Ranking and Comparing Speakers based on Crowdsourced Pairwise Listener Ratings

Timo Baumann

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, generalizations are useful and there is often a broad intersubjective consensus about whether one speaker is more likable than another. We present a methodology to efficiently create a likability ranking for many speakers from crowdsourced pairwise likability ratings which focuses manual rating effort on pairs of similar quality using an active sampling technique. Using this methodology, we collected pairwise likability ratings for many speakers (>220) from many raters (>160). We analyse listener preferences by correlating the resulting ranking with various acoustic and prosodic features. We also present a neural network that is able to model the complexity of listener preferences and the underlying temporal evolution of features. The recurrent neural network achieves remarkably high performance in estimating the pairwise decisions and an ablation study points towards the criticality of modeling temporal aspects in speech quality assessment.

1 Introduction

Speaker traits (such as age or gender), emotional coloring (such as anger or distress), socio-cultural aspects (such as accent or dialects), conscious or subconscious coloring towards the addressee (such as friendliness or positivity), and other paralinguistic aspects (such as clarity and comprehensibility) are expressed through various prosodic, suprasegmental, segmental and non-segmental aspects of one's speech and voice, where the combination of features and their temporal evolution

Timo Baumann
Universität Hamburg, Language Technology Group, e-mail: mail@timobaumann.de

are far from trivial. Intermittent deficiencies (e. g. a lisp) or deviations limited to a few features (e. g. nasalisation) can already have strong influences on the perceived quality. Together, they form the 'quality' of speech. It is important to note that no one 'best' combination of all features exists that would constitute 'ideal' speech.

Voice is a highly personal and subjective matter such that a multitude of combinations of these features result in a 'good' voice. This often makes likability comparisons hard and inherently subjective. Despite subjective preferences, *intersubjective* agreement on the preferences can often be found by-and-large, making generalizations useful. Generalisations are also necessary, for example to cast news speakers, readers or other speaking roles that need to approximate an intersubjective consensus. Such castings are typically performed by small expert jurys (potentially limiting the universality of decisions) and for small numbers of speaker candidates (for practical reasons).

In our work, we use rankings to analyze the influencing factors of speaker likability for broad speaker populations, or to eventually 'score' a voice sample along a range of speakers. Hence, we are interested in full rankings rather than in who is the best speaker for a task. Our aim is to create rankings for large speaker populations, by large and diverse jurys, and while keeping the effort as low as possible.

To simplify the human effort involved in creating the ranking, we have participants take many pairwise decisions on which of two stimuli is better. We then create a ranking from the pairwise comparisons (see below). 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 → 45 rating pairs), these are infeasible for large speaker groups (225 speakers → 25200 rating pairs), in particular when relying on volunteer raters. Thus, our method must be able to build rankings from incomplete comparisons. Note, however, that many of the ratings will have predictable outcomes if one known-strong and one known-weak speaker are paired. It will be helpful to not waste too much human effort on such pairs; in contrast, human input on speakers of similar (or unknown) quality is most informative.

The main idea is to start from an initial ranking (based on some initial ratings) which is iteratively revised as more evidence becomes available with more ratings. 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 the most informative pairs; this will be described in detail in Section 2.

Section 3 describes the corpus developed via crowdsourcing and based on the iterative method, both in terms of the stimuli used, as well as the resultant preference ratings. Section 4 examines the overall preference ranking derived from all pairwise ratings and finds some explaining factors in terms of high-level properties of the speech stimuli (and their speakers) via linear correlations.

As outlined above, however, prosody is a highly non-linear phenomenon and we hence build a recurrent neural network-based model that successfully identifies listener preferences using non-linear (but opaque) aggregation functions. Via an ablation study we find that the tunes in to phone-specific prosodic aspects given phonetic identity as additional features. Section 5 describes model for estimating the preferences of raters and analyses the importance of features for modeling speaker preference. We conclude that modeling the *temporal aspects* of speech is critical for preference estimation.

