INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Large-scale Speaker Ranking from Crowdsourced Pairwise Listener Ratings

Timo Baumann

Language Technology Institute, Carnegie Mellon University, Pittsburgh, USA
Natural Language Systems Group, Department of Informatics, Universität Hamburg, Germany
tbaumann@cs.cmu.edu

Abstract

Speech quality and likability is a multi-faceted phenomenon consisting of a combination of perceptory features that cannot easily be computed nor weighed automatically. Yet, it is often easy to decide which of two voices one likes better, even though it would be hard to describe why, or to name the underlying basic perceptory features. Although likability is inherently subjective and individual preferences differ frequently, generalizations are useful and there is often a broad intersubjective consensus about whether one speaker is more likable than another. However, breaking down likability rankings into pairwise comparisons leads to a quadratic explosion of rating pairs. We present a methodology and software to efficiently create a likability ranking for many speakers from crowdsourced pairwise likability ratings. We collected pairwise likability ratings for many (>220) speakers from many raters (>160) and turn these ratings into one likability ranking. We investigate the resulting speaker ranking stability under different conditions: limiting the number of ratings and the dependence on rater and speaker characteristics. We also analyze the ranking wrt. acoustic correlates to find out what factors influence likability. We publish our ranking and the underlying ratings in order to facilitate further research.[1]
Index Terms: ranking, speech quality, likability, found data, ratings, crowd-sourcing

1. Introduction

Ordering speakers by likability (or other perceptory aspects of their speech quality such as comprehensibility, positiveness, coolness, ...) is an important yet controversial task and inherently subjective. Generalisations are still useful, as *intersubjective* agreement on the abovementioned criteria can often be found by-and-large. Generalisations are also necessary, for example to cast news speakers, readers or other speaking roles. Such castings are typically performed by small expert jurys and for small numbers of speaker candidates. In our work, we intend to use rankings to analyze the influencing factors of speaker likability for broad speaker populations. Hence, we are interested in full rankings rather than in who is the best speaker for a task. We aim to create rankings for large speaker populations, by large and diverse jurys, and while keeping the effort as low as possible.

We use recordings from the Spoken Wikipedia[2] as a broad sample of read *speech in the wild*. The Spoken Wikipedia project unites volunteer readers who devote significant amounts of time and effort into producing read versions of Wikipedia articles as an alternate form of access to encyclopaedic content. It can thus be considered a valid source of speech produced by ambitious but not always perfect readers. The data has been prepared as a corpus [1] and the German subset of the corpus, which we use here, contains ~300 hours of speech read by ~300 speakers.

To simplify the human effort involved in creating a ranking, we have participants take pairwise decisions on which of two stimuli is better. We then create a ranking from the pairwise comparisons. The number of possible pairs grows quadratically with the number of the stimuli compared. Thus, while full comparisons for each rater are possible for small speaker groups (10 speakers $\rightarrow$ 45 rating pairs), these are infeasible for large speaker groups (225 speakers $\rightarrow$ 25000 rating pairs), in particular when relying on volunteer raters. Thus, we need a method that is able to build rankings from incomplete comparisons. Note, however, that many of the ratings (with one strong and one weak speaker) will have predictable outcomes and human input on speakers of similar quality is most informative.

The main idea is to start from an initial ranking hypothesis which is iteratively revised as more evidence becomes available. Once the initial ranking is available, rating outcomes can be predicted and human effort can be directed away from comparisons with clear outcomes and towards more informative pairs.

We extend the methodology introduced by Sakaguchi et al. [2] who created rankings for machine translation systems from pairwise comparisons using Microsoft TrueSkill™ [3]. TrueSkill uses a Bayesian estimation of rankings from pairwise comparisons originally developed for ranking players of online games (based on their win/loss performance). The metaphor views each rating as a match in which the preferred stimulus wins against the dispreferred stimulus. We then compute the 'skill' of stimuli and their ranking. TrueSkill also provides *match making* capabilities that, given one player, select an opponent that has the most similar skill (which leads to interesting matches). We use match making to select stimulus pairs for human rating. In comparison to [2], which ranked 13 translation systems for which complete evaluation data had already been collected, we rank a total of 223 speakers, thus well over an order of magnitude more, in a live setting without external reference ranking.

The remainder of this paper is structured as follows: we detail our methodology for rating pair selection in Section 2 and describe our crowd-sourcing experiment in Section 3. We then validate our method and examine the derived rankings in Section 4 including an analysis of how rater properties influence their preferences. We briefly touch on acoustic correlates of perceived likability in Section 5 and conclude in Section 6.

2. Rankings from pairwise comparisons

Rankings have a long history in competitive sports, where individuals or teams play against each other in order to determine who's best. Two common forms, elimination and round-robin tournaments both require a high degree of control over who plays who and may lead to only partial rankings. In chess, Elo's system [4] was designed to overcome these issues: a player's

[1] Find our results at http://islrn.org/resources/684-927-624-257-3/: stimuli, ranking, pairwise ratings, the iPython notebook for analyses, and links to the software used.

[2] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Spoken_Wikipedia

http://dx.doi.org/10.21437/Interspeech.2017-1697

skill is estimated based on prior match outcomes, and skills are updated after each match, and the changes correspond to the surprisal of the system by the match outcome. A ranking can be derived by ordering players by their skill.

Microsoft TrueSkill™ models skill as a normal distribution, i. e., it makes the system's uncertainty about skill explicit, which enables smoother updates and more robust results when few match outcomes are available. TrueSkill also provides for match making: given one player, it finds the player that is most similar in skill and where uncertainty is low (technically, TrueSkill estimates the probability of a draw and prefers matches with high draw probability). This is meant to lead to interesting matches with similarly skillful opponents.

In our application, the abovementioned strategy for match making is flawed: as scores tend to get more certain with more data, stimuli are preferred that already participated in many comparisons. As a result, the number of comparisons is not balanced on all stimuli but accumulates on few, well known anchor points.[3] We use an approach that better balances the number of ratings per stimulus: We (1) pick a first stimulus based on the system's uncertainty about its ranking and (2) compute the match quality for all opponents and pick the opponent based on the predicted match quality with a dampening factor for the number of comparisons that the opponent has played so far. As a result, we (a) favour little-tested stimuli over well-tested ones and (b) select informative games over predictable ones. We randomly select pairs weighted by the criteria mentioned above which enables us to sample multiple 'interesting' pairs at once.

3. Stimuli and crowd-sourcing experiment

To avoid rating preferences based on *what* is spoken rather than how, we choose as stimuli the opening that is read for every article in the Spoken Wikipedia, which is (supposed to be) identical for all articles except for the article lemma.[4] We extract that stimulus for every speaker in the German subset of the Spoken Wikipedia Corpus using the alignment information given in [1]. As some alignment information was missing or clearly wrong, our stimulus pool is reduced to 227 speakers. We then masked the article lemma with noise in a length that matches the average reading speed of the stimulus. For every rating pair, participants were asked to rate which of the two voices they would prefer for having a Wikipedia article read out to them.

We realized the web-based rating experiment on the basis of BeaqleJS [5] which we extended to allow for an open number of pairwise ratings for each participant. The experiment operated with a mini-batch cache of 1000 rating pairs from which clients sampled randomly. The cache was updated manually whenever more than 200 ratings had been submitted by re-creating a new best ranking and selecting stimulus pairs as outlined above. We opted against an active backend with immediate update and selection of the next most relevant rating pair to ensure availability in times of high system usage (e. g. during the minutes after a mailing list advertised our experiment).

We solicited participants to our experiment via the German Wikipedia 'off-topic bulletin board' and various open mailing lists of student organizations (particularly CS students), as well as the Chaos Computer Club in Germany, Austria and Switzerland in order to reach a wide variety of dialect and age groups.

[3]This may not be a problem when using TrueSkill for match assignment, as not all players would be available at all times.

[4]Expected reading: "Sie hören den Artikel *article lemma* aus Wikipedia, der freien Enzyklopädie." (You are listening to the article *article lemma* from Wikipedia, the free encyclopedia.)

Table 1: *Breakdown of self-reported meta information of participants and their rating counts.*

		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

We deliberately did not explicitly invite the Spoken Wikipedia community to participate, as they could be particularly biased.

Statistics of the participants' self-reported meta data are shown in Table 1. As can be seen, Northern Germans, males, and 20-30 years olds (presumably computer science students at Universität Hamburg) are over-represented in our data. However, almost all other demographic groups are included as well, at least to some extent. In total, we collected 5440 ratings from 168 participants, none of whom received any compensation.

Although participants could perform as many ratings as they liked, they were instructed that 10 ratings are sufficient, 30-50 preferable, and that they should take a break after 100 ratings (and possibly return the next day). We excluded participants who submitted a single rating only. The median ratings per participant were 26 with half the participants between 11 and 43 ratings and 5/95 % quantiles at 4 and 101 ratings, respectively.

Participants were asked to always state a preference, even if unsure. It is more informative for our setup to get contradicting preferences than to explicitly invite the participants to omit a decision. As our method steers towards 'difficult' comparisons, many omitted decisions could otherwise have been expected. Our software, however, did allow to skip ahead without a decision and sometimes participants did not provide a decision (accidentally or on purpose). These instances were ignored in further processing, as no rating has been recorded.

