living room, and random environment) we receive a reduced prediction error of RMSE = 0.86.

Discussion and Conclusion

The study reported in this paper intended to answers questions about influence of environmental noise on the speech quality assessment task performed in the crowdsourcing approach. In the two phase study, participants performed speech quality rating task in different environmental conditions including quite laboratory, simulated noisy in laboratory and in the field Results show that ratings obtained within laboratory quite condition strongly agreed with ratings reported by a standard laboratory test (according to ITU-T P.800) although participants used smartphone as a listening device and different headset in current study ($r_s = .96$). Meanwhile, environmental noise influences the judgment of participant and

Tabelle 2: Comparison between MOS values obtained in each study condition and the MOS values reported from standard laboratory.

Location	Environment	Listening through	Standard Laboratory Test	
			$r_s(p)$	RMSD
Laboratory	Quiet	Own	.972 ($p < .001$)	0.24
		PRO.	.964 ($p < .001$)	0.25
	Simulated Cafeteria	Own	.738 ($p < .001$)	0.7
		PRO.	.54 (.014)	1.01
	Simulated Crossroad	Own	.657 (.002)	0.83
		PRO.	.472 (.036)	1.91
Crowdsourcing	Cafeteria	Own	.765 ($p < .001$)	0.62
	Living room	Own	.903 ($p < .001$)	0.42
	Not specified	Own	.907 ($p < .001$)	0.39

the type of headphone they wear may reduce that effect. Ratings in simulated noisy environment were more inline with standard laboratory when participants used their own headphone (mostly were in-ear) rather than open-back laboratory headphone. Furthermore normal crowdsourcing situation (when workers were not structured to be in specific place) strongly agree with laboratory test with same group of participants ($r_s = .96$, RMSE= 0.25). Further investigations shows that loudness of environment, perceived loudness of stimulus, and over all quality of it predict the difference between ratings obtained in this study and ratings reported by standard laboratory test.

For future works, we would like to investigate usage of environment-recording for predicting the appropriateness of environment for performing speech quality assessment task. Although in current study participants recorded 10 seconds of environmental noise before and after each rating session, considerable number of recordings were invalid (participant touched the phone during recording, talked, or performed different gestures).

Literatur

[1] S. Möller and A. Raake, Eds., *Quality of Experience*. Cham: Springer International Publishing, 2014.

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

[3] Naderi, B., Polzehl, T., Wechsung, I., Köster, F., Möller, S.: *Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm*. In: 16th Ann. Conf. of the Int. Speech Comm. Assoc. (Interspeech 2015). ISCA, pp. 2799–2803 (2015)

[4] Naderi, B. *Motivation of Workers on Microtask Crowdsourcing Platforms*. Springer International Publishing, 2018.

[5] ETSI TS 103 224, *Speech and multimedia Transmission Quality (STQ); A sound field reproduction method for terminal testing including a background noise database*. FR–Sophia Antipolis: European Telecommunications Standards Institute, 2014.