Voice Attributes Affecting Likability Perception

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

[1]Quality & Usability Lab, DT. Laboratories, TU Berlin, [2]Deutsche Telekom Laboratories

[BWeiss, Felix.Burkhardt]@telekom.de

Abstract

Ratings of voices' likability were collected in two successive studies. A single scale seems to be sufficient for assessing such ratings. Based on limited but controlled data, spectral parameters as well as f0 and articulation rate correlate with the ratings obtained. An automatic classification confirms the relevance of spectral features for the perception of likability. As a simple method of collecting more data for further studies, the single scale was validated within the bounds of the small data set. Both the spectral parameters and items from a comprehensive questionnaire indicate the relevance of timbre for the likability perception.

Index Terms: likability, voice, timbre, speaking style

1. Introduction

Prosodic features are known to be evaluated by listeners in order to recognize attributes of speakers like gender, age, emotion, or mood with quite some reliability, e.g. [1]. This is not necessarily the case for physical characteristics like weight or height (e.g. [2]). For example, women have proven to be consistent in their estimation of pleasantness of men's voices, but also height, weight, and age, although height was not correctly estimated [3].

Listeners can also ascribe personality traits – like the introversion-extraversion opposition – and attractiveness to a speaker purely based on short samples of his/her voice (e.g. [4]). Voice timbre is said to play a crucial role for the appraisal of personality [5]. However, detailed analysis of correlates of timbre are conducted only infrequently. Mostly, other prosodic features, namely pitch and duration are analyzed instead. Those seem to be strongly correlated with the expertise in speaking (e.g. acoustic measures: f0, rate, pauses, fluency [6, 7]), which is more often addressed in literature than what makes a good voice. Also, differences in pronunciation can result in negative evaluation [8, 9]. Regarding different aspects of voice and speaking, Fährmann classifies habitual qualities (pitch, loudness, timbre, sonority), individual qualities (rate, rhythm, accentuation) and accessory qualities (speaking style, syntax, wording) [5].

Ketzmerick studied acceptability of voices for e-learning applications. In her study low pitch, "clear" voice and constant voice "capacity" – assessed perceptually – are relevant for likable voices [10]. From a different analysis she finds correlations of these features with acoustic parameters – instrumental analysis – especially minimal f0 and f0 frequency distribution. In [11], speech rate, pausing, and pitch were analyzed on their effects in advertising showing that higher rate and lower pitch significantly correlate with a positive attitude towards the advertisement.

In summary, the topic of what aspects in voice and speaking are relevant for likability perception are not satisfactorily analyzed. Mostly, pitch related measures have been analyzed. We are particularly interested in the effect of voice timbre and its spectral correlates, but of course other factors like intonation have to be considered as well. We present two basic experiments and a classification procedure aiming at identifying (1) if voice likability is rated consistently and (2) what kind of acoustic parameters contribute to the likability perception.

In the first study, speech stimuli are rated subjectively for how likable the speaker sounds on a single scale in order to test this construct for stability. Based on these results, a subset of stimuli was rated again on a comprehensive questionnaire to test for the internal validity of the single scale. Perceptually assessed aspects of voice related to likability ratings are presented thereafter, along with a basic classification analysis.

2. First Study: Assessing Voice Likability

By selecting emotionally neutral sentences of read speech from the emoDB [12], possible influence of content, dialect, mood, situational adequacy and disfluencies were controlled. The speaking style was factual and speakers showed no signs of pathological voice quality. Nine different sentences (four short, five longer) for each[1] of the 10 speaker (gender balanced) were randomly presented to 20 subjects (7 female, 13 male, aged 21–37, M=27.7, SD=4.34, paid for their contribution). Headphones were used (Sennheiser HD 485, mono on both ears). Each stimulus was rated on a seven-point scale "sympathisch – unsympathisch"[2] with a mouse on a computer-screen. The tenth's sentence was used for training ("a07"). The study took about 30 minutes including instruction.

2.1. Results

As shown in Fig. 1, there is not much variation between the ratings of the individual sentences of each speaker. On the basis of the averaged ratings for each stimulus (averaged over each listener), repeated ANOVA confirms that there is no significant influence of the sentence ($F(8, 71) = 1.75$, $p = 0.1029$), but that there is a significant ranking of the speakers ($F(9, 71) = 25.12, p < 2e-16$***). Although, the variation in the subjects agreement is quite high (Fig. 2), a Kruskal Wallis test shows that the ranking of the speakers is not significantly different for the listeners ($\chi^2(19) = 24.05$, $p = 0.19$). Neither gender of the subjects nor gender of the speakers does have a significant impact on the ratings averaged over sentences.