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, which is not always possible). In addition, they may lead to only partial rankings. In chess, Elo's system [11] was designed to overcome these issues: a player's skill is estimated based on prior match outcomes, and skills are updated after each match. Skill 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™ [14] uses a Bayesian estimation of rankings from pairwise comparisons originally developed for ranking players of online games (based on their win/loss performance). 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.

Most work in speech quality estimation has used direct scalar ratings of individual stimuli [7] or required each subject to assign a complete ranking for all stimuli. Fernández Gallardo [13] feeds paired comparisons into a Bradley-Terry-Luce model [5] and finds similar results to direct scaling. Both of these methods have been limited to few raters and/or few stimuli. We extend the methodology introduced by Sakaguchi et al. [21] who created rankings for machine translation systems from pairwise comparisons using Microsoft TrueSkill™. In our metaphor, we view 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 and where uncertainty of the skill difference 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. We use match making to select stimulus pairs for human rating in an iterative fashion which uses the ratings collected so far to steer our *active sampling* approach to select among the possible stimulus pairs to be compared. We actively select stimulus pairs that are expected to be informative for the full ranking based on a preliminary ranking of all ratings performed so far.

In our application, we found the abovementioned strategy for match making to be 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.[1]

[1] This may not be a problem when using TrueSkill for match assignment, as participation in games is limited by the players' availability.

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.

In comparison to [21], 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.

3 Stimuli and rating collection via crowd-sourcing

We limit our likability judgements to one specific reading genre: the reading of encyclopaedic entries in Wikipedia. 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 [2] and the German subset of the corpus, which we use here, contains ~300 hours of speech read by ~300 speakers.

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.[3] We extract that stimulus for every speaker in the German subset of the Spoken Wikipedia Corpus using the alignment information given in [2]. 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. The mean/median duration per stimulus is 4.7/4.57 s with 5/95 % quantiles at 3.74/6.03 s.

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 a web-based rating experiment on the basis of BeaqleJS [16] 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

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

[3] 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

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. We deliberately did not explicitly invite the Spoken Wikipedia community to participate, as they could have been 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 are over-represented in our data (presumably computer science students at Universität Hamburg). However, almost all other demographic groups are included as well, at least to some extent. In total, we collected 5440 ratings from 168 participants. Participation was strictly voluntary and without compensation and hence the resulting ratings are unlikely to be prone to vandalistic behaviour.

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, and did not explicitly have the option to state that they could not decide. 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 making 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.[4]

The stimulus ordering was randomized. Participants have a slight tendency for stimulus B over A (2784 vs. 2656, n.s.: sign test, $p = .09$), which could be interpreted as a recency effect.

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 [10] 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).

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.[5] As proposed by [21], we use the fact that rankings depend on the rating order to validate our method: 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.[6]

[4] We substitute the median for the slowest 2.5 % of ratings, as participants were obviously sidetracked who spent more than 55 s for a single rating.

[5] 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.

[6] 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.

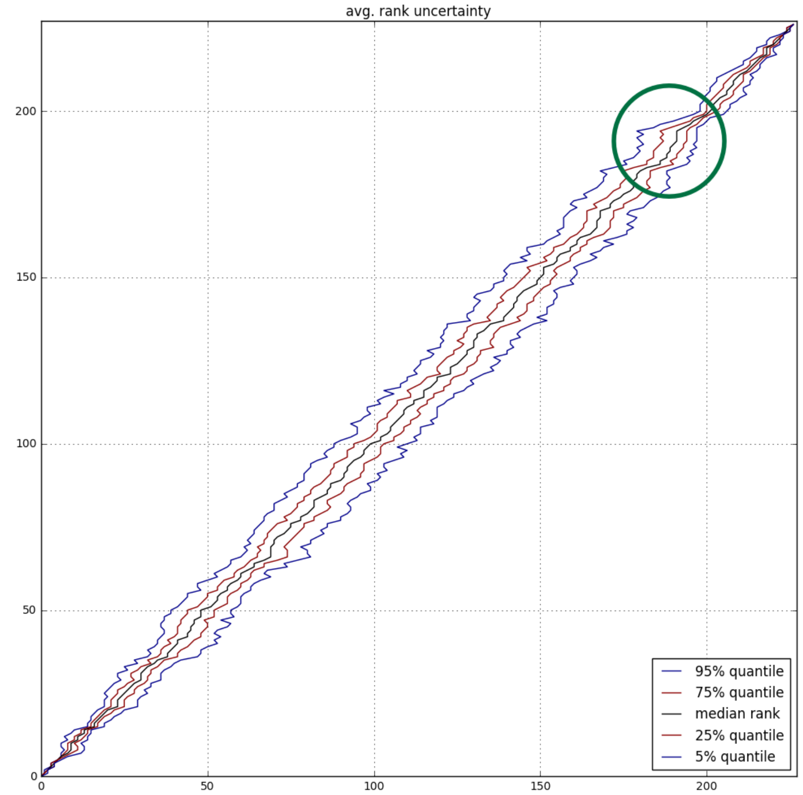

Fig. 1 Ranking results (both axes ordered by median ranking) including rank confidence on the x-axis. The circled area is further discussed in the text.

Rankings can be compared using correlation coefficients like Kendall's Tau [17, 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 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 the 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.

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-

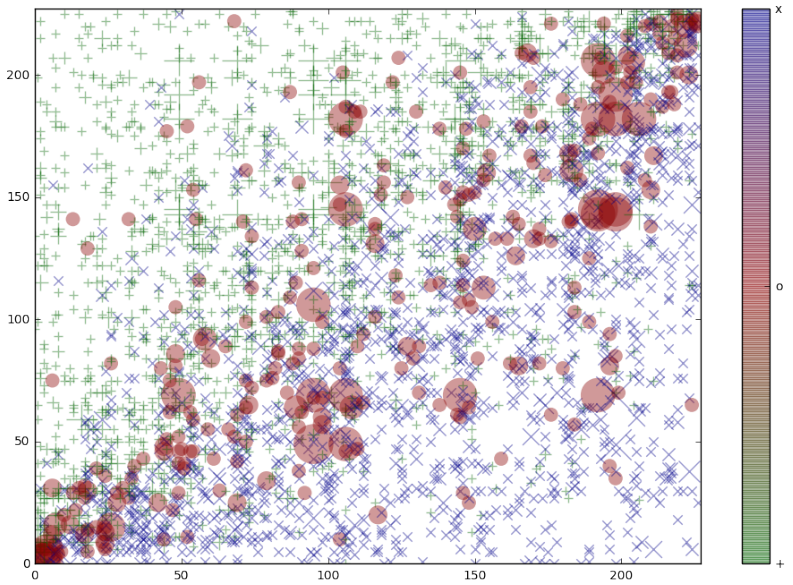

Fig. 2 Scatter plot of pairs compared (axes ordered by median ranking, color-coding indicates the avg. outcome of comparisons). The plot is more dense along the diagonal, as stimuli are compared more often when they are of comparable rank.

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 Section 2 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 controversial 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 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.

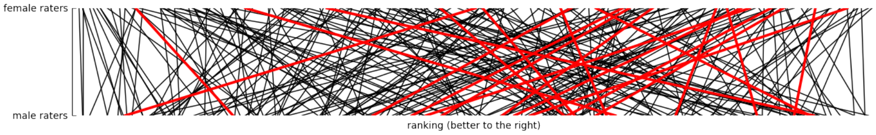

Fig. 3 Line graph comparison of median rankings for female (top) and male (bottom) raters. Stimuli spoken by females are shown in red.

4.1 The influence on rater population on ranking outcome

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.[7] 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. [9] find that male speakers respond more truthfully to questions posed by female voices, yet they seem to disprefer them in our data. The results highlight the importance of gender-appropriate voice selection for reading encyclopaedic, and possibly other 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).

4.2 Acoustic correlates of ranking quality

We finally 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 [19]. We find a low (but significant) correlation ($\tau = 0.14, p < .002$) of achieved median ranking and

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

estimated MOS for the audio transmission quality.[8] 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 [7] 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.

