Likability of human voices: a feature analysis and a neural network regression approach to automatic likability estimation

Florian Eyben, Felix Weninger, Erik Marchi, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Eyben, Florian, Felix Weninger, Erik Marchi, and Björn Schuller. 2013. "Likability of human voices: a feature analysis and a neural network regression approach to automatic likability estimation." In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), 3-5 July 2013, Paris, France*. Piscataway, NJ: IEEE. https://doi.org/10.1109/wiamis.2013.6616159.

LIKABILITY OF HUMAN VOICES: A FEATURE ANALYSIS AND A NEURAL NETWORK REGRESSION APPROACH TO AUTOMATIC LIKABILITY ESTIMATION

Florian Eyben, Felix Weninger, Erik Marchi, Björn Schuller

Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, GERMANY

ABSTRACT

Recently, the automatic analysis of likability of a voice has become popular. This work follows up on our original work in this field and provides an in-depth discussion of the matter and an analysis of the acoustic parameters. We investigate the automatic analysis of voice likability in a continuous label space with neural networks as regressors and discuss the relevance of acoustic features. We provide results on the Speaker Likability Database for comparison with previous work and a subset of the TIMIT database for validation.

1. INTRODUCTION

People have preferences which types of voices they prefer and which they dislike (cf. [1]). For many technical tasks it is desirable to select voices that appear pleasant or seem likable to a large number of people. Examples include the voice of a dialogue system, a chatbot, announcements in public places, and TV or radio advertisements. The automatic assessment of voice likability would enable automatic tuning of voices in such applications towards more pleasant voices (cf. [2]) while requiring very little human feedback and supervision. Moreover, it could enable robots and chatbots to build a profile of their dialogue partners and develop a certain stance (positive or negative) towards them - in the end making the bot more human like. While this is not desirable for all applications (e. g., service robots should always be friendly and not like or dislike a person), it brings an added value to other areas of application like toy robots, entertaining chatbots, and automated characters in computer games.

In previous research [3], it was investigated how "pleasant" a speaker's voice appears based on EMO-DB. This is a database of 10 actors, which simulate emotional categories. One of the biggest drawbacks of that study is the very low number of speakers. A larger database was introduced in [4] and later made available to the public as the Speaker Likability Database (SLD) in the INTERSPEECH 2012 Speaker Trait Challenge [5]. The first attempts for systematic automatic classification of likability were encouraged in the scope of the Speaker Trait Challenge. Likability thereby was assessed as a binary classification task (likable and not likable). The results of the challenge show, that automatic labelling performance based on a standard set of spectral and prosodic acoustic descriptors is above chance level, however, only barely significantly. This indicates the highly challenging nature of this task.

In [6] the authors showed that clear differences exist between the 30 most likable and the 30 most unlikable speakers in the SLD (Agender), which suggests that a higher performance is possible. The two class approach taken in the Challenge, discards information about how likable a voice is. In this paper we thus investigate if a continuous modelling of the likability ratings with neural networks can bring an improvement or if new acoustic features or even higher level features are necessary.

This paper is structured as follows: In Section 2 we briefly describe the Agender database and the subset of the TIMIT speech database which we have analysed for this paper. In Section 3 the the acoustic feature set used is described and the results of a correlation based feature value analysis is discussed.

2. DATABASES

Previous work was based on the "Speaker Likability Database" (SLD) [7, 4] and the EMO-DB database [3]. In order to validate our approach on a new data-set, we chose to annotate a subset of the TIMIT database, which contains speech of over 600 speakers. More details on this set are found in Section 2.2, while SLD is described in Section 2.1.

2.1. Speaker Likability Database

As first evaluation database, the SLD is used [4]. SLD is a subset of the German Agender database [7], which was originally recorded to study automatic age and gender recognition from telephone speech. The speech is recorded over fixed and mobile telephone lines at a sample rate of 8 kHz. The database contains 18 utterance types taken from a set listed in detail in [7]. An age and gender balanced set of 800 speakers is selected. For each speaker, we used the longest sentence consisting of a command embedded in a free sentence, in order to keep the effort for judging the data by many listeners as low as possible. Likability ratings of the data were established by presenting the stimuli to 30 participants (17 male, 13 female, aged 20–42, mean=28.6, standard deviation=5.4). To control for effects of gender and age group on the likability ratings, the stimuli were presented in six blocks with a single gender / age group. To mitigate effects of fatigue or boredom, each of the participants rated only three out of the six blocks in randomised order with a short break between each block. The order of stimuli within each block was randomised for each participant as well. The participants were instructed to rate the stimuli according to their likability, without taking into account sentence content or transmission quality. The rating was done on a seven point Likert scale. All participants were paid for their service. A preliminary analysis of the data shows no significant impact of participants' age or gender on the ratings, whereas the samples rated are significantly different (mixed effects model, $p < .0001$). 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). As intended by the instruction, there is no significant correlation between the averaged ratings and quality estimates (Spearman's $\varrho = .04, p = .27$). To establish a consensus from the individual likability ratings (16 per instance), the evaluator weighted estimator (EWE) [8] was used. As a first step, we calculated the agreement (reliability) of rater

$k = 1, \dots, K$ ($K = 16$) with respect to the arithmetic mean likability rating $\bar{l}_n$ for each instance n,

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

where $l_{n,k} \in \{-3, -2, -1, 0, 1, 2, 3\}$ is the likability rating assigned by rater k to instance n. As a measure of reliability for each k, we computed the cross-correlation CC_k between $(l_{n,k})$ and $(\overline{l_n})$, $n = 1, \dots, N$. Results are shown in Table 2. It can be seen that the the reliability in terms of CC_k considerably differed, ranging from .079 ($k = 29$) to .668 ($k = 4$). Hence, as a robust estimate of the desired rater-independent likability of each instance n, we also considered the evaluator weighted estimator (EWE) [8], denoted by l_n, besides the mean rating $\bar{l}_n$ in further analyses:

$$l_n = \frac{1}{\sum_{k=1}^{K} \mathrm{CC}_k} \sum_{k=1}^{K} \mathrm{CC}_k l_{n,k}. \qquad (2)$$

The SLD was used as the official evaluation corpus in the INTERSPEECH 2012 Speaker Trait Challenge in the likability sub-challenge [5].

2.2. TIMIT

The TIMIT corpus [9] is widely used in the Speech Community. It contains 630 speakers (438 female, and 192 male) from 8 US regions with distinctly different dialect. The database contains two so-called *sa* sentences for each speaker, where the linguistic content is fixed. These sentences constitute the subset used herein (1260 utterances in total). Four annotators (1 female, 3 males) annotated the likability/pleasantness of the voices for each utterance on a scale from 1 to 4 using integer numbers. The utterances were presented to the raters in a randomized order. As there are two sentences for each speaker, we can compute the consistency of the ratings on these (making the assertion that ideally both sentences should receive the same rating as they are from the same speaker) in terms of MLE (Table 1). The rater consistency is significantly better than the expected consistency for random rating equally distributed over the labels 1, 2, 3, and 4 (expected MLE 1.25).

Rater	MLE
1	0.87
2	0.39
3	0.87
4	0.48

Table 1. Rater consistency on TIMIT likability set. MLE between ratings of first (sa1) and second (sa2) sentence of the same speaker. Expected MLE value for two random ratings is 1.25.

3. ACOUSTIC FEATURES

We based our analysis on the ComParE acoustic feature set [10], which is an improved version of the INTERSPEECH 2012 Speaker Trait Challenge baseline feature set [5]. The features are a brute-force set of 6 373 acoustic features where numerous functionals (such as mean, standard deviation, regression coefficients) are applied to a large set of commonly used low-level descriptors (LLDs) and their delta coefficients. The functionals are applied over one utterance resulting in one 6 373 dimensional feature vector for every utterance, regardless of its length. The features are extracted with our open-source audio and paralinguistics analysis toolkit openSMILE [11].

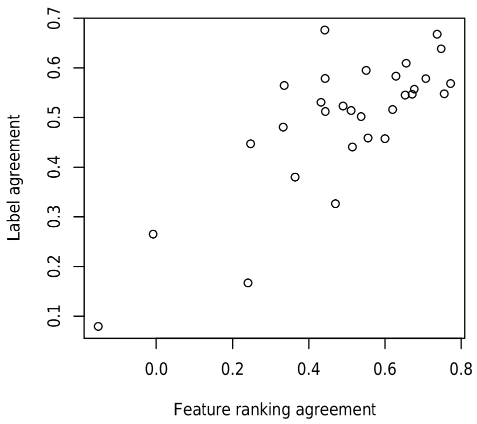

Fig. 1. Correlation between label agreement (CC of likability ratings with mean rating) and feature ranking agreement (rank correlation of per-annotator and average acoustic feature relevance) for each annotator: Spearman's ϱ = .61 ($p < .001$). Evaluation on entire SLD.

In Table 2, we show the feature ranking agreements and label agreements for the 30 annotators. For each annotator, the feature ranking agreement is obtained as follows: First, the acoustic features are ranked by their CC with the annotators' normalised rating (rater-specific ranking). Then, the acoustic features are ranked by their CC with the mean rating (average ranking). Finally, for each annotator the correlation between rater-specific ranking and average ranking constitutes the annotator's feature ranking agreement.

The label agreement of an annotator is simply the correlation coefficient of the annotator's rating versus the mean rating, as used in the EWE calculation (cf. above). Feature ranking agreements and label agreements are depicted in Figure 1 as a scatter-plot. It can be seen that they are significantly ($p < .001$) correlated ($\varrho = .61$), indicating that generally those raters who agree with the majority also agree in the choice of their acoustic features.

Next, we consider the median of the absolute CCs of the acoustic features per rater as a heuristic measurement of how strongly they weigh acoustic cues in their decision ('acoustic feature importance'). The relation of this measurement with the label agreement is depicted in Figure 2. It can be seen that there is a significant correlation (Spearman's $\varrho = .51$, $p < .005$). This provides evidence that generally those raters who agree more with the majority generally weigh the acoustic cues from the ComParE feature set higher than those who agree less.

We also computed Pearson Correlation Coefficients (CC) for each of the 6 373 acoustic features with the mean likability label and the EWE likability label on SLD and the mean likability label on TIMIT. In order to summarise and interpret the results, we combined the CCs for each low-level descriptor over all functionals by choosing the maximum absolute CC value per LLD (*max*) and the average absolute CC value (*avg*) per LLD. We did the same averaging for the functionals to find relevant functionals. Looking at the top 10 LLDs

Rater	1	2	3	4	5	6	7	8	9	10
CC (Feature ranking)	.364	.432	.538	.737	.707	.629	.672	.656	.653	.620
CC (Label)	.380	.531	.502	.668	.578	.583	.547	.609	.545	.516
Rater	11	12	13	14	15	16	17	18	19	20
CC (Feature ranking)	.600	.241	.444	-.008	.772	.556	.490	.442	.515	.247
CC (Label)	.457	.167	.512	.265	.568	.459	.523	.676	.441	.447
Rater	21	22	23	24	25	26	27	28	29	30
CC (Feature ranking)	.677	.748	.333	.511	.336	.470	.756	.551	-.152	.443
CC (Label)	.557	.638	.481	.514	.564	.326	.548	.595	.079	.578

Table 2. Label agreement and feature ranking agreement (cf. above) for individual raters. Evaluation on entire SLD.

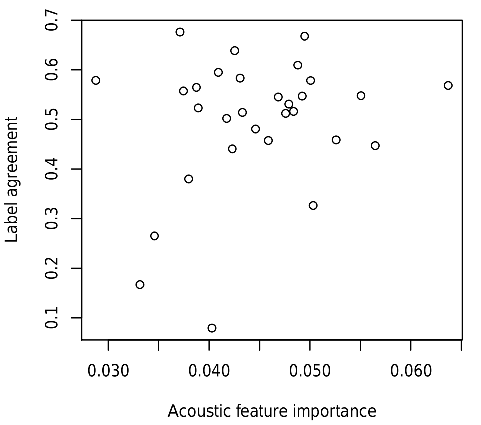

Fig. 2. Correlation between label agreement (CC of likability ratings with mean rating) and acoustic feature importance (median CC of acoustic features with single rating) for each annotator: Spearman's $\varrho = .51$ ($p < .005$). Evaluation on entire SLD.

and functionals, we selected the best LLD and functionals which occur highly ranked in both max CC and avg CC groups. For SLD, the features ranked among the top 10 were nearly the same for mean and EWE as targets. Therefore we don't discriminate between these two cases but instead only report features that were highly ranked both for mean and EWE likability. The result can be seen in Table 3.

Interestingly the relevant LLD for both corpora (SLD and TIMIT) are very different. For SLD, the overall signal energy features (RMS energy and loudness) are clearly dominating the top 10, followed by specific energies in various (auditory) spectral bands. In conjunction with the functionals (means, inter quartile ranges, means of peaks, and standard deviation) it becomes clear that for SLD the mean loudness and the distribution of loudness and energies in certain speech formant related bands (1–4 kHz) is relevant. As most of the individual mean energy features are negatively correlated with the target, it seems that voices which are too loud or have too much energy variation while speaking are considered as unpleasant. Besides the energy related LLD, spectral flux (the change in spectral distribution from one frame to another) and the spectral slope (low-pass, white, or high-pass spectrum) are also relevant. These findings are in line with [4], where critical band energy related features (auditory spectral bands) were also found to be the best performing ones for a binary classification task of likability on SLD.

LLDs	**Functionals**
SLD	
Loudness, RMS energy, Energy 1 kHz–4 kHz, Energy 250–650 Hz, Spectral flux and slope, Energy 400–500 Hz, Energy in critical bands	Means, Inter quartile ranges, Linear regression offset, Standard deviation, Mean of peaks,
TIMIT	
Probability of voicing, Harmonicity, F_0, 75% and 90% spectral roll-off freq., Zero-crossing rate, Psychoacoustic sharpness (= weighted spectral centroid)	Means, Linear regression and quadratic regression offsets, Range (max - min), Maximum value

Table 3. Best ranked acoustic LLD and functionals by CC with the mean likability on TIMIT and both mean and EWE likability on SLD (note: both mean and EWE have near the same rankings).

For TIMIT mostly LLD which describe the spectral quality of the signal are relevant. Energy related features never appear among the top 10 features. Instead the pitch, voicing probability, and harmonicity, as well as spectral distributions (centroid/sharpness) and roll-off points are relevant. In conjunction with the functionals this suggests that pleasantness is correlated with the pitch and perceived sharpness of the voice.

4. AUTOMATIC REGRESSION

According to the analysis of acoustic features a weak correlation has been found between several features and the likability target. We will now describe experiments on the feasibility of automatic regression of the likability scores. The goal of these experiments is to find out whether a regressor which takes all features jointly into account can improve the correlation with the likability target over the single best acoustic features.

In this study we use neural networks configured as a multi-layer perceptrons with one hidden layer containing neurons with sigmoid transfer functions. We investigated hidden layer sizes of 5, 10, 20, 40, 80, 120, and 160 neurons. The networks' input layers are of size 6 373, which corresponds to the number of features. The networks have one linear output neuron corresponding to the EWE or mean likability target. Networks were trained with standard gradient de-

net. size:	20	40	80	120	160
IS12 Challenge test partition					
CC	.19	.25	.30	.18	0.21
UAR [%]	59.5	57.7	64.6	57.6	58.3
p-val.	0.15	0.14	0.02	0.09	0.14
AUC	.585	.634	.654	.622	.604
IS12 Challenge development partition					
CC	0.32	0.31	0.36	0.29	0.32
UAR [%]	59.8	54.2	59.1	63.9	59.8
p-val.	0.09	0.30	0.16	0.03	0.10
AUC	.638	.605	.643	.645	.621

Table 4. Regression and classification results.

scent and output error back-propagation. The initial weights in the network were initialized with random weights uniformly distributed in the range $[-0.1; 0.1]$ In order to smooth out variations in the results due to different initialisations and local minima in the search space, we repeat each network training 5 times with different initial random weights (generated by using different random seeds in the random number generator for the weights).

The networks were trained on the INTERSPEECH 2012 Challenge training partition of the SLD. As training target the EWE likability score is used. The development set was used for early stopping of the network training to avoid over-fitting to the training data. If no improvement of the quadratic error on the development set was observed for 10 training iterations the training is aborted on the best network on the development is evaluated on the test set. The resulting Pearson correlation coefficient (CC) for predicted likability output with the ground truth EWE ratings is shown in Table 4.

A binary classification result was produced from the continuous outputs and the ground truth by applying a decision threshold of 0 to both. Unweighted Average Recall (UAR) and the corresponding p-values of a paired T-test where the predictions are compared to a set of random predictions are given in Table 4.

The results in Table 4 show that the best results are achieved with networks between 80 and 120 hidden units. The UAR of the best network (80 hidden units for the test partition and 120 hidden units for the development partition) is significantly over random guess at a level of $\alpha = 0.05$, while all other UAR results are not. We see a similar trend as in the Challenge baseline paper [10] that a better CC can be achieved on the development partition than on the test partition, while UAR is lower.

5. CONCLUSION AND OUTLOOK

We have presented two publicly available databases for research of speaker likability. An in-depth analysis of acoustic features was given, showing that between prosodic and spectral acoustic features and likability a rather low correlation exists. The best correlated features suggest that the energy level and energy modulations affect the likability as well as the spectral distribution of the speaker (wrt. high/low frequencies, white or harmonic spectrum). The automatic classification results reflect the challenging nature of the problem and are in line with the baseline results of the INTERSPEECH 2012 Speaker Trait Challenge. The best result is 64.6% Unweighted Average Recall on the test partition.

The work presented here undermines the findings of previous work ([6]) that spectral features are less important than higher level features such as dialect and speaking style. For the Speaker Likability Database [6] suggests that the amount of disfluencies and thus the speech rhythm has an influence on likability. This, in turn, concurs with our finding that standard deviation and the quartiles of loudness and signal energy are best correlated with likability on the Speaker Likability Database.

Future work will focus on higher level acoustic and linguistic analysis, such as automatic retrieval of disfluencies, advanced descriptors of speech rhythm, and acoustic analysis with knowledge of the phonetic context.

6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 289021 (ASC-Inclusion).

7. REFERENCES

[1] K. Scherer, H. London, and J. Wolf, "The voice of condence: Paralinguistic cues and audience evaluation," *Journal of Research in Personality*, vol. 7, pp. 31–44, 1973.

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

[3] Benjamin Weiss and Felix Burkhardt, "Voice attributes affecting likability perception," in *Proc. of INTERSPEECH*. 2010, ISCA.

[4] Felix Burkhardt, Björn Schuller, Benjamin Weiss, and Felix Weninger, ""Would You Buy A Car From Me?" – on the Likability of Telephone Voices," in *Proc. of INTERSPEECH*. 2011, ISCA.

[5] B. Schuller, S. Steidl, A. Batliner, E. Nth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 speaker trait challenge," in *Proc. of INTERSPEECH, Portland, Oregon, USA*. 2012, ISCA.

[6] Benjamin Weiss and Felix Burkhardt, "Is 'not bad' good enough? aspects of unknown voices' likability," in *Proc. of INTERSPEECH, Portland, Oregon, USA*. 2012, ISCA.

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

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

[9] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "TIMIT acoustic-phonetic continuous speech corpus CD-ROM," 1990.

[10] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. of INTERSPEECH 2013, Lyon, France*. 2013, ISCA.

[11] 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*. 2010, pp. 1459–1462, ACM.