Timo Baumann

Carnegie Mellon University
Language Technologies Institute

Large-scale Speaker Ranking from Crowdsourced Pairwise Listener Ratings

ABSTRACT:

Sakaguchi et al. [1] proposed to use Microsoft TrueSkill™ [2] to reduce the effort of human evaluation for machine translation. We extend this method and apply it to ranking speakers for their reading quality.

We decomposed and crowdsourced the ranking problem into pairwise ratings: “Which one do you like better, A or B?”

We selected rating pairs in an online fashion so that human input was maximally informative, based on the preliminary rankings.

We generate a **ranking of 227 speakers** of the Spoken Wikipedia Corpus [3] and analyse influencing factors to explain the ranking.

→ methodology to create ranking from pairwise ratings
- efficient: only a small subset of pairs
- flexible: users can provide as few/many ratings
- adaptive: the algorithm selects most informative pairs

→ result: speaker ranking for German Spoken Wikipedia
- diverse & large set of speakers
- diverse & large set of raters
- lots of additional material available for every speaker

→ analysis:
- acoustic quality (little influence)
- speaker livelihood as measured on additional material
- same-gender preference of raters

Data Collection

Most articles in the Spoken Wikipedia start with the identical sentence “You are listening to the article XYZ from Wikipedia, the free encyclopedia.” We extract this sentence for all 227 speakers in the SWC where text/audio alignment is avaiable.

We solicited participants via various mailing lists and the Wikipedia off-topic board in the German-speaking countries. We did not offer any compensation: the only incentive was to help research and improve open-source speech techonology. As a result, there was no need for data cleansing.

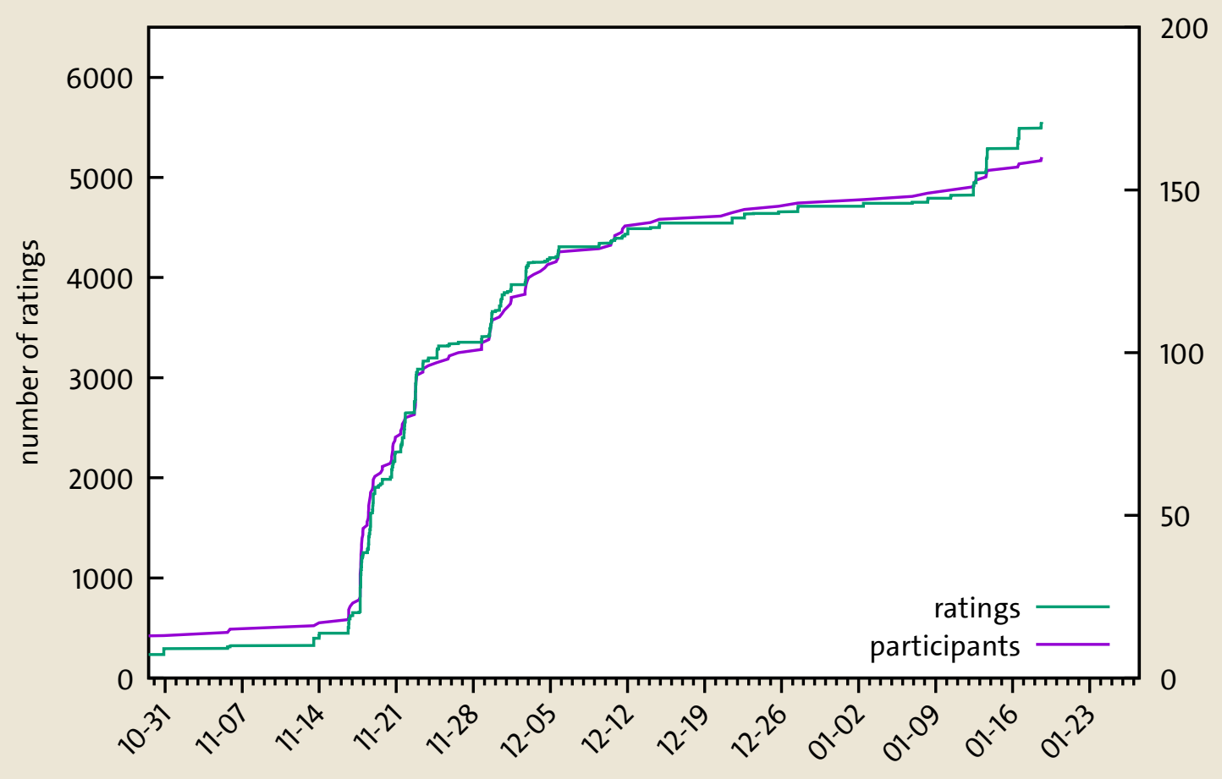
We collected demographics of the raters and got data for most strata.

However, male raters from northern German, and 20-30 years old (presumably university students in CS) are over-represented.

In total, participants donated about 26 hours of their time.

		participants	ratings
	total	168	5440
gender	female	41	1665
	male	109	3221
	unreported	18	554
age	<20	18	358
	20-30	78	2593
	30-40	34	1030
	40-60	24	886
	>60	6	418
	unreported	8	155
dialectal origin	Northern Germany	83	2656
	Berlin/Brandenburg	8	128
	Northrhine-Westphalia	11	464
	Middle Germany	9	443
	Rhine-/Saarland	3	82
	Baden-Wurttemberg	15	432
	Bavaria	8	405
	Austria	5	179
	Switzerland	0	0
	unsure/other	26	651

Experiment progress in winter 2016



References

[1]: Sakaguchi, Keisuke, Matt Post and Benjamin Van Durme: "Efficient Elicitation of Annotations for Human Evaluation of Machine Translation". In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014.
[2]: Herbrich, Ralf, Tom Minka and Thore Graepel: "TrueSkill A Bayesian Skill Rating System". In: *Advances in Neural Information Processing Systems*, 2007.
[3]: Köhn, Arne, Florian Stegen and Timo Baumann: "Mining the Spoken Wikipedia for Speech Data and Beyond". In: *Proceedings of the Language Resource and Evaluation Conference*, 2016.

Method and Results

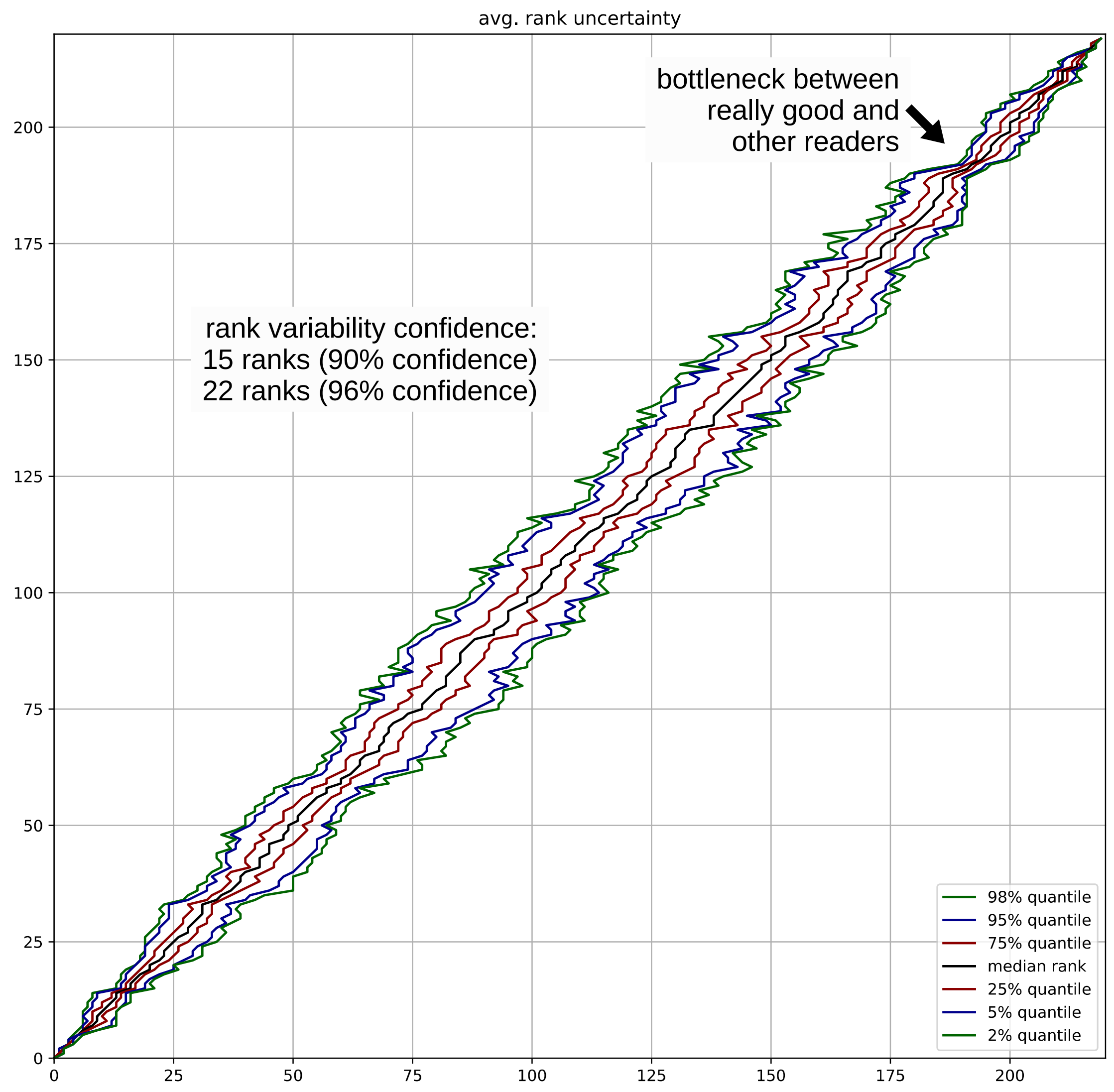
Used as intended, TrueSkill™ matches online players of equal strength (for maximally interesting games) and computes player rankings.

We handle stimuli as “players” that compete in many rating “games”.

After initialization, we compute a ranking and select rating pairs weighed by the following factors:
- stimuli that have not been rated often yet (addition to TrueSkill)
- are maximally informative for the model

We recomputed rankings after mini-batches of about 100-200 ratings.

Final ranking after 5440 ratings:



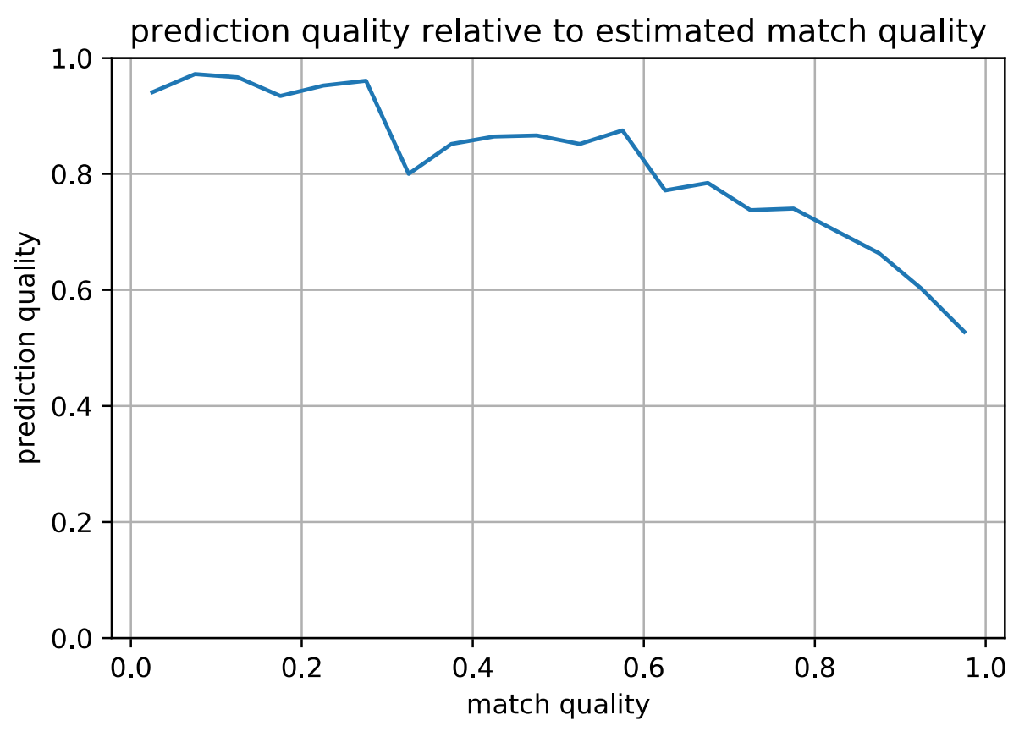
reliability test: randomly permute ratings, measure rank correlation
- median ranking is highly reliable (Kendall's τ > 0.95)

ranking confidence: deviations from median ranking
- on avg. 90% confidence for 15 ranks (within 7% of speakers)
- on avg. 96% confidence for 22 ranks (within 10% of speakers)
- some rankings are unanimous (best/worst reader)

rating consistency: proportion of feedback arcs among all ratings
- strong disagreement among raters: 29% feedback arcs
- even disagreement within individual raters

Ranking can be used to predict rating outcome and quality of prediction is well estimated.

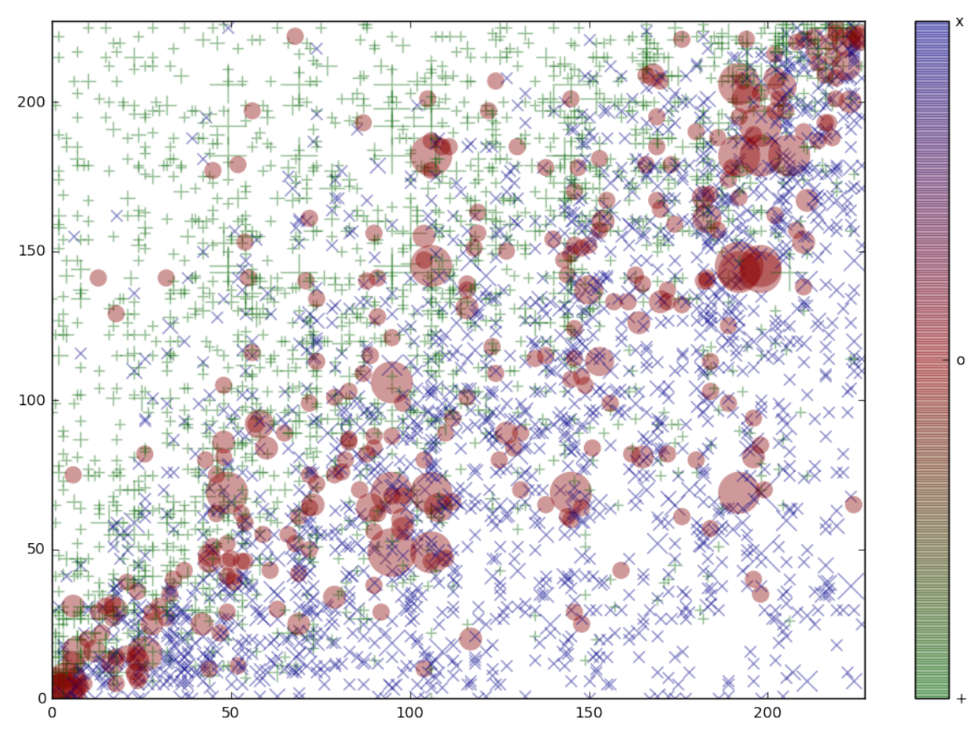
- 10-fold cross-validation
- 68% prediction performance
- quality and performance correlate strongly: τ= -0.81



Stimuli pair selection is efficient!

- green: wins
- blue: losses
- red: unclear along the diagonal
- size: #ratings

Only 7.7% of possible rating pairs needed to be rated to derive the consistent ranking.



Scatter plot of pairs compared.

Explaining Factors

We try to identify explaining factors for the ranking through correlation analysis.

Acoustic Quality

We compute the perceptual quality of the audio using ITU-T P.563 reference software and find a slight but significant correlation: Kendall's τ = 0.14, p < .002.

Implication: actual influence, or: careful readers also care about noise. Future: include measured quality in ranking algorithm.

Pitch-range of Speaker

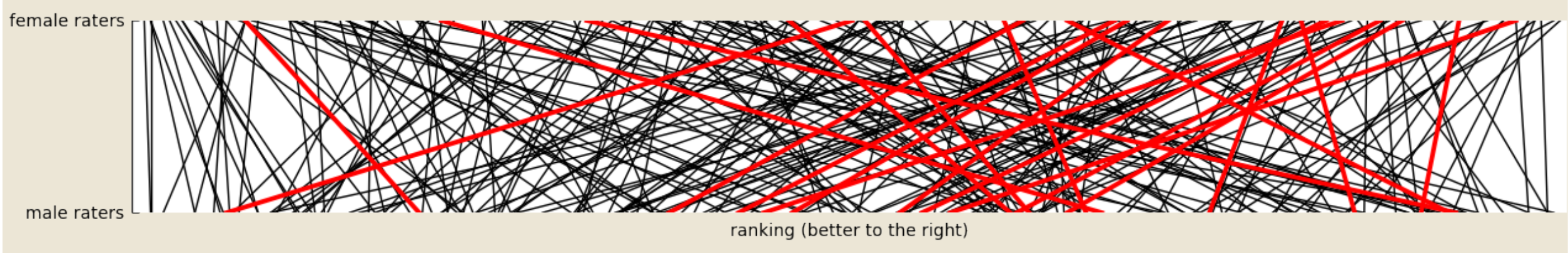
We compute the 90%-pitch range of the speaker on the complete recording (rather than just the short stimulus played to raters). We find a slight but significant correlation: τ = 0.10, p < .03.

Additional audio is helpful to explain human judgements, as humans are extremely good in judging overall performance based on small samples.

Gender Effects

separate rankings for female/male raters:
- only moderate correlation, τ=0.44, between female/male ranking

analysis by speaker gender:
- females listeners like female speakers (on avg. 12.7 ranks better)



Comparison of median rankings for female (top) and male (bottom) rankings. Female stimuli shown in red.

Gender is more influential than rater age or dialect:
- for both splits stronger correlations across groups

Is there a preference for one's own dialect?
- investigate speaker/listener dialect matching (future work)

Please e-mail for further information:
→ *baumann@informatik.uni-hamburg.de*

The ratings, ranking, software, paper and poster can be found on
→ *www.timobaumann.de/work/pub*

this work was conducted while at Universität Hamburg and partially sponsored by Daimler-Benz Foundation.

Daimler und Benz Stiftung

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG