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, Technische Universität Berlin, Germany

{laura.fernandezgallardo, rafael.zequeira, sebastian.moeller}@tu-berlin.de

Introduction

- Purpose: collection of human perceptions of voice likability:

  To what extent provides crowdsourcing valid subjective ratings as in a laboratory testing?

- In-lab listening tests:
  - control over the background environment and equipment
  - supervision of participant's behavior
  - confirmation of participant's understanding of test instructions

- Crowdsourcing listening tests:
  - micro-task rewarded with micro-payments
  - large and diverse pool of participants
  - scalable, fast, and low cost
  - test performed on the user's device

Speech Material and Listening Tests

- Same sentence (mean = 4.4s) from 30 German speakers (15 males, 15 females) from the Nautilus Speaker Characterization (NSC) Corpus

  http://www.qu.tu-berlin.de/?id=nsc-corpus

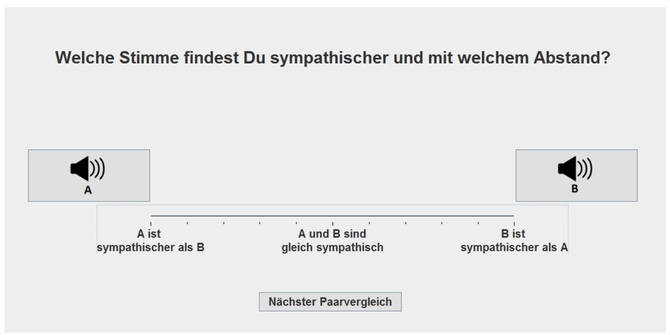
- Pair Comparison experiments: **Lab-PC**, **CS-PC**
  - 15 male stimuli combined in $\binom{15}{2} = 105$ unique pairs
  - Preference selected for voice A or for voice B

- Direct SCAling experiments: **Lab-SCA**, **CS-SCA**
  - Male and female speech stimuli
  - Likability rating indicated on a continuous slider for each stimulus

Pair Comparison

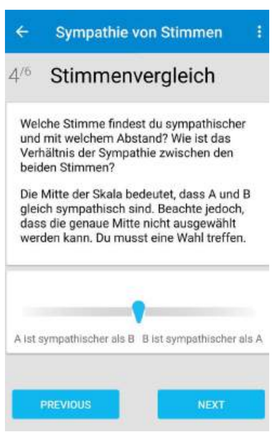
Lab-PC

- 13 German female listeners
- Test session took 30 minutes



CS-PC

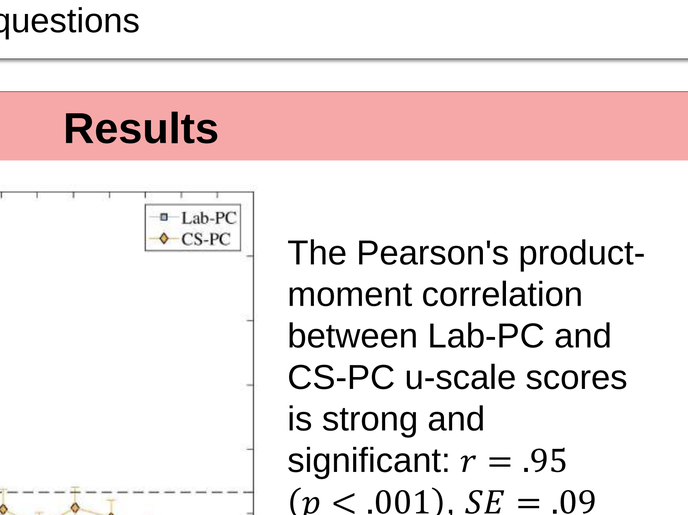
- Executed in the Crowdee mobile-CS platform
- 1365 (105 x 13) micro-tasks
  - one pair-comparison each
  - 32.4s on average (range: 11-209s)
- Qualification micro-task for the users to earn access to the study
- 92 German users
- Controlled:
  - users' environmental noise
  - use of two-eared headphones
  - *trapping* and *control* questions



Direct SCAling

Lab-SCA

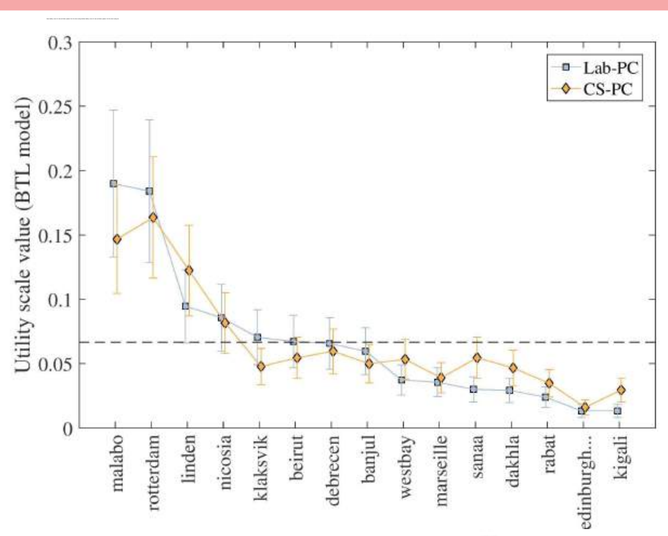
- 29 German listeners
- Test session took 20 minutes



CS-SCA

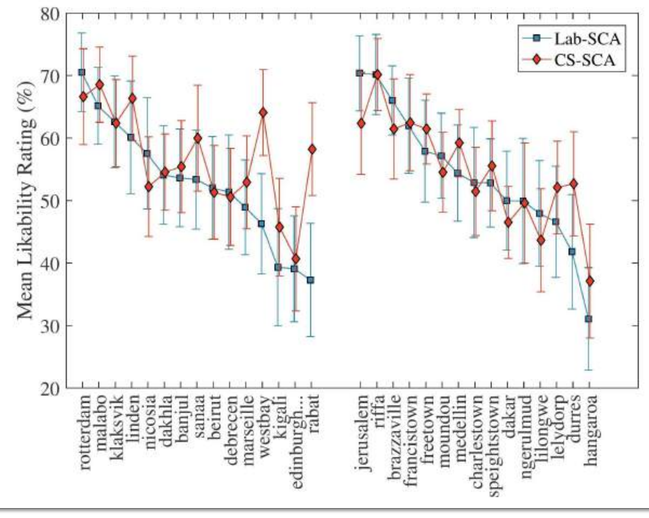
- 120 micro-tasks, divided into two (female and male stimuli)
- 8 stimuli per micro-task + one *trapping* question
  - 95.5s on average (range: 39-236s)
- Qualification micro-task for the users to earn access to the study
- 69 German users
- Controlled:
  - users' environmental noise
  - use of two-eared headphones
  - *trapping* and *control* questions

Results



The Pearson's product-moment correlation between Lab-PC and CS-PC u-scale scores is strong and significant: $r = .95$ $(p < .001)$, $SE = .09$

The mean scores of Lab-SCA and CS-SCA were also correlated: Pearson $r = 0.68$ $(p < 0.005)$ and $SE = 0.20$ and Pearson $r = 0.89$ $(p < 0.001)$ and $SE = 0.13$ for male and for female speakers, respectively



Conclusions

- Strong and statistically significant Pearson correlations between voice likability scores obtained in the lab and via crowdsourcing
- CS-PC can offer more reliable likability scores than CS-SCA. The drawback of increased test length is not as critical in CS as in the lab
- We have indicated appropriate control questions and mechanisms to manage the trustworthiness of users and their answers