Perception of Synthetic Voices in Human-Agent Interaction

Sarah Warchhold
Daniel Duran
sarah.warchhold@germanistik.uni-freiburg.de
daniel.duran@germanistik.uni-freiburg.de
Deutsches Seminar – Germanistische Linguistik
Albert-Ludwigs-Universität
Freiburg, Germany

ABSTRACT

We present a preliminary study on auditory categorization of synthetic voices by naïve human listeners. Different re-synthesized speech stimuli are rated in a perception test. These include linguistic variations which are not commonly employed in spoken dialog systems. The results of the conducted experiments are important for further research and development on human-agent interaction which aims to offer a more natural verbal interaction.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Synthetic voices, human speech processing, individual differences

ACM Reference Format:
Sarah Warchhold and Daniel Duran. 2020. Perception of Synthetic Voices in Human-Agent Interaction. In *Proceedings of the 8th International Conference on Human-Agent Interaction (HAI '20), November 10–13, 2020, Virtual Event, NSW, Australia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3406499.3418756

1 INTRODUCTION

Speech-based human-agent interaction made huge progress by imitating human verbal interaction. One goal in the development of spoken dialog systems (or interactive voice response systems) is to provide a means of interaction which is as natural as possible. Thus the (explicit or implicit) model for verbal human-agent interaction is verbal interaction between humans using natural language.

Recent research on human language has shown, how individual differences in psychological (personality related) or cognitive features (related to mental processing capabilities) affect the perception, processing and production of speech. Speech and language sciences also increasingly focus on the factors which influence human-human verbal interaction with respect to dynamics, variability or situation-dependent accommodation. Additionally, human communication is affected by various non-linguistic factors. For example, studies have shown that speech perception in noisy conditions increases cognitive load and that cognitive load in turn affects production and perception. Also, it has been shown that irrelevant background noise impairs memory access, or that lexical neighborhood density affects word recognition in certain subjects [5, 11, 19, 28, 32, 40]. Speech signal processing still has unsolved technical problems in everyday situations like background noises, reverberation, crosstalk and other irrelevant sound sources. The verbal interaction itself is of particular relevance, as the development of spoken dialog systems aims for a more and more natural way of speaking.

The artificial generation of human-like speech has a long history [6, 7, 10, 15, 20, 24, 35] (see [18] for a review). It was assumed for a long time that listeners female voices are not suitable for synthesis [13]. Since human listeners *perceive* gender from acoustic cues in a voice, they automatically attribute gender to the owner of a voice. Even robots are subject to gender stereotyping [17, 38]. Early concepts of intelligent agents already envisaged spoken natural language as one important modality in human-computer interaction [14, 36]. A recent review on the state of the art in research on spoken dialog systems showed that most of the reviewed studies explored "usability" or "concepts and theories from research in human communication" [4]. The authors identify the development of theories on spoken human-agent interaction as one important research challenge.

Technology, especially in speech synthesis, increasingly achieves human-like performance. Hence, phenomena like the *uncanny valley* effect [23] may become an issue in human-agent interaction. If an agent uses human-like language with a human-like voice, the listener will attribute human-like cognitive abilities to the owner of that voice [21, 22]. Speech does not only convey a linguistic message encoding some propositional information. It does also convey sociolinguistic information about the social and regional background of the speaker, their attitudes, etc, as well as personal information about the identity of the speaker [16]. Listeners expect this dense stream of information, when they hear a human voice. Also, human listeners *perceive* personality traits (like social attractiveness, competence or trustworthiness), social status and background from the speech of their interlocutors [1, 29, 30]. The perception of speaker traits affects the behavior of listeners [27].

Humans accommodate to the ways of speaking of other humans by becoming more similar in their speech. This kind of accommodation reduces perceived distances and improves mutual understanding [12]. A speaker may, however, also chose to diverge linguistically from their interlocutor in order to increase the social distance between them. The interaction between multilingual

HAI '20, November 10–13, 2020, Virtual Event, NSW, Australia

ACM ISBN 978-1-4503-8054-6/20/11.
https://doi.org/10.1145/3406499.3418756

speakers or speakers of different dialects or accents is subject to complex processes of convergence and divergence. Although humans accommodate in interaction with artificial agents as well, acoustic differences between human-directed and device-directed speech were found [2, 3, 8, 26, 31]. These (potentially opposing) aspects need to be adressed critically.

Trustworthiness is one of those features that listeners attribute to the owner of a voice which is particularly important in the context of spoken language human-agent interaction. Apart from an agent's actual behavior, trust is affected by the voice of an agent, for example, its pitch or temporal organization [9, 34]. It has been shown that physical (visual) attractiveness of an embodied intelligent agent affects users' trust in the system [39]. Findings like these in conjunction with findings from work on vocal attractiveness suggest that the perceived characteristics of an agent's voice are important factors in verbal human-agent interaction.

Speech synthesis in systems like Alexa or Siri produce standard, i.e. non-dialectal, non-accented speech. However, most people around the world use several, situation-specific non-standard varieties in everyday communication. Furthermore, most of the world's population is bilingual with code-switching (i.e. the switching from one language to another within or across utterances) being an integral part or their everyday communication [37]. Only recently (in comparison to its long history) are regional or dialectal varieties considered in speech synthesis [e.g. 25, 33]. Hence, it is important to take non-standard synthetic speech in artificial agents into account.

In our ongoing research project we focus on the human side of speech-based human-agent interaction, taking psychological, cognitive and social factors into account. Our main research question in this study adresses the interaction between the perceived gender of the agent (artificial speaker) and the quality of synthetic speech. Additional questions are for example if there is an interaction between the linguistic distance between the agent and the human listener (user) and how user expectations towards an agent affect the perception of more natural (i.e. more colloquial, accented or dialectal) synthetic speech. As we earlier discussed, humans have several expectations from voices. Refering to those expectations, we hypothesize that linguistic variation. cause irritations. Probably those irritations will show up in prolonged reaction times (indicating higher cognitive load) or even in significantly different assigned categorizations. We present work in progress where these issues are addressed from a psycholinguistic perspective.

2 EXPERIMENTS

2.1 Participants and Material

German native speakers between 20 and 35 have been recruited as participants in a perception study. All participants can be regarded as *naïve listeners* as they are not trained in human-agent interaction, dialog systems or speech synthesis. Due to the present ubiquity of spoken dialog systems, it is hard to find participants without any experience with synthetic voices. We assess the amount of prior experience with spoken dialog systems with a questionnaire, filled out by the participants after the perception tests.

The speech stimuli which have been prepared for the perception tests were based on samples from natural speech recordings of human speakers. Applying re-synthesis, the samples have been manipulated and *degraded* in order to simulate properties of synthetic speech, such as spectral discontinuities, unnatural pitch contours, unnatural rhythmic patterns, wrong or missing para-linguistic signals like pauses, breathing noises etc. Speech from three model speakers was recorded: one adult female speaker, one adult male speaker, and one child speaker. All speech samples are short utterances between two and seven seconds long. All utterances for the first experiment were produced in Standard German. The speech stimuli were prepared with five levels of artificiality: from natural (i.e. the original recordings) to most artificial (i.e. the highest degree of manipulation). The combination of the three model speakers and the different manipulations results in different sets of speech stimuli, which we refer to as different *agents* – although they are just *voices*, in this preliminary study.

2.2 Method and Evaluation

A series of two consecutive experiments was conducted. In a first perception test, participants are presented with stimuli from all three speakers in randomized order. Speech samples are presented only auditorily without any visual cues of the identity of the agent. The task is to listen to each sample and rate their naturalness on a seven-point Likert scale (from 1="natural human voice" to 7="definitely synthetic computer voice"). Ratings as well as the reaction times (i.e. the time required to make a decision) are recorded.

In a second test, participants are presented with stimuli which are prepared in the same way, except for one additional variable: the speech samples are produced in either Standard German, a non-standard dialect of German or with a foreign accent. The latter two conditions contradict the participants' expectations, since synthetic speech is commonly realized in Standard German. The same rating procedure is applied, as in the first experiment.

We evaluate the listener rating on how natural the voice appears by fitting mixed-effects regression models with the ratings as dependent variable and reaction time, the speaker type (female, male, child), the linguistic variety, experience with Synthetic voices and the various acoustic manipulation parameters as fixed effects.

Note that the experiments were still running as of writing this paper. Thus, we cannot report on the results, yet.

3 CONCLUSION AND OUTLOOK

There is still a need to develop a theory of spoken human-agent interaction in contrast to – or along with – human speech communication and interaction. The results of the study are important for further research and development on human-agent interaction regardless of the outcome. Possible implications of the results offer a more natural verbal interaction. Furthermore, they contribute to the investigation of more superior research questions such as if a more natural and spontaneous speech improves human-machine interaction or if there is a limit from which a higher naturalness of the synthetic language no longer improves or even worsens the interaction.

ACKNOWLEDGMENTS

This work is supported by *Vector Stiftung*, https://vector-stiftung.de (grant to Daniel Duran).

REFERENCES

[1] Pascal Belin, Bibi Boehme, and Phil McAleer. 2017. The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLOS ONE* 12, 10 (Oct. 2017), e0185651. https://doi.org/10.1371/journal.pone.0185651

[2] Štefan Beňuš. 2014. Social Aspects of Entrainment in Spoken Interaction. *Cognitive Computation* 6, 4 (Dec. 2014), 802–813. https://doi.org/10.1007/s12559-014-9261-4

[3] Štefan Beňuš, Marian Trnka, Eduard Kuric, Lukáš Marták, Agustín Gravano, Julia Hirschberg, and Rivka Levitan. 2018. Prosodic entrainment and trust in human-computer interaction. In *9th International Conference on Speech Prosody 2018*. ISCA, Poznań, Poland, 220–224. https://doi.org/10.21437/SpeechProsody.2018-45

[4] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (June 2019), 349–371. https://doi.org/10.1093/iwc/iwz016

[5] Herbert A. Colle and Alan Welsh. 1976. Acoustic masking in primary memory. *Journal of Verbal Learning and Verbal Behavior* 15, 1 (Feb. 1976), 17–31. https://doi.org/10.1016/S0022-5371(76)90003-7

[6] Franklin S. Cooper. 1961. Speech Synthesizers. In *Proceedings of the Fourth International Congress of Phonetic Sciences*, Antti Sovijärvi and Pentti Aalto (Eds.). Mouton & Co., The Hague, 3–13.

[7] Franklin S. Cooper, Pierre C. Delattre, Alvin M. Liberman, John M. Borst, and Louis J. Gerstman. 1952. Some Experiments on the Perception of Synthetic Speech Sounds. *The Journal of the Acoustical Society of America* 24, 6 (Nov. 1952), 597–606. https://doi.org/10.1121/1.1906940

[8] Benjamin R. Cowan, Holly P. Branigan, Mateo Obregón, Enas Bugis, and Russell Beale. 2015. Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human–computer dialogue. *International Journal of Human-Computer Studies* 83 (Nov. 2015), 27–42. https://doi.org/10.1016/j.ijhcs.2015.05.008

[9] Aaron C. Elkins and Douglas C. Derrick. 2013. The Sound of Trust: Voice as a Measurement of Trust During Interactions with Embodied Conversational Agents. *Group Decision and Negotiation* 22, 5 (Sept. 2013), 897–913. https://doi.org/10.1007/s10726-012-9339-x

[10] Zsuzsanna Fagyal. 2001. Phonetics and speaking machines: On the mechanical simulation of human speech in the 17th century. *Historiographia Linguistica* 28, 3 (2001), 289–330.

[11] Lisa A. Farley, Ian Neath, David W. Allbritton, and Aimée M. Surprenant. 2007. Irrelevant speech effects and sequence learning. *Memory & Cognition* 35, 1 (01 Jan 2007), 156–165. https://doi.org/10.3758/BF03195951

[12] Howard Giles (Ed.). 2016. *Communication Accommodation Theory: Negotiating Personal Relationships and Social Identities Across Contexts*. Cambridge University Press, Cambridge.

[13] Inger Karlsson. 1991. Female voices in speech synthesis. *Journal of Phonetics* 19, 1 (Jan. 1991), 111–120. https://doi.org/10.1016/S0095-4470(19)30306-7

[14] Fakhreddine Karray, Milad Alemzadeh, Jamil Abou Saleh, and Mo Nours Arab. 2008. Human-Computer Interaction: Overview on State of the Art. *International Journal on Smart Sensing and Intelligent Systems* 1, 1 (2008), 137–159. https://doi.org/10.21307/ijssis-2017-283

[15] Jens-Peter Köster. 1973. *Historische Entwicklung von Syntheseapparaten zur Erzeugung statischer und vokalartiger Signale nebst Untersuchungen zur Synthese deutscher Vokale*. Number 4 in Hamburger Phonetische Beiträge: Untersuchungen zur Phonetik und Linguistik. Helmut Buske Verlag, Hamburg.

[16] Peter Ladefoged and D. E. Broadbent. 1957. Information Conveyed by Vowels. *Journal of the Acoustical Society of America* 29, 1 (Jan. 1957), 98–104. https://doi.org/10.1121/1.1908694

[17] Robin C. Ladwig and Evelyn C. Ferstl. 2018. What's in a name?: an online survey on gender stereotyping of humanoid social robots. In *Proceedings of the 4th Conference on Gender & IT - GenderIT '18*. ACM Press, Heilbronn, Germany, 67–69. https://doi.org/10.1145/3196839.3196851

[18] Tiina Männistö-Funk and Tanja Sihvonen. 2018. Voices from the Uncanny Valley: How Robots and Artificial Intelligences Talk Back to Us. *Digital Culture and Society* 4, 1 (2018), 45–64. https://doi.org/10.14361/dcs-2018-0105

[19] Sven L. Mattys and Lukas Wiget. 2011. Effects of cognitive load on speech recognition. *Journal of Memory and Language* 65, 2 (Aug. 2011), 145–160. https://doi.org/10.1016/j.jml.2011.04.004

[20] Werner Meyer-Eppler. 1949. *Elektrische Klangerzeugung*. Ferd. Dümmlers Verlag, Bonn.

[21] Roger K. Moore. 2012. A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Scientific Reports* 2, 864 (2012), Article number: 864. https://doi.org/10.1038/srep00864

[22] Roger K. Moore. 2017. Is Spoken Language All-or-Nothing? Implications for Future Speech-Based Human-Machine Interaction. In *Dialogues with Social Robots*, Kristiina Jokinen and Graham Wilcock (Eds.). Lecture Notes in Electrical Engineering, Vol. 427. Springer Singapore, Singapore, 281–291. https://doi.org/10.1007/978-981-10-2585-3_22

[23] Masahiro Mori. 2012. The Uncanny Valley. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100. https://doi.org/10.1109/MRA.2012.2192811 Number: 2 Translators: _:n1360.

[24] John J. Ohala. 2011. Christian Gottlieb Kratzenstein: Pioneer in Speech Synthesis. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS)*. International Phonetic Association, Hong Kong, 156–159. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/SpecialSession/Session7/Ohala/Ohala.pdf

[25] Michael Pucher, Dietmar Schabus, Junichi Yamagishi, Friedrich Neubarth, and Volker Strom. 2010. Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis. *Speech Communication* 52, 2 (Feb. 2010), 164–179. https://doi.org/10.1016/j.specom.2009.09.004

[26] Eran Raveh, Ingmar Steiner, Ingo Siegert, Iona Gessinger, and Bernd Möbius. 2019. Comparing phonetic changes in computer-directed and human-directed speech. In *Elektronische Sprachsignalverarbeitung: Conference proceedings of the 30st conference in Dresden (Studientexte zur Sprachkommunikation, 93)*, Peter Birkholz and Simon Stone (Eds.). TUDpress, Dresden, Germany, 42–49.

[27] Astrid M. Rosenthal-von der Pütten, Nicole C. Krämer, Stefan Maderwald, Matthias Brand, and Fabian Grabenhorst. 2019. Neural Mechanisms for Accepting and Rejecting Artificial Social Partners in the Uncanny Valley. *The Journal of Neuroscience* 39, 33 (Aug. 2019), 6555–6570. https://doi.org/10.1523/JNEUROSCI.2956-18.2019

[28] Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. 2014. The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*. ISCA Archive, Singapore, 427–431. http://www.isca-speech.org/archive/interspeech_2014/i14_0427.html

[29] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, and Benjamin Weiss. 2015. A Survey on perceived speaker traits: Personality, likability, pathology, and the first challenge. *Computer Speech & Language* 29, 1 (Jan. 2015), 100–131. https://doi.org/10.1016/j.csl.2014.08.003

[30] Antje Schweitzer, Natalie Lewandowski, and Daniel Duran. 2017. Social Attractiveness in Dialogs. In *Interspeech 2017*. ISCA, 2243–2247. https://doi.org/10.21437/Interspeech.2017-833

[31] Ingo Siegert and Julia Krüger. 2018. How do we speak with ALEXA: Subjective and objective assessments of changes in speaking style between HC and HH conversations. *Kognitive Systeme* 2018, 1 (2018). https://doi.org/10.17185/duepublico/48596 Publisher: DuEPublico: Duisburg-Essen Publications Online, University of Duisburg-Essen, Germany.

[32] Vanessa Taler, Geoffrey P. Aaron, Lauren G. Steinmetz, and David B. Pisoni. 2010. Lexical Neighborhood Density Effects on Spoken Word Recognition and Production in Healthy Aging. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 65B, 5 (Sept. 2010), 551–560. https://doi.org/10.1093/geronb/gbq039

[33] Markus Toman, Michael Pucher, Sylvia Moosmüller, and Dietmar Schabus. 2015. Unsupervised and phonologically controlled interpolation of Austrian German language varieties for speech synthesis. *Speech Communication* 72 (Sept. 2015), 176–193. https://doi.org/10.1016/j.specom.2015.06.005

[34] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. 2018. Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience. In *Proceedings of the Technology, Mind, and Society – TechMindSociety '18*. ACM Press, Washington, DC, USA, 1–6. https://doi.org/10.1145/3183654.3183691

[35] Jürgen Trouvain and Fabian Brackhane. 2011. Wolfgang von Kempelen's 'Speaking Machine' as an Instrument for Demonstration and Research. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS)*. International Phonetic Association, Hong Kong, 164–167. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/SpecialSession/Session7/Trouvain/Trouvain.pdf

[36] Wolfgang Wahlster (Ed.). 2006. *SmartKom: Foundations of Multimodal Dialogue Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/3-540-36678-4

[37] Li Wei (Ed.). 2000. *The bilingualism reader*. Routledge, London ; New York.

[38] Melanie Weirich and Adrian P. Simpson. 2018. Gender identity is indexed and perceived in speech. *PLOS ONE* 13, 12 (Dec. 2018), e0209226. https://doi.org/10.1371/journal.pone.0209226

[39] Beste F. Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or Beauty: How to Engender Trust in User-Agent Interactions. *ACM Transactions on Internet Technology* 17, 1 (March 2017), 1–20. https://doi.org/10.1145/2998572

[40] Adriana A. Zekveld and Sophia E. Kramer. 2014. Cognitive processing load across a wide range of listening conditions: Insights from pupillometry: Processing load across a wide range of listening conditions. *Psychophysiology* 51, 3 (March 2014), 277–284. https://doi.org/10.1111/psyp.12151