We also measured the time taken for each rating. The median time per rating is 14.3 seconds with half the ratings between 11.3 and 21.3 s and 5/95 % quantiles at 6.3 and 39.7 s respectively. 6.3 seconds can still be considered a reasonable lower bound for listening to both stimuli and then taking the decision quickly. In total, participants spent ~26 hours on rating stimulus pairs.[5]

The stimulus ordering was randomized. Participants have a slight tendency for stimulus B over A (2784 vs. 2656, n.s.: sign

[5]We substitute the median for the slowest 2.5 % of ratings, as participants were obviously side-tracked who spent more than 55 s per rating.

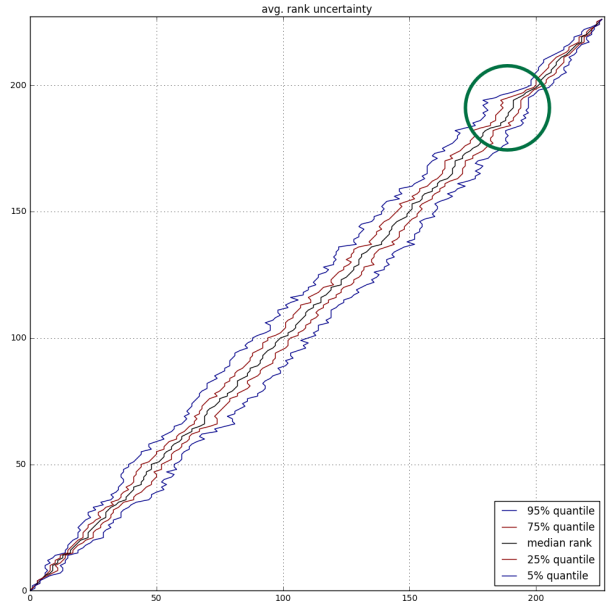

Figure 1: *Ranking results (both axes ordered by median ranking) including rank confidence on the x-axis. The circled area is further discussed in the text.*

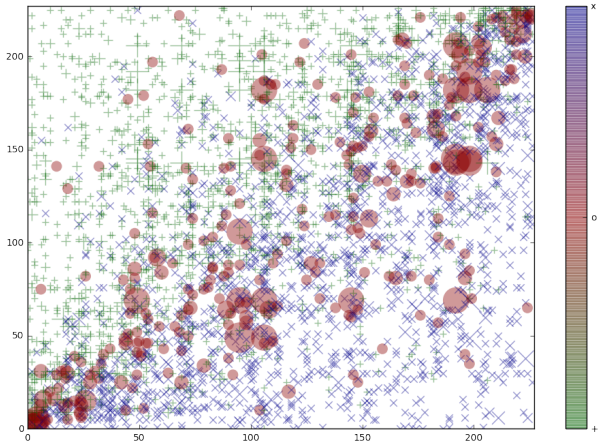

Figure 2: *Scatter plot of pairs compared (axes ordered by median ranking, color-coding indicates the avg. outcome of comparisons). The plot is more dense along the diagonal, as stimuli are compared more often when they are of comparable rank.*

test, $p = .09$), which could be interpreted as a recency effect.

4. Ranking analyses

We feed all pairwise ratings into TrueSkill™ to derive rankings. In TrueSkill, more recent ratings are more influential for the final ranking due to the iterative update mechanism.[6] As proposed by [2], we use the fact that rankings depend on the rating order to validate the ranking: we permute the ratings and create many rankings for the same set of ratings (below: N=300). We then take the median ranking as the final decision. Thus, we are also able to report ranking confidence levels.[7]

Rankings can be compared using correlation coefficients like Kendall's Tau [6, Ch. 16]. We find that pairwise correlations of the 300 rankings result in $\tau > 0.92$ and that each ranking against the median ranking gives $\tau > 0.95$. Thus, we conclude that TrueSkill leads to consistent rankings (within some confidence bounds) and that the median ranking is a meaningful middle ground for all rankings.

The final median ranking with quartile and 5/95 % confidence ranks is shown in Figure 1. As can be seen in the figure, there is no one clear ranking of all speakers. While there is a best and worst stimulus shared among all rankings, variability is larger in the middle. Overall, the average rank variability is 6.7 ranks within the 25-75 % confidence interval and 16.4 ranks within a 90 % confidence interval. Interestingly, some clusters of similarly 'good' stimuli emerge, e. g. as highlighted in the green circled area where 11 stimuli share similar ranks with a high variability that are delimited with high confidence to higher ranks (upper right of circled area) and slightly less to lower ranks.

