STUDYING VOCAL SOCIAL ATTRACTIVENESS BY RE-SYNTHESIS – RESULTS FROM TWO STUDENT PROJECTS APPLYING ACOUSTIC MORPHING WITH TANDEM-STRAIGHT

Benjamin Weiss, Anabell Hacker, Cleopatra Moshona, Frederic Rudawski, Matthias Ruhland

Quality and Usability Lab, Technische Universität Berlin
benjamin.weiss@tu-berlin.de

Abstract: Acoustic analysis and synthesis experiments provide meaningful methods to study vocal processes of social cognition. Common attributions and evaluations of person's speech include personality, emotions, and social attitudes. Recent advances in speech coding allow improving the synthesis methodology by defining valid parameter target values without much manual analysis effort. In two student projects, the de-factor standard for speech coding/morphing, Tandem-Straight, was applied to conduct experiments on vocally manifested social cognition processes. Hypothesis-based stimuli were produced on four, respectively five steps in a bi-directional fashion. The evaluations took about 30 min. for each participant. Each group conducted a qualitative analysis for stimulus selection and a quantitative pre-test to obtain social evaluations for hypothesis formulation. During this preparation, both groups decided to select female speakers, as only these provided sufficient variations in acoustics and social ratings. Results show an impact of a darker spectrum on liking ratings for 47 participants, as well as an effect of different, eventually lower, F0 on dominance attribution for 51 participants.

1 Introduction

The analysis-by-synthesis paradigm aims at verifying observed relationships between acoustic measures and target conditions or questionnaire values, such as attributed speakers' traits (personality, competence, physical characteristics), states (emotions), or attitudes (liking, interest). For the universal two-factor model comprising valence (warmth, benevolence, rapport, communion) and competence (dominance, agency) [1][11], studies identified tempo as one relevant acoustic parameter: While the relationship between speech rate and the attribution of competence seem to be linear, an ideal-point was found for benevolence [1][4][7][21][22]. For the fundamental frequency (F0), however, there is a negative relationship detected and confirmed by resynthesis for warmth [3][8] and competence or dominance [1][2][4][17]. Others find a positive correlation with warmth [17]. Acoustic analysis in both sexes also showed a negative relationship between F0 and social attractiveness [15][23][26].

Social attractiveness is in this case the rater's liking of the speaker [12] and is often identified as the warmth dimension [4][10][13][17]. Conceptually, however, this attractiveness might also comprise physical attractiveness and competence [18], and questionnaire items assessing liking do not only load on warmth, but in some studies on both dimensions warmth and competence in a factor analysis [9][22][23][27]. Spectral characteristics have also been analyzed in respect to social cognitive attributions, but not in a similar extent and not as often with (re-)synthesis. Parametrizations, like the harmonics-to-noise-ratio (HNR) [15][25], spectral tilt [25], or spectral center of gravity (CoG) [25] sometimes show no significant relationship with rating data, but do in others (CoG [24], HNR [17]).

In order to lead on research of these universal two socio-cognitive concepts, the morphing paradigm was chosen that is recently applied for studying vocal person identification [20]. By

high quality coding of speech samples from real recordings, morphs of all coding parameters, or just a selection, can be manipulated in several steps towards the values of a selected target sample. There is a benefit in studying the warmth and competence dimensions by acoustic morphing in comparison to systematically varying a selected parameter in usual (re-)synthesis experiments: Typically, additional recordings have not been used to obtain precise target values. But with morphing, both, original and target voices have naturally occurring values, as these stem from real recordings. The acoustic quality of stimuli produced in such a way seem to be quite similar, so re-synthesizing original stimuli is not necessary, as the coded version of the original already exhibits comparable degradations.

In addition, real recordings can be selected as starting point or target for morphing by previous assessments in order to control or test context effect. For example, the first experiment on the effect of brightness in timbre used likable and dislikable recordings as origin to test the hypothesis in context of other factors. This potential of increasing validity comes with a practical advantage, as the graphical tool of the chosen speech coder Tandem-Straight [16] requires only initially relevant manual work to prepare the starting and target representations, whereas the morphing and selection of (bundles of) morphing parameters is time effective. These features make this approach attractive for university teaching, especially for interdisciplinary groups with the regular need for theoretic and methodological discussions. The experiments presented here were prepared and conducted in the summer semester 2016.

2 Methodology

Two student groups conducted hypothesis-based experiments of social-cognitive concepts. Each group conducted a qualitative analysis for stimulus selection and a quantitative pre-test to obtain social evaluations for hypothesis formulation. During this preparation, both groups decided to select female speakers, as only these provided sufficient variations in acoustics and social ratings. Stimuli were produced on 4, respectively 5 steps in a bi-directional fashion with the Matlab morphing tool Tandem-Straight. The evaluations took about 30 min. for each participant and among all participants of each experiment, a 15€ voucher was raffled off. The Tandem-Straight morphing is supported by a Matlab graphical user interface to code the source and target stimuli and manually optimize mapping between the coded features. It allows for morphing in mapping of temporal axis, frequency axis, time-frequency indexed spectrographic level as well as aperiodicity, and F0.

3 Spectral morphing to study likability

Spectral energy distribution is known to correlate with the perceived timbre of voices. According to literature, "darker" timbre is often associated with warmth, which is potentially related to a preference of "darker" female voices in regards to likability [27], in concordance with [23][24]. However, also the opposite, a preference for "younger" and "brighter" voices in females could be possible [13].

3.1 Material

The stimuli used for the experiment were taken out of the Phondat 1 corpus [5], for which likeability rankings were available. The group selected stimuli of 4 female speakers (*mxbd, w02a, w04a, w06a*), uttering the German sentence S03 *"Heute ist schönes Frühlingswetter",* to represent each morphing end condition of "likeable/unlikable" and "dark/bright". The audio test was conducted by means of the online software SoSci-Survey, for which the stimuli had to be converted to mp3 format (192 kBit/s).

3.2 Procedure

In order to exclude potential influences on the likability rating and ensure comparability, the stimuli were selected with the following criteria: emotionally neutral utterances without remarkable sociolects, idiolects or dialects. To minimize morphing artifacts, a short sentence that did not contain any speech errors was chosen. A pre-selection of the stimuli in respect to their "brightness/darkness" was conducted by means of auditory analysis and visual comparison of spectral energy distribution.

The utterances were annotated with Praat to provide reliable and precise temporal anchor points for all speakers. The annotated files were imported to Tandem-Straight via a script that was specifically developed for this purpose. For morphing, the standard Tandem-Straight F0 extractor was used. The extracted F0 contour and the recognition of voiced-unvoiced segments was corrected manually to ensure better morphing conditions and results. The stimuli were morphed bi-directionally in 5 steps (0%, 25%, 50%, 75%, 100%), between the supposedly "dark" and "bright" speakers. In order to analyze whether likability perception is linear or categorical, the 3 steps in between were included. Stimuli morphing was conducted by only manipulating the spectrum parameter in Tandem-Straight to ensure that no further aspects influenced the likeability rating. In addition, it was found that F0 inclusion resulted in morphs that were too similar to the original target stimuli.

Group A recruited 47 participants (27 female, 20 male) for the audio test, which ran from 28.06. to 05.07.2016. The audio test was conducted in laboratory conditions in 40 cases. In 7 cases the experiment took place in home environment with similar conditions using the same headphones. The questionnaire consisted of 3 items, assessing likability, gender and age. Likability was rated on a scale of 1 (very dislikable) to 7 (very likeable).

The morphed nature of the stimuli was not revealed to the participants, in order to avoid potential bias in regards to the naturalness of the voices and the resulting likability ratings. Gender and age were not evaluated but instead used as a means to conceal the aim of the experiment. To distract from unavoidable morphing artifacts, participants were told that the stimuli were partially of poor recording quality. During the evaluation, randomization errors in SoSci survey were noticed, which resulted in duplicated stimuli ratings for some participants and missing values in others. In total however, the stimuli received an equal number of evaluations.

3.3 Results

Due to the missing values, two-sided t-tests (Bonferroni-corrected to $p<.00625$) were conducted for each of the four pairs comparing only the original and target stimuli testing for main effects (Table 1). There are five significant results and three non-significant results. All significant results reveal a positive effect of the spectral values from the presumably "darker" voices (Figure 1). Visual inspection of all the ratings, including partially morphed stimuli, revealed a rather linear pattern.

3.4 Discussion

The results of the listening experiment support the initial assumption that the two voices subjectively labelled as "dark" provide more likable spectral information. Morphing to target values taken from the two "bright" voices result in less positive ratings, whereas morphing towards the "darker" spectral values result in more positive ratings. In all cases the increase/decrease of likeability follows a rather linear than categorical pattern. An expectation of 50% morphs being lower in signal naturalness and thus lower likeability was not supported. However, not all stimuli pairs show significant differences. The pair of voices *w04a* and

w06a, where the "brighter" voice is originally more likable, show no significant difference when morphed in either direction, which is not different from comparing the original stimuli directly ($t(109)= 1.29$, $p=.20$).

"dark" voices	"bright" voices	dark → bright	bright → dark
w06a	*mxbd*	***t*(47.88)=3.54, *p*=.0009**	$t(57.90)=1.34$, $p=.1851$
w06a	*w04a*	$t(50.84)=0.85$, $p=.4016$	$t(55.61)=2.23$, $p=.0299$
w02a	*mxbd*	***t*(52.93)=7.33, *p*<.0001**	***t*(50.75)=-3.12, *p*=.0029**
w02a	*w04a*	***t*(47.42)=4.27, *p*<.0001**	***t*(59.00)=-3.05, *p*=.0035**

Table 1 – Results of the statistical analysis comparing 0% and 100% morphed spectra (sig. results in bold)

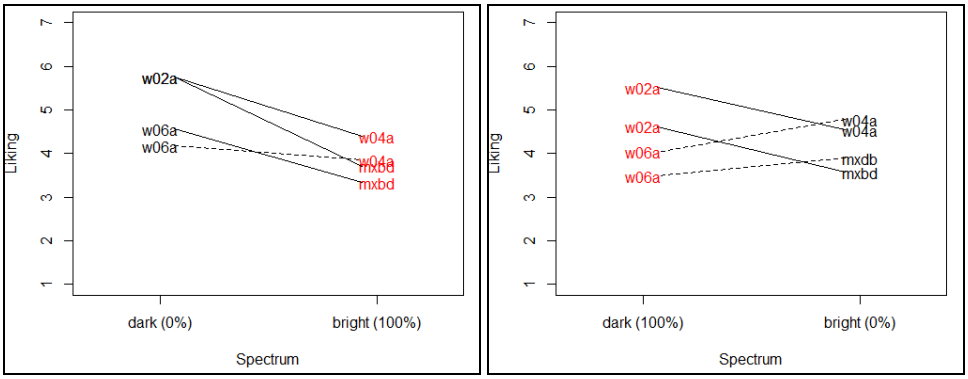

Figure 1 – Mean ratings of the 0% (black) and 100% (red) morphed spectra; solid lines indicate significant differences (w02a on the left is printed two time with nearlly identical values)

Unexpectedly, voice *mxbd* ("bright" spectrum) was rated more likeable during the audio test than in the original rating. This could be explained by the fact that this speaker was estimated to be much older than the other speakers by audio test participants, which may imply a stronger tolerance of bright spectra in older age groups or another factor of, e.g., intonation.

Limitations in the interpretation of the findings in this study may arise due to a number of reasons. Firstly, it cannot be ruled out that not only spectral information but also other factors that may have an influence on likability rating are morphed when manipulating the spectra of two natural stimuli in Tandem-Straight. A possible solution could be to perform a synthesis, which focuses solely on the spectrum. However, this often results in a loss of naturalness.

4 F0 morphing to study trust and dominance

Although the influence of on perceived speaker traits tends to be one of the best researched acoustical measures, results regarding the universal two-factor model are inconsistent [1][2][3][4][8][17]. There is evidence suggesting a positive relationship between F0 and dominance while rating female speakers. In contrast, lower-pitched male voices are associated with dominance, while a higher fundamental frequency leads to higher ratings in valence [17], indicating an effect of the speaker's gender. In an attempt to get a better understanding of the relationship between the average pitch and perceived social traits, the second study group focused on testing systematic variations of female speaker's F0.

4.1 Material

For the purpose of speaker selection, a pretest was conducted in which 22 participants had to rate a variety of male and female speakers on the two dimensions of the two-factor model. In concordance with Group A, the recordings of a German sentence („*Es ist 8 Uhr morgens.*“, s12) were taken from the from the Phondat 1 corpus [5]. Results showed that only the female samples provided enough variation in the ratings of dominance and valence to suffice as starting point for the hypothesis-based morphing.

In an attempt to reduce participants' fatigue during the experiment a second German sentence (“*Nachts haben Meiers gut geschlafen.*”, s10) was selected to provide some text variation. Like in Group A, all recordings were pre-selected to not contain dialects or speech errors. Using a bi-directional 4-step morphing method (0%, 33%, 66%, 100%) applied on 12 morphing pairs, a total of 72 stimuli were created that featured F0 values in the range of naturally occurring values of the respective base stimuli. Using the same methodology, 72 additional samples were created that utilized the 4 remaining morphing dimensions of Tandem-Straight besides F0 manipulation to serve as control condition. Every presented stimulus was converted to mp3 format (192 kBit/s) and normalized to an average level to avoid the natural bias for loudness. The 4-step paradigm was opted for to avoid the middle step (50%) where acoustical artifacts seem to occur frequently as well as to keep the overall count of stimuli as low as possible due to the mixed design method explained in the following.

4.2 Procedure

Dominance and trust assessment were separated in order to obtain more stable results [17]. However, this comes with the cost of presenting each stimulus twice during a trial. To counter this, a mixed design was applied where participants were separated into two versions of the study using the survey tool LimeSurvey. In both studies every morphing pair and morphing step was present, but different sentences were used for the 6 blocks of questioning, alternating between repeated blocks assessing dominance and trust. While block 1 used sentence 10 for the first and last two blocks of the questionnaire, block 2 used the stimuli of sentence 12 for the majority of the blocks. Stimuli were presented randomized within the blocks and the order of dominance—trust assessment was reversed for block 2.

Just like in Group A, the experiments were conducted under laboratory conditions using AKG K-601 headphones for each testing. In the timeframe between 11.07.2016 and 17.08.2016 a total of 27 participants (13 female, 14 male, aged 20—35, M=28, SD=3.8) was recruited for study A, while 24 participants (11 female, 13 male, aged 17—51, M=29.6, SD=7.95) completed the second version of the study. During the experiment, both dominance and trust were rated on a scale of 1 (not dominant/not trustworthy) to 9 (very dominant/very trustworthy). Additionally, the subject's age, gender and native language was assessed.

4.3 Results

In both dimensions the speakers *gisn* and *w04a* are rated lower compared to *w64a* and *w02a* (Figure 2). Due to the blocked design of the second experiment, single, univariate mixed-models (subjects nested within sentence) were conducted (Bonferroni-corrected to $p<.003125$) for each social concept, trustworthiness and dominance. The pitch morphing from and to *gisn* does not have an effect, but from and to *w04a* to the more dominantly regarded speakers improves, respectively decreases, the ratings (Table 2).

The control condition of morphing all parameters except for pitch resulted in seemingly complementary effects (Table 3), as *gisn* shows consistent lower ratings compared to morphed stimuli with target values from *w02a* and *w64a*, but *w04a* does not. A similar

analysis for trust does not result in systematic effects (Table 3), as there is only one significant result (*w064* decreases to *w04a*; F(1,100)=43.63, p<.0001, but not in the other direction).

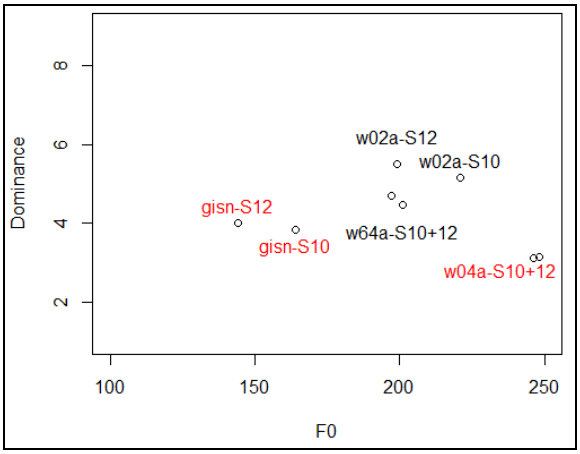

Figure 2 – Selected Stimuli: Dominance ratings and average F0 (black for high trust, red for low trust)

"negative" voices	**"positive" voices**	**neg. → pos.**	**pos. → neg.**
gisn	w02a	*F*(200)=5.88, *p*=.0162	*F*(200)=3.39, *p*=.067
gisn	w64a	*F*(200)=4.62, *p*=.0328	*F*(200)=2.53, *p*=.114
w04a	w02a	***F(200)=40.04, p<.0001***	**F(200)=24.04, p<.0001**
w04a	w64a	***F(200)=25.64, p<.0001***	**F(200)=24.18, p<.0001**

Table 2 – Results of the statistical analysis comparing 0% and 100% morphed F0 for dominance (sig. results in bold)

"negative" voices	**"positive" voices**	**neg. → pos.**	**pos. → neg.**
gisn	w02a	**F(200)=41.38, p<.0001**	**F(200)=45.18, p<.0001**
gisn	w64a	**F(200)=9.45, p=.0024**	**F(200)=21.46, p<.0001**
w04a	w02a	**F(200)=9.27, p=.0026**	*F*(200)=7.64, *p*<.0062
w04a	w64a	*F*(200)=1.05, *p*=.3080	*F*(200)=0.06, *p*=.8120

Table 3 – Results of the statistical analysis comparing 0% and 100% morphed stimuli except F0 for dominance (sig. results in bold).

4.4 Discussion

The morphing of fundamental frequency in a bi-directional way from two female speakers with high values of dominance to speakers with lower values resulted in a systematic confirmation for only one of the speakers with lower dominance (*w04a*). Interestingly, the difference in dominance for the other speaker (*gisn*) seems to originate from other acoustic factors not related to F0. No such consistent effect was found for trust. However, an informal survey among the project members assessing stimulus quality suggested that there is a considerable negative relationship between artificiality of a sample and the respective trust rating. This seems plausible due to the personal nature of the factor, which is why future studies in this domain should be especially cautious while using morphed stimuli of varying

naturalness. For identifying the origin of the dominance ratings, additional information is required, as the missing results for *gisn* might also be caused by the small difference in dominance to the two "positive" speakers. However, it is also likely that average pitch is not relevant here, but instead the intonation contour, which is also affected by the F0 morphing. Therefore, separating average pitch from the intonation contour might be a next step in identifying the dominance related speaker differences.

5 Conclusion

Both experiments give initial results on identifying acoustic effects on personality attribution for female speakers. The re-synthesis with natural acoustic parameter values could confirm hypotheses formulated from literature on analysis. However, only a few speakers were chosen which limits generalization of the effects found. Follow up experiments should study more detailed aspects of linearity of the acoustic-attribution relationship, or the interplay between single features, e.g. spectral vs. F0 effects. During this study, no subgroups (e.g. age groups, gender, social background) were taken into consideration during the evaluation of the results. This aspect could be addressed in future research.

From a methodological point of view, Tandem-Straight allowed even non-experts to prepare stimuli for the experiments. To save time, up to 3 participants took the audio test simultaneously, resulting in background noise due to open headphones. This may have further influenced the focus of the participants and could have been avoided through the use of closed headphones. In the standard configuration of the Tandem-Straight GUI interface not all control buttons are readable on a Windows system. Therefore, the Matlab resize function was enabled. In addition, the *F0ExtractorGUI.m* file was adapted, in order to enable subsequent 192fine-tuning of F0 extraction, by preventing the closing of the *F0ExtractorGUI* in the *finishButton_Callback* function. In case of unsatisfying synthesis results, the F0 extraction has to be restarted otherwise.

6 Acknowledgment

We want to thank all members of the study project for their commitment and collaboration and all participants of the two experiments.

References

[1] ABELE, A.E., CUDDY, A.J.C., JUDD, C.M., YZERBYT, V.Y.: Fundamental dimensions of social judgment. editorial to the special issue. *European Journal of Social Psychology* 38 (7), 2008, 1063–1065.

[2] APICELLA, C.L., FEINBERG, D.R.: Voice pitch alters mate-choice-relevant perception in hunter-gatherers. *Proc. of the Royal Society B-Biological Science* 276, 2009, 1077–1082.

[3] APPLE, W., STREETER, L.A., KRAUSS, R. M.: Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology* 37(5), 1979, 715–727.

[4] ARGYLE, M.: *Bodily Communication* (2nd ed.). Methuen, New York, 1988.

[5] BAYERISCHES ARCHIV FÜR SPRACHSIGNALE: *PhonDat 1*. München, 1995.

[6] BROWN, B.L., STRONG, W.J., RENCHER, A.C.: Fifty-four voices from two: the effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech. *Journal of the Acoustical Society of America* 55(2), 1974, 313–318.

[7] BROWN, B.L., STRONG, W.J., RENCHER, A.C.: Acoustic determinants of perceptions of personality from speech. *Linguistics* 13(166), 1975, 11–32.

[8] BRUCKERT, L., LIENARD, J., LACROIX, A., KREUTZER, M., LEBOUCHER, G.: Women use voice parameter to assess men's characteristics. *Proc. Biological Sciences* 237, 2006, 83–

89.
[9] CUDDY, A.J., FISKE, S.T., GLICK, P.: Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in Experimental Social Psychology* 40, 2008, 62–149.
[10] DEPAULO, B.M., KENNY, D.A., HOOVER, C.W., WEBB, W., OLIVER, P.V.: Accuracy of person perception: Do people know what kinds of impressions they convey? *Journal of Personality and Social Psychology* 52 (2), 1987, 303–315.
[11] FELDSTEIN, S., DOHM, F., CROWN, C.: Gender and speech rate in the perception of competence and social attractiveness. *Journal of Social Psychology* 141, 2001, 785–806.
[12] FERGUSON, M. J., FUKUKURA, J.: Likes and dislikes: A social cognitive perspective on attitudes. In: Fiske, S. T., Macrae, C. N. (Eds.), *The SAGE Handbook of Social Cognition*. SAGE Publications Ltd, London, 2012, 165–186.
[13] FIEDLER, C.: Sprechwirkung: Ästhetische Urteile über weibliche Sprechstimmen. Prosodische Merkmale einer attraktiven und sympathischen Frauenstimme. Universität Salzburg, 2005.
[14] FISKE, S.T., CUDDY, A.J., GLICK, P.: Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11(2), 2007, 77–83.
[15] GRAVANO, A., LEVITAN, R., WILLSON, L., BEŇUŠ, Š., HIRSCHBERG, J., NENKOVA, A.: Acoustic and prosodic correlates of social behavior. In: *Proc. Interspeech*, 2011, 97–100.
[16] KAWAHARA, H., MORISE, M., TAKAHASHI, T., NISIMURA, R., IRINO, T., BANNO, H.: A temporally stable power spectral representation for periodic signals and applications to inference-free spectrum, F0, and aperiodicity estimation. *Proc. ICASSP*, 2008, 3933–3936.
[17] MCALEER, P., TODOROV, A., BELIN, P.: How do you say 'Hello'? Personality Impressions From Brief Novel Voices. *PLOS ONE* 9(3), 2014.
[18] MCCROSKEY, J., MCCAIN, T.: The measurement of interpersonal attraction. *Speech Monographs* 41, 1974, 261–266.
[19] RAY, G.B.: Vocally cued personality proto-types: An implicit personality theory approach. *Communication Monographs*, 53, 1986, 266–276
[20] SCHWEINBERGER, S.R., SCHNEIDER, D.: Wahrnehmung von Personen und soziale Kognition. *Psychologische Rundschau* 65(4), 2014, 212–226.
[21] SMITH, B.L., BROWN, B.L., STRONG, W.J., RENCHER, A.C.: Effects of speech rate on personality perception. *Language and Speech*, 18, 1975, 145–253.
[22] STREET, R.L., JR., BRADY, R.M.: Speech rate acceptance ranges as a function of evaluative domain, listener speech rate, and communicative context. *Communication Monographs*, 49, 1982, 290–308.
[23] WEIRICH, M.: *Die attraktive Stimme: Vocal Stereotypes. Eine phonetische Analyse anhand akustischer und auditiver Parameter*. Saarbrücken: Verlag Dr. Müller, 2010.
[24] WEISS, B., BURKHARDT, F.: Voice Attributes Affecting Likability Perception. *Proc. Interspeech*, 2010, 1934–1937.
[25] WEISS, B., BURKHARDT, F.: Is ‘not bad’ good enough? Aspects of unknown voices’ likability. In: *Proc. Interspeech*, 2012.
[26] WEISS, B., MÖLLER, S.: Wahrnehmungsdimensionen von Stimme und Sprechweise. In: *Proc. Elektronische Sprachsignalverarbeitung (ESSV)*, Aachen, KRÖGER, B.J. & BIRKHOLZ, P. (Eds.), Studientexte zur Sprachkommunikation 61, Dresden: TUDpress, 2011, 261–268.
[27] Weiss, B.: Voice Descriptions by Non-Experts: Validation of a Questionnaire”. *Proc. Phonetics&Phonology*, 2016, 228–231.
[28] WORTMAN, J., WOOD, D.: The personality traits of liked people. *Journal of Research in Personality* 45, 2011, 519–528.