The nine sentences used were the same for all speakers. Thus, basic spectral analyses for each stimulus could be conducted without taking into account linguistic content. To capture a suitable range of prosodic metrics, measures of intona-

[1] There is one sentence "a01" lacking for speaker 14, so instead of 90, there were only 89 stimuli.

[2] "Sympathisch" should be translated as "likable" or "pleasant".

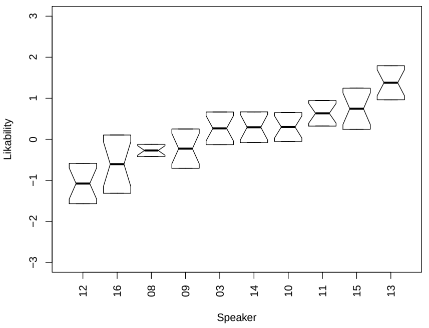

Figure 1: *Mean ratings for the speakers. CI and SD for the sentence differences, averaged over listener.*

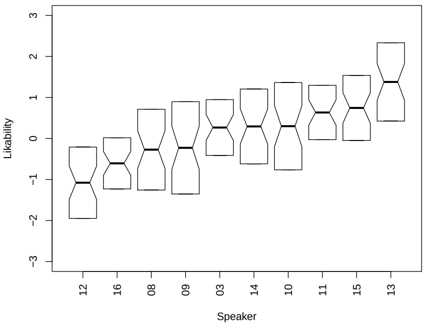

Figure 2: *Mean ratings for the speakers. CI and SD for the listener differences, averaged over sentences.*

tion (pitch related parameters), duration (in this case articulation rate), quantity (intensity) and timbre related parameters (simple parameters to assess the energy distribution in the spectral domain) were extracted from the speech recordings. The following list presents acoustic parameters assessed with Praat [13] for each stimulus:

- spectrum averaged for each stimulus: center of gravity (*S_cog*), standard deviation (*S_sd*), skewness (*S_skw*; the 3th central moment normalized by the standard deviation), kurtosis (*S_kts*), 3rd central moment (*S_cm3*), all to the power of 2
- fundamental frequency ($f0$), standard deviation ($f0$*_sd*)
- articulation rate as number of syllables per second (*rate*)
- intensity (*INT*), standard deviation (*INT_sd*)

Pronunciation or other sophisticated metrics (e.g. accentuation strategy or rhythm events) were not considered here, as in this basic investigation the general relation between different categories of acoustic parameters and likability are in the focus.

From the list of parameters used, only some measures significantly contribute to listers' ratings, as results of a repeated measures ANCOVA[3] reveal (cf. Table 1).

Of course, most significant results of the acoustic analysis are interaction effects with the gender of the speaker. Only syllable rate is a main effect. *Listeners' ratings of likability are higher with:*

- *higher articulation rate*,
- *lower spectral center of gravity*,
- *lower* $f0$ (only male speakers),

Table 1: *Significant results of ANCOVA for acoustic data and mean ratings of each stimulus (averaged over listeners).*

parameter	DF	F	p
rate	1,59	14.35	0.0003577***
gender:*S_cog*	2,59	8.50	0.0005696***
gender:*S_sd*	2,59	7.83	0.0009658***
gender:*S_skw*	2,59	4.51	0.0150783*
gender:*S_cm3*	2,59	2.97	0.0592371.
gender:$f0$	2,59	3.74	0.0294425*

- *higher spectral standard deviation and skewness* (only female speakers),
- but with *lower 3rd central moment* (only female speakers).

A model of linear regression with significant factors (additionally including gender:*S_cm3* but not "sentence") explains 45% of the data ($R^2_{adjust} = .37$, $RSE = .64$[4], $p < .001$; with "sentence" as factor: $R^2 = .54$, $R^2_{adjust} = .41$, $RSE = .62$, $p < .001$).

2.2. Discussion

In this small listening test, subjects did differentiate their rating of likability for the 10 speakers independently from the linguistic material and gender of both, speakers and listeners. The scale was found to be used consistently. Acoustic parameters assessed on sentence level correlate well with the subjects' ratings, including spectral parameters related to timbre, pitch and rate.

Lower $f0$ and *S_cog* are found to positively correlated with "likability" for men, which is consistent with a male stereotype of lower and darker voice and verified e.g. by results in [14]. However, this finding might be limited to neutral settings, as in other context, showing emotional arousal resulting in higher $f0$ is more adequate and therefore regarded as better [15]. Interestingly, a lower *S_cog* – likely corresponding to a darker perception of voice – is also positively perceived for female voices. The opposite – a bright voice as possible indication of youth and attractiveness – would have also been reasonable. Women are rated more positively with higher *S_sd*, i.e. more energy spread over the spectrum. Effects of the other parameters, 3th central moment and skewness, might directly result from those of the lower center of gravity and standard deviation, and thus have to be analyzed in more detail with other spectral measures.

Before additional voices can be assessed or additional acoustic parameters can be analyzed comprehensively, the validity and reliability of the rating-scale has to be evaluated. For this purpose, an additional test with a subset of stimuli was conducted.

3. Second Study: Assessing Voice and Speaking Attributes

In this listening test, only two sentences from each speaker were used in order to limit the duration of the study to one hour at most: Those stimuli, which have been rated not better or worse than speaker's mean ($+-1$ SD) were chosen to represent the individual voice. All nine sentences could be included about 2–

[3] Repeated measures factor is "sentence".

[4] R^2_{adjust}: R-squared adjusted for the number of explanatory terms in a model; RSE: the Residual Standard Error

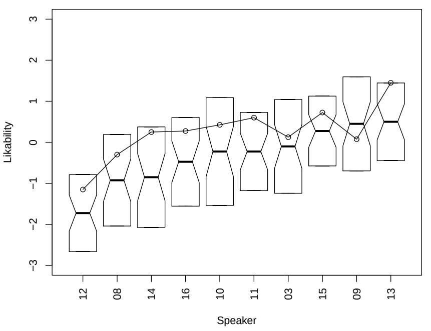

Figure 3: *Mean ratings for the speakers. CI and SD for the listener differences, averaged over sentences. For comparison, a line with means from the same stimuli from the 1st study*

3 times to avoid fatigue. But this time, a set of 7-point antonym-pairs – chosen by 10 experts based on literature [16, 17, 5, 3, 10] – was used to assess various impressions of speaker (13 items, including the sub-scale used in the first study) as well as characteristics of voice and speaking style (25 items). By this method, the internal validity of “likability” should be tested when assessing several aspects of voice impression.

The subset of stimuli is covering a similar range of ratings as the whole set and the ratings should thus span comparably.

The so-called big five (OCEAN: openness, conscientiousness, extraversion, agreeableness, neuroticism) were not included explicitly, as it was found that the dimension of extraversion is the only valid scale comparing self-description and external (i.e. friends’) description [18]. As this study aims at assessing the effect of voice on listeners’ rating of likability, but not the attribution of speakers personality, the big five were not explicitly assessed. Although, some items can be attributed to some of the 5 dimensions.

20 different subjects rated the randomly presented stimuli on the questionnaire (11 women, 9 men, aged from 21 to 42, M=28.7, SD=5.58). Instead of using the mouse for assessment, the questionnaire was printed on paper to improve subjects’ comfort and speed. Other aspects of the procedure were similar to the first study.

3.1. Results

The results for the likability sub-scale are depicted in Fig. 3. Apart from a general tendency of lower ratings in the second study, speaker “03” and “09” stand out in comparison. The standard deviation of the subjects’ ratings (averaged over sentences) is comparable ($SD_1 = 1.20$, $SD_2 = 1.55$). Despite these two outliers, the likability scales seem to measure at least strongly related constructs, if not the same, as the raking is otherwise identical. Further data from more subjects is needed for final consideration.

Using the 25 items describing voice and speaking style, a parallel analysis reveals seven factors. In the subsequent factor analysis (57% cumulative variance explained, Kaiser-Meyer-Olkin measure of sampling adequacy =.84) all items were eliminated with loadings $< .5$ and all cross-loadings, resulting in finally six factors with 19 items.

Using the means of the items as representation of the factors, a linear model shows, that 3 of the 6 factors significantly contribute to the likability ratings ($R^2_{adjust} = .39$, $RSE = 1.19$). These factors include the following items[5]: 1: “warm”, “edgeless”, “soft”, “comforting”, “relaxed”, 2: “not hoarse”, “high”, “not dark”, 4: “resonant”, “voluminous”, “powerful”. Other items are not related to the likability scale (3: “nasal”, “throaty”, “aspirated”, “creaky”“fluent”, 5: “professional”, “practiced”, and excluded: “regular” , “varying”, “slack”, “clear”, “slow”, “articulate”, “natural”).[6]

3.2. Discussion

Despite the small number of participants in both studies, the usage of the “likability” scale seems to be comparable for both studies. All of the qualitative ratings correlated with this scale can be described as habitual qualities, two even as aspects of timbre (factors 1 and 3). The listeners rated warm, resonant and not dark voices positively. The discrepancy between the sub-scale “high” on the one hand and “relaxed” and the lower f0 values for men in the first study on the other hand may be caused by the whole factor construct of factor 2: It comprises aspects of too low, too dark, and not healthy voices, but not necessarily low pitch.

These results fit well with the correlation of acoustic parameters (i.e. S_cog in the first study, that could effect the impression of a warm and resonant voice. Unfortunately, there are too few stimuli used in the second study to correlate the subjects’ rating directly to acoustic parameters, as the linguistic material varies with the stimulus selection.

4. Prediction of Voice Likability

Acoustic parameters of pitch, spectrum, and rate could be identified to correlate well with ratings of voice likability. In order to evaluate the automatic predictability for the data, we conducted a further classification experiment. Instead of selecting a small number of possibly relevant parameters, an already established bigger set of parameters was used instead. To extract these acoustic features, we used the open-source Emotion and Affect Recognition Toolkit openEar [19]. It extracts the following feature groups: Signal energy (Root Mean-Square and logarithmic), FFT-Spectrum (Phase, magnitude), Autocorrelation Function (ACF via square of FFT magnitude), Mel-Spectrum, Cepstral, Pitch Fundamental Frequency (F0 via ACF in Hz, Bark and closest semitone, Probability of Voicing, Voice Quality (Harmonics-To-Noise Ratio), Time Signal (Zero-Crossings, Max./Min. value, DC), Spectral (Energy in bands 0–250 Hz, 0–650 Hz, 250–650 Hz, 1k–4k Hz, Centroid, Flux, and rel. pos. of spectral max. and min), LPC (Coefficients and Residual), Formants (Center frequency and bandwidth) and Musical (Semitone spectrum, Chroma, Chroma Energy distribution Normalized Statistics (CENS)).

The aim is to compare the set of relevant parameters from the first study with the resulting features from this classification approach. For a preliminary test we therefore simply used the configuration file delivered with the distribution “emobase.conf”, which is a configuration to extract a base set of 988 features for emotion research, 1st level functionals of low-level descriptors such as MFCC (Mel Frequency Cepstral Coefficients), Pitch or LSP (line spectral pairs).

As a classification engine, the LADT (LogitBoost Alternating Decision Trees) classifier implemented by the WEKA toolkit [20] was used. It is is a classification technique that

[5] Just one of the two poles are stated.

[6] Please note, that the subjects were layman concerning phonetics, so we could not expect consistent results for very specialized terms.

combines decision trees with the predictive accuracy of boosting into a set of interpretable classification rules. As a boosting technique the LogitBoost strategy is used. The number of boosting iterations was set to ten.

Classification trees have the advantage over rivaling classifications techniques, that the rules that are learned during the training phase can be used to gain knowledge on the underlying connection between features and category membership.

4.1. Results

In order to transfer the numeric "likability" value (-3 – +3) into a nominal class value (which is a prerequisite for most classifier architectures), we decided for a binary classifier ("likable" vs. not "likable") and defined 0.5 as a threshold to account for the slightly positive tendency in evaluation. This resulted in 34 likable samples and 56 "dislikable" ones. The baseline for this experiment (a trivial classifier always deciding for the most frequent category) would thus be 62.9 % accuracy.

Because we are interested in the "likability" of individual speaker's voices, a leave-one-speaker-out test procedure conducted by testing 10 times one speaker against a training set formed from the other nine speakers. It resulted in 69.663% accuracy and .769 F-measure, a measure that combines Precision and Recall as the harmonic mean of precision and recall. The most prominent features used in the tree construction are the tenth, sixth and fifth MFCC mean value as well as the second, third and fourth formant center frequency.

4.2. Discussion

With this automatic classification, a significant result above the chance level could be obtained. In contrast to the limited set of parameters from the first study, only spectral features are dominant for prediction. Some of the parameters used are comparable, namely MFCC coefficients and spectral measures like *S_cog* and *S_sd*.

The inclusion of frequencies of the 2nd, 3rd, and 4th formant in this classification motivates for a more detailed analysis of relevant acoustic features. A classification of the features used in section 2 with the LADT approach resulted in 62.92% accuracy, obviously some important features are missing in the hand-selected set.

We also tried to repeat the described experiments with gender-separated data but did not get clear results, which most probably is a result of data sparsity. We plan to continue experiments on larger data sets.

5. Conclusion and Outlook

How likable a speaker sounds can be assessed relatively reliably and valid on a single scale. Prosodic acoustic parameters including spectral metrics, f0, and articulation rate correlate well with listeners' ratings. Some of the relevant metrics, e.g. spectral center of gravity, as well as factors from the detailed questionnaire correspond to the concept of timbre. Based on this first but encouraging results, further studies are planned to analyze the impact of correlates of timbre and speaking style on the perception of likability more closely.

As the single scale has proven to be useful, it should be possible to gather sufficient data – the number of speakers would have to exceed the hundred – for building a reliable model to predict subjective ratings of likability. We plan to extend our studies on larger datasets, ideally one that is publicly available and can be used to compare results from different research sites, e.g. [21].

6. References

[1] F. Burkhardt, R. Huber, and A. Batliner, "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.

[2] W. van Dommelen and B. Moxness, "Acoustic parameters in speaker height and weight identification: Sex-specific behaviour," *Language and Speech*, vol. 38, pp. 267–287, 1995.

[3] L. Bruckert, J. Lienard, A. Lacroix, M. Kreutzer, and G. Leboucher, "Women use voice parameter to assess men's characteristics," *Proc. Biological Sciences*, vol. 237, no. 1582, pp. 83–89, 2006.

[4] 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.

[5] R. Fährmann, "Elemente der Stimme und Sprechweise," in *Vokale Kommunikation*, K. Scherer, Ed., Gießen, 1982, pp. 228–252.

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

[7] H. Monitz, A. I. Mata, I. Trancoso, and M. C. Viana, "How can you use disfluencies and still sound as a good speaker?" in *Proc. INTERSPEECH*, 2008, p. 1678.

[8] R. Eklund and A. Lindström, "Xenophones: An investigation of phone set expansion in Swedish and implications for speech recognition and speech synthesis," *Speech Communication*, vol. 35, no. 2, pp. 81–102, 2001.

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

[10] B. Ketzmerick, *Zur auditiven und apparativen Charakterisierung von Stimmen*. Dresden: TUDpress, 2007.

[11] A. Chattopadhyay, D. W. Dahl, R. J. Ritchie, and K. N. Shahin, "Hearing voices: The impact of announcer speech characteristics on consumer response to broadcast advertising," *Journal of Consumer Psychology*, vol. 13, no. 3, pp. 198–204, 2003.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.

[13] P. Boersma, "Praat, a system for doing phonetics by computer." *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[14] S. Thiel, "Die mittlere Sprechstimmlage als Wirkungsfaktor," in *Ergebnisse der Sprechwirkungsforschung*, E.-M. Krech, J. Suttner, and E. Stock, Eds. Abteilung Wissenschaftspublizistik d. Universität Halle-Wittenberg, 1987, pp. 202–208.

[15] 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," *ZfAL*, vol. 27, pp. 5–39, 1997.

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

[17] S. Singh and T. Murry, "Multidimensional classification of normal voice qualities," *JASA*, vol. 64, no. 1, pp. 81–87, 1978.

[18] K. Scherer, "Stimme und Persönlichkeit – Ausdruck und Eindruck," in *Vokale Kommunikation*, K. Scherer, Ed. Gießen: Neltz, 1982, pp. 188–210.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proc. 4th Int. HUMAINE Association Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2009.

[20] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

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