The fact that no one clear rating emerges can also be attributed to disagreement in pairwise ratings. We measure the degree of disagreement by constructing a directed acyclic graph of the preference relation expressed through all ratings (i. e., the stimuli are nodes and one edge is introduced per rating). If ratings were consistent, there would not be any rating circles (a < b, b < c but c < a) and the proportion of feedback arcs can be taken as a measure of consistency. We heuristically compute the minimum feedback arc set of all ratings [7] and find the proportion to be 29 %. In a preliminary experiment using only 10 stimuli and all 45 possible comparisons, only one rater was 'perfect' in not producing any circles. Hence, we know that both within-rater and across-rater inconsistencies occur. In addition, our stimulus selection process is tailored towards choosing pairs that are expected to be hard to rate (and the disagreeing proportion grew over the runtime of the experiment).

Finally, we use rankings to predict the outcome of ratings as another way of testing the ranking validity. We assume that a rating will be 'won' by the better-ranked stimulus (although similarly ranked stimuli could easily have any outcome). We use 100-fold cross-validation and find that on average, the prediction performance is 68 %. Given that 29 % of ratings can be expected to be mis-predicted due to the rating inconsistencies, the rankings have a high level of predictive value. As described above, TrueSkill can compute match quality, effectively describing how likely a rating will lead to disagreement among raters. We find that prediction performance highly correlates with that score (Kendall's $\tau = -0.81, p < .001$).

We investigate which stimulus pairs have been selected for comparison to find out whether the method proposed in the preceding section works effectively. The rated pairs are presented in Figure 2. We find that pairs along the diagonal (i. e., with similar ranks) have been tested more densely than pairs further apart. Furthermore, the plot shows that 'better' stimuli (as per the ranking) win more often against inferior stimuli (green/blue division of the plot) and multiple contradictive ratings (red) mostly occur along the diagonal. Overall, our 5440 ratings spread over 4000 different pairs, that is, 7,7 % of all possible comparisons. 3057 pairs have been tested once, 666 pairs twice, and the remaining pairs up to 9 times (which seem to be artefacts of older versions of pair selection). Overall, the average stimulus has been rated

[6] This is a feature when ranking human players, as their true performance may change over time – but this is not the case in our experiment.

[7] The confidence is about TrueSkill producing a preference ordering given another permutation of ratings. We cannot make any guarantee with respect to some 'gold' ranking, which does not exist for our data.

Figure 3: *Line graph comparison of median rankings for female (top) and male (bottom) rankings. Female stimuli are shown in red.*

46 times with the 5/95 % quantiles at 39 and 56 ratings. Thus, our rating pair selection strategy successfully balances stimulus selection and opponent assignment.

Finally, we analyze the rankings wrt. to gender. We produce one median ranking each for ratings from female and male listeners (randomly subsampling the male ratings to the number of female ratings; see Table 1). We find only a moderate correlation ($\tau = 0.44, p \ll .001$) between female and male listener rankings, which indicates different preferences between these listener groups. We further analyze the ranking wrt. to speaker gender of the stimuli.[8] The rank assigned to a female speaker is on average 12.7 ranks better for female than for male listeners (half of the stimuli between -32 and +60 ranks), indicating that one major difference between female and male listeners is their preference towards female voices.

Figure 3 compares the gender-dependent rankings (each line corresponds to a stimulus, female stimuli in red). The less inclined a line, the more similar the rank for female/male listeners. As can be seen, preferences differ both in ranking female speakers as for male speakers. It is interesting to note that Dykema et al. [8] find that male speakers respond more truthfully to questions posed by female voices, yet they seem to dispref er them in our data. The results highlight the importance of gender-appropriate voice selection for reading factual information.

We also divide our data by age (<30 vs. >30) and dialect (Northern German vs. all other dialects as there is insufficient data to further differentiate among dialects). In both cases, correlation between the groups is stronger (age: $\tau = 0.50$, dialect: $\tau = 0.54$) than in the gender partition. No age or dialect information is available for the speakers, hence we cannot compare within/across-group effects (e. g. we would expect matched dialects of speaker and listener being preferred).

5. Acoustic correlates of ranking quality

We briefly experiment with acoustic factors that could explain the speaker likability expressed by the median ranking shown in Figure 1. First, we compute the perceptual quality of audio stimuli as standardized by ITU-T P.563 [9]. We find a low (but significant) correlation ($\tau = 0.14, p < .002$) of achieved median ranking and estimated MOS for the audio transmission quality.[9] We conclude that carefully arranged recording conditions could coincide with better speech quality, or that listener judgements are influenced by encoding quality – in contrast to [10] where no such influence was found in a similar task.

