**Understanding Voice Naturalness**

Christine Nussbaum, Sascha Frühholz, and Stefan R. Schweinberger

**Abstract** (120 words)

Perceived naturalness of a voice is a prominent property emerging from vocal sounds, which affects our interaction with both human and artificial agents. Despite its importance, a systematic understanding of voice naturalness is elusive. We argue this is due to (a) conceptual underspecification, (b) inconsistent operationalization, (c) lack of exchange between research on human and synthetic voices and (d) insufficient anchoring in voice perception theory. Here we reflect on current insights into voice naturalness by pooling evidence from a wider interdisciplinary literature. Against that backdrop, we develop a concise definition of naturalness and propose a conceptual framework rooted both in empirical findings and theoretical models. We identify gaps in current understanding of voice naturalness and sketch perspectives for empirical progress.

Word Limit: 3500 words

Inhalt

[1. Introduction – voice naturalness (450) 3](#_Toc160791725)

[2. Current Problems (800) 3](#_Toc160791726)

[2.1. Conceptual Underspecification (300) 3](#_Toc160791727)

[2.2. Inconsistent Operationalization (250) 4](#_Toc160791728)

[2.3. Lack of exchange between different research domains (150) 5](#_Toc160791729)

[2.4. Insufficient anchoring in voice perception theory (150) 5](#_Toc160791730)

[*3.* Proposition of a concise framework for voice naturalness (900*)* 6](#_Toc160791731)

[3.1. Definitions of naturalness (500) 6](#_Toc160791732)

[3.2. Differentiation from other concepts (400) 7](#_Toc160791733)

[4. Progressing in conjunction (400) 7](#_Toc160791734)

[5. Naturalness research rooted in voice perception theory (400) 8](#_Toc160791735)

[6. Open questions and future/outlook (400) 9](#_Toc160791736)

# Introduction – voice naturalness (450)

Human behavior is influenced by the perceived quality of objects and organisms that are encountered in our natural, social, and virtual environments. An important quality dimension concerns perceived “naturalness”. Assessing naturalness has an evolutionary meaning, as it influences interactions, food choice, and social trust (Quelle). Naturalness, from a biological perspective, can be understood as the adaptive norm, with extreme deviations supposedly being rather “unnatural” instances (Quelle). Besides the biological context, the recent emergence of AI-generated digital and virtual contexts has brought human-machine interactions to everyday life, and therefore questions of naturalness to the forefront of scientific research.

A domain where features of (un)naturalness are of particular importance is the voice, as one of the prime channels for communication for humans [1] and beyond: Synthetic voices increasingly emerge as major carriers of communicative interactions, such as in customer service calls, gaming environments, or support platforms [2,3]. When listening to voices, we form an instant impression about them [4]. Crucially, listeners seem to be very sensitive to (un-)natural voice features, which affects communicative quality. ToDo: Satz, wie eine unnatürliche Stimme klingen könnte. For **human voices**, consistent evidence from different speech-language pathologies shows that impairments in speech naturalness affect everyday interaction to a degree that can result in social isolation, reduced quality of life, and even depression [5,6]. Similarly, deliberate acoustic manipulations and distortions disrupt effective communication [7–10]. For **synthetic voices**, one can hardly keep up with the rapid developments which make indefatigable efforts to resemble human vocal expression [11,12]. However, as of today, synthetic voices are consistently rated as less natural than human voices, which simultaneously affects perceived likeability, trustworthiness, and pleasantness [13–17].

Given the widespread practical importance, it is crucial to put the role of voice naturalness into scientific focus. But although many recent studies provide useful empirical insights, we are currently looking at a rug rag rather than a research field. This has motivated us to take a step back and reflect on four problems in the present literature: (a) conceptual underspecification, (b) inconsistent operationalization, (c) lack of exchange between research domains and (d) insufficient anchoring in voice perception theory. We argue that these problems have so far precluded a systematic understanding of vocal naturalness, impeded the visibility to a wider readership, made us overlook crucial research questions, and led to a divergence between theory and practice. In what follows, we will elaborate on each of these problems, before proposing concrete measures to address them, starting with the development of a concise conceptual framework for voice naturalness. To this end, we aim to provide a useful basis for systematic and theory-driven research on voice naturalness in the future.

# Current Problems (800)

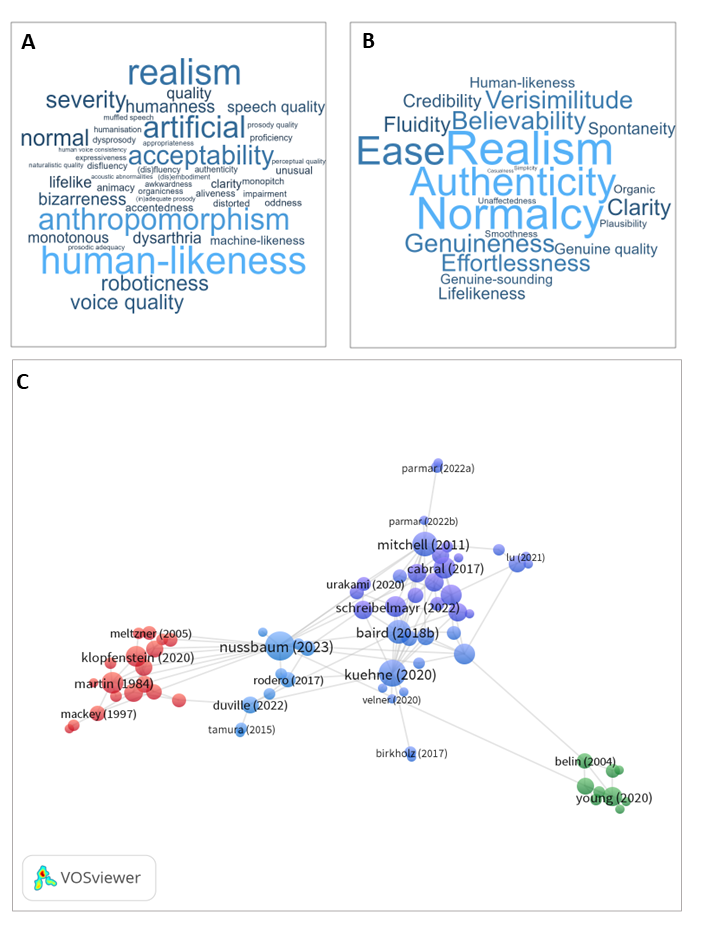
## Conceptual Underspecification (300)

Voice naturalness lacks a consistent definition and terminology in the literature (see Figure 1, A-B). In fact, the majority does not even provide an explicit definition of naturalness at all (see Box 1). In these studies, the conceptualization of naturalness can only be drawn implicitly from the empirical design. If definitions are provided, they vary tremendously across research contexts. In speech-language pathology, several researchers refer to the definition provided by Yorkston and collegues (1999): “*Naturalness is defined as conforming to the listener’s standards of rate, rhythm, intonation, and stress patterning and to the syntactic structure of the utterance being produced*” [6,18]. In contrast, research on synthetic and non-human voices usually defines naturalness as “*speech most closely perceived as a human voice*“ [19] or “*the degree to which a user feels a certain technology or system is human-like*” [20]. Accordingly, many studies using synthetic voices do not refer to naturalness but to human-likeness or anthropomorphism of voices.

Interestingly, these definitions seem to share two important assumptions: First, voice naturalness is a perceptual and subjective measure. Second, listener’s naturalness perception is the result of a complex multifactorial impression formation, presumably based on the integration and weighting of many acoustic cues [21]. Beyond that, however, the conceptualizations are very heterogeneous because they are tailored to the respective empirical focus. Unfortunately, despite covering relevant aspects, these prevailing inconsistencies alongside the heterogeneous terminology make it very challenging to compare and integrate different insights into voice naturalness. We therefore see a strong need to unite them under a concise conceptual framework, which we provide in Section 3.

**Figure 1**

Overview over 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 (for details, see Box1). Word size represents number of occurences.* ***B)*** *A similar word cloud but generated by ChatGPT (*[*https://chatgpt.com/?oai*](https://chatgpt.com/?oai)*, 29.04.2024), when prompted to generate 10 synonyms each for pathological, synthetic/manipulated, and healthy voices, together with relative occurrence frequency. The full prompt and the generated response is accessible on OSF.* ***C)*** *A bibliographic network visualization using VOSviewer (Quelle), 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. Size of the dots indicate the number of links to other publications. Clustering (depicted by different dot colors) is performed automatically in VOSviewer. Upon closer inspection, we inferred that the green ones are the basic voice papers, red correspond predominantly to publications on pathological voices and blue ones to synthetic/manipulated ones. A full documentation and an interactive version of the bibliographic network can be found on OSF.*

## Heteregeneous Operationalization (250)

A common consequence of inconsistent conceptualization is heterogeneous operationalization. Primarily, this concerns the studied vocal categories and features, which include human vs. synthetic voices [22–24]; cartoon voices [25], pathological voices such as in individuals with Parkinson’s’ disease [26,27], tracheoesophageal speech [28,29], dysarthria [30], or stuttering [31]; acoustically manipulated human voices [32], vocal fry [33], as well as different accents [34,35], dialects [36], age groups [37–39], and gender identities [40–42]. In addition, it concerns the experimental designs and measurements, especially rating scales, which differ in the number of levels and denominations of endpoints. In principle, such empirical heterogeneity can be a powerful source of insight. However, an insufficient report of empirical details impedes a meaningful integration of findings. Specifically, it is often not stated how naturalness and the related experimental task was explained to the listeners – but precise instructions can be crucial determinants of study outcome. Further, the precise acoustic properties of the voice material often remain elusive, bearing a risk for potential undetected confounds. Finally, few studies only provide measurements on reliability [43]. To address these issues, we collected some practical recommendations as a guidance for future research in Box 2. ToDo: vllt Beispielstudie/Design oder Instruktion für die Probanden

## Lack of exchange between different research domains (150)

Research on voice naturalness is inherently interdisciplinary, with two main domains: speech-language pathology and synthetic voices. However, while the scientific output is well-received within disciplines, they are remarkably poorly interconnected. Figure 1C illustrates this via a cross-citation analysis using VOSViewer [44], showing several distinct clusters of studies reminiscent of echo chambers which are frequently discussed in social media [45]. One may argue that this is not problematic, because the different disciplines simply have different interests and readerships. However, some intriguing commonalities and systematic patterns only emerge when pooling evidence from all available angles. For example, across synthetic, pathological and acoustically manipulated voices, converging evidence emerges for a strong effect of pitch variation on perceived naturalness [8,15,46]. Further, while several studies failed to find an uncanny valley effect for synthetic voices [13,47], a recent study suggest it might exist for pathological ones [48]. In fact, we argue that the lacking exchange between research fields has not only precluded relevant insights but has impeded the visibility and impact of voice naturalness research as a whole.

## Insufficient anchoring in voice perception theory (150)

The majority of naturalness research comes from applied fields, aiming to optimize artificial agents and improving the quality of life in patients with voice disorders. These findings equip us with valuable practical knowledge, but they are insufficiently anchored in voice perception theory. As an illustration, we added ten influential, theory-building voice perception publications to the VOSViewer analysis (Figure 1C), with the outcome suggesting that these tend to be ignored by most previous naturalness research. This leaves us with an intriguing divergence between increasing applied knowledge in rapidly developing branches (especially synthetic voices) on the one hand, but a simultaneous lack of understanding of basic mechanisms on the other. To fully understand how naturalness affects our perception and response to voices, this void needs to be filled.

# Toward a concise framework for voice naturalness (900*)*

After identifying key problems that impede a systematic understanding of naturalness in voices, we now propose concrete measures to address them, starting with a conceptual framework for the explicit definition of naturalness in voices.

## Definitions of naturalness (500)

We propose a taxonomy with two distinct types: Deviation-based naturalness and human-likeness-based naturalness **(**Fig 2 ToDo**)**. In **deviation-based naturalness**, naturalness is defined as the deviation from a reference that represents maximum naturalness. Example instructions for raters could be “Does this voice sound distorted?”, “Does this voice sound unusual”, or just “Does this voice sound natural?”. This conceptualization needs two important specifications: the **reference** representing maximum naturalness, and the **type of deviation**. In some cases, the reference is explicitly provided e.g. through a comparison or baseline stimulus (see [49]). However, in many studies, raters are instructed to use an inner implicit reference which is based on their experience and expectations, e.g. whether “*the voice stimulus is perceived as a plausible outcome of the human speech production system*“ [8]. The type of deviation is specified through the vocal material. It can virtually cover all acoustic features, ranging from specific manipulations (e.g. spectral features or speech rate [7,50,51]) to complex multivariate vocal patterns (e.g. in distorted or pathological voices [52]).

**Human-likeness-based naturalness** defines naturalness by its resemblance to a real human voice. An instruction for raters could be “Does this voice sound like a real human speaker?” or “How human-like does the voice sound to you?” Compared to the deviation-based definition, it comes with an important additional assumption: the existence of a non-human voice category, and hence a categorical boundary to human voices (although the transition between categories can be continuous). In other words, a definition of human-likeness is only meaningful if we assume that voices can be non-human in principle. Apart from this important distinction, human-likeness-based naturalness can be seen as a special case of deviation-based naturalness: the reference is a human voice (or listeners´ representation of a human voice),, and the deviation lies on the human/non-human spectrum.

With this taxonomy, we provide a flexible and intuitive reference for the explicit definition of naturalness alongside with its underlying assumptions. With future research committed to one conceptual framework, systematic integration and comparison of findings could be greatly facilitated. In fact, both conceptualizations seem already prevalent, but often remain implicit through certain design choices only (see Box 1). For example, comparing human to synthetic voices typically implies human-likeness based naturalness, whereas assessment of pathological voices often employs the deviation-based approach. One study deserves particular mention: Diel and Lewis [48] studied the uncanny valley effect in different types of unnatural voices. They found that impressions of uncanniness resulted from “deviation from familiar categories” rather “categorical ambiguity”. This could reflect initial empirical observations in line with our proposed conceptual distinction.

## Delimiting distinctiveness and authenticity (400)

In the following, we briefly discuss the demarcation of the proposed naturalness definitions from two established concepts in perception research, starting with distinctiveness. **Distinctiveness,** as opposed to typicality**,** has been defined as the degree to with faces or voices stick out due to rare or unusual features and this concept is commonly used to refer to voice identity (Quelle). According to face or voice space models, individual instances are represented along multiple perceptual dimensions, and they appear as distinctive if they deviate substantially from a central tendency or norm in that space [54]. Our deviation-based definition of naturalness is closely related to the concept of distinctiveness, as both share two critical features, a norm/reference and a deviation. However, we understand distinctiveness as a much broader term which captures many forms of perceptual deviations beyond naturalness. Thus, we assumed that unnatural voices would always be perceived as somewhat distinctive, but natural voice can b distinct or typical. Impressions of human-based naturalness, however, could potentially be quite independent from impressions of distinctiveness, e.g. a person who is very accustomed with a smart-speaker device may not rate synthetic voices as very distinctive but still clearly non-human. In that vein, the link between distinctiveness and naturalness may not primarily be a conceptual but an empirical matter.

The other concept that deserves particular consideration is **authenticity**. When prompted for synonyms of naturalness, this was ChatGPT´s first reply (**Figure 1 B**), which serves only to suggest a semantic link between these two terms in openly accessible online sources. In the scientific literature, however, authenticity is an established concept with a somewhat different meaning. Importantly, instead of the holistic impression of voices, it usually targets specific voice cues. Emotional authenticity, for example, refers to the distinction between a posed and a “real” or spontaneous emotional expression, which leads to differential behavioral and neural outcomes [55–57]. In the context of voice cloning and the now very prevalent danger of deepfakes [58], identity authenticity is assessed with regard to a specific speaker. Likewise, voice gender cues can be rated for gender authenticity, which is closely related to judgement of gender conformity [59,60]. In principle, it can be argued that authenticity is just a special form of deviation-based naturalness, with a more specific reference. E.g. “Does this sound like a natural voice?” is converted into “does this sound like a natural emotional expression?”. However, since these are two very different research questions, we tend to keeping the concepts of naturalness and authenticity rather separate.

# Progressing in conjunction (400)

In our view, understanding of voice naturalness requires pooling evidence from all relevant fields. Even when these may nurture different perspectives on voice naturalness, they are united by overarching questions: How do we form an impression on voice naturalness? Which acoustic features affect this impression? How does naturalness impact perception, interaction, and communication? Can we understand differences across individuals and listening contexts?

We hold that conceptual progress for disintegrated – but also highly interdisciplinary – naturalness research can be achieved by two steps: (a) converting, via an integrative perspective, empirical heterogeneity (Section 2.2) from an impediment into an advantage and (b) fostering mutually beneficial exchange between fields. Awareness for the interdisciplinary nature of the field is crucial for implementing both steps: First, publications need to be findable and accessible for others, preferably through the establishment of common terminology that converts into common keywords. Second, findings need to be communicated inclusively for readerships from diverse backgrounds. This entails providing explicit definitions, avoiding technical jargon, incorporating scientific standards from other fields where appropriate, and discuss own findings against a wider interdisciplinary naturalness literature. Finally, conceptual and empirical aspects need to be reported with sufficient detail to promote comparability. In Box 2, we converted these suggestions into practical recommendations.

We believe progress along these lines will not only enhance mutual inspiration between clinicians and engineers but could also foster innovative health technology. For instance, voice naturalness is a key objective for cochlear implant (CI) research, where a sensory prosthesis restitutes hearing in people with sensorineural deafness by resynthesizing auditory signals for direct electrical stimulation of the cochlea [e.g. 61], and real-time synthesis in CI sound processors could be modified to achieve better perceptual outcomes, ultimately benefiting of quality of life [62]. ToDo: add voice banking technology (Hyppa-Martin).

# Naturalness research rooted in voice perception theory (500)

Several authors have pointed out that research on naturalness is not sufficiently rooted in theory [6,11]. As discussed in section 2.4, the strongly applied orientation of the field comes at the expense of basic research, although several influential models on voice perception offer good staring points: The voice-space model proposed by Quelle represents voices in terms of their acoustic deviation from one another or a potential reference. The functional model by Belin et al. [63] assumes that an initial structural analysis of voices is followed by dissociable pathways processing vocal speech analysis, vocal affect analysis and voice recognition. Recently, Lavan and McGettigan [4] integrated these previous models in a unifying framework, explaining how listeners form multiple impressions about both familiar and unfamiliar voices. Commonly studied person characteristics include identity, gender, age, emotion and personality of speakers.

Although voice naturalness is in principle covered by these models, it is never explicitly mentioned. This is particularly intriguing against the backdrop of a questions that has prompted extensive debate and empirical efforts in basic voice research: Are voices special (Belin 2011)? In other words, do voices recruit network and resources in the brain that are not recruited by other types of acoustic stimuli? Voices with varying degrees of naturalness provide a powerful tool to shed new light on this debate. What makes human voices special? What makes natural voices special? In a nutshell, trying to understand the impact of naturalness on voice perception means trying to answer these questions.

This is not all. Rooting naturalness research in voice perception theory prompts further crucial questions that are not fully answered yet. First, to which degree is naturalness a threat to ecological validity (Nussbaum 2023)? Many voice researchers use acoustic manipulation such as voice morphing which could have unintended side effects on perceived naturalness. If this cannot be avoided, perceived naturalness should be at least quantified, and where possible be considered as a moderating variable. Second, how does naturalness interact with the processing of other voice characteristics? For example, first insights into the interplay of naturalness and emotionality suggest that […] (Quellen, shall I go into detail?). Third, [ToDo, brain data?, or “is naturalness always better than unnaturalness”, role of experience and learning history] Note that all of these questions are of relevance beyond the vocal modality. For faces, several of these aspects are covered in recent meta-analysis (Miller 2023)

(grade noch ein ziemlicher Flickenteppich…,this is one of the sections where I hope for substantial refinement from collaborators – Stefan:\_ok can do this, but for now, let´s wait for Sascha´s thoughts)

# Open questions and outlook (400)

Topics to cover:

- link to multimodal and visual research

- putting the conceptualizations to the test and compare whether they lead to different outcomes  
- systematic comparison of human-pathological, human-distorted, and synthetic voices  
- categorical perception between human- and non-human voices?  
- naturalness implications for ecological validity?  
- in naturalness always better?  
- individual differences  
- neurocognitive insights (aaaaaall kinds of brain data)

Box 1 (400 words): A field in numbers

For a more systematic overview on scientific insights into naturalness in voices, we conducted a literature search on Web of Science on 26 April 2023 using the search terms “naturalness AND voice” and “human-likeness AND voice”, which was repeated on 28 May 2024 to detect the most recent papers. This initial search resulted in 339 articles, to which we applied the following inclusion criteria: (1) Language of publication was English. (2) Papers were published in peer-reviewed journals or as a conference contribution. (3) Voice naturalness/human-likeness was either measures or manipulated. (4) Papers reported either a quantitative empirical analysis of human performance/perception data or a literature integration of such works. Thus, we excluded works on automatic naturalness classification and mere descriptions of toolboxes or datasets. (5) Finally, we focused on spoken utterances, excluding singing voices and non-linguistic vocalizations. Following these criteria, we also screened the reference lists of the identified articles for relevant publications. For a full documentation of the literature search process and all included papers, please refer to OSF.

In total, we identified 72 articles, covering a time range from 1984 to 2024. 38 (53%) were published in the last 5 years. 61 report behavioral empirical data, of which 47 are solely rating data. Three are literature reviews, and two used neurophysiological measures. Regarding voice category, 33 used synthetic, 18 human-pathological, 6 human-manipulated and 5 human-healthy voices. 10 used more than one of these voice categories. In only 33 papers, we could identify an explicit definition of naturalness. The full compilation of extracted definitions can be accessed on OSF. We noticed that the articles presented a large variability in wording and vocabulary. In an attempt to capture this verbal space, we scanned all articles for synonyms and closely related concepts of naturalness. The output is captured in the word cloud in Figure 1, A. Subsequently, we compared these to the articles’ keywords: 57 papers provided keywords, but only 13 had keywords related to naturalness or any of its synonyms. Finally, we coded the conceptualization of naturalness according to the taxonomy proposed in Section 3. In case no definition of naturalness was provided, we inferred the ‘implicit’ conceptualization from the research design. With this approach, we concluded that 27 employed a deviation-based conceptualization, 35 used human-likeness, and 10 used a combination of both.

Box 2 (400 words): Practical recommendations for voice naturalness research

Research on voice naturalness is a highly interdisciplinary field. To make future research accessible to a wide readership across disciplines, and allow comparability and integration of findings, a sensible awareness for this interdisciplinarity is crucial. Here, we compiled a number of practical recommendations as a tentative roadmap for future research:

* Offer a concise definition of voice naturalness to both readers and participants. With our taxonomy of naturalness in section 3, we offer a conceptual framework that can be tailored to any empirical design, e.g. by specifying the reference and the type of deviation. If used consistently in the literature, this taxonomy offers quick orientation for readers and forster comparability across findings.
* Use consistent keywords to make the research findable across disciplines. We recommend “naturalness”, “human-likeness” or, in cases discussed in section 3.2, “authenticity”.
* Include full reports on all methodological details, including acoustic manipulations, measurements (i.e. rating scales), instructions to readers and report on reliability.
* Wherever possible, provide stimulus examples. Often, the direct auditory impression is way more insightful than a list of acoustic measures and descriptions. In some cases, differences in audio material may offer a straight-forward explanation for different empirical outcomes.
* Finally, findings need to be communicated inclusively enough for readerships from very diverse backgrounds. This entails to provide some explicit definitions, avoid technical jargon, incorporate scientific standards from other fields where deemed fit, and discuss one’s findings against the backdrop of a wider interdisciplinary literature

Glossary:

* Synthetic/artificial voices
* Uncanny valley
* Anthropomorphism
* acoustic cues/features
* (operationalization)
* tracheoesophageal speech
* dysarthria
* pitch
* ChatGPT
* (ecological validity)
* deepfakes

References

1. Young, A.W.; Frühholz, S.; Schweinberger, S.R. Face and voice perception: Understanding commonalities and differences // Face and Voice Perception: Understanding Commonalities and Differences. *Trends Cogn Sci* **2020**, *24*, 398–410, doi:10.1016/j.tics.2020.02.001.

2. Rodero, E.; Lucas, I. Synthetic versus human voices in audiobooks: The human emotional intimacy effect. *New Media & Society* **2023**, *25*, 1746–1764, doi:10.1177/14614448211024142.

3. Rodero, E. Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices. *Computers in Human Behavior* **2017**, *77*, 336–346, doi:10.1016/j.chb.2017.08.044.

4. Lavan, N.; McGettigan, C. A model for person perception from familiar and unfamiliar voices. *Commun Psychol* **2023**, *1*, doi:10.1038/s44271-023-00001-4.

5. *The SAGE Encyclopedia of Human Communication Sciences and Disorders;* Damico, J.S.; Ball, M.J., Eds.; SAGE Publications, Inc: 2455 Teller Road, Thousand Oaks, California 91320, 2019, ISBN 9781483380834.

6. Klopfenstein, M.; Bernard, K.; Heyman, C. The study of speech naturalness in communication disorders: A systematic review of the literature. *Clin. Linguist. Phon.* **2020**, *34*, 327–338, doi:10.1080/02699206.2019.1652692.

7. Moore, B.C.J.; Tan, C.-T. Perceived naturalness of spectrally distorted speech and music. *J. Acoust. Soc. Am.* **2003**, *114*, 408–419, doi:10.1121/1.1577552.

8. Nussbaum, C.; Pöhlmann, M.; Kreysa, H.; Schweinberger, S.R. Perceived naturalness of emotional voice morphs. *Cogn. Emot.* **2023**, 1–17, doi:10.1080/02699931.2023.2200920.

9. Birkholz, P.; Drechsel, S. Effects of the piriform fossae, transvelar acoustic coupling, and laryngeal wall vibration on the naturalness of articulatory speech synthesis. *Speech Commun* **2021**, *132*, 96–105, doi:10.1016/j.specom.2021.06.002.

10. Birkholz, P.; Martin, L.; Xu, Y.; Scherbaum, S.; Neuschaefer-Rube, C. Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis. *Computer Speech & Language* **2017**, *41*, 116–127, doi:10.1016/j.csl.2016.06.004.

11. Seaborn, K.; Miyake, N.P.; Pennefather, P.; Otake-Matsuura, M. Voice in Human–Agent Interaction. *ACM Comput. Surv.* **2021**, *54*, 1–43, doi:10.1145/3386867.

12. Triantafyllopoulos, A.; Schuller, B.W.; \.Iymen, G.; Sezgin, M.; He, X.; Yang, Z.; Tzirakis, P.; Liu, S.; Mertes, S.; André, E.; et al. An overview of affective speech synthesis and conversion in the deep learning era. *Proceedings of the IEEE* **2023**.

13. Kühne, K.; Fischer, M.H.; Zhou, Y. The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. *Front. Neurorobot.* **2020**, *14*, 593732, doi:10.3389/fnbot.2020.593732.

14. Schreibelmayr, S.; Mara, M. Robot Voices in Daily Life: Vocal Human-Likeness and Application Context as Determinants of User Acceptance. *Front. Psychol.* **2022**, *13*, 787499, doi:10.3389/fpsyg.2022.787499.

15. Baird, A.; Parada-Cabaleiro, E.; Hantke, S.; Burkhardt, F.; Cummings, N.; Schüller, B. The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech. In *Interspeech 2018.* Interspeech 2018, 2-6 September 2018; ISCA: ISCA, 2018; pp 2863–2867.

16. Lee, E.-J. The more humanlike, the better? How speech type and users’ cognitive style affect social responses to computers. *Computers in Human Behavior* **2010**, *26*, 665–672, doi:10.1016/j.chb.2010.01.003.

17. Lu, L.; Zhang, P.; Zhang, T. Leveraging “human-likeness” of robotic service at restaurants. *International Journal of Hospitality Management* **2021**, *94*, 102823, doi:10.1016/j.ijhm.2020.102823.

18. Yorkston, K.M.; Beukelman, D.R.; Strand, E.A.; Hakel, M. *Management of motor speech disorders in children and adults;* Pro-ed Austin, TX, 1999.

19. Mawalim, C.O.; Galajit, K.; Karnjana, J.; Kidani, S.; Unoki, M. Speaker anonymization by modifying fundamental frequency and x-vector singular value. *Computer Speech & Language* **2022**, *73*, 101326, doi:10.1016/j.csl.2021.101326.

20. Hu, P.; Lu, Y.; Gong, Y. Dual humanness and trust in conversational AI: A person-centered approach. *Computers in Human Behavior* **2021**, *119*, 106727, doi:10.1016/j.chb.2021.106727.

21. Mayo, C.; Clark, R.A.J.; King, S. Listeners’ weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis. *Speech Commun* **2011**, *53*, 311–326, doi:10.1016/j.specom.2010.10.003.

22. Abdulrahman, A.; Richards, D. Is Natural Necessary? Human Voice versus Synthetic Voice for Intelligent Virtual Agents. *MTI* **2022**, *6*, 51, doi:10.3390/mti6070051.

23. Urakami, J.; Sutthithatip, S.; Moore, B.A. The Effect of Naturalness of Voice and Empathic Responses on Enjoyment, Attitudes and Motivation for Interacting with a Voice User Interface. In *Human-Computer Interaction. Multimodal and Natural Interaction*; Kurosu, M., Ed.; Springer International Publishing: Cham, 2020; pp 244–259, ISBN 978-3-030-49061-4.

24. Velner, E.; Boersma, P.P.; Graaf, M.M. de. Intonation in Robot Speech. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction.* HRI '20: ACM/IEEE International Conference on Human-Robot Interaction, Cambridge United Kingdom, 23 03 2020 26 03 2020; Belpaeme, T., Young, J., Gunes, H., Riek, L., Eds.; ACM: New York, NY, USA, 2020; pp 569–578, ISBN 9781450367462.

25. Ko, S.; Barnes, J.; Dong, J.; Park, C.H.; Howard, A.; Jeon, M. The Effects of Robot Voices and Appearances on Users’ Emotion Recognition and Subjective Perception. *Int. J. Human. Robot.* **2023**, *20*, doi:10.1142/S0219843623500019.

26. Abur, D.; Subaciute, A.; