"Would You Buy A Car From Me?" – On the Likability of Telephone Voices

Felix Burkhardt[1], *Björn Schuller*[3], *Benjamin Weiss*[1,2], *Felix Weninger*[3]

[1]Deutsche Telekom Laboratories, Berlin, Germany
[2]Quality & Usability Lab, Technische Universität Berlin, Germany
[3]Institute for Human-Machine Communication, Technische Universität München, Germany
[Felix.Burkhardt | BWeiss]@telekom.de, [schuller | weninger]@tum.de

Abstract

We researched how "likable" or "pleasant" a speaker appears based on a subset of the "Agender" database which was recently introduced at the 2010 Interspeech Paralinguistic Challenge. 32 participants rated the stimuli according to their likability on a seven point scale. An Anova showed that the samples rated are significantly different although the inter-rater agreement is not very high. Experiments with automatic regression and classification by REPTree ensemble learning resulted in a cross-correlation of up to .378 with the evaluator weighted estimator, and 67.6 % accuracy in binary classification (likable / not likable). Analysis of individual acoustic feature groups reveals that for this data, auditory spectral features seem to contribute the most to reliable automatic likability analysis.

Index Terms: speaker traits, likability, classification

1. Introduction

How much we like a speaker based on the sound of her/his voice and manner of speaking is a fascinating topic. It is part of a higher family of problems: speech based classification. Speech based classification attempts to categorize people based solely on their voice and way of speaking. The categories may be relatively invariant like age, gender or dialect, or time changing like emotional state. These features differ strongly with respect to the extent they can be detected in a speech signal, for example the sex of a person can be found out with high probability due to the fact that women have shorter vocal folds and therefore speak with a higher pitch than men. Women have proven to be consistent in their estimation of pleasantness of men's voices, and also height, weight and age, although height was not correctly estimated [1]. Listeners can also ascribe personality traits – like the introversion-extroversion opposition – and attractiveness to a speaker purely based on short samples of his/her voice, e. g., [2].

In previous research [3], we have already investigated how "likable" or "pleasant" a speaker appears based on the Emo-DB, a database of 10 actors simulating emotional arousal. One of the biggest drawbacks of this study was the limited number of speakers, which is by far not enough to generalize over speaker specific differences. In search for a larger database, ideally publicly available in order to facilitate comparison studies, we decided to use the Agender database, [4]. The Agender database was recorded originally to study automatic speaker age and gender detection in voice portals and it contains about 940 speakers of mixed age and gender recorded over landline and mobile telephone network. The fact that this data is of limited bandwidth and the single utterances consist only of a few words is a disadvantage with respect to likability rating, but corresponds to the constraints given in a real application scenario, e. g., if a call center agent would like to test the likability of his/her own voice. Generally the research on likable voices has many applications, e. g., to enhance text-to-speech synthesizers or for self-assessment, further discussed in [5].

The article is structured as follows. In section 2, the selection of the audio data is explained. Then, section 3 reports on the procedure to judge the samples by human listeners. The next section analyses the results with respect to consistency between listeners and hidden dependencies. Section 5 finally attempts an automatic classification and regression of the perceived likability. The last section concludes the article.

2. Data selection

The spoken content of the database is based on 18 utterances taken from a set of utterances listed in detail in [4]. The topics of these were *command words*, *embedded commands*, *month*, *week day*, *relative time description*, *public holiday*, *birth date*, *time*, *date*, *telephone number*, *postal code*, *first name*, *last name*, *yes/no* with according free or preset inventory and corresponding 'eliciting' questions as *"Please tell us any date, for example the birthday of a family member."*.

The database contains at least 100 German speakers for each of seven age/gender groups acquired from all German Federal States without perfect balance of German dialects. The age sub-clusters (7-14, 15-24, 25-54, 55-80 years) are of equal size: to account for the different age intervals of the groups, CHILDREN and YOUTH are uniformly distributed within 2 year clusters and ADULTS and SENIORS in 5 year clusters. This means, for example, that 25 children from seven to eight years and 20 young-aged females between 17 to 18 years participated. All age groups, including the CHILDREN, have equal gender distribution. For the experiments described in this paper we excluded the children with the aim to reduce data. It is probably hard to judge likability of a child's voice because one tends to find children 'cute' in any case.

