Collecting Subjective Ratings of Voice Likability: Laboratory vs. Crowdsourcing

Rafael Zequeira Jiménez, Laura Fernández Gallardo, Sebastian Möller
Quality and Usability Lab, TU-Berlin Berlin, Germany
Email: {rafael.zequeira, laura.fernandezgallardo, sebastian.moeller}@tu-berlin.de

***Abstract*—Crowdsourcing has become a powerful approach for rapid collection of user input from a large set of participants at low cost. While previous studies have investigated the acceptability of crowdsourcing for obtaining reliable perceptual scores of audio or video quality, this work examines the suitability of crowdsourcing to collect voice likability ratings. We describe our conducted tests based on direct scaling and on paired-comparisons, that were executed in crowdsourcing using micro-tasks and in the laboratory under controlled conditions. Design considerations are proposed for adapting the laboratory listening tests to a mobile-based crowdsourcing platform to obtain trustworthy listeners' answers. The likability scores obtained by the different test approaches are highly correlated. This outcome motivates the use of crowdsourcing for future listening tests investigating e.g. speaker characterization, reducing the costs involved in engaging participants and administering the test on-site.**

I. Introduction

The micro-task crowdsourcing (CS) paradigm offers small tasks to anonymous users on the Internet that normally require human intelligence for being resolved. The users can carry out those micro-tasks from their computer or from their mobile device, and they get rewarded after completion. This approach is being adopted in multiple domains to collect human input for data acquisition and labeling. Experiments conventionally executed in a laboratory (lab) setup can now be addressed to a wider and diverse audience. However, it remains the question of whether the CS outcomes are valid and reliable, that is, comparable to those obtained in a constrained and quiet environment.

This work investigates the validity of CS for collecting non-expert subjective voice likability scores contrasting the results with in-lab conducted listening tests. Voice likability, or voice pleasantness, can be viewed as a speaker social characteristic that can determine the listener's attitudes and decisions towards the speaker and their message. The collection of valid voice likability labels is crucial for a successful automatic prediction of likability from speech features. This work presents different auditory tests that collect likability ratings of a common set of 30 voices. Four experiments were conducted, two in the lab and two via CS. In the first in-lab test, a continuous scale was presented to the listeners, on which to indicate the degree of likability of each of the utterances presented. This study is described in [1]. The second in-lab test, detailed in [2] adopted a paired-comparison (PC) approach, in which the speech samples were presented in pairs and the listeners were asked to decide which one was more likable. The first and the second lab tests will be referred in the following to as Lab-SCA and Lab-PC, respectively.

II. Crowdsourcing Experiments and Results

The CS experiments were conducted using the mobile-based Crowdee platform [3]. Inspired by the work in [4], [5], we designed a qualification micro-task for the users to earn access to the study and we included also trapping and control questions for quality control. Details about the considerations taken to adapt the Lab-PC and the Lab-SCA to CS can be found in [6] and in [7], respectively.

The results of [6] presents a strong and statistically significant correlation (Pearson $r = 0.95$, $p < 0.001$, standard error (SE) $= 0.09$) of the CS scores with the voice preference results gathered in the lab. Fig. 1(a) (c) shows screenshots of the scales presented to the users in the lab and in CS for the PC study. A preference choice matrix was built and the Bradley-Terry-Luce (BTL) probabilistic choice model [8], [9] was applied to derive the ratio scale measures of voice likability, using the R package 'eba'. The BTL model were successfully created for the lab and for the CS results and a meaningful ordering of listeners' preferences could be derived in the form of utility scale (v-scale) values by probabilistic choice modeling, shown in Fig. 2. The same tendency was observed among the listeners' preferences for the CS and for the lab experiment.

Furthermore, the mean scores results presented in [7] about SCA (lab vs. CS), were also correlated: (Pearson $r = 0.68$, $p < 0.005$, SE$= 0.20$ and Pearson $r = 0.89$, $p < 0.001$, SE$= 0.13$ for male and for female speakers, respectively). Fig. 1(b) (d) shows screenshots of the scales employed by the users for rating in lab and in CS. Fig. 3 represents the mean likability ratings obtained for each speaker.

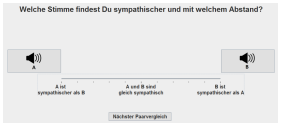
References

[1] L. Fernández Gallardo and B. Weiss, "Speech Likability and Personality-Based Social Relations: A Round-Robin Analysis over Communication Channels," in *INTERSPEECH*, 2016, pp. 903–907.

[2] L. Fernández Gallardo, "A Paired-Comparison Listening Test for Collecting Voice Likability Scores," in *Speech Communication; 12. ITG Symposium*, oct 2016, pp. 185–189.

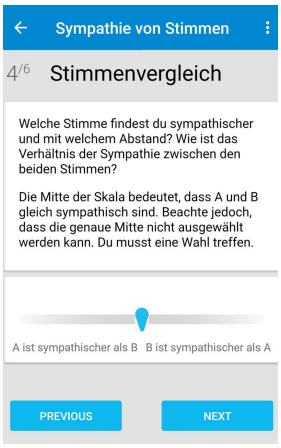
[3] B. Naderi, T. Polzehl, A. Beyer, T. Pilz, and S. Möller, "Crowdee: mobile crowdsourcing micro-task platform for celebrating the diversity of languages," in *INTERSPEECH*, 2014, pp. 1496–1497.

[4] T. Polzehl, B. Naderi, F. Köster, and S. Möller, "Robustness in speech quality assessment and temporal training expiry in mobile crowdsourcing environments," in *INTERSPEECH*, 2015, pp. 2794–2798.

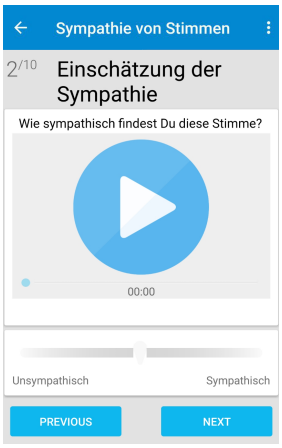


(a) Lab-PC

(b) Lab-SCA



(c) CS-PC



(d) CS-SCA

Fig. 1: Scaling and paired-comparison tasks from the laboratory and crowdsourcing experiments.

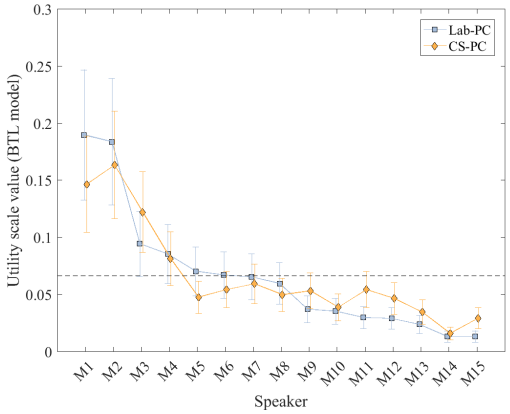


Fig. 2: Ratio scale preferences estimated by the Bradley-Terry-Luce model for the paired-comparison tests conducted in laboratory and via CS. Error bars show 95% confidence intervals. The indifference line is plotted as $y = 1/Nspeakers$.

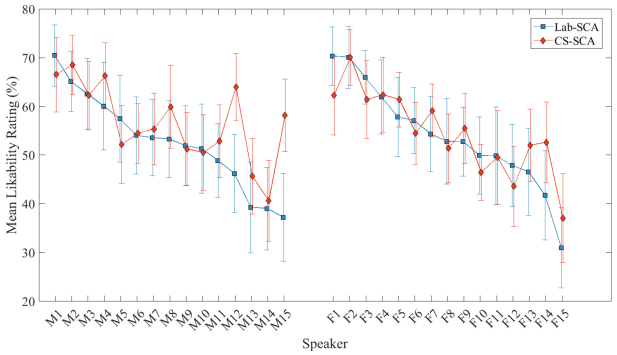


Fig. 3: Mean likability scores for male and female speakers from the direct scaling test in laboratory and using crowdsourcing. Error bars show 95% confidence intervals.

[5] B. Naderi, T. Polzehl, I. Wechsung, F. Köster, and S. Möller, "Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm," in *INTERSPEECH*. ISCA, 2015, pp. 2799–2803.

[6] R. Zequeira Jiménez, L. Fernández Gallardo, and S. Möller, "Scoring Voice Likability using Pair-Comparison: Laboratory vs. Crowdsourcing Approach," in *accepted for: Int. Conf. on Quality of Multimedia Experience (QoMEX)*, 2017.

[7] L. Fernández Gallardo, R. Zequeira Jiménez, and S. Möller, "Perceptual Ratings of Voice Likability Collected through In-Lab Listening Tests vs. Mobile-Based Crowdsourcing," in *submitted to: INTERSPEECH*, 2017.

[8] R. A. Bradley and M. E. Terry, "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[9] R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.