



Outliers Detection vs. Control Questions to Ensure Reliable Results in Crowdsourcing. A Speech Quality Assessment Case Study

Rafael Zequeira Jiménez, Laura Fernández Gallardo, Sebastian Möller

Quality and Usability Lab, Technische Universität Berlin

HumL@WWW2018 - 1st International Workshop on Augmenting Intelligence with Humans-in-the-Loop





Motivation

Speech quality is important for the Quality of Experience (QoE) in:



audio books



virtual or robotic
conversational agents

* The collected ratings can be used to train AI systems to predict the speech quality automatically



Motivation



Speech quality experiments
traditionally conducted in
Laboratory

- Professional audio equipment
- Soundproof room
- Limited number of participants



Crowdsourcing Study

- Conducted a speech quality assessment experiment
- Crowd-workers were presented with 20 speech stimuli
- Opinion about overall quality gathered in a 5-point scale





Speech Material:

- Database number 501 from ITU-T Rec. P.863
- 4 Germans were recorded per condition
- 200 speech stimuli (9s long on avg.)
- 50 degradation conditions:
 - narrow- & wide- band
 - temporal clipping
 - signal-correlated noise,
 - combinations of these degradations
- The database contains quality ratings to the 200 stimuli made by 24 different native German listeners, in accordance to ITU-T Rec. P.800



Study Conditions:

- Address the study to native Germans
- Collect 24 ratings per stimulus from different listeners
- Experiment in accordance with the ITU-T Rec. P.800





Crowdsourcing Platform:



- German based CS platform
- Reported 1 million global users in September 2017
- Most of their users are from German speaking countries



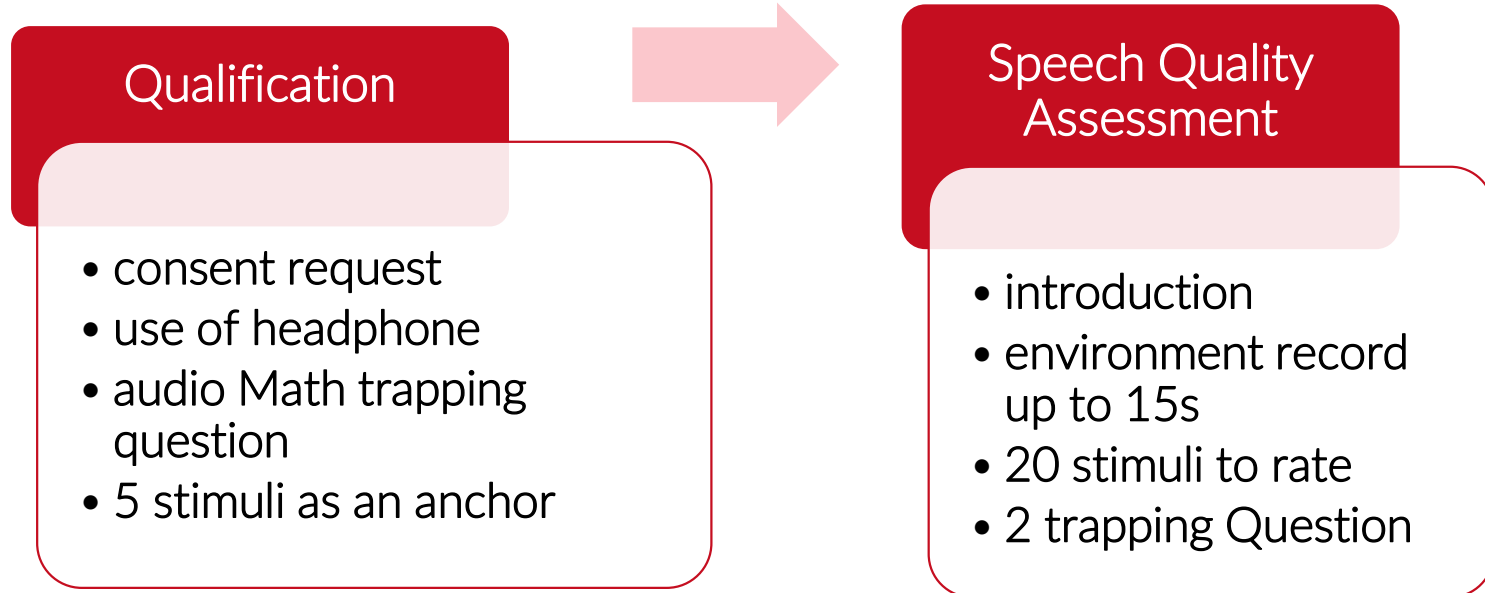
Crowdsourcing Experiment



- Screening task to recruit German listeners
- Speech quality assessment task:
 - Qualification phase
 - Speech quality assessment



Crowdsourcing Experiment





Crowdsourcing Experiment

Qualification

- consent request
- use of headphone
- audio Math trapping question
- 5 stimuli as an anchor

Speech Quality Assessment

▶ 0:08 / 0:08



Sprachqualität:

Bewertung

- | | |
|-------------------------------------|---|
| <input type="radio"/> Ausgezeichnet | 5 |
| <input type="radio"/> Gut | 4 |
| <input type="radio"/> Ordentlich | 3 |
| <input type="radio"/> Dürftig | 2 |
| <input type="radio"/> Schlecht | 1 |

Zurück

Weiter



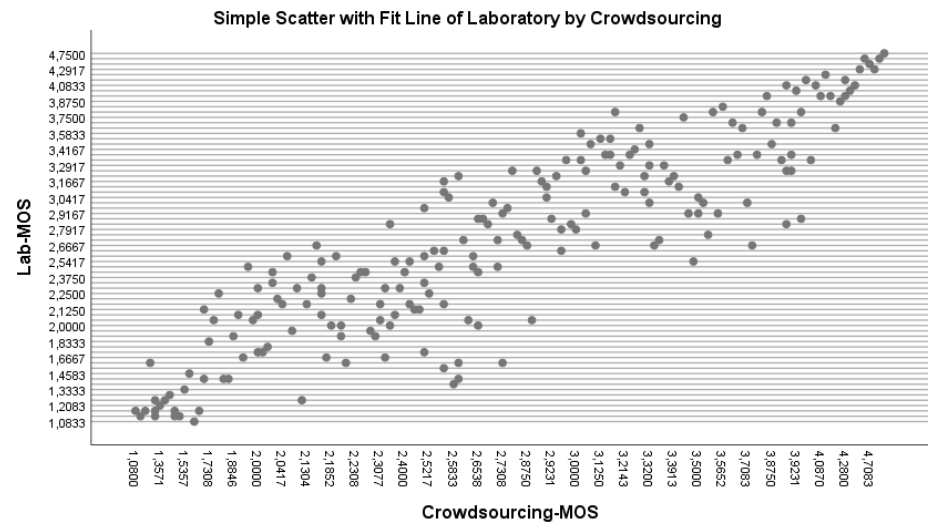
Results

- 87 workers participated in the study
- 8 workers failed the Qualification phase
- 53 unique listeners:
 - 60,4% males
 - 96,2% native Germans
 - provided 4840 ratings
- the collected ratings account for 24 to 26 assessment from different listeners per file



Crowdsourcing vs. Laboratory

- Spearman's rank-order correlation:
 - $\rho = 0,864$ ($p < 0,001$)
- Monotonic relationship between Lab- and CS- MOS
- Root Mean Square Error:
 - $RMSE = 0,474$





Filtering from unreliable workers

- Work in [1] and [2] recommends:
 - the use of trapping question, to catch inattentive users
 - when the user fail, then all of their ratings are discarded

This approach was effective in [1] and improved slightly the results in [2]

[1] 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.

[2] R. Zequeira Jiménez, L. Fernández Gallardo, and S. Möller, “Scoring Voice Likability using Pair-Comparison: Laboratory vs. Crowdsourcing Approach,” in *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–3.



Filtering from unreliable workers

A worker is unreliable or untrustworthy when:

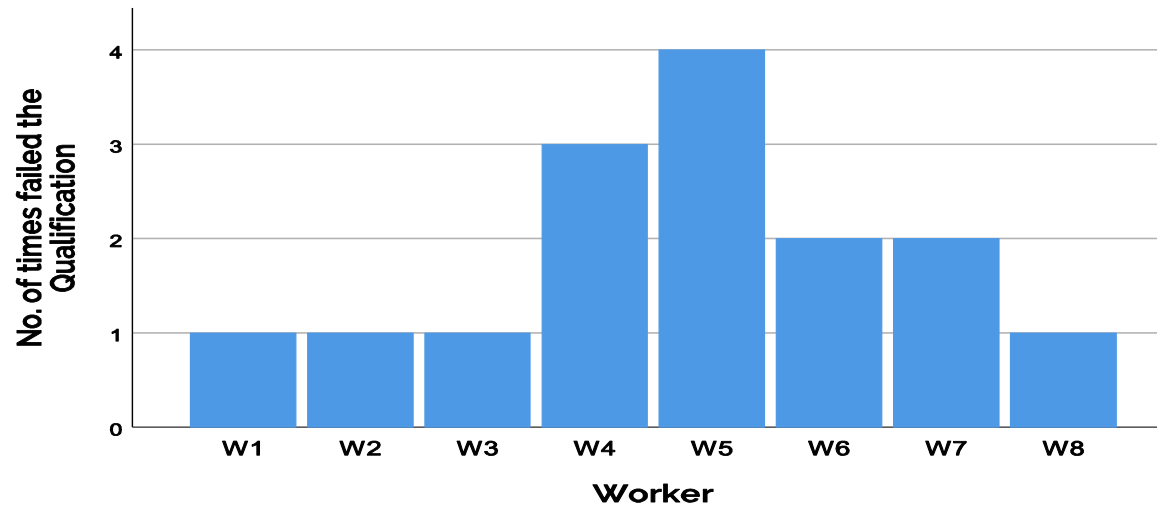
- s/he fails the trapping question in the SQAT
- s/he fails the Qualification more than once



Filtering from unreliable workers

A worker is unreliable or untrustworthy when:

- s/he fails the trapping question in the SQAT
- s/he fails the Qualification more than once





Filtering from unreliable workers

- Discarded 320 ratings in total from W4, W5, W7
- W6 did not conduct the SQAT

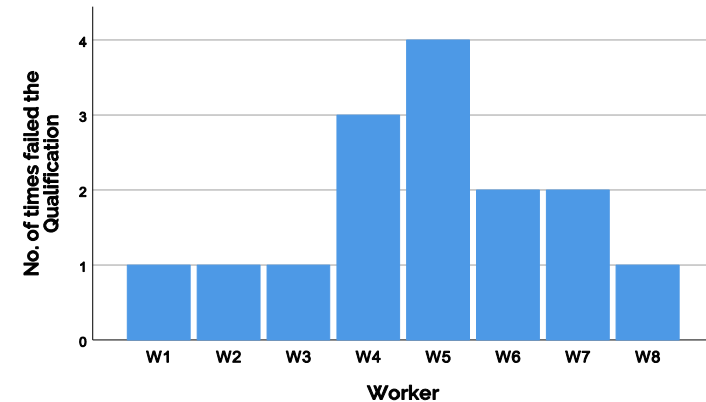
Method:

“filtering by trapping question” (F-TQ)

- Spearman’s rank-order correlation on 4520 ratings:
 - $\rho = 0,862$ ($p < 0,001$)

When discarding all the workers (F-TQ) :

- $\rho = 0,854$ ($p < 0,001$)





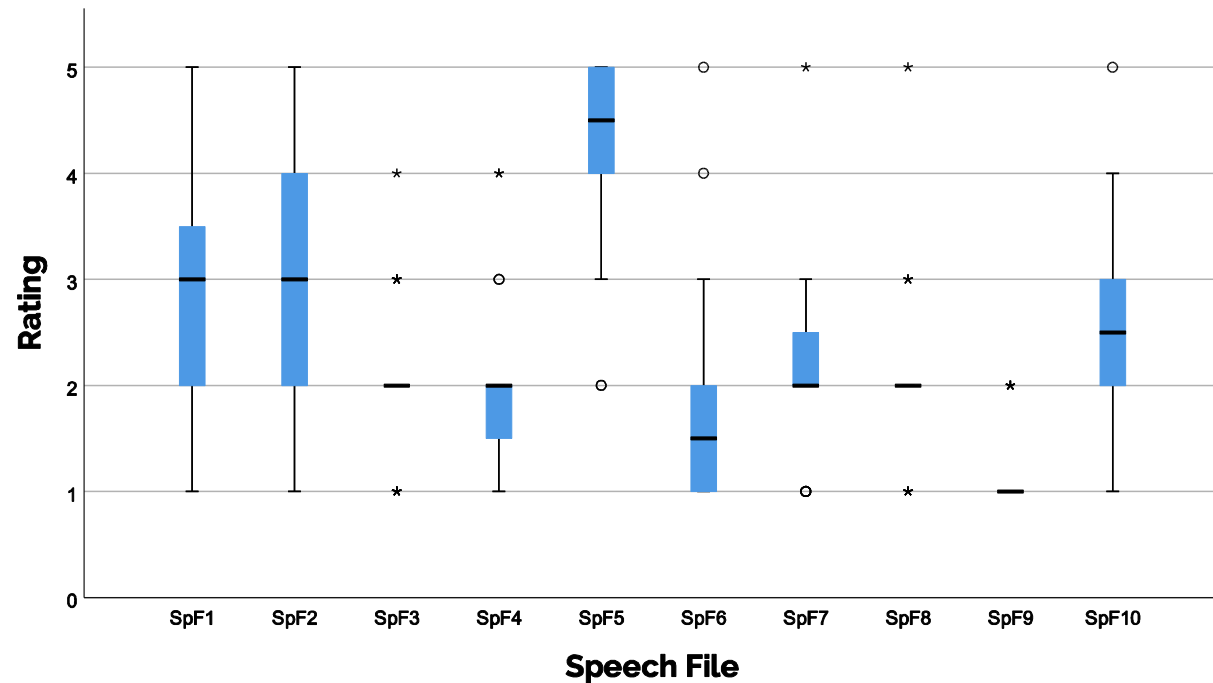
Outlier Detection

outliers:

- ratings above 1,5 interquartile range (IQR)
- depicted by circles

extreme outliers:

- ratings at 3,0 IQR or above
- depicted by asterisks





Outlier Detection

- Discarded 122 ratings identified as extreme outliers

Method:

“filtering by outlier detection 1” (**F-OD1**)

- Spearman’s rank-order correlation:
 - $\rho = 0,863$ ($p < 0,001$)

still not better than the first coefficient when no data was discarded



Outlier Detection 2

- Discarded 1480 ratings from 12 workers that were outliers or extreme outliers three times or more [5].

Method:

“filtering by outlier detection 2” (**F-OD2**)

- Spearman’s rank-order correlation:
 - $\rho = 0,867$ ($p < 0,001$)



Alternative Approach

- Applied **F-OD1** and **F-OD2** and discarded 1529 ratings in total.
- Identify the outliers made by all the workers that failed the trapping questions. Then removed 17 ratings.

Method:

F-TQ-OD

- Spearman's rank-order correlation on 3294 ratings:
 - $\rho = 0,868$ ($p < 0,001$)



Results Overview

Approach	Ratings discarded	rho	RMSE
-	0	0,864*	0,474
F-TQ	320	0,862*	0,476
F-TQ'	780	0,854*	0,480
F-OD1	122	0,863*	0,477
F-OD2	1480	0,867*	0,474
F-TQ-OD	1546	0,868*	0,479

* $p < 0,001$



Results Comparison

Approach	Method	Workers Discarded	Ratings Discarded
[6]	gold standard questions	25%	75%
[7]	verification questions	-	34,3%
F-TQ-OD	trapping question + outliers detection	22%	31,9%

[6] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, “Quantification of YouTube QoE via Crowdsourcing,” in 2011 IEEE International Symposium on Multimedia, 2011, pp. 494–499.

[7] J. Redi and I. Pova, “Crowdsourcing for Rating Image Aesthetic Appeal: Better a Paid or a Volunteer Crowd?,” in International ACM Workshop on Crowdsourcing for Multimedia, 2014, pp. 25–30.



Discussion


Approach	Ratings discarded	rho	RMSE
-	0	0,864*	0,474
F-TQ	320	0,862*	0,476
F-TQ'	780	0,854*	0,480
F-OD1	122	0,863*	0,477
F-OD2	1480	0,867*	0,474
F-TQ-OD	1546	0,868*	0,479

- We recommend to employ F-OD1 in case “high correlation” is not a priority. This is the most cost effective approach.
- We recommend to use F-TQ-OD for more accurate results.



Conclusion

- Adapted successfully a Laboratory listening test to Crowdsourcing
- Obtained a strong and statistically significant Spearman correlation: $r=0.868$
- Tested outliers detection and trapping question to filter the data from unreliable ratings
- Proposed a combination of outlier detection and trapping question that leads to more accurate results
- Further testing is required to determine for which type of experiment our approach can be applied.

An aerial photograph of a city street during sunset. The sky is filled with soft, golden light, and the sun is low on the horizon, creating a hazy atmosphere. The street is filled with cars, and buildings line both sides. A red text overlay is positioned in the upper left quadrant.

Thank you for your Attention!

Rafael Zequeira Jiménez
rafael.zequeira@tu-berlin.de
@zequeiraj