Excluding the children, we came up with the age and gender distribution shown in Table 1. Because this approach still leaves 800 speakers, we used only one sentence of the available data per speaker, in order to keep the effort for judging the data by many listeners as low as possible. To select the sentence, we looked at the phrases that consist of a command embedded in a free sentence (*s4* and *s5* from the database) and searched for the longest sentence available for each participant, based on the number of word tokens. This resulted in sentences with maximum eight words length (mean: 4.4 tokens). Typical sentences would include "*mach weiter mit der Liste*" ("*continue with the list*") or "*ich hätte gerne die Vermittlung bitte*" ("*I'd like an operator please*"). We're aware of the fact that the meaning of the

words might affect the perceived likability and it would have been better to have the same text spoken by all test speakers, but the database does not include longer texts of same wording for all speakers.

Table 1: *Distribution of age (Y: young, A: adult, S: senior) and gender (F, M) groups in the data.*

# YF	# YM	# AF	# AM	# SF	# SM	sum
121	112	135	129	147	155	800

3. Judging the likability

To control for effects of gender and age group on the likability ratings, the stimuli were presented to the participants in the following six blocks: male and female youths, adults and seniors, respectively. To mitigate effects of fatigue or boredom, each of the 32 participants (17 male, 15 female, aged 20–42, mean=28.6, standard deviation=5.4) rated only three out of the six blocks in randomized order with a short break between each block. The order of stimuli within each block was randomized for each participant as well. One rating session took about one hour. In other words, the whole data set was rated 16 times by a pair of raters, and 16 ratings from different individuals are available per instance. The participants were instructed to rate the stimuli according to their likability, without taking into account sentence content nor transmission quality. The rating was done on a seven point scale. For playing back Sennheiser HD 485 headphones were used. No participant reported hearing loss. All participants were paid for their service.

4. Data analysis

A preliminary analysis of the data shows no significant impact of participants' age or gender on the ratings (mixed effects model; gender: $F(1, 28) = 1.44$, $p = .24$; age: $F(1, 28) = 1.62$, $p = .44$), whereas the samples rated are significantly different ($F(799, 11970) = 4.94$, $p < .0001$).

All ratings are normalised by the evaluator weighted estimator (EWE) [6]. Informally, the EWE is a weighted mean likability rating, with cross-correlations as weights (see Eqn. 2, Section 4). Controlling for significant effects of variation in the transmission quality on the ratings is done with the instrumental method recommended by the ITU for no-reference cases (ITU-T Rec. P.563) [7]. As intended by the instruction there is no significant correlation between the averaged ratings and quality estimates (Spearman's $rho = .04$, $p = .27$).

Table 2: *60 low-level descriptors (LLD).*

4 energy related LLD
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate
50 spectral LLD
RASTA-style filt. auditory spectrum, bands 1–26 (0–8 kHz)
MFCC 1–12
Spectral energy 25–650 Hz, 1 k–4 kHz
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90
Spectral Flux, Entropy, Variance, Skewness, Kurtosis, Slope
5 voice related LLD
F_0
Probability of voicing
Jitter (local, delta)
Shimmer (local)

Table 3: *33/6 applied functionals.*

33 base functionals
quartiles 1–3
3 inter-quartile ranges
1 % percentile ($\approx$ min), 99 % percentile ($\approx$ max)
percentile range 1 %–99 %
arithmetic mean, standard deviation
skewness, kurtosis
mean of peak distances
standard deviation of peak distances
mean value of peaks
mean value of peaks – arithmetic mean
linear regression slope and quadratic error
quadratic regression a and b and quadratic error
contour centroid
duration signal is below 25 % range
duration signal is above 90 % range
duration signal is rising/falling
gain of linear prediction (LP)
LP Coefficients 1–5
6 F_0 functionals
percentage of non-zero frames
mean, max, min, std. dev. of segment length
input duration in seconds

5. Automatic analysis

5.1. Acoustic feature set

We use the baseline feature set of the INTERSPEECH 2011 Speaker State Challenge, which was extracted by the open-source feature extractor openSMILE [8] that also provided the features for the Challenge, to ensure compatibility of results. It consists of 4 368 acoustic features comprising features built from three sets of low-level descriptors (LLD) and one corresponding set of functionals for each LLD set. The LLD sets are given in Table 2: A major focus is on auditory features, including an auditory spectrum derived loudness measure and the use of RASTA-style filtered auditory spectra instead of conventional Mel-spectra. Further, a base set of 33 functionals is introduced as shown in Table 3. To the 54 energy and spectral LLD and their first order deltas, the base functional set and the mean, max, min, and the standard deviation of the segment length are applied, resulting in 3 996 features. To the 5 pitch and voice quality LLD and their first order deltas, the base functional set as well as the quadratic mean and the rise and fall durations of the signal are applied only to voiced regions (probability of voicing greater 0.7). This adds another 360 features. Another 12 features are obtained by applying a small set of six functionals to the F_0 contour (including non-voiced regions where F_0 is set to 0) and its first order derivative as also shown in Table 3. Segments in this case correspond to continuous voiced regions, i. e., where F_0 is > 0.

5.2. Reliability and performance bounds

We designed systems for automatic analysis with the goal of recognizing those instances that seem generally likable, as de-

Table 4: *Reliability analysis of rater pairs ($k = 1, \ldots, 16$): Cross-correlation with the mean rating of all raters (CC_k), and the EWE of other raters ('leave-one-out', CC_k^{LOO}).*

k	1	2	3	4	5	6	7	8
CC_k	.57	.44	.51	.59	.64	.53	.52	.55
CC_k^{LOO}	.48	.32	.43	.51	.56	.40	.42	.48
k	9	10	11	12	13	14	15	16
CC_k	.59	.41	.55	.40	.45	.52	.14	.55
CC_k^{LOO}	.51	.28	.47	.28	.34	.42	.01	.46

termined by a variety of raters. Thus, a consensus has to be derived from the individual ratings. As a first step, we calculated the agreement (reliability) of rater pair $k = 1, \ldots, K$ ($K = 16$) with respect to the arithmetic mean likability rating $\bar{l}_n$ for each instance n,

$$\bar{l}_n = \frac{1}{K} \sum_{k=1}^{K} l_{n,k} \tag{1}$$

where $l_{n,k} \in \{-3, -2, -1, 0, 1, 2, 3\}$ is the likability rating assigned by rater pair k to instance n. As a measure of reliability for each k, we computed the cross-correlation CC_k between $(l_{n,k})$ and $(\overline{l_n})$, $n = 1, \ldots, N$. Results are shown in Table 4. It can be seen that the the reliability in terms of CC_k considerably differed, ranging from .14 ($k = 15$) to .64 ($k = 5$). Hence, as a robust estimate of the desired rater-independent likability of each instance n, we used the evaluator weighted estimator (EWE) [6], denoted by l_n, instead of $\bar{l}_n$ in all further analyses:

$$l_n = \frac{1}{\sum_{k=1}^{K} CC_k} \sum_{k=1}^{K} CC_k l_{n,k}. \tag{2}$$

Next, to derive a rough estimate of the performance that can be expected from an automatic regression system trained on the EWE, we analyzed for each k the cross-correlation CC_k^{LOO} between $(l_{n,k})$ and the EWE $l_n^{\setminus k}$ of all rater pairs except k – in other words, a 'leave-one-out' reliability analysis. The results (Table 4) suggest that a regression on the exact EWE will be challenging – the maximum CC_k^{LOO} is .56 for $k = 5$, while the minimum is near zero ($CC_{15}^{LOO} = .01$), and the average CC_k^{LOO} is .40.

5.3. Binary classification

With possible applications in mind, it can be argued that the exact EWE is not needed – rather, a decision such as 'likable or not?' seems sufficient. Thus, we assigned each recording n a binary class label L (likable) whenever its EWE l_n was above the median (0.149) of all l_n, and NL (not likable) otherwise. Consequently, both, the L and NL classes, contain 400 recordings in total, thus enforcing balanced training as a side-effect.

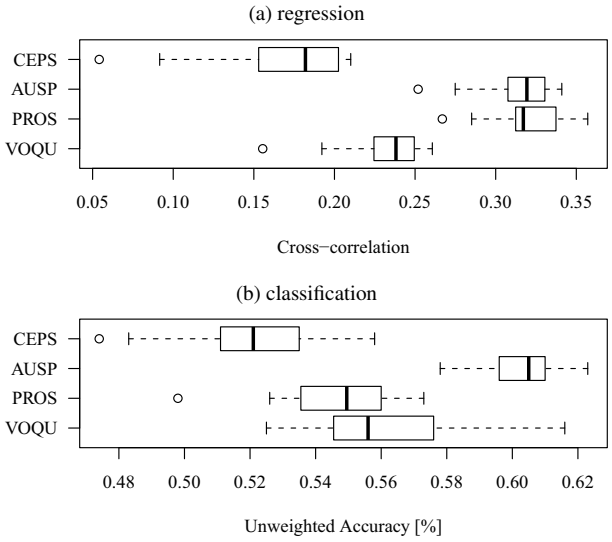
5.4. Performance evaluation

For automatic regression, we trained ensembles of REPTrees on random feature sub-spaces, using the open-source implementation of the Weka toolkit [9], as this meta-learning technique seems particularly suited to large, brute-forced acoustic feature sets. In all subsequent analyses, the set of 800 recordings was split into a training (394), development (258), and test (148) set – this corresponds to the subdivision in the INTERSPEECH 2010 Paralinguistic Challenge [10], enforcing stratification by age and gender, and compatibility of results.

Table 5: *Performance of automatic analysis: regression and classification by REPTree ensemble learning (2 000 trees) with random feature subspaces (2 % for regression / 10 % for classification).*

	Regression		**Classification**	
	CC	MLE	% UA	% WA
Train vs. Develop	.378	.618	65.4	65.5
Train vs. Test	.256	.637	67.6	67.6

Figure 1: *Performance of LLD feature groups in regression and classification of the development set. Ranges correspond to different classifier parameters for random sub-space learning with REPTrees (cf. Fig. 2).*



The feature sub-space size as well as the number of trees in the ensemble were tuned by a two-dimensional grid search, evaluating on the development set. Thereby trees were not pruned, but the maximum tree size was limited to 25, as in the Challenge baseline [10]. As shown in Figure 2a, best performance on the development set in terms of CC (.378), as shown in Table 5, is obtained for a sub-space size of 2 % and 2 000 trees. The mean linear error (MLE) in that case was .618 (on a scale from -3 to 3).

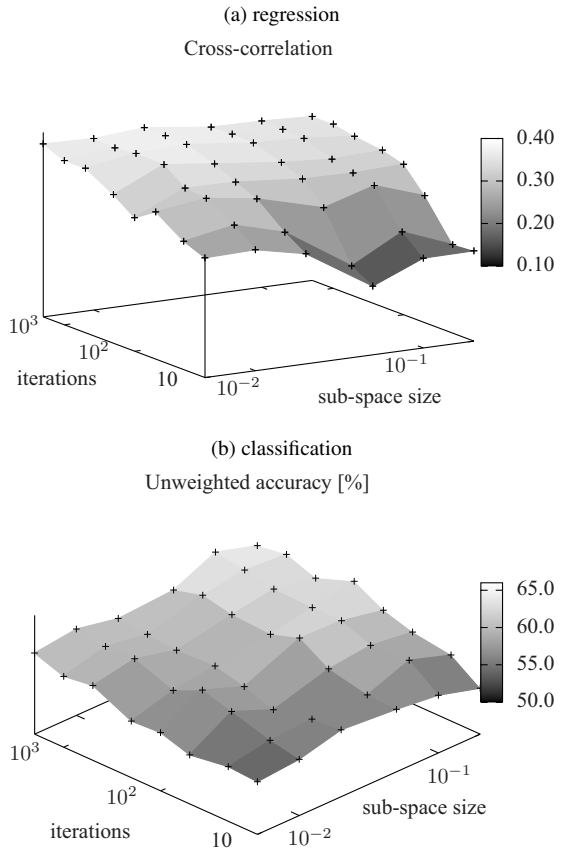
We performed an analogous grid search for binary classification (Figure 2b), optimizing on the unweighted accuracy (UA) on the development set. Best performance (65.4 % UA) is obtained for 2 000 trees, with a sub-space size of 10 %.

After that, we evaluated the tuned classifiers on the test set (Table 5). On the one hand, a very remarkable UA of 67.6 % is obtained, which is highly significantly above chance level ($p < .001$ according to a z-test, N=148). On the other hand, regression on the EWE seems to be even more challenging on the test set, resulting in a CC of only .256, and also higher MLE (.637).

5.5. Relevance of feature types

We conclude our investigation of automatic analysis by assessing which types of LLD contribute the most to regression, and classification, performance. To this end, we subdivided the IS11 feature set into four groups: cepstral (CEPS), auditory spectral (AUSP), prosodic (PROS), and voice quality (VOQU) features. Then, we evaluated regression and classification performance

Figure 2: *Optimization of REPTree parameters on development set: sub-space size* $\in \{5{\cdot}10^{-3}, 10^{-2}, 2{\cdot}10^{-2}, 5{\cdot}10^{-2}, 10^{-1}, 2{\cdot}10^{-1}\}$*, number of trees* $\in \{10, 20, 50, 10^2, 2{\cdot}10^2, 5{\cdot}10^2, 10^3\}$.



using REPTree ensembles on the development set, using only one of these LLD groups, and for varying feature sub-space size / number of trees. Results are shown in Figure 1 as box-and-whisker plots – boxes range from the first to the third quartile, and all values that exceed that range by more than 1.5 times the width of the box are considered outliers, depicted by circles. Interestingly, it can be seen that cepstral features do not enable robust regression or classification - in fact, the mean UA for classification is near chance level (52 %). In contrast, auditory spectral features seem to contribute the most to reliable automatic likability analysis for regression as well as classification, followed by prosodic and voice quality features.

5.6. Rater-Dependent classification

The observed performances in rater-independent likability recognition are all the more remarkable as the inter-rater agreement on likability is fairly low – it can be argued that the annotation is highly subjective. This suggests that a rater-dependent classification might be more robust than relying on an estimate of ground truth likability. Such classifiers are tied to a variety of possible applications, including suggestion of likable persons in a social network, based on voice recordings. Thus, we performed additional experiments for each rater pair on the development set, dividing the instances into NL and L instances, using the median rating by one rater pair as threshold. Note that a true single-rater-based classification would actually halve the training and test data per experiment, due to our rating procedure. Using the classifier configuration that performed best in the rater-independent case, however, results varied strongly, and were inferior to rater-independent classification, and not even significantly above chance in the majority of cases (min: 48.9 %, max: 63.0 %, mean: 53.2 % UA). Thus, it seems that the correlation of acoustic features significantly increases when using a likability estimate from multiple raters.

6. Conclusions and outlook

We have demonstrated that although inter-rater agreement on likability is not very high, automatic analysis based on the evaluator weighted estimator is robust, paving the way for a multitude of interesting applications in human-machine and human-human communication, such as in voice portals or social networks. In binary classification, 67.6 % unweighted accuracy have been obtained on a subset of the "Agender" database, consisting of real-life telephony speech without pre-selection of prototypical instances. As such, our results are line with typical results of paralinguistic audio analysis 'in the wild' (e. g., [10]).

Future work might include ratings of multiple samples per speaker: This way, one could assess the consistency of individual raters, and possibly derive an even more solid ground truth estimate for automatic learning. In that context, one might also consider generation of additional training material by including a variety of utterances and assigning them the EWE likability of a subset assessed by human listeners.

7. References

[1] L. Bruckert, J. Lienard, A. Lacroix, M. Kreutzer, and G. Leboucher, "Women use voice parameter to assess mens characteristics," in *Proc. Biological Sciences*, vol. 237, no. 1582, 2006, p. 8389.

[2] J. Trouvain, S. Schmidt, M. Schröder, M. Schmitz, and W. Barry, "Modelling personality features by changing prosody in synthetic speech," in *Proceedings of the Conference on Speech Prosody Dresden*, 2006.

[3] B. Weiss and F. Burkhardt, "Voice attributes affecting likability perception," in *Proc. Interspeech*, Makuhari, Japan, 2010.

[4] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," *Proc. LREC (Language Resources Evaluation Conference), Valetta*, 2010.

[5] F. Burkhardt, R. Huber, and B. Anton, "Application of speaker classification in human machine dialog systems," in *Speaker Classification I: Fundamentals, Features, and Methods*, C. Müller, Ed. Springer, 2007, pp. 174–179.

[6] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. of ASRU*. IEEE, 2005, pp. 381–385.

[7] ITU-T Rec. P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," International Telecommunication Union, Geneva, 2004.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*. Florence, Italy: ACM, October 2010, pp. 1459–1462.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[10] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proc. Interspeech*, Makuhari, Japan, 2010.