Pair-Comparison for Collecting Voice Likability Ratings: 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 established itself as a powerful tool being currently adopted in multiple domains as a means to collect human input for data acquisition and labelling. Experiments conventionally executed in a laboratory setup can now be addressed to a wider audience while controlling its diversity. However, it remains the question of whether the crowdsourcing outcomes are valid and reliable, that is, comparable to those obtained in a constrained and quiet environment. This paper presents a study performed both in a laboratory and on a mobile-crowdsourcing platform, adopting a paired-comparison setup to obtain ratings of voice likability. We show considerations taken to adequately adapt the laboratory-based test to the remote-labour approach. Once all pair-comparison answers were collected, preference choice matrices were built and the Bradley-Terry-Luce probabilistic choice model was applied to estimate a ratio scale of preferences, reflecting the voice likability scores. Our results show a strong correlation between the scores obtained by the two approaches considered, which indicates the validity of crowdsourcing for the acquisition of voice likability ratings. This is of great benefit when datasets need to be quickly and reliably labeled for speech applications relying on detection or on synthesis of speaker and voice characteristics.**

I. Introduction

In the context of Quality of Experience (QoE), research efforts concentrate on understanding subjective user preferences in terms of expectations fulfillment, perceived quality and enjoyment of the multimedia content. Due to the high subjective dimension of QoE and its user-centric nature, the interest in using crowdsourcing (CS) as a fast, low cost, and scalable tool for QoE studies has increased considerably in the recent years [1].

This work proposes the use of CS for subjective evaluations of speech. Specifically, we focused on user preferences towards voices, i.e. voice likability, which we consider to play an important role on QoE judgments in scenarios involving e.g. virtual or robotic conversational agents, personal assistance, advertisements in public places or interactive voice response systems [2]. Appealing or likable voices in such systems are desired to attract the listeners' attention and to generate favorable attitudes towards a voice message. In this paper, we explore the suitability of CS for collecting reliable subjective voice likability ratings in contrast to typical data labeling by a panel of listeners in laboratory (lab) environments [3]. Research on speech features can be based on the collected labels and contribute to "make a voice likable" [4] or to detect likable voices automatically.

II. Related Work

Previous work has shown that CS can provide reliable measures of QoE for image [5], [6], video [7], [8], [9], speech [10] and audio [11] applications. In [6] the author introduces CS for collecting ground truth on image appeal, and in [5] image quality evaluations are addressed. On the other hand, work in [7], [8] uses CS to evaluate users' perceived quality in video streaming, and [9] presents the results of an experiment in subjective video quality judgment. Research in [11] compares the results of conducting perceptual audio evaluations task in lab and CS environments. And work on [12] presents different techniques to achieve reliable results in speech quality assessment tasks.

The use of pair-comparison settings in CS for QoE judgments of multimedia contents has been also of interest in multiple studies. The work in [13] introduces a pair-comparison framework for quantifying QoE in multimedia as a more convenient approach for CS, due to the easiness of the task in contrast to 5-point scale rating. Also, the videos in [7], [8], [9] to be evaluated by the users were presented in pairs.

Multiple research has explored the challenges of presenting to an online-crowd experiments analogous to previously conducted lab studies [13], [8]. The main challenges highlighted by these previous works were: adequating the experiment duration, ensuring users' appropriateness and trustworthiness for the study, and developing mechanisms to gather meaningful outcomes. Still, most of this work considers web-based CS platforms. To the best of our knowledge, it has not yet been addressed whether the same considerations apply to mobile-based CS environments. In contrast to web CS, in mobile CS users use a mobile application to find and perform micro-tasks.

There are challenges specific to user studies over mobile CS. Users might execute QoE assessment tasks wherever and whenever they want. Due to this flexibility, users could happen to perform a CS task in a noisy environment that may affect the results, e.g. public transportation, crowded rooms, open places [14]. Internet coverage also needs consideration, as it is sometimes unstable in mobile networks and the media content for the study might not be available all the time. Our research considers only mobile CS since it offers more realistic consideration of real-life influences and the possibility to reach the users faster than web-CS. We will disclose guidelines to overcome the presented challenges in Section III-B.

III. Experiment settings: lab and CS

We designed two experiments to be conducted in the lab and via CS in which participants had to indicate whether they liked one voice over the other in a pair-comparison setup. As speech material, we considered a short German sentence (mean duration 4.4 s) uttered by 15 male German speakers [15]. They were 26.3 years old on average (range: 21-31) and did not present strong deviations from High German dialect, which according to [4], might influence the perceptions of voice likability. The sentences were combined in $\binom{15}{2} = 105$ unique stimuli pairs and employed in our lab and CS studies as described in the following.

A. Lab Experiment

A total of 13 female listeners participated in the paired-comparison listening test. They were 27.8 years old on average (range: 20-34) and were all speakers of standard High German dialect. This study examined only cross-gender likings and no male listeners performed a similar test with female voices.

Each participant judged only one of the two orderings (”A”, ”B”) or (”B”, ”A”) of a given stimulus pair. All 105 pairs were presented randomly. Fig. 1 presents a screenshot of the graphical user interface presented to the listeners. The question to the listeners was (translated from German): ”Which voice do you find more likable, and to which extent?”. For listening to each stimulus the participants were asked to click on the buttons ”A” and ”B”. They could listen to the speech sample as many times as they wished, and they could not listen to ”B” before having listened to ”A”. The participants could resolve their preference for ”A” or ”B” by selecting a value on a slider (close to ”A” or close to ”B”, respectively) only after having listened to both voices at least once. Leaving the knob on the exact middle of the slider was not possible. After selecting a value on the slider and clicking on (translated from German): ”Next pair comparison”, the next stimulus pair was presented.

Since each of the test session took about 30 minutes to complete, a short pause was included after ending the first half of the pair-comparisons to avoid the listeners' fatigue. The participants were rewarded with 6 €, the overall cost being 78 €. More details on the laboratory conditions, equipment and the procedure followed for the experiment are given in [15].

B. Crowdsourcing Experiment

1) Mobile-based CS platform: The CS experiment was conducted using the Crowdee mobile-CS platform, which provides the opportunity for app-based CS micro-tasks assessments [16]. The application (app) to be used by the users (or workers) to perform micro-tasks (or jobs) is freely available in the Google Play Store with the name ”Crowdee”.

2) Transitioning the listening test to the crowd: For our study we set a German language filter to select only German speakers. Our goal was to use Crowdee to obtain the same number of responses to the 105 pair-comparisons as in the test conducted in the lab. Since the test was performed by 13 listeners, 1365 (105 x 13) ratings were to be collected from the crowd. Evidently, the same speech files as in the lab experiment were employed.

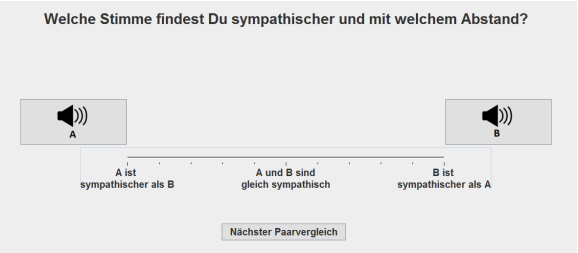


Fig. 1: Slider used in the lab for the paired-comparison test. The texts of the labels from left to right are ”A is more likable than B”, ”A and B are equally likable”, ”B is more likable than A” (written in German).

To transfer the study from the lab to a CS environment it was necessary to deal with:

- dividing the entire test into CS micro-tasks, each of them not lasting longer than 5 minutes [1] to avoid fatigue and boredom,
- user's trustworthiness: users work on experiments without supervision, and may thus give erroneous answers carelessly or dishonestly. This is one of the major challenges of QoE assessment using CS,
- controlling the use of two-eared headphones for the listening test instead of the device speakers,
- controlling the user's environment quietness.

To overcome these challenges some adaptations of the lab experimental protocol was needed. First of all, instead of 13 long tasks with 105 speech pairs each, our CS study consisted of 1365 micro-tasks with one pair-comparison each. We restricted that each user could perform up to 25 micro-tasks and they were never confronted with the same pair more than once.

Inspired by the work in [12], [17] we designed a qualification micro-task for the users to earn access to the pair-comparison assessment. We asked the users to perform the task in a quiet room and to wear two-eared headphones. Both the qualification and the pair-comparison micro-task could only be started when connecting headphones.

To check the user's obedience and commitment, we included a question in the qualification micro-task asking the users to record the environment for a minimum of seven seconds. These data were used to filter out those users who were located in a noisy environment (a loudness threshold was set empirically). Also, on the next screen we validated the use of the two-eared headphones by a short math exercise with digits panning left to right in stereo, they could select the answer from within 5 options that were randomized each time. After this, a control question about the speaker gender was added. The rest of the qualification micro-task included screens to gather socio-demographic information about the user.

Only after successfully executing the qualification micro-task, the users were assigned with a time frame of 60 minutes,

in which they could perform the pair-comparison micro-task at wish up to 25 times. When timeout was reached, the access to the speech comparisons was revoked so they needed to go through the qualification task again. Other control questions were included as well in the pair-comparison micro-task, detailed in the next subsection.

3) Collecting CS users' likability ratings: Each pair-comparison micro-task consisted of six screens: screens 1 and 2 presented text introducing the task. Screen 3 included the stimulus pair to be listened to and a single choice question about the speaker's gender in the stimulus (this answer was used later as a quality control check). Screen 4 presented the slider used to indicate to which extent one voice was preferred over the other. Screen 5 showed the single question (translated from German): "If you are reading this question, please choose the answer "agree more"", to answer this, seven mutually exclusive options were presented and randomized every time. This question was added for controlling the user focus on the study. Finally, Screen 6 displayed a goodbye message.

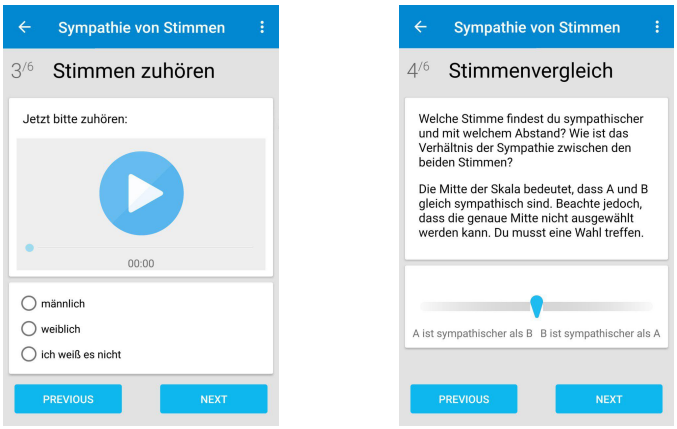
The time spent to perform one pair-comparison task was 32.4 s on average (range 11–209 s). Users were rewarded with 0.14 € after completion only if their recorded environmental noise was determined to be low and if they answered correctly to the control questions.

It was important to keep the CS experiment as close as possible to the lab study, since small changes in the test setting might lead to different or unwanted results [18]. Therefore, we used the same labels for the micro-task screens. One main difference between the test presentations was that two buttons permitted listening to the speech corresponding to each speaker given one pair, while only one play button was available in Crowdee due to technical limitations, Fig. 2(a). When the users clicked on that button, they listened to the concatenated two speakers' speech. Screenshots in Fig. 2 shows the screen presented to the user for listening to the speech pair and the screen with the slider for rating.

The Crowdee platform created the 1365 (105 x 13) micro-tasks to be performed by users. When answers from the users were received, our check and controls were applied to possibly reject untrustworthy answers. Each time an answer was rejected, the corresponding pair-comparison became automatically available (unsolved) for other users to provide a new answer. There were 317 micro-task answers rejected. Therefore, 1682 micro-tasks were created in total for this study. Those 1682 micro-tasks were performed by 92 unique users, 69 of them providing only accepted answers, 15 providing only rejected answers, and 8 providing both accepted and rejected answers.

IV. Results and Discussion

Ratio-scale likability scores, accounting for the listeners' preferences, are computed for the lab and for the CS tests independently. From the answers to the pair-comparison tasks, a preference choice matrix was built and the Bradley-Terry-Luce (BTL) probabilistic choice model [19], [20] was applied



(a) Speech Pair (b) Slider for rating

Fig. 2: Screenshots of the pair-comparison micro-task. (a) Play button to listen to the concatenated speech. (b) slider on which to indicate the preference for the first or for the second voice.

to derive the ratio scale measures of voice likability, using the R package 'eba'.

The BTL model implies a very strong form of stochastic transitivity. Given a triple of voices x, y, z for which the preference is determined as $x > y$ and $y > z$, a transitivity is violated if $x < z$. Transitivity violations reflect individually inconsistent choice behavior or disagreement between raters. The BTL model can only be fit if no systematic violations of the stochastic transitivity occur.

Weak (WST), moderate (MST), and strong (SST) stochastic transitivity violations were computed as follows. Let P_{xy} denote the empirical probability that voice x is chosen over voice y. The stochastic transitivity variants imply that if $P_{xy} \geqslant 0.5$ and $P_{yz} \geqslant 0.5$, then

$$P_{xz} \geqslant \begin{cases} 0.5 & \text{(WST)}, \\ min\{P_{xy}, P_{yz}\} & \text{(MST)}, \\ max\{P_{xy}, P_{yz}\} & \text{(SST)}, \end{cases} \quad (1)$$

for all unique triads x, y, z in the test.

Table I presents the number of violations encountered for the lab and the CS test. A greater number of violations can be observed for the CS study compared to the lab experiment. Even greater is the number of violations found when all CS answers are included in our analysis.

While WST violations may hinder the BTL representation, violations of the other stochastic transitivity variants are less severe [21]. The violations can originate from disagreement among raters or from inconsistencies in the answers from one particular rater. It was determined in [15] that all participants of the laboratory listening test were consistent, according to a computed Kendall's ζ coefficient [22]. This coefficient can only be calculated if all stimuli are presented to the listeners, which is not the case in CS. The violations for the lab test are therefore mainly due to disagreement among listeners. This disagreement could be expected given the subjectivity of indicating voice likability.

TABLE I: weak, moderate and strong stochastic transitivity violations for the lab and for the CS experiments. Number of tests = $\binom{15}{3} = 455$

	Laboratory	Crowdsourcing (only trustworthy answers)	Crowdsourcing (all answers)
WST	10	28	22
MST	28	64	100
SST	134	219	245

The R package 'eba' was employed for the analyses of the choice frequencies and computation of the BTL probabilistic choice model for the lab and the CS experiments. The likelihood ratio tests of the BTL models' fits were not significant ($p > 0.05$) in any case, which indicated that the transitivity violations encountered were random for the two experiments, i.e. not systematic [21]. Hence, the restrictive BTL models could be successfully fit for the lab and for the CS results. Therefore, a meaningful ordering of listeners' preferences could then be derived in the form of utility scale (v-scale) values by probabilistic choice modeling, shown in Fig. 3.

The same tendency can be observed among the listeners' preferences for the CS and for the lab experiment. The Pearson's product-moment correlation between the two score series resulted to be strong and significant, with $r = 0.95$ ($p < 0.001$) and standard error (SE) $= 0.09$. The Spearman correlation, which accounts for the strength and direction of the monotonic relationship between the two score series, yields $\rho = 0.89$ ($p < 0.001$) and $SE = 0.13$.

When considering all answers and not only those given by trustworthy users, the correlations between lab and CS likability scores were slightly worse, but still strong and significant: Pearson's $r = 0.92$ ($p < 0.001$), $SE = 0.11$, and Spearman's $\rho = 0.87$ ($p < 0.001$), $SE = 0.14$. This manifests that our proposed check and control questions have only been marginally useful. The minor decrease in the correlation might be due to the low number of untrustworthy users that participated in the CS study (15 out of 92 unique users). Still, regardless the dimension of the CS study, we strongly recommend the use of quality control mechanisms to control the trustworthiness of users, as also pointed out in [1].

Distance matrices were built from the values indicated by the slider answering the question "Which voice do you find more likable, and to which extent?" for lab and for CS, respectively. The non-parametric Mantel test [23] was applied to assess the correlation between the distance matrices created with the lab and the CS results (only trustworthy answers). For this, the R package 'ade4' was used. The distance matrices were first transformed into matrices with the positive eigenvalues of the Euclidean representation.

Whereas strong correlations have been found between lab and CS likability scores (derived from the preference of one voice over another), the distance matrices did not correlate ($r = 0.008$, $p = 0.46$). This suggests that lab and CS results can be similar for a task considered simple such as pair-comparisons [9], [8]. However, the lab results in terms of indicating the extent of liking one voice over the other on a slider could not be replicated in CS with our approach.

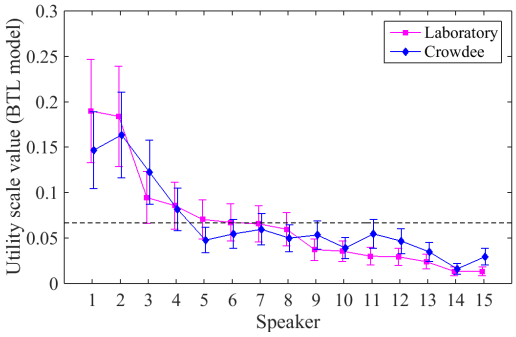


Fig. 3: Normalized v-scale values assigned to each speaker. Error bars show 95% confidence intervals and the indifference line is plotted as $y = 1/Nspeakers$.

As pointed out before, the cost of the experiment in the lab was 78 €, which is less than the amount expended in the Crowdee platform: 195.92 €. This resulted from 0.14 € payed to users who provided accepted answers (1365), plus 0,02 € for the completion of the qualification micro-task (241 times). When calculating the lab expenses we did not consider the costs of the previous preparation to the experiment, neither the expenses of paying the person administering the study in the lab. If these considerations are taken, we can assume that the costs of the lab experiment would have been higher than those of the CS experiment.

V. Conclusion & Future Work

In this paper, we have compared laboratory and crowdsourcing-based pair-comparison experiments for collecting scores of voice likability. We have found that users in crowdsourcing can be quite consistent with the lab participants, despite the lack of control compared to the laboratory experiment, particularly regarding test environment, equipment and user behavior.

We have proposed mechanisms to adapt pair-comparison tests using speech material, traditionally performed in controlled laboratory settings, to crowdsourcing without compromising the derived likability scores. Our approach allowed us to obtain a strong and statistically significant Pearson's correlation with $r = 0.95$ and standard error $= 0.09$ to the voice preference results gathered in the laboratory. However, the extent of preferring one voice over the other in each pair-comparison was not indicated via crowdsourcing compared as done by the participants of the laboratory experiment. This outcome indicates the suitability of crowdsourcing to perform paired-comparison tasks, and motivates further research into the promising capabilities of crowdsourcing for other kinds of auditory tests.

References

[1] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best Practices for QoE Crowdtesting: QoE Assessment

With Crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, feb 2014.

[2] F. Burkhardt, R. Huber, and B. Anton, *Application of Speaker Classification in Human Machine Dialog Systems*. Berlin, Heidelberg: Springer, 2007, pp. 174–179.

[3] F. Burkhardt, B. W. Schuller, B. Weiss, and F. Weninger, ""Would You Buy a Car from Me?" - On the Likability of Telephone Voices," in *INTERSPEECH*, 2011, pp. 1557–1560.

[4] B. Weiss and F. Burkhardt, "Is 'not bad' good enough? Aspects of unknown voices' likability," in *INTERSPEECH*, 2012, pp. 510–513.

[5] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *18th IEEE International Conference on Image Processing*, sep 2011, pp. 3097–3100.

[6] E. Siahaan, A. Hanjalic, and J. Redi, "A Reliable Methodology to Collect Ground Truth Data of Image Aesthetic Appeal," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1338–1350, jul 2016.

[7] M. Shahid, J. Søgaard, J. Pokhrel, K. Brunnström, K. Wang, S. Tavakoli, and N. Gracia, "Crowdsourcing based subjective quality assessment of adaptive video streaming," in *Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, sep 2014, pp. 53–54.

[8] J. Søgaard, M. Shahid, J. Pokhrel, and K. Brunnström, "On subjective quality assessment of adaptive video streaming via crowdsourcing and laboratory based experiments," *Multimedia Tools and Applications*, pp. 1–22, 2016.

[9] D. Saupe, F. Hahn, V. Hosu, I. Zingman, M. Rana, and S. Li, "Crowd workers proven useful: A comparative study of subjective video quality assessment," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[10] J. Parson, D. Braga, M. Tjalve, and J. Oh, *Evaluating Voice Quality and Speech Synthesis Using Crowdsourcing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 233–240.

[11] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman, "Fast and easy crowdsourced perceptual audio evaluation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, mar 2016, pp. 619–623.

[12] 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.

[13] K.-T. Chen, C.-J. Chang, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "Quadrant of Euphoria: A Crowdsourcing Platform for QoE Assessment," *IEEE Network*, vol. 24, no. 2, pp. 28–35, mar 2010.

[14] S. Möller, T. Hoßfeld, and B. Naderi, "Comments on P.CROWD," full, International Telecommunication Union, CH-Geneva, ITU-T contribution COM 12 C 386 E, jun 2016.

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

[16] 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.

[17] 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.

[18] J. Redi, E. Siahaan, P. Korshunov, J. Habigt, and T. Hoßfeld, "When the Crowd Challenges the Lab: Lessons Learnt from Subjective Studies on Image Aesthetic Appeal," *Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia*, pp. 33–38, 2015.

[19] 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.

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

[21] S. Choisel and F. M. Wickelmaier, "Evaluation of Multichannel Reproduced Sound: Scaling Auditory Attributes Underlying Listener Preference," *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 388–400, 2007.

[22] M. G. Kendall, *Rank Correlation Methods*, 4th ed. Charles Griffin, 1970.

[23] N. Mantel, "The Detection of Disease Clustering and a Generalized Regression Approach," *Cancer Research*, vol. 27, no. 2, pp. 209–220, 1967.