Presubmission enquiry - *Trends in Cognitive Sciences*

Authors: Christine Nussbaum[1,2], Sascha Frühholz[3,4] and Stefan R. Schweinberger[1,2,5]

[1]Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Germany

[2]Voice Research Unit, Friedrich Schiller University, Jena, Germany

[3]Department of Psychology, University of Oslo, Oslo, Norway

[4]Cognitive and Affective Neuroscience Unit, Department of Psychology, University of Zurich, Zurich, Switzerland.

[5]Swiss Center for Affective Sciences, University of Geneva, Switzerland

Understanding Voice Naturalness

When humans hear voices, they form intuitive impressions about the speaker/agent within just a few hundred milliseconds [1]. Perceived naturalness of a voice [2] is a prominent property within these impressions when we hear vocal sounds – and one which affects social interactions, both in a purely human context and in scenarios with human and artificial agents.

Despite its intuitive appeal, a systematic understanding of voice naturalness is elusive. Here we argue that this is due to four factors – (a) conceptual underspecification, (b) inconsistent operationalization, (c) lacking exchange between research on human and synthetic voices and (d) insufficient anchoring in voice perception theory. We show (using quantitative bibliometric analysis) that voice naturalness research is situated within different research domains that focus either on voice naturalness in health-related voice changes [3,4], or in synthesized, computer-generated voices [5-8]. Importantly, our analysis reveals that these domains resemble echo chambers within science, in that both neither cross-refer to one another nor to current voice perception theory (Figure 1). Accordingly, they use heterogeneous measurements of voice naturalness, with poor consideration of psychological theory.

In this article, we integrate current insights into voice naturalness by pooling evidence from a wider interdisciplinary literature. We then develop a concise definition of naturalness, and propose a conceptual framework which is rooted both in theoretical models and current empirical findings across fields. In short, we propose a taxonomy of naturalness with two distinct types - human-likeness-based naturalness and deviation-based naturalness – and clarify the demarcation of naturalness from established concepts of distinctiveness and authenticity. Subsequently, we identify core gaps in our current understanding of voice naturalness, and discuss a tentative roadmap for future research that includes practical recommendations.

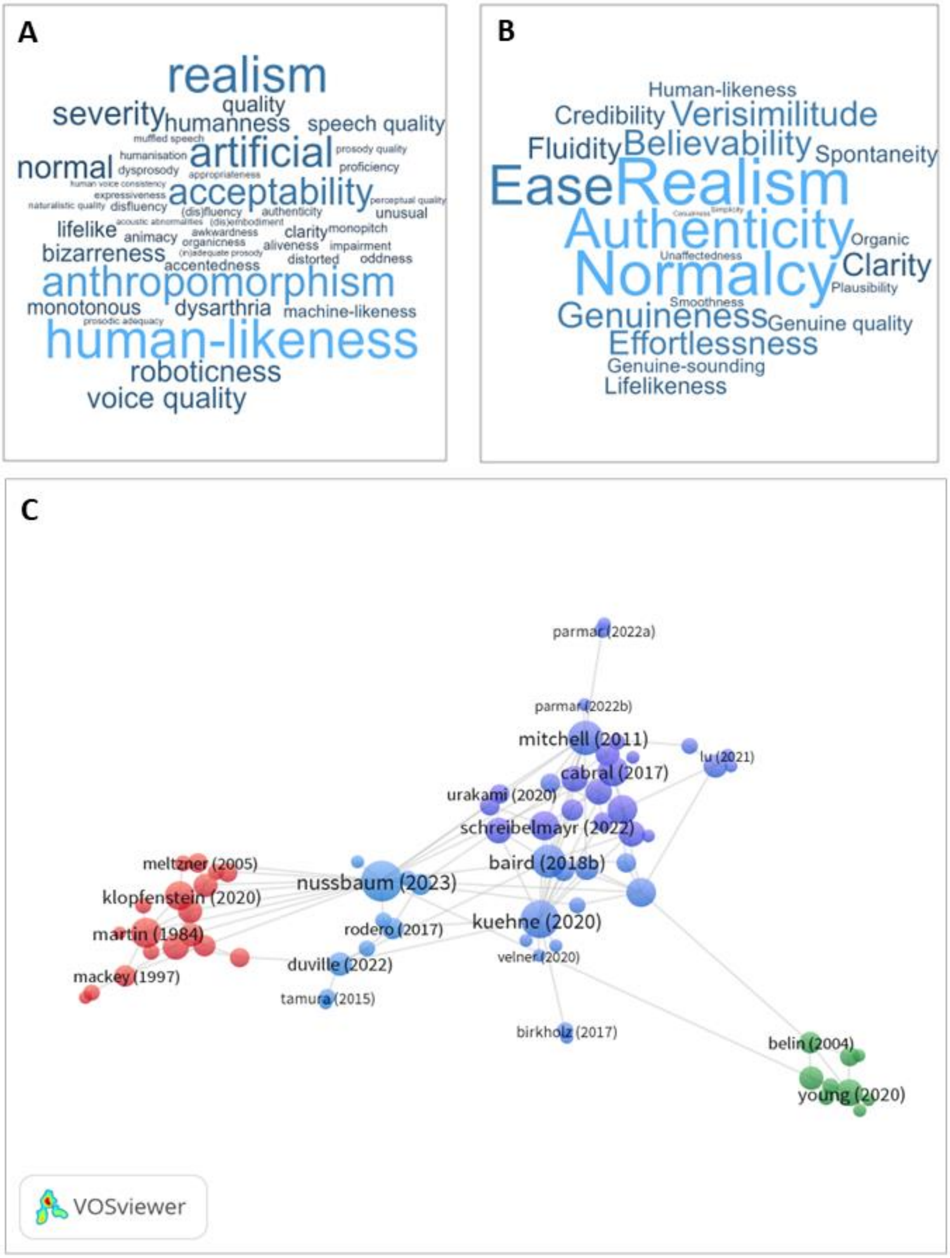
Our article has implications for a range of fields. To give examples, voice naturalness can be a priority in evaluating speech pathology in children [3] or desired gender attribution in adult transgender speakers [4], but naturalness of synthesized voices is also a powerful factor to promote positive (or uncanny) feelings [5] and user experience in audio-books [6]. Moreover, human-like voices may attract attention to service robots [7] and influence consumer decisions [8]. Emphasizing the importance of integrating clinical and technical approaches, perceived voice naturalness also is a key objective for cochlear implant research, when a sensory prosthesis restitutes hearing in people with sensorineural deafness by resynthesizing auditory signals for direct electrical stimulation of the cochlea [e.g., 9].

For reasons of space and stringency, our article will put a clear focus on developing the first conceptual framework for voice naturalness. Nevertheless, we will briefly consider multisensory interactions between voice and visual information in the context of perceived naturalness. Recent research demonstrated conditions in which synthesized faces can be perceived as more human than genuine human faces, and has begun to identify the visual features that can trigger such a paradoxical facial "hyperrealism" effect [10]. Although research regarding interactions between vocal and visual aspects of naturalness is in its infancy [7,8], we will briefly sketch future perspectives for identifying commonalities and differences between voice and face/body naturalness and their combinations. Our paper will also consider the acoustic characteristics of voices that influence both subjective perceptions and voice processing in the brain [11,12].

Regarding the accelerating spread of misinformation via social media [13], next-generation deepfakes likely will be even more efficient vehicles of misinformation, by instrumentalizing person-related trust via high-level perceptual deception. We expect that our conceptualization of voice naturalness will promote better understanding of characteristics of "successful" vocal deepfakes and their neuronal processing [14], and we believe this work will be of great interest for a wide readership of *TiCS*.

Figure 1

Terminology and interconnectivity of voice naturalness research



Note. ***A)*** *Word cloud depicting synonyms and closely related concepts from 66 publications that target naturalness in voices. Word size represents number of occurrences.* ***B)*** *A similar word cloud but generated by ChatGPT (https://chatgpt.com/?oai, 29.04.2024), when prompted to generate 10 synonyms each for pathological, synthetic/manipulated, and healthy voices, together with relative frequency of occurrence.* ***C)*** *A bibliographic network visualization using VOSviewer (https://www.vosviewer.com/), covering publications related to voice naturalness across different domains and 10 basic voice theory papers. Each colored dot represents a publication and grey links represent citations between papers. Size of dots indicates the number of links to other publications. Clustering (depicted by different dot colors) is performed automatically in VOSviewer, depending on citation patterns. Closer inspection reveals that green refers to basic voice theory papers, red corresponds predominantly to papers on pathological voices and blue refers to synthetized/manipulated voices. This paper addresses the emerging relevance of interconnectivity between those areas.*

References:

1. Lavan, N., Rinke, P., & Scharinger, M. (2024). The time course of person perception from voices in the brain. *Proceedings of the National Academy of Sciences of the United States of America, 121*(26), e2318361121. https://doi.org/10.1073/pnas.2318361121
2. Nussbaum, C., Pöhlmann, M., Kreysa, H., & Schweinberger, S. R. (2023). Perceived naturalness of emotional voice morphs. *Cognition and Emotion*, *37*(4), 731-747. https://doi.org/10.1080/02699931.2023.2200920
3. Schölderle, T., Haas, E., & Ziegler, W. (2023). Speech naturalness in the assessment of childhood dysarthria. *American Journal of Speech-Language Pathology*, *32*(4), 1633-1643. https://doi.org/10.1044/2023_AJSLP-23-00023
4. Hardy, T. L., Rieger, J. M., Wells, K., & Boliek, C. A. (2020). Acoustic predictors of gender attribution, masculinity–femininity, and vocal naturalness ratings amongst transgender and cisgender speakers. *Journal of Voice*, *34*(2), 300-e11. https://doi.org/10.1016/j.jvoice.2018.10.002
5. Kühne, K., Fischer, M. H., & Zhou, Y. (2020). The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in Neurorobotics*, *14*, 593732. https://doi.org/10.3389/fnbot.2020.593732
6. Rodero, E., & Lucas, I. (2023). Synthetic versus human voices in audiobooks: The human emotional intimacy effect. *New Media & Society, 25*(7), 1746-1764. https://doi.org/10.1177/14614448211024142
7. Li, M., Guo, F., Wang, X., Chen, J., & Ham, J. (2023). Effects of robot gaze and voice human-likeness on users' subjective perception, visual attention, and cerebral activity in voice conversations. *Computers in Human Behavior*, *141*, 107645. https://doi.org/10.1016/j.chb.2022.107645
8. Lu, L., Zhang, P., & Zhang, T. C. (2021). Leveraging "human-likeness" of robotic service at restaurants. *International Journal of Hospitality Management*, *94*, 102823. https://doi.org/10.1016/j.ijhm.2020.102823
9. Von Eiff, C.I., Frühholz, S., Korth, D., Guntinas-Lichius, O., & Schweinberger, S.R. (2022). Crossmodal benefits to vocal emotion perception in cochlear implant users. *iScience, 25,* 105711. https://doi.org/10.1016/j.isci.2022.105711
10. Miller, E. J., Steward, B., Witkower, Z., Sutherland, C. A. M., Krumhuber, E. G., & Dawel, A. (2023). AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychological Science*, *34*(12), 1390-1403. https://doi.org/10.1177/09567976231207095
11. Nussbaum, C., Schirmer, A., & Schweinberger, S.R. (2022). Contributions of Fundamental Frequency and Timbre to Vocal Emotion Perception and their Electrophysiological Correlates. *Social, Cognitive and Affective Neuroscience, 17*(12), 1145-1154. https://doi.org/10.1093/scan/nsac033
12. Staib, M., & Frühholz, S. (2021). Cortical voice processing is grounded in elementary sound analyses for vocalization relevant sound patterns. *Progress in Neurobiology*, *200*, 101982. https://doi.org/10.1016/j.pneurobio.2020.101982
13. Kauk, J., Kreysa, H., Scherag, A. & Schweinberger, S.R. (2024). The adaptive community-response (ACR) method for collecting misinformation on social media. *Journal of Big Data, 11:*35. https://doi.org/10.1186/s40537-024-00894-w
14. Roswandowitz, C., Kathiresan, T., Pellegrino, E., Dellwo, V., & Frühholz, S. (2024). Cortical-striatal brain network distinguishes deepfake from real speaker identity. *Communications Biology, 7*(1), 711. https://doi.org/10.1038/s42003-024-06372-6