5 Listener preference classification

In previous work [7], speaker likability has been modeled using OpenSmile [12] features based on linear and non-linear aggregation functions (such as means and medians) to aggregate over the duration of the stimulus. Features were used to train classifiers such as SVMs which resulted in moderately high (better than chance) performance in classifying speakers as being above or below median likability [7]. Like the analyses in Section 4.2, the abovementioned aggregation functions cannot take into account the context of feature characteristics in the stimulus, and are unlikely to accurately express more fine-grained details relevant for speech quality (such as where and how a pitch accent is realized, beyond mean pitch). In this section, we experiment with neural sequence-learning methods (RNNs) to encode the complex temporal evolution of features of speech quality into a latent feature space and use the difference in these for pairs of speech stimuli to train our classifier.

5.1 Model architecture

The task of preference ranking is *asymmetric* in the sense that if the two stimuli to be compared are swapped, then the comparison result is the opposite. This has two consequences: (a) parameters for sequence analysis of both stimuli can be shared which is called a *siamese model* [6] and reduces the degrees of freedom of the model, making learning more efficient, and (b) the outputs from sequence analysis of each stimulus can simply be subtracted and the difference be subjected to a final decision layer.

[8] We must mention that all speech in the Spoken Wikipedia is distributed as OGG/Vorbis, with varying bit rates and under diverse recording conditions.

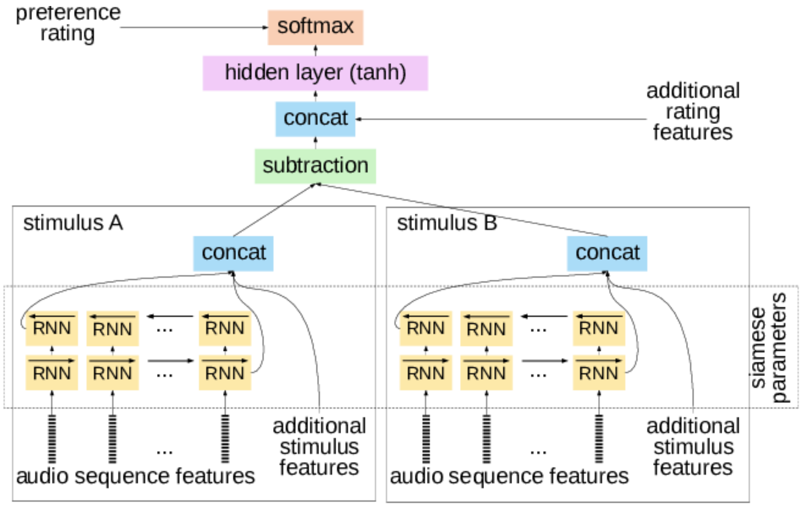

Fig. 4 Diagram of the neural architecture for speech likability preference. The task is symmetric (whether a stimulus is A or B is irrelevant) and hence the parameters for the RNNs can be shared (siamese network). Additional features about stimuli and the rating can be concatenated in.

In our model and as shown in Figure 4, we use two layers of bidirectional RNN (LSTMs [15] or GRUs [8]) to model the feature sequence of each stimulus and concatenate the outputs of the forward and backward pass. We can also concatenate additional stimulus-level features into the representation at this time, e. g. measures of signal quality such as ITU-T P.563 [19] (cmp. Section 4.2), or meta information about the speaker or the audio recording (such as gender or bitrate, cmp. Section 4.1).

Given that our final decision is based on the quality *difference* alone, not the overall quality, we subtract both stimuli's vectors.[9] We then pass the difference to one hidden layer and a final binary softmax layer that models the preference decision. We can also optionally include additional meta features of the rating (such as identity, age or dialectal region of the rater). These can easily be concatenated in before the hidden layer, in order to model the relative preferences of individual raters or rater groups. As we found that preferences differ, this could be useful information.

5.2 Data and evaluation

The original purpose of the rating collection reported in Section 3 was to create a ranking and effort was put into maximizing the efficiency of human annotation by focusing the human effort on 'difficult' pairs using *active sampling* of stimulus pairs that focus human annotation effort on 'similarly good' speakers. As a result, the

[9] We found, in initial experimentation, that this performs much better than concatenating the outputs of each speaker.

stimulus pairs that were rated by participants are much more similar in their quality than randomly selected stimulus pairs would be.

In addition, inconsistency in the data set is high, as are pairs of stimuli that have been rated multiple times. Above, we have computed the minimum feedback arc set, i. e., the subset of ratings that lead to a fully consistent ranking [10]. We found the proportion of conflicting arcs to be 29 %. This can act as an indicator of the proportion of ratings that are inconsistent (where potentially different raters have different preferences, or simply cannot reliably tell the difference). In addition, we here compute an oracle correctness for all pairs that have been rated more than once, by checking for each rating, if it is the majority rating for this pair (deciding randomly to resolve draws). We find that such an *oracle classifier* reaches a correctness of only 65 % for those pairs that have been rated more than once. Pairs that were rated just once *potentially* are easier to classify, which makes it possible to beat this oracle.

For evaluations, we report multiple settings below. The settings are meant to counter-balance the difficulties introduced by the data elicitation technique and to test different aspects of listener preference classification:

naïve we sample randomly among the evaluation instances from the corpus of human-rated pairs; as the corpus focuses on difficult pairs, we cannot expect a spectacular performance;

easy based on the median ranking derived in Section 4, we sample instances with 'large' ranking differences (distance on the ranking scale > 0.25 or > 0.5), in order to test if our classifiers fare better with stronger preference differences (and hence easier to identify differences in speech quality).

Given that stimuli were presented in random order, the data set is balanced in terms of which stimulus outperforms the other. Thus, we focus on accuracy as the only evaluation metric.

5.3 Features and conditions

Using a sliding window, we derive a multitude of local features from the audio stream that might capture aspects of speech quality. All features use a frame shift of 10 ms. In particular, we measure Mel-frequency cepstral coefficients (MFCCs, 12+1 energy) to capture voice and recording characteristics, f_0 (measured using `Snack`'s `esps` implementation) as a first measure of speech melody, and Fundamental Frequency Variation (FFV) features [18] as these are more robust (and might contain more valuable information) than single f_0. Using Praat [4], we compute jitter (PPQ5), shimmer (APQ5), and harmonics-to-noise ratio [3]. We do not perform z-scale normalization on the feature streams.

The Spoken Wikipedia Corpus also contains phonetic alignments that were computed using the MAUS tool [22]. The alignments allow us to assign phone annotations to every frame. With this information, the model is informed explicitly that different phones have different phonetic characteristics (as expressed in the MFCCs) and can condition its learning of speech quality on these characteristics. In other words: the model can learn to focus on a phone's quality aspects (e. g. nasalization) without needing to learn to differentiate phones.

One frame of features for every 10 ms may overwhelm the model with very large amounts of parameters, reducing training efficiency as well as effectiveness. In order to keep training tractable, we subsample the feature frames with various values (see *seq. step size* in Table 2). When we do so, we use mean aggregation for numeric values (ignoring missing values for pitch and HNR).

5.4 Experiments and results

We separate out about 1/10th of the 5440 ratings as the test data: the **naïve** test set contains 400 ratings, and we sample among ratings with 'large' differences 100 ratings each for the > 0.25 and > 0.5 **easy** test sets.

We implemented our network in dynet [20]. In the experiments reported below, we train for 50 epochs using AdamTrainer and no dropout. We concatenate the various audio features that are computed for every frame. We use embeddings to characterize the phonetic labels.

5.4.1 Meta parameter optimization

As originally reported in [1], we have performed an optimization to find good sizes for various meta parameters of the model:

- To reduce the length of the sequence that need to be learned by the LSTMs (and to avoid the problem of vanishing gradients through long sequences), we subsample the audio features by mean-aggregating values over a number of frames (5, 10, or 15).
- To represent the discrete phonetic labels, we use embeddings of varying sizes (8, 16, or 24), in order to allow the model to cluster similar phones.
- The sequential LSTM state size determines how many dimensions can be considered during the sequence analysis and we experiment with various sizes (24, 32, 48, or 64).
- The output from concatenation of both forward and backward LSTMs doubles the size of the next layer's input. For the hidden layer size, we hence consider scaling factors (2, 3, or 4) over the size of the sequential state size.

