Speak to me and I tell you who you are! A language-attitude study in a cultural-heritage application

Stephanie Schreitter, Brigitte Krenn, Friedrich Neubarth, and Gregor Sieber

Austrian Research Institute for Artificial Intelligence, 1010 Vienna, Austria,
firstname.lastname@ofai.at

Abstract. When developing artificial companions, social attribution significantly influences the attitudes of humans towards the agents. We present results from a language-attitude study based on three synthetic varieties of Austrian German (a standard and two Viennese varieties) in the context of a cultural-heritage application. We show that language variety together with voice quality elicit attributions of different personas and influence the attitudes of the listeners toward the speakers.

1 Introduction

Virtual agents and robots gain increasing interest as future companions of humans for entertainment and leisure activities or as assistants in the virtual as well as the physical world. Common to both virtual and robotic companions is that they are perceived as social actors by their human partners, and they need to be capable of communicative interaction. In this regard, an understanding of how behavioural cues are perceived and evaluated by humans is essential.

Based on results from previous natural language-attitude studies comparing standard versus dialectal natural speech, we explore the hypothesis that similar attributions can be found for synthetic speech: Humans attribute personas[1] to synthetically generated language varieties and synthetic voices, which has a strong impact on human social perception and evaluation. To the best of our knowledge, our work is the first to investigate social effects of language variety using synthetic voices. Additionally, we analyze effects of standard and non-standard language varieties of synthetic speech in the context of a cultural-heritage setting. In particular, we investigate the effects on social evaluation of three synthetic voices (male standard Austrian German, female colloquial Viennese, male dialectal Viennese) representing three virtual tour-guides in a 3D-animation of the *State Hall of the Austrian National Library.* For details on speaker selection and the design of the voices see [8], and on the cultural heritage application see [2].

In the following sections, we introduce related work on language-attitude studies and the assessment of virtual agents. We then present the experimental

[1] In this context we use persona as those aspects of a character that are perceived by others and thus are relevant for the (social) evaluation of the character.

design and the evaluation methods employed, followed by an analysis of the data and a discussion of the results.

2 Related Work

Since the 1980s, research has been carried out showing that the degree an artificial entity resembles a human correlates with the likeliness that the entity will evoke social and psychological processes in humans, e.g. [11]. In order to develop personalized companions accompanying and supporting users, a better understanding of gender, language variety, and social as well as ethnical effects on users are of increasing importance. Results of experiments reported in [6] support findings in the field of gender linguistics, e.g. that on the one hand social identification and proximity to communication partners of the same sex is rated higher, and on the other hand that male agents are rated as more competent by both men and women. Crowell et al. [1] conducted an experiment comparing sex-related differences in reactions towards gendered synthetic voices that are either physically embodied within a robot or disembodied. Their results have shown that both men and women found the disembodied female voice and the male embodied voice to be more reliable.

Findings in the field of language-attitude studies using synthetic voices have shown that women consistently rated both male and female voices more positive than men [7, 9]. The same effect is visible when subjects evaluate natural human voices [10]. Past research in language-attitude studies based on natural voices also has shown that standard speakers are frequently rated higher for competence than non-standard speakers [3, 4], and unlike other Austrian dialects, the Viennese dialect is perceived as characteristic for a lower social class [5, 8]. Moosmüller [4] has shown that upper, middle and lower class informants in Vienna do not attach prestige to dialect usage. In her study, speakers of Viennese dialect were rated as not very intelligent, tolerant, kind-hearted, friendly, likeable or honest. Findings on dialect usage in Linz (Upper Austria), on the contrary, have shown higher social acceptance and appreciation [10].

3 Methods and Participants

Methods: In our study, we employ the most commonly applied method for speaker evaluation, a version of the matched guise technique [3], where different speakers are evaluated for different language varieties, e.g. [10]. First, text containing information about the statues in the *State Hall* was read to the participants by the facilitator. This was followed by three videos of guided tours, with the same text being synthesised in three different language varieties. Following each video, the test persons rated the three disembodied tourist-guides on a 5-point bipolar semantic differential containing 19 adjective pairs such as 'likeable - unlikeable', 'educated - uneducated', etc. (See Table 1 for the list of adjectives.) The adjective pairs and the rating dimensions of our scale reflect past research on language attitudes in various contexts, cf. [3, 10].

Additionally, the subjects responded to a set of open questions regarding their impressions concerning the personas behind the voices and their assessment in the cultural-historic context of the particular cultural heritage site.

Participants: The study was conducted during two interdisciplinary lectures at the University of Vienna. 54 Austrian and German students (50 female, 4 male) participated in the study, with an average age of 21.7 years. All participants have German mother tongue and 39 state that they use dialectal/colloquial varieties. We take evidence from [10] that the factor 'sex' has only a limited and quite predictable effect on the ratings in language-attitude studies. We will therefore treat sex as an independent variable in our analysis. Therefore, no conclusions will/can be drawn about differences in perception of male and female participants.

4 Data Analysis and Discussion of Results

Semantic Differential: We conducted an ANOVA comparing the speakers' mean scores for each of the adjectives in the semantic differential. With an exception of 'likeable', 'friendly' and 'not arrogant', the mean differences for the three agents (voices) was statistically very significant for the other 16 adjectives, at $p \leq 0.01$ (see Table 1, rightmost column). The results support the hypothesis that different synthetic voices elicit different social evaluations from the participants. Similar evidence for human voices is provided by [3, 10], and by [1, 6] for synthetic voices. Comparing the three voices, the male voice representing standard Austrian German (TG1) is evaluated as significantly more 'educated', 'trustworthy', 'polite', 'competent', 'serious' and 'refined', whereas the other male voice, representing a dialectal variety of Viennese (TG3), is perceived as most 'self-confident', 'natural', 'relaxed', 'open minded', with the highest 'sense of humor' and 'least strict' (see Fig. 1). These results are comparable to those in Soukup's study [10] on Upper Austrian dialect using natural voices.

Pairwise ANOVA[2] analyses show: The rating of the female voice representing a colloquial Viennese standard (TG2) is closer to the male Austrian standard voice (TG1), as for 11 of the 19 attributes in the semantic differential there is no significant difference between TG2 and TG1. TG1 scores significantly higher for 'educated', 'trustworthy', 'polite', 'competent', 'serious' and 'refined', whereas TG2 is rated significantly higher for 'sense of humor' and 'emotional'. TG1 and TG3 significantly differ in almost all attributes, except for 'likeable', 'friendly' and 'not arrogant'. Similarly, TG2 and TG3 significantly differ in all but four attributes: 'likeable', 'friendly', 'emotional' and 'not arrogant'.

In summary, the results highlight the importance of language variety for the social evaluation of virtual agents, and provide evidence for similar social

[2] We additionally conducted a Wilcoxon test to specifically account for the ordinal scale of our data and the dependencies between the evaluations. The results showed the same significant differences between the three voices, except for 'likeable' regarding TG1 and TG3, which is significant in the Wilcoxon test($p = 0.024$).

evaluation of language varieties irrespective of whether natural human voices or synthetic voices are used.

	Mean Values			ANOVAs p - Values			
Adjective item	TG 1	TG 2	TG 3	TG 1 & TG 2	TG 2 & TG 3	TG 1 & TG 3	TG 1 & TG 2 & TG 3
likeable	3.89	3.54	3.52	0.08441	0.93629	0.08611	0.15789
educated	4.11	3.61	2.5	0.00527	0.00000	0.00000	0.00000
trustworthy	4.17	3.72	3.56	0.02307	0.03817	0.00002	0.00010
polite	4.33	3.85	3.06	0.00404	0.00005	0.00000	0.00000
intelligent	3.78	3.57	2.74	0.28891	0.00002	0.00000	0.00000
friendly	3.78	3.94	3.59	0.39752	0.11445	0.37637	0.24457
self-confident	3.67	3.37	4.3	0.12414	0.00000	0.00090	0.00000
competent	4.17	3.39	2.93	0.00005	0.02613	0.00000	0.00000
natural	2.31	2.67	4.35	0.18783	0.00000	0.00000	0.00000
sense of humor	1.91	2.85	3.81	0.00002	0.00009	0.00000	0.00000
emotional	1.67	3.26	3.2	0.00000	0.80293	0.00000	0.00000
relaxed	2.54	2.65	4.3	0.63910	0.00000	0.00000	0.00000
serious	3.96	3.31	2.41	0.00156	0.00008	0.00000	0.00000
not aggressive	4.48	4.46	3.61	0.90383	0.00004	0.00002	0.00000
not strict	2.87	3.11	3.8	0.32470	0.01052	0.00006	0.00059
open-minded	2.37	2.28	2.94	0.65660	0.00569	0.01117	0.00605
gentle	3.54	3.89	2.52	0.05077	0.00000	0.00000	0.00000
not arrogant	3.33	3.37	3.74	0.87681	0.13712	0.06475	0.16234
refined	3.91	3.54	1.81	0.02884	0.00000	0.00000	0.00000

Table 1. The mean scores per adjective item and the p-values (grey indicates significance) rounded to five decimal places of the three speakers TG1 (male voice, standard Austrian German), TG2 (female voice, colloquial Viennese) and TG3 (male voice, Viennese dialect).

Persona Characteristics: With a number of open questions we aimed at assessing a) whether and which persona aspects are conveyed through the voices, and b) how the personas are evaluated in the specific cultural context of the *State Hall* scenario.

Persona aspects: To summarise, 24 of the participants assume that TG1 lives in a city, 7 of which believe is Vienna; for 13 participants, TG2 lives in a city, 6 of which say is Vienna; while for 16 TG3 lives in Vienna (no one states 'city') and for 22 he lives on the country side. These results show, on the one hand, a perceived connection between synthesised Viennese language variety and place of residence, with increasing percentage of mentions of Vienna, the more dialectal the voice appears. On the other hand, they show a perceived connection between dialect and rural origin which has been attested also in previous research on natural voices, e.g. [5]. As regards other factors, 13 participants expressed the opinion that TG1 works for broadcast media. In [10], the standard speaker was also believed to work in public media. Additionally, 16 participants believe TG1 is an academic. 28 agree that TG2 is an elderly, retired person, 8 refer to her as a grandmother. For TG3, 12 participants speculate that he likes to go to the pub. Thus we see a clear distinction between the personas. While for TG1 the characteristic attributions to a standard speaker are most prominent, it is age

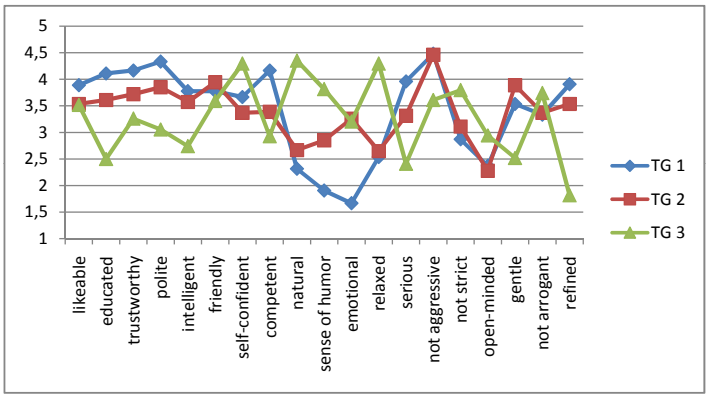

Fig. 1. Line diagram of mean scores for each speaker. 5 indicates 'very likeable', 'educated', 'trustworthy' etc.; 1 indicates 'unlikeable', 'uneducated', 'not trustworthy' etc.

for TG2 and regional attribution for TG3. In other words, not only language variety but also other vocal characteristics such as age are relevant for social interpretation and attribution.

Personas and cultural context: Referring to the question which of the tour-guides is the preferred one, 42 participants agree on TG1, only 8 subjects prefer TG2 and 5 TG3. Equally, 31 participants rate TG3 as least preferable as tour-guide because they find him difficult to understand and they consider dialect to be inappropriate in the particular cultural context. 19 do so for TG2 (difficult to understand), but only 4 for TG1. These results may reflect two issues: On the one hand, the voice of TG1 is better developed than the voices of TG2 and TG3. On the other hand, local or dialectal varieties tend to be less comprehensible than the standard variety. In [10], for instance, 70 out of 213 participants brought up the issue of comprehension in relation to the use of dialectal varieties. Regarding the questions 'How appropriate is the Viennese variety of TG3 for the *State Hall* application' and 'Which language variety would be best suited', 40 participants agree that TG3 is rather inappropriate to very inappropriate. 33 claim that standard Austrian German would be best suited.

5 Conclusion

Regarding the two synthetic male voices, we found similar results to Soukup's study on natural voices [10]. In both studies, the voice representing Austrian standard was evaluated as most 'educated' and 'refined', while the voice representing a dialectal variety was evaluated most 'natural', 'emotional', 'relaxed' and with the highest 'sense of humor'. Additionally, the analysis of the open questions covering persona aspects also shows a clear distinction between the three personas. In future work, we plan to put additional effort into assessing whether female and male listeners evaluate gendered voices of the same language variety similarly or differently (as evidenced in [10]). To cover this aspect, additional synthetic voices are required. Furthermore, the group of participants

needs to be well balanced between males and females. Important lessons from our results are that language variety together with voice quality elicit attributions of different personas and influence the attitudes of the listeners towards the speakers. The presented work is novel as it shows that this holds for synthesized speech, and thus is in accordance with previous findings in the context of natural speech. Moreover, our results provide evidence that the standard variety is favoured over local varieties in a cultural-heritage scenario as represented by the *State Hall* application.

Acknowledgments: This work has been conducted as part of the project "Companions für Userinnen" (C4U) funded by the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) under the research programme "FEMtech women in research and technology".

References

1. Crowell, C., Scheutz, M., Schermerhorn, P., Villano, M.: Gendered voice and robot entities: perceptions and reactions of male and female subjects. Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems. St. Louis, Missouri (2009)
2. Krenn, B., Sieber, G., Petschar, H.: Metadata Generation for Cultural Heritage: Creative Histories - The Josefsplatz Experience. Proceedings of EVA (Electronic Information, the Visual Arts and Beyond) 2006, Vienna, Austria , 27–34 (2006)
3. Lambert, W.: A Social Psychology of Bilingualism. Journal of Social Issues. 23(2), 91–109 (1967)
4. Moosmüller, S.: Dialekt ist nicht gleich Dialekt. Spracheinschätzung in Wien. Wiener Linguistische Gazette 40-41, 55–80 (1988)
5. Moosmüller, S.: Hochsprache und Dialekt in Österreich. Soziophonologische Untersuchungen zu ihrer Abgrenzung in Wien, Graz, Salzburg und Innsbruck. Wien, Köln, Weimar: Böhlau (1991)
6. Nass, C., Brave, S.: Wired for Speech. MIT Press, Cambridge, MA (2005)
7. Nomura, T., Kanda, T., Suzuki, T.: Experimental investigation into influence of negative attitudes toward robots on human-robot interaction. AI & Society, 20(2), 138–150 (2006)
8. Pucher, M., Neubarth, F., Strom, V., Moosmueller, S., Hofer, G., Kranzler, C., Schuchmann, G., Schabus, D.: Resources for speech synthesis of Viennese varieties. Proceedings of LREC, 2010, Malta, 105–108 (2010)
9. Schermerhorn, P., Scheutz, M., Crowell, C.: Robot social presence and gender: Do females view robots differently than males? In Proceedings of the Third ACM IEEE International Conference on Human-Robot Interaction, Amsterdam, NL, 263–270 (2008)
10. Soukup, B.: Dialect use as interaction strategy: A sociolinguistic study of contextualization, speech perception, and language attitudes in Austria. Wien: Braumüller (2009)
11. Quintanar, L., Crowell, C., Moskal, P.: The interactive computer as a social stimulus in human-computer interactions. In Salvendy, G., Sauter, S., Hurrell, J. (eds.), Social ergonomic and stress aspects of work with computers. Elsevier, Amsterdam, 303–310 (1987)