Perceptual Ratings of Voice Likability Collected through In-Lab Listening Tests vs. Mobile-Based Crowdsourcing

Laura Fernández Gallardo, Rafael Zequeira Jiménez, Sebastian Möller

Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany

(laura.fernandezgallardo|rafael.zequeira|sebastian.moeller)@tu-berlin.de

Abstract

Human perceptions of speaker characteristics, needed to perform automatic predictions from speech features, have generally been collected by conducting demanding in-lab listening tests under controlled conditions. Concurrently, crowdsourcing has emerged as a valuable approach for running user studies through surveys or quantitative ratings. Micro-task crowdsourcing markets enable the completion of small tasks (commonly of minutes or seconds), rewarding users with micro-payments. This paradigm permits effortless collection of user input from a large and diverse pool of participants at low cost. This paper presents different auditory tests for collecting perceptual voice likability ratings employing a common set of 30 male and female voices. These tests are based on direct scaling and on paired-comparisons, and were conducted in the laboratory and via crowdsourcing using micro-tasks. Design considerations are proposed for adapting the laboratory listening tests to a mobile-based crowdsourcing platform to obtain trustworthy listeners' answers. Our 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 efforts involved in engaging participants and administering the tests on-site.

Index Terms: speech corpus annotation, voice likability, crowdsourcing

1. Introduction

Speaker characterization has emerged as an important area of research to develop systems able to detect, predict, or synthesize human behavior, e.g. conversational agents for human-machine communications [1]. The present work focuses on human perceptions of voice likability, or voice pleasantness, which can be viewed as a speaker social characteristic that can determine the listener's attitudes and decisions towards speakers and their message. Gathering valid and precise ratings of speech likability perceptions [2] is crucial for the later automatic likability detection from speech features. These ratings are generally collected by conducting in-lab listening tests that permit the control, among other factors, over the room, equipment, and the listeners' behavior and understanding of the test instructions. Another possibility is to employ crowdsourcing (CS) techniques to outsource the test to a large number of individuals over the Internet.

The crowdsourcing paradigm offers small tasks (micro-tasks) to anonymous users that normally require human intelligence for being resolved. The users (also called workers) can perform the micro-tasks using their computer (web-based CS) or their mobile Internet-connected device (mobile-based CS), and they might get rewarded after completion. In the last decade, researchers have found in CS a fast, low cost, and scalable method to collect more data than traditional approaches. While CS can induce real-life environment conditions for performing user studies, there is loss of control to supervise the participant, and often little information on their playback system and background environment. To guarantee the quality of the result, considerations needs to be taken to monitor these factors to the extent possible.

This work investigates the validity of CS for collecting non-expert subjective voice likability scores contrasting the results with in-lab conducted listening tests. Four experiments are presented (Sections 3 and 4); two of them were conducted in the laboratory and two of them 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 [3]. The second in-lab test, detailed in [4], involved partly the same speech stimuli and the same listeners from the first in-lab test, and adopted a paired-comparison approach, in which the speech utterances were presented in pairs and the listeners were asked to decide which one was more likable. The first and the second laboratory tests will be referred in the following to as Lab-SCA and Lab-PC, respectively.

The same speech stimuli were used for the two CS experiments, which also examined direct scaling and paired-comparison for gathering voice likability scores. We will refer to these tests as CS-SCA and CS-PC, respectively, in the remaining of this paper. While the Lab-PC and CS-PC tests are contrasted in [5], this paper focuses on describing CS-SCA and the considerations for adapting the in-lab tests to CS. Checks and control questions are included to detect and discard untrustworthy (inaccurate or malicious) submissions, inspired by methods in [6, 7]. We then examine correlations of likability scores among the four experiments and discuss the potential of CS for speech data annotation.

2. Previous work

The speech likability database (SLD) [2] contains utterances of different contents from 800 speakers. These data were labeled in terms of likability on a 7-point Likert scale by 16 listeners who performed an in-lab listening test. The ratings were later binarized and the database was proposed for the Interspeech likability sub-challenge organized in 2012 [8]. Also Likert scales were used in [9] employing the *emoDB* data to assess voice likability, and in [10] and in [11] to measure vocal attractiveness using internally collected corpora. Continuous scales, suited for posterior prediction of likability ratings using regression techniques, are adopted in [12, 3]. In all these investigations, the subjective scores were obtained in controlled listening environments and carefully instructing listeners on the rating task.

The work in [13] used an online survey (this can be viewed as web-based CS) with a 5-point scale to gather, among other factors, voice pleasantness ratings. The survey was completed by 112 listeners, a greater number of participants compared to

those generally recruited for on-site tests (32 participants in [2], 20 in [9], 33 in [10], 30 in [11] and in [3], and 39 in [12]). However, the population of [13] is less controlled and hence their results subject to biases introduced by different participants' ages and language proficiencies. The authors of [14] also employed a web survey in which each participant, out of a total of 320, rated one speaker trait for one of the 64 voices under test. The validity of their results relies on the honesty of the participants, who were asked to perform the test in a quiet room using headphones or speakers.

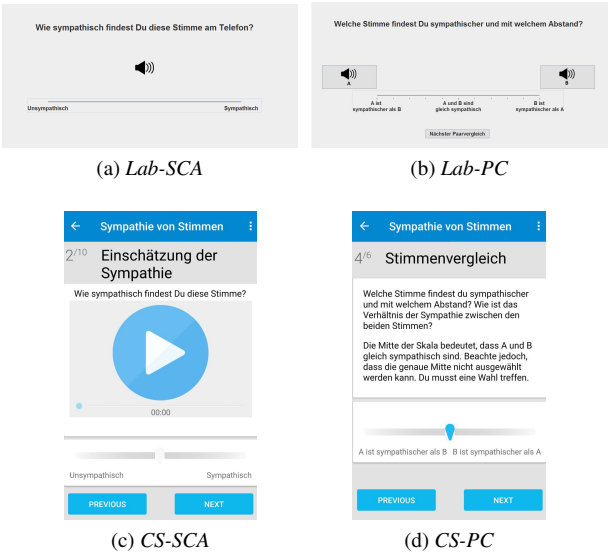
A recent analysis using web-CS to collect voice likability ratings on the SLD database [2] is presented in [15]. The authors employed test questions to determine participants' reliability and to derive weights to compute utterances' likability scores. Differently, in our work, we present experiments via mobile-CS to obtain ratings indicated on continuous scales and derived by pairwise comparisons and discuss the correlation of scores of in-lab and CS approaches. In [15], no correlation is shown between in-lab (SLD ratings) and CS results. Instead, the authors perform automatic likability classification experiments with the gathered annotations.

Other studies using CS for collecting rates on a Likert scale concentrate on audio quality assessments [16, 7] employing the discrete 5-point absolute-category-rating (ACR) for subjective Mean Opinion Scores), on voice naturalness [17], and on perceived Quality of Experience (QoE) in a teleconference system [18]. Different to using a Likert scale, the approach in [19] was to ask CS participants to rate correct or incorrect realizations of the /r/ sound in words. The study in [20] investigated the use of different rating scales (5-point ACR, continuous with the ACR labels, continuous with visual anchors, and binary) for assessments of image aesthetic appeal. In laboratory, the most reliable ratings were obtained for the 5-point ACR scale and for the continuous scale with anchors, and in CS, the highest user agreement was found for the discrete ACR scale, yet lower reliability compared to laboratory was detected for all scales. Also the analyses in [18] revealed lower participant reliability in CS compared to laboratory, and [7] and [19] showed strong agreement between CS and in-lab test participants.

Paired-comparison approaches have been widely adopted in CS for quality assessment of image [21], video [22], audio [23] and synthetic speech [24, 17]. The study in [23] introduces a pair-comparison framework for quantifying QoE of multimedia content as a more convenient approach compared to 5-point scale ratings. Also the authors of [17] claimed that employing an ACR scale on which to rate speech naturalness "might have been overwhelming" for the raters and therefore conducted later a paired-comparison test, which provided more substantial differences among the tested voices.

3. Laboratory experiments

As speech data, we employed a set of utterances from 30 speakers (15 males, 15 females) of standard High German dialect, recorded in clean conditions as specified in [3, 25], and aged 27.2 years on average (range: 20-34). The same sentence was extracted from the speakers' recordings: "Ich würde auf die SMS gern verzichten und meine Frei-Minuten dafür erhöhen" (In English: "I would like to give up the SMS and increase my free minutes in return"). These segments were transmitted through a wideband communication channel involving the codec G.722 at 64 kbit/s [3]. We stick to this set of utterances also for the CS experiments.



(a) *Lab-SCA* (b) *Lab-PC*

(c) *CS-SCA* (d) *CS-PC*

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

3.1. Scaling test in laboratory (Lab-SCA)

The same 30 participants whose voices were recorded participated in this listening test following a round-robin design with the aim of analyzing interpersonal perceptions [3]. The male and the female voices were presented in two separate blocks and each stimulus was only played once. In all, every utterance was rated by 29 listeners on a continuous likability scale.

The test was administered in a quiet 83 m^3 acoustically damped room with RT60=0.23 s at 2 kHz. The participants employed the headphones AKG K601 (frequency response 12–39,500 Hz) with diotic listening to listen to the stimuli and used a mouse to indicate their answers. The test graphical user interface (GUI) was presented on a Fujitsu SCENICVIEW LCD-Monitor P19-2, 48,3 cm (19-inch). A screenshot of the GUI can be seen in Figure 1a.

While narrowband- and wideband-transmitted speech was employed for this test and also personality perceptions were collected, only the likability ratings to the wideband voices are considered further for the rest of experiments in this paper.

3.2. Paired-comparison test in laboratory (Lab-PC)

The motivation for conducting this test [4] was to assess whether pairwise comparisons were suitable to derive a ratio scale of the listeners' preferences and whether higher agreement among participants could be achieved compared to Lab-SCA. A total of 13 out of the 15 female participants of the Lab-SCA test were recruited again for this experiment. Only male speech was used in the test, involving a total of $\binom{15}{2} = 105$ pair comparisons presented to the listeners. They were asked to click on the buttons "A" or "B" to listen to the stimuli and then select their degree of preference for any of the two (Figure 1b). Analogously as for Lab-SCA, this experiment was conducted with the same headphones, same monitor, and in the same room.

A preference ratio scale could be derived from the listeners' responses by applying the Bradley-Terry-Luce (BTL) probabilistic choice model [26, 27]. It was observed in [4] that, while these ratio scores and the ratings in Lab-SCA correlated with Pearson $r = 0.81$, $p < 0.001$, standard error (SE) $= 0.20$, the

paired comparison test provided higher agreement among raters and allowed for better discriminability between likable and non-likable speakers compared to the direct scaling approach. One major inconvenient is that the number of pairs $\binom{N}{2}$ in the test grows quadratically ($\Theta(N^2)$) with the number of voices to be scaled. Mechanisms have been proposed in the literature to reduce the pairs to be presented [28], which can be considered for future work.

4. Crowdsourcing experiments

We used the Crowdee platform to conduct our mobile-CS experiments based on micro-tasks [29]. The Crowdee application is freely available in the Google Play Store and most of its users are German young adults.

Our CS experiments aimed at obtaining a similar number of responses as for the in-lab experiments, using the same speech material. In contrast to the study in [15], we used a significantly smaller set of voices (wideband-quality speech from 30 speakers of similar age who uttered the same text, unlike the narrowband-quality SLD segments of different contents from 800 German speakers aged in the range 15–80 years [2]). Our CS participants (German, similar ages) can be seen as a more controlled group of assessors compared to those in [15] (original from ten different countries, participants from Romania, India, and United Kingdom were predominant). Also, we conducted mobile-CS campaigns, where other factors (e.g. street noises, interruptions, unstable Internet connection) different to those in web-CS can distract the participants and hence affect the results. Our considerations taken to guarantee the quality of the results are detailed in the following subsections.

4.1. Scaling test via crowdsourcing (CS-SCA)

4.1.1. Adapting the listening tests to the CS environment

We needed to address different issues in order to transfer our in-lab studies to a CS environment:

- dividing the complete in-lab test into CS micro-tasks, each of them not longer than a few minutes to avoid workers' loss of focus [6],
- controlling for workers' trustworthiness: detecting erroneous answers given sloppily or maliciously,
- controlling for workers' background noise (critical in mobile-CS, also related to workers' trustworthiness),
- controlling for the use of two-eared headphones for the completion of micro-tasks.

Instead of offering lengthy annotation tasks to workers, we partitioned the test into two CS jobs, one for male and one for female speakers, each job comprising 60 micro-tasks. Each micro-task consisted of 8 speech stimuli, presented on separated Crowdee screens (Figure 1c). In each screen, the workers were asked to listen to the stimulus and to indicate the degree of liking the voice on a continuous slider with the labels "likable" and "non-likable" at its ends. Each worker could only perform one micro-task out of each job. This way each worker could rate the likability of up to 8 male and 8 female speakers, and we assured that they were not confronted with the same speech sample more than once. The stimuli were randomized for different workers. On average, the time spent on completing each micro-task was 95.5 s, with standard deviation 38.0 s.

To control workers' trustworthiness, we created a qualification micro-task for the workers to earn access to the test micro-tasks. This qualification micro-task introduced the study and advised workers to perform it in a quiet room and to wear two-eared headphones. Besides, workers were asked to record their environment during at least seven seconds. We verified whether workers were located in noisy environments (e.g. street or crowded places) based on these recordings. Afterwards, workers were requested to adjust the device volume to a comfortable level by listening to a speech sample (not from the test). A single-choice question about the speakers' gender heard was used to control the worker's focus (trustworthiness). Plugged-in headphones were detected automatically to permit the workers to start the micro-tasks. Additionally, workers were presented a simple sum exercise with digits being played left to right in stereo. The response to this question was also considered for determining workers' trustworthiness.

If the workers provided correct answers to the two control questions, they were granted with a time frame of 25 minutes in which they could complete the test micro-tasks previously described. The workers were assigned randomly to any of the jobs, and if they wished and were still within the allowed time frame, they could continue and perform the test micro-task with voices of the other gender. One *trapping question* was inserted among the screens of the test micro-task, inspired by the work in [7]. It consisted of a speech stimulus that started as the rest of stimuli of the test but that was interrupted after two seconds and included a new voice (not from the considered speakers) that asked the worker to move the sliders' knob completely to the right or to the left. We also analyze these answers to ascertain workers' trustworthiness.

4.1.2. CS-SCA test results

We obtained answers from a total of 69 unique workers who responded correctly to the questions of the qualification micro-task. The trapping question responses were wrong for seven micro-tasks, and the corresponding test answers were hence deemed as untrustworthy. The Crowdee system continued offering micro-tasks until we collected 30 trustworthy likability ratings to each of the 30 utterances, a similar amount of ratings as for the Lab-SCA test. In all, 68 unique workers were deemed as trustworthy, 52 of them performed both test micro-tasks. According to information they provided, they were aged 30.4 years on average, standard deviation 9.8 years, and 44.1% female. Among these workers, only the background of one female was detected as somewhat noisy (television was on). Her answers were, however, considered for the trustworthy set.

The mean likability ratings obtained for each speaker, only considering trustworthy answers, can be seen in Figure 2. Every speaker was assigned a city name as pseudonym. They are sorted from the most to the least likable for each gender separately according to the Lab-SCA results. The Lab-SCA and the CS-SCA mean scores correlated highly (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). The correlation was found to be identical when the answers from the seven micro-tasks that were rejected were included in the CS-SCA results.

We currently do not have an explanation for the lower correlation found for male voices compared to that for female voices. Specially for male speakers, it can be observed that the workers tended to provide higher likability ratings than the laboratory listeners. The t-test indicated statistically significant differences ($p < 0.001$) between Lab-SCA and CS-SCA mean ratings for the male speakers "westbay" and "rabat". A posterior listening analysis did not reveal particularities in these voices that explain

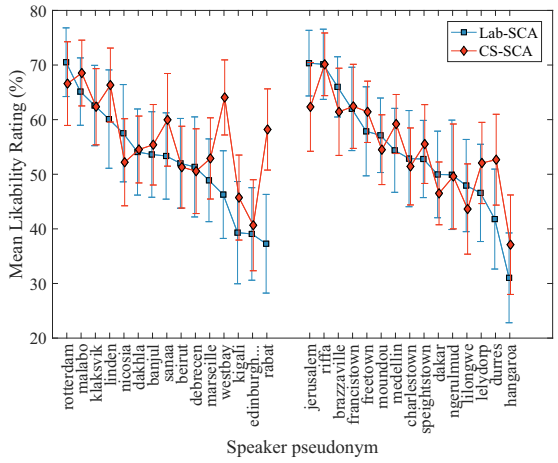


Figure 2: *Mean likability scores for male and female speakers from the direct scaling test in laboratory and using crowdsourcing. Error bars show 95% confidence intervals.*

the differences in the mean ratings.

We evaluated the inter-rater reliability by calculating the intraclass correlation coefficient (ICC) over the trustworthy likability ratings, using the *ICCest* function of the R package 'ICC'. We obtained for CS-SCA ICC = 0.11 with 95% confidence interval 0.06–0.18, and for Lab-SCA the ICC was 0.14 with 95% confidence interval 0.08–0.24. In view of this result, the level of agreement across workers was somewhat lower compared to the in-lab listeners. It has to be noted that each worker provided 13.2 ratings on average, while each in-lab listener provided 29 ratings. Lower participant reliability in CS compared to in-lab tests was also reported in [20, 18].

4.2. Paired-comparison test via crowdsourcing (CS-PC)

The same speaker pairs as for Lab-PC were presented in the CS-PC test micro-tasks, each of them including one pair comparison. 1365 micro-tasks (105 comparisons x 13 responses from unique workers) were offered in order to gather the same amount of responses as in the Lab-PC test. The workers were allowed to perform different test micro-tasks up to 25 times.

The test adaptation from lab to Crowdee was done similarly as described in Subsubsection 4.1.1 (additional details in [5]). A similar qualification micro-task as for CS-SCA was created, and two trapping questions were included to control for worker focus on the test micro-task. One asked for the gender of the speaker heard and the other one asked (translated from German): "If you are reading this question, please choose the answer X". We received a total of 1682 submissions, from which 317 were rejected because the workers background was determined to be noisy, or because the workers responded incorrectly to the trapping questions of the qualification or of the test micro-tasks. 77 unique workers provided the 1365 trustworthy answers. The paired comparison task was divided into 2 screens, the first one enabled playing the two voices (concatenated and separated by a beep sound), and the second one, shown in Figure 1d, presented a slider with which to indicate the degree of preference for the first or for the second voice.

The ratio scale preferences estimated by the BTL model are displayed in Figure 3 for Lab-PC and for CS-PC. A strong and significant correlation was found between the two score sets (Pearson $r = 0.95$, $p < 0.001$, SE = 0.09). The correlation decreased only to $r = 0.92$, $p < 0.001$, SE = 0.11 when all answers (trustworthy and untrustworthy) were computed. However, no correlation between the two distance matrices (built from the slider values from the Lab-PC and the CS-PC tests, respectively) was found. Our analyses also showed higher disagreement among raters or rater inconsistencies for CS-PC compared to Lab-PC [5].

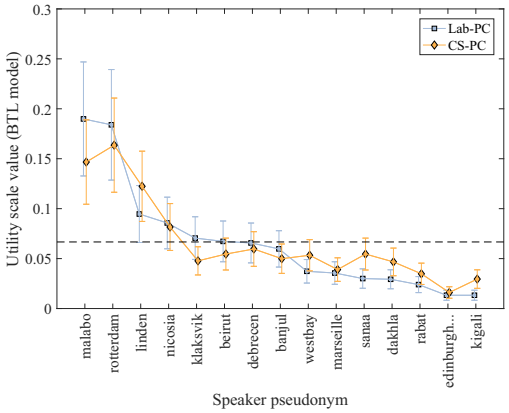


Figure 3: *Ratio scale preferences estimated by the BTL model for the paired-comparison tests conducted in laboratory and via crowdsourcing (male speakers only). Error bars show 95% confidence intervals. The indifference line is plotted as* $y = 1/15$.

Despite the different speaker ranking, similar likability tendencies can be seen for the male speakers of Figures 2 and 3. The scores from the two CS tests correlated with Pearson $r = 0.73$, $p = 0.002$, SE = 0.19.

5. Conclusions

In-lab and CS listening tests based on direct scaling and on paired comparisons have been presented to gather annotations of voice likability. For the scaling tests, strong correlations between in-lab and CS ratings ($r = 0.68$ and $r = 0.89$) were evinced for male and for female speakers, respectively. An even higher correlation was found for the paired-comparison in-lab and CS tests ($r = 0.95$). Our outcomes suggest that, while both proposed CS tests were valid for obtaining meaningful speech annotations in terms of voice likability, the CS paired-comparison test can offer more reliable likability scores than the CS scaling test. This may be due to the fact of having a reference voice and to the simplicity of the task, as also observed in [23, 17]. The drawback of the increased test length is not critical in CS as opposed to lab, and can be overcome by adopting techniques such as those proposed in [21, 28].

Since the great majority of workers were deemed reliable, only marginal decreases in the correlations between CS and in-lab ratings were found. A larger study would be needed to carefully examine how results can be influenced by a considerable number of untrustworthy workers. Still, we consider our qualification micro-tasks and control checks to be crucial to make workers aware of the listening test requirements (quiet room and two-eared headphones), to control background noise and workers' focus, and to possibly detect hearing impairments.

It also remains to be explored in future work whether crowdsourcing annotations can be modeled to derive ground truth utterance labels for speaker characterization (e.g. in terms of likability, personality, or emotions).

6. References

[1] F. Burkhardt, H. R., and A. Batliner, "Application of Speaker Classification in Human Machine Dialog Systems," in *Speaker Classification I - Fundamentals, Features, and Methods*, C. Müller, Ed. Berlin, Heidelberg: Springer, 2007.

[2] F. Burkhardt, B. 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.

[3] 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.

[4] L. Fernández Gallardo, "A Paired-Comparison Listening Test for Collecting Voice Likability Scores," in *Informationstechnische Gesellschaft im VDE (ITG) Conference on Speech Communication*, 2016, pp. 185–189.

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

[6] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force "Crowdsourcing"," 2014, cOST Action IC1003 European Network on Quality of Experience in Multimedia Systems and Services (QUALINET).

[7] 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*, 2015, pp. 2799–2803.

[8] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "A Survey on Perceived Speaker Traits: Personality, Likability, Pathology, and the First Challenge," *Computer Speech & Language*, vol. 29, no. 1, pp. 100–131, 2015.

[9] B. Weiss and F. Burkhardt, "Voice Attributes Affecting Likability Perception," in *Interspeech*, 2010, pp. 1934–1937.

[10] M. Zuckerman, K. Miyake, and C. S. Elkin, "Effects of Attractiveness and Maturity of Face and Voice on Interpersonal Impressions," *Journal of Research in Personality*, vol. 29, no. 2, pp. 253–272, 1995.

[11] M. Babel, G. McGuire, and J. King, "Towards a More Nuanced View of Vocal Attractiveness," *PLoS One*, vol. 9, no. 2, pp. 1–10, 2014.

[12] B. Weiss and K. Schoenenberg, "Conversational Structures Affecting Auditory Likeability," in *Interspeech*, 2014, pp. 1791–1795.

[13] L. Pinto-Coelho, D. Braga, M. Sales-Dias, and C. Garcia-Mateo, "On the Development of an Automatic Voice Pleasantness Classification and Intensity Estimation System," *Computer Speech & Language*, vol. 27, no. 1, pp. 75–88, 2013.

[14] P. McAleer, A. Todorov, and P. Belin, "How Do You Say Hello? Personality Impressions from Brief Novel Voices," *PLoS One*, vol. 9, no. 3, p. e90779, 2014.

[15] S. Hantke, E. Marchi, and B. Schuller, "Introducing the Weighted Trustability Evaluator for Crowdsourcing Exemplified by Speaker Likability Classification," in *10th Language Resources and Evaluation Conference (LREC)*, 2016.

[16] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "CrowdMOS: An Approach for Crowdsourcing Mean Opinion Score Studies," in *ICASSP*, 2011, pp. 2416–2419.

[17] E. Cooper, Y. Levitan, and J. Hirschberg, "Data Selection for Naturalness in HMM-based Speech Synthesis," in *Speech Prosody*, 2016.

[18] T. Volk, C. Keimel, M. Moosmeier, and K. Diepold, "Crowdsourcing vs. Laboratory Experiments - QoE Evaluation of Binaural Playback in a Teleconference Scenario," *Computer Networks*, vol. 90, pp. 99–109, 2015.

[19] T. M. Byun, P. F. Halpin, and D. Szeredi, "Online Crowdsourcing for Efficient Rating of Speech: A Validation Study," *Journal of Communication Disorders*, vol. 53, pp. 70–83, 2015.

[20] 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, 2016.

[21] Q. Xu, Q. Huang, and Y. Yao, "Online Crowdsourcing Subjective Image Quality Assessment," in *ACM International Conference on Multimedia*, 2012, pp. 359–368.

[22] 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.

[23] C. C. Wu, K. T. Chen, Y. C. Chang, and C. L. Lei, "Crowdsourcing Multimedia QoE Evaluation: A Trusted Framework," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1121–1137, 2013.

[24] S. Buchholz and J. Latorre, "Crowdsourcing Preference Tests, and How to Detect Cheating," in *Interspeech*, 2011, pp. 3053–3056.

[25] L. Fernández Gallardo, "Recording a High-Quality German Speech Database for the Study of Speaker Personality and Likability," in *12. Tagung Phonetik und Phonologie im deutschprachigen Raum*, 2016, pp. 43–46.

[26] 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.

[27] R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis*. New York, USA: Wiley, 1959.

[28] A. Eichhorn, P. Ni, and R. Eg, "Randomised Pair Comparison - An Economic and Robust Method for Audiovisual Quality Assessment," in *International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*, 2010, pp. 63–68.

[29] 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, Show & Tell Session*, 2014, pp. 1496–1497.