We performed a grid search over the possible meta parameter values as summarized in Table 2 and focusing on the naïve data set. We found an optimum for sequence step size of 5 (i. e., one feature frame for every 50 ms of speech), phone embedding with 16 dimensions, sequence state size of 48, and hidden layer size of $3 \times 48 = 144$ (sequence state size of 32 and $4 \times 32 = 128$ was a close contender).

At these settings, our model yields an accuracy of 67.25 % on the naïve test set, 93 % on the **easy-0.25** test set and 97 % on the **easy-0.5** test set. The accuracy on the naïve test set is close to what we estimated as the upper limit for the harder part of our training data.

5.4.2 Ablation study on phonemic alignments

We hypothesized above that our performance gain over previous work may be largely due to the model being able to perform prosodically meaningful aggregations and could, for example, relate prosodic parameters to the phones spoken. To test this hypothesis, we perform an ablation study and remove the phoneme embeddings from the input features. We perform this experiment with the other meta parameters set to their optima as found in the previous subsection. As shown in Table 3, we find performance to drop substantially when the phone identity feature is removed. We believe this is because the model is unable to make maximum sense of features such as MFCCs given speech quality is obviously just a secondary feature, far behind phone identities. If the model is not informed about the phonetic identities, it needs to resolve whether input has good quality, whereas the full model only needs to resolve the quality of a feature given the particular speech sound.

6 Conclusions and future work

We have presented a method 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 [13] which uses Bradley-Terry-Luce models, our method does not require a complete comparison of all pairs, and works on a small subset (in our case: 7 % of possible comparisons) 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. In fact, we found in Section 4.2 that extracting the 50 % pitch

Table 2 Meta parameters considered in grid search. Best values are shown in boldface.

meta parameter	values
sequence step size	**5**, 10, 15
phone embed size	8, **16**, 24
sequence state size	24, 32, **48**, 64
hidden layer size	2, **3**, 4 × sequence state size

Table 3 Accuracy (in percent) of full and reduced feature set (without phone alignment).

setting	accuracy **naïve**	**easy-0.25**	**easy-0.5**
full mode	67.25	93	97
without phone alignment	58.75	73	80

range as an estimate of liveliness significantly correlates with likability, at least if liveliness is extracted from the full speech, rather than just the one sentence used in human ratings, potentially because this circumvents effects from faulty fundamental frequency extraction.

We have also presented a neural architecture for determining which of two speech stimuli is rated as the better of the two in noisy human annotations. Our model yields good performance most likely because the RNN provides for complex aggregations of the (conventional) feature sequences. Our model's aggregations are able account for sequential information, in particular it is able to relate acoustic features to the phones spoken, unlike more coarse-grained aggregation functions as have been used before.

In [7], the authors train classifiers to differentiate whether a stimulus is better/worse than average and reach a classification accuracy of 67.6 %. Their setup is comparable to our decisions for stimuli that are relatively far apart on the rating scale, in which case the neural aggregation and classification yields a classification accuracy of 93–97 %. We believe this to be caused by the better temporal modeling of our approach and the use of phonetic identities during aggregation.

Despite the relatively good results, our method is still basic in terms of the neural architecture employed. In particular, our method does not yet employ an attention mechanism that could help to better weigh the speech quality encoding. Given that all speakers in our corpus speak (more or less) the same content, we envision that our model would profit greatly if the comparison between both stimuli could attend to particular differences rather than only the comparison of the final BiLSTM output vectors. An attention model would also help the analysis of *why* a speaker is rated as better than another, as it would indicate the relative importance of parts of the stimuli in the comparison. Another venue, at least for comparisons on shared text would be connectionist temporal classification to temporally relate the feature streams before comparison for a better notion of timing differences between the stimuli. Finally, it might be worthwhile to pre-train the intermediate representations of the model.

In the end, our model could weigh slight mis-pronunciations against voice quality or prosodic phrasing, and we intend to use analysis techniques to ultimately understand the relative weights of these aspects in comparisons.

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 their speech.

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

References

[1] Timo Baumann. “Learning to Determine Who is the Better Speaker”. In: *Proceedings of Speech Prosody*. 2018.

[2] Timo Baumann, Arne Köhn, and Felix Hennig. “The Spoken Wikipedia Corpus Collection: Harvesting, Alignment and an Application to Hyperlistening”. In: *Language Resources and Evaluation* (2018). Special Issue representing significant contributions of LREC 2016. ISSN: 1574-0218. DOI: 10.1007/s10579-017-9410-y.

[3] P. Boersma. “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound”. In: *Proceedings of the Institute of Phonetic Sciences* 17 (1993), pp. 97–110.

[4] P. Boersma. “Praat, a system for doing phonetics by computer”. In: *Glot international* 5.9/10 (2002), pp. 341–345. ISSN: 1381-3439.

[5] Ralph Allan Bradley and Milton E Terry. “Rank analysis of incomplete block designs: I. The method of paired comparisons”. In: *Biometrika* 39.3/4 (1952), pp. 324–345.

[6] Jane Bromley et al. “Signature verification using a" siamese" time delay neural network”. In: *Advances in Neural Information Processing Systems*. 1994, pp. 737–744.

[7] Felix Burkhardt et al. ““Would You Buy a Car from Me?”-On the Likability of Telephone Voices”. In: *Proceedings of Interspeech*. ISCA. 2011.

[8] Kyunghyun Cho et al. “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. URL: http://www.aclweb.org/anthology/D14-1179.

[9] Jennifer Dykema et al. “ACASI gender-of-interviewer voice effects on reports to questions about sensitive behaviors among young adults”. In: *Public opinion quarterly* 76.2 (2012), pp. 311–325.

[10] Peter Eades, Xuemin Lin, and William F Smyth. “A fast and effective heuristic for the Feedback Arc Set Problem”. In: *Information Processing Letters* 47.6 (1993), pp. 319–323.

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

[12] Florian Eyben, Martin Wöllmer, and Björn Schuller. “Opensmile: the munich versatile and fast open-source audio feature extractor”. In: *Proceedings of the*

18th ACM international conference on Multimedia. ACM. 2010, pp. 1459–1462.

[13] Laura Fernandez Gallardo. "A Paired-Comparison Listening Test for Collecting Voice Likability Scores". In: *Speech Communication; 12. ITG Symposium; Proceedings of*. VDE. 2016, pp. 1–5.

[14] Ralf Herbrich, Tom Minka, and Thore Graepel. "TrueSkill™: A Bayesian Skill Rating System". In: *Advances in Neural Information Processing Systems 20*. MIT Press, Jan. 2007, pp. 569–576.

[15] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[16] Sebastian Kraft and Udo Zölzer. "BeaqleJS: HTML5 and JavaScript based Framework for the Subjective Evaluation of Audio Quality". In: *Linux Audio Conference*. 2014.

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

[18] Kornel Laskowski, Mattias Heldner, and Jens Edlund. "The fundamental frequency variation spectrum". In: *Proceedings of FONETIK 2008*. 2008.

[19] L. Malfait, J. Berger, and M. Kastner. "P.563 — The ITU-T Standard for Single-Ended Speech Quality Assessment". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.6 (Nov. 2006), pp. 1924–1934. ISSN: 1558-7916. DOI: `10.1109/TASL.2006.883177`.

[20] Graham Neubig et al. "DyNet: The Dynamic Neural Network Toolkit". In: *arXiv preprint arXiv:1701.03980* (2017).

[21] Keisuke Sakaguchi, 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*. Baltimore, Maryland, USA: ACL, June 2014, pp. 1–11.

[22] Florian Schiel. "MAUS goes iterative". In: *Proceedings of the LREC*. 2004.