We estimate the liveliness of the speaker's prosody as it might be a relevant factor of likability. We compute the pitch range in semi-tones and take the 50 % (25-75 %) and 90 % (5-95 %) ranges of the measured pitch. On average, the 50 %/90 % ranges are 4.3/12.8 semi-tones for all speakers. We find very slight but non-significant correlations between either liveliness measure and the ranking. As this could be due to very little data from each short stimulus, we also extract pitch from the full articles. This allows us to estimate each speaker's liveliness *in general*, not just in the opening of the article. Here we find that the inter-quartile (50 %) pitch range correlates somewhat ($\tau = 0.10, p < .03$) with the ranking.

[8] Unfortunately, just 20 of 227 stimuli (9 %) were spoken by females.

[9] We must mention that all speech in the Spoken Wikipedia is distributed as OGG/Vorbis, with varying bit rates.

6. Discussion and future work

We have presented a method and experimental software for creating crowd-sourced speaker likability rankings from pairwise comparisons. The material that we base our ratings on is freely available and we likewise publish the ratings and the software to derive rankings from those ratings under the same terms.

Unlike [11] which uses Bradley-Terry-Luce models, our method does not require a complete comparison of all pairs, and works on a small subset jointly provided by many participants.

One advantage of the Spoken Wikipedia corpus is the availability of much more data from each speaker beyond the short stimuli that are used in the ranking experiment. Thus, more complex characteristics of a speaker such as accentuation and other prosodic idiosyncrasies (which listeners presumably would be able to judge in one sentence) can be derived from up to an hour of (closely transcribed and aligned) speech. We intend to use this wealth of data to extend our current acoustic analyses to those mentioned in [10] and beyond.

We have limited our study to one identical stimulus sentence in order to exclude contextual differences, and to one stimulus per speaker. We plan to extend the study to other stimulus pairs where the sentences (or sentence fragments) are spoken by different speakers across the Spoken Wikipedia. In this way, we hope to get a better judgement of the speakers, based on more than (on average) 4.7 seconds of speech.

Finally, multiple factor ranking systems [12] are able to account for external influences on the rating outcomes (features pertaining to the listener, recording, or stimulus rather than the speaker's voice itself) and generate more precise rankings and a better understanding of likability.

7. Acknowledgements

We thank our listeners/raters as well as the volunteers of the Spoken Wikipedia for donating their time and voice. This work is supported by a Daimler and Benz Foundation PostDoc grant.

8. References

[1] A. Köhn, F. Stegen, and T. Baumann, "Mining the Spoken Wikipedia for speech data and beyond," in *Proceedings of the Language Resource and Evaluation Conference*, 2016.

[2] K. Sakaguchi, M. Post, and B. Van Durme, "Efficient elicitation of annotations for human evaluation of machine translation," in *Pro-*

ceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, Maryland, USA: ACL, June 2014, pp. 1–11.

[3] R. Herbrich, T. Minka, and T. Graepel, "Trueskill™: A Bayesian skill rating system," in *Advances in Neural Information Processing Systems 20*. MIT Press, January 2007, pp. 569–576.

[4] A. E. Elo, *The rating of chessplayers, past and present*. Arco Pub., 1978.

[5] S. Kraft and U. Zölzer, "BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality," in *Linux Audio Conference*, 2014.

[6] A. N. Langville and C. D. Meyer, *Who's #1? The Science of Rating and Ranking*. Princeton University Press, 2012.

[7] P. Eades, X. Lin, and W. F. Smyth, "A fast and effective heuristic for the feedback arc set problem," *Information Processing Letters*, vol. 47, no. 6, pp. 319–323, 1993.

[8] J. Dykema, K. Diloreto, J. L. Price, E. White, and N. C. Schaeffer, "ACASI gender-of-interviewer voice effects on reports to questions about sensitive behaviors among young adults," *Public opinion quarterly*, vol. 76, no. 2, pp. 311–325, 2012.

[9] L. Malfait, J. Berger, and M. Kastner, "P.563 — The ITU-T standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, Nov 2006.

[10] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, ""would you buy a car from me?"-on the likability of telephone voices," in *Proceedings of Interspeech*. ISCA, 2011.

[11] L. F. Gallardo, "A paired-comparison listening test for collecting voice likability scores," in *Speech Communication; 12. ITG Symposium; Proceedings of*. VDE, 2016, pp. 1–5.

[12] M. Stanescu, "Rating systems with multiple factors," Master's thesis, School of Informatics, University of Edinburgh, 2011.