ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Is 'not bad' good enough? Aspects of unknown voices' likability

Benjamin Weiss[1], *Felix Burkhardt*[2]

[1]Quality & Usability Lab, Telekom Innovation Laboratories, TU, Berlin, Germany
[2]Telekom Innovation Laboratories, Berlin, Germany
benjamin.weiss@tu-berlin.de, felix.burkhardt@telekom.de

Abstract

From the DTAG likability database the 30 most and 30 least likable ones have been selected for further phonetic analysis and expert questionnaire. In contrast to dislikable speakers, likable ones exhibit almost no perceivable accent, command style or disfluencies and were rated very high on all six questionnaire scales. Dislikable speakers are only moderately rated. However, dislikable speakers display also lower pronunciation precision, different amounts of jitter, lower articulation rate and higher pitch, indicating that just the absence of negatively perceived characterizations is not enough to become a likable speaker.
Index Terms: likability, speaker traits, speaking style

1. Introduction

Likability of unknown voices can be evaluated unproblematically and with high reliability, if the social situation is controlled [1]. Even automatic classification may be attempted [2, 3]. There are a lot of potential factors already identified, which might constitute likability ratings of voices and speaking style (cf. [4]): For example, appropriateness of speaking style [5], proficiency of speaking [6], and several auditory markers used for attribution processes like attractiveness [7], social background [8], or personality [9].

But which of those potential factors constitute specific likability ratings? For clear speech, it could be shown, that from several descriptive factors especially *warm/relaxed* correlates significantly with likability, and with acoustics parameters of less pressed, more breathy voice quality and lower spectral center of gravity [10, 1].

The aim of this paper is to present results how phonetic aspects affect likability for telephone quality speech. It is structured as follows: Section 2 describes the database. Section 3 presents additional analysis of the likability database, followed by Section 4 which explains the expert questionnaire and Section 5 talking about the parametric analysis. The paper finishes with a discussion in Section 6 and conclusion in Section 7.

2. Data description

The DTAG likability database is a sub part of the Agender database [11]. The spoken content is based on 18 utterances spoken over telephone taken from a set of utterances listed in detail in [11]. The topics of these were prompted text like typical voice portal commands as well as 'eliciting' questions like *"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 (7-14, 15-24, 25-54, 55-80 years) groups acquired from all German Federal States without perfect balance of German dialects.

All age groups have equal gender distribution.

For the likability database 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. 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, 3.05 sec.). This is about the length used in other rating experiments with highly reliable ratings (e.g. [1]).

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.

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 situation of the recordings was mentioned. The rating was done on a single seven point scale ("likable – not likable"). For playing back Sennheiser HD 485 headphones were used. No participant reported hearing loss. All participants were paid for their service.

This database was already used in [3] and is also used as part of the Speaker Trait Challenge 2012 [12] presented in this conference.

3. Preliminary Analysis

According to [3], there is no significant impact of participants' age or gender on the likability ratings, whereas the samples rated are significantly different. Also, the transmission quality does not correlate with likability.

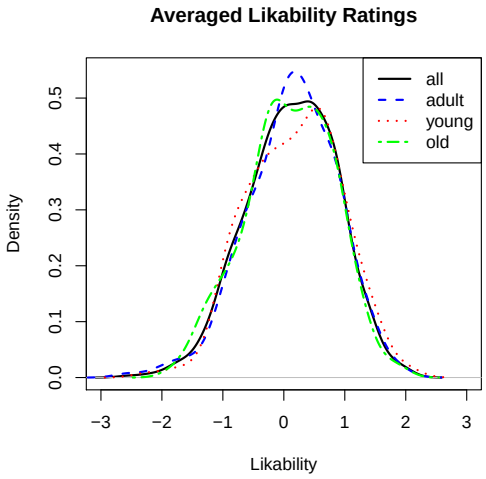

Figure 1: *Likability density distribution.*

In addition to this, we observe no significant effect of the gender of the speakers ($F(1, 12765) = 0.57$, $p = 0.45$), nor of the interaction between gender of the raters and speakers ($F(2, 12765) = 1.06$, $p = 0.35$). However, age groups are rated differently (age $F(2, 12765) = 3.49$, $p < 0.05$). Speakers from the younger group are more positively assessed than those from the older group ($alpha = .05$), although the effect is very small (effect size $part.\eta^2 = 0.039$). As the participants were only controlled for gender, not for age group, we did not test for interaction between age of speaker and gender. Also, the age result reported may be specific to the participants group, as the majority of them would belong to the age group of adults.

The participants of the listening test provide reliable ratings of the stimuli in both groups of 16 votes (Intra Class Correlation: $ICC(3, 2) = 0.77, 0.75$, respectively).

All values are normalized by the evaluator weighted estimator (EWE) [13], which is a weighted mean likability rating, with cross-correlations as weights.

The distribution of voices concerning likability is only roughly normal (significantly different by means of an Anderson-Darling test $A = 0.93, p < .05$) and quite symmetric (cf. Figure 1). This might explain, why the automatic classification in [3] performs only weak for the binary case [3]. In order to obtain more information about phonetics affecting the judgments, the five best and five worst rated samples for each of the six age- and gender groups, resulting in 60 stimuli, were analyzed.

4. Subjective phonetic description

In a pre-test with three experts (the two authors and a third), a questionnaire to assess phonetic description [1] was reduced to three items, namely *"sonorous–flat", "relaxed–tensed"* and *"articulately–inarticulately"*. This was done using the LisTEn tool by RWTH Aachen [14]. As the original database did not include hesitations, false starts, repairs or even regional varieties, the three items *"fluent–rugged", "accent–no accent"* and *"pronunciation unremarkable–pathologic"*, have been added. On these six items, each of the 60 stimuli have been rated by ten experienced subjects (5m, 5f, aged 22 to 54, M=30.3) with sufficient reliability ($ICC(3, 2) = 0.85, 0.77, 0.91, 0.89, 0.89, 0.77$).

On each of the six scales, the likable and dislikable speakers differ from each other significantly (repeated t-test, all p-values $< .001$), see Figure 2.

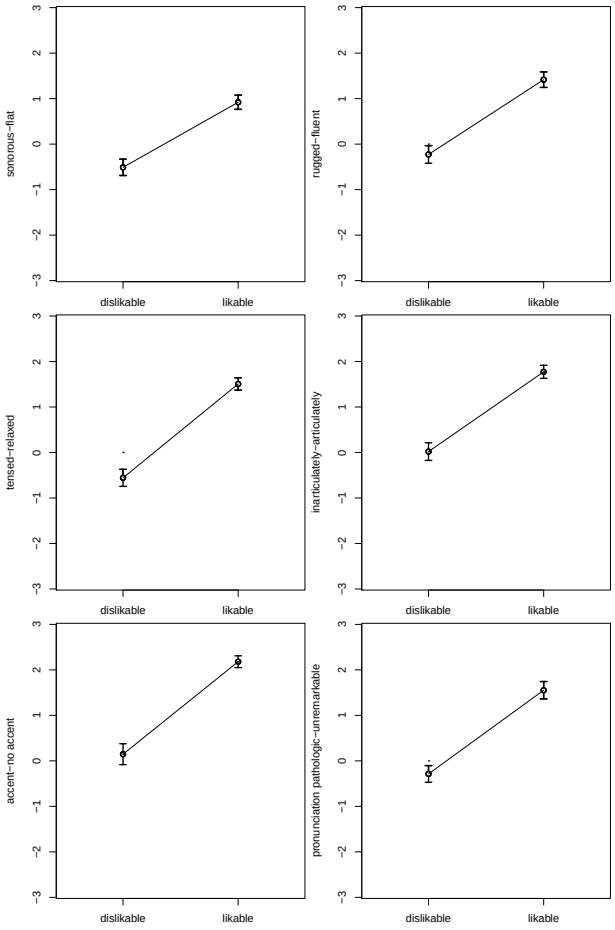

Figure 2: *Means and confidence intervals for the 30 most and least likable speakers on six rating scales.*

Noteworthy is the absolute distribution of ratings, as the most likable speakers are rated very positive, whereas the worst speakers not necessarily exhibit negative ratings.

In order to parameterize the descriptive results, all sentences have been phonetically transcribed and commented by the authors (canonic transcription and phonetic deviation from this canonic transcription). Comments appearing more than once are considered as binary annotation of each stimulus (cf. Table 1). Accent comprises both, regional and foreign accent. The impression of a command style compared to a normal command, request or even question was induced during transcription by noticeable vocal effort, flat intonation, pauses and/or hyper-articulation (fricative or even trill variant of typically vocalized /r/ or "-en" realization in unstressed syllables) emphasizing single words. This is indicated by *italics*. Interestingly, a rising pitch indicating a question was solely present in likable sentences. Table 1 displays the number of occurrences of annotations. However, not every dislikable stimulus was actually characterized by such labels. When excluding "glottalization" and "rising pitch", there are 5 out of 30 dislikable stimuli without any label in comparison to 26 out of 30 for the likable ones.

Table 1: *Number of occurrences of various labels for likable and dislikable stimuli. Aspects of command style in italics.*

Label	"likable"	"dislikable"
accent	2	12
disfluencies	0	4
vocal effort	*0*	*9*
pauses	*1*	*5*
flat intonation	*0*	*3*
hyper-articulation	*0*	*4*
rising pitch	3	0
glottalization	1	2

5. Parametric phonetic description

Acoustic parameter extraction was done automatically with Praat (with different settings for each group of speakers). Global parameters measured are:

- articulation rate (syllables per second based on manual syllable segmentation)
- measures of pitch (mean and variation)
- global spectral distribution (tilt and center of gravity based on the averaged spectra)
- measures of intensity (variation)
- harmonic-to-noise ratio
- jitter of voices parts (mean absolute difference betw. consecutive periods / average period) [15]
- global pronunciation accuracy (divergence from canonic pronunciation: elision, epenthesis, and consonantal difference)

Pronunciation accuracy is measured as number of segments in the canonic transcription minus consonantal segments deleted, inserted or clearly pronounced non-canonically as well as number of non-central vocalic segments pronounced non-canonically and number of schwa realized unnecessarily in percent.

The results for the ANOVAS are summarized in Table 2 indicating the direction of the effect for likable speakers. There was no interaction effect observed for likability with gender:

Table 2: *Results for the* ANOVAS. *Parameter direction representing differences of likable vs. not-likable speakers, sig. results bold.*

Parameter	F(DF)	p
higher rate	(1,56)=30.62	**<.000001**
lower f0 (average)	(1,56)= 6.82	**=.0115**
f0 (SD)	(1,56)= 0.00	=.9519
spectral tilt (voiced)	(1,56)= 0.22	=.642
center of gravity (average)	(1,56)= 2.40	=.127
intensity (SD)	(1,56)= 0.23	=.635
harmonic-to-noise ratio	(1,56)= 2.04	=.158
higher jitter	(1,56)= 9.05	**=.00393**
higher pronunciation accuracy	(1,56)= 7.98	**=.00649**

The significant parameters are depicted in Figure 3, separated for gender in cases of a significant gender main effect (and thus the confidence intervals displayed will be too big).

The four significant effects are for the acoustic parameters *f0* and *jitter* as well as *rate* and *pronunciation accuracy* which are extracted from the transcription labels. In contrast to [1] there is no significant result for spectral measures tilt or center of gravity.

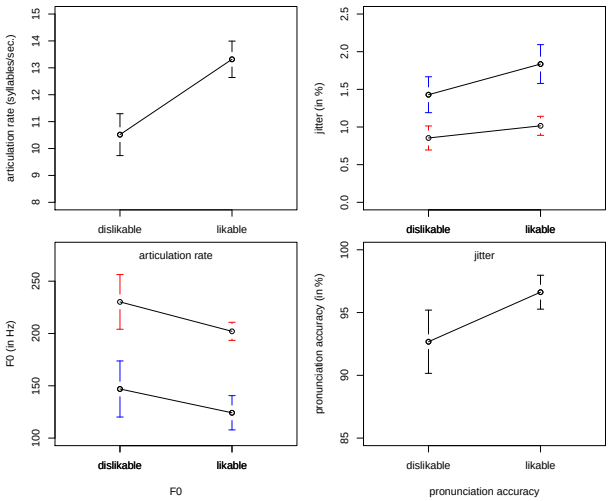

Figure 3: *Means and 95% confidence intervals in significant global acoustic parameters (if gender is significant, male=blue, female=red).*

6. Discussion

On all six scales of the short questionnaire, the 30 most likable adult speakers of the Agender database are rated significantly better than the 30 least likable ones. In accordance to the annotations of each sample, likable speakers are characterized by the absence of a noticeable regional accent, disfluencies (repairs, false starts, hesitations), and command style (interpret on basis of noticeable vocal effort, flat intonation, pauses and/or hyper-articulation). While dislikable speakers do exhibit such characteristics with the stimuli, they are not necessarily rated negatively on the ratings scales, but just significantly more negative than the likable speakers. These are matching results for aspects of *fluency, pronunciation, relaxation* and *accent*.

Likability ratings as subjective validations in concrete situations have to reflect situation appropriateness. Accordingly, the three instances of question intonation in the likable group of speakers are most likely special for the human-computer interaction of the corpus, in which such questions seem to be appropriate. Interestingly, the command style quite often used in interaction with spoken dialog systems is considered unlikable, which seems plausible for humans. But this may be evaluated differently if providing the raters more context of the dialog than single stimuli. In the same light the accent of speakers has to be considered, as an regional background is not considered bad per se, but certainly dependent on strength of the accent and identity of both, speaker and listener.

Moreover, parameterizations of *articulation* and other aspects known from literature display significant differences for likable/dislikable speakers as well: A higher articulation rate is perceived as more competent [16] and a lower pitch is typically rated more positively in laboratory settings [6, 1]: Therefore both significant effects do not appear unexpected. However, the impression of a flat command-like intonation could not be confirmed with acoustic data. Also, spectral parameters related to *timbre* did not show any significant results despite the promising effects for the questionnaire items *sonorous* and *relaxed*. This might be caused by the quite crude approach of measuring over

whole stimuli with varying linguistic content and varying transmission degradations. A closer look on individual phones might be more suitable here.

The significant result found for jitter is hard to interpret, as in lack of balanced material all voiced parts were used for the measurement. This means that jitter values include pitch variation and noise. Consequently, the values are above the range of 0.1% to 1.0% typically considered healthy and thus can only motivate to search for proper material in order to analyze jitter affecting likability.

At last, the measure of pronunciation accuracy based on the transcription shows a significant relation between high pronunciation accuracy and likable voices. Such results are already known for attribution of e.g. intelligence [16] but not for likability. Together with the significant effect for the questionnaire scale *"articulately–inarticulately"*, the known aspect of *pronunciation* could be parameterized here to some degree.

As the significant effects observed here seem to be plausible and therefor valid for other non-laboratory situations, as they are consistent with other attributes based on vocal information. For example, the appearance of command style is caused by the recording situation and might have been considered appropriate, but was not. Only for prediction of individual likability ratings, the relevant parameters found here will have different power: E.g., pronunciation accuracy will probably be a valid parameter for many situations, whereas the effect of regional accent definitely depends on the background of the raters. Therefor, it was essential to separate the groups of speakers to exclude age and gender preferences of the raters.

The original likability ratings are not separated for age groups. The general systematics of this analysis should not be concerned, however, as age is not known to affect likeability ratings in a complex way, just that younger listeners seem to be more critical than older ones [17].

7. Conclusion

From unknown speakers, recorded with telephone quality, the 30 most and 30 least likable ones have been selected for further analysis. In contrast to dislikable speakers the likable ones exhibit almost no perceivable accent, command style or disfluencies and rated very high on the questionnaire scales. Dislikable speakers do exhibit such characteristics, but are still moderately rated.

However, the absence of such markers as a bad speaker, inappropriate style or regional stereotype, seems not sufficient to be characterized as member of the likable group. In addition, unlikable speakers show higher pitch, different amounts of jitter, lower articulation rate and lower pronunciation precision indicating that just the absence of negatively perceived characterizations is not enough to become a likable speaker, but other phonetic aspects measured here have also to appear. In summary, five of the six significant aspects assessed with the questionnaire could be confirmed by subjective labels or even parametric data. Unfortunately, effects for parameters of *timbre* are not found to further strengthen this hypothesis. This might be caused by the overall restricted and varying quality of the signals, but also by a lack of well established parameterizations of aspects of *timbre*.

It is still open, how valid these results are for non-laboratory situations. Certainly, the importance of the hearers' traits will increase in more natural situations, especially for cross-gender ratings. Building on these results, the prospective parameterizations have to be validated experimentally in order to explain and predict likability of unknown voices. For example, a possible interrelation between speaker proficiency (e.g. pronunciation accuracy or disfluencies) and speaker's voice (e.g. pitch, rate, timbre) should be studies with synthesis experiments.

8. Acknowledgements

We thank Magnus Schäfer for adjusting the LisTEn tool [14]. This work was supported by the DFG (WE 5050/1-1).

9. References

[1] B. Weiss and F. Burkhardt, "Voice attributes affecting likability perception," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 1485–1488.

[2] L. Coelho, D. Braga, M. Sales-Dias, and C. Garcia-Mateo, "An automatic voice pleasantness classification system based on prosodic and acoustic patterns of voice preference," in *Proceedings of Interspeech*, Florence, Italy, 2011, pp. 2457–2460.

[3] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, "Would you buy a car from me? – on the likability of telephone voices," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 1557–1560.

[4] K. Scherer and H. Giles, *Social markers in speech*. Cambridge University Press, 1979.

[5] A. Paeschke and W. F. Sendlmeier, "Die Reden von Rudolf Scharping und Oskar Lafontaine auf dem Parteitag der SPD im November 1995 in Mannheim – Ein sprechwissenschaftlicher und phonetischer Vergleich von Vortragsstilen," *Zeitschrift für angewandte Linguistik*, vol. 27, pp. 5–39, 1997.

[6] E. Strangert and J. Gustafson, "What makes a good speaker? subjective ratings, acoustic measurements and perceptual evaluation," in *9th INTERSPEECH*, Brisbane, Australia, 2008, pp. 1688–1691.

[7] M. Zuckermann, H. Hodgins, and K. Miyake, "The vocal attractiveness stereotype: Replication and elaboration," *Journal of Nonverbal Behaviour*, vol. 14, pp. 97–112, 1990.

[8] R. van Bezooijen, "Approximant /r/ in Dutch: Routes and feelings," *Speech Communication*, vol. 47, no. 1, pp. 15–31, 2005.

[9] J. Trouvain, S. Schmidt, M. Schröder, M. Schmitz, and W. Barry, "Modelling personality features by changing prosody in synthetic speech," in *Proc. Speech Prosody*, Desden, Germany, 2006, pp. 4–7.

[10] B. Weiss and S. Möller, "Wahrnehmungsdimensionen von Stimme und Sprechweise," in *Proc. ESSV*, Berlin, Germany, 2011, pp. 261–268.

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

[12] B. Schuller, S. Steidl, A. Batliner, E. Noth, 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 *13th INTERSPEECH*, 2012.

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

[14] S. Schäfer, M., B. C., Geiser, and P. Vary, "A listening test environment for subjective assessment of speech and audio signal processing algorithms," in *Proceedings Elektronische Sprachsignalverarbeitung (ESSV)*, 2011, p. 237–244.

[15] P. Boersma, "Should jitter be measured by peak picking or by waveform matching?" *Folia Phoniatrica et Logopaedica*, vol. 61, pp. 305–308, 2009.

[16] J. Kreiman and D. Van Lancker Sidtis, *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Chichester: Wiley-Blackwell, 2011.

[17] L. Deal and H. Oyer, "Ratings of vocal pleasantness and the aging process," *Folia Phoniatrica*, vol. 43, pp. 44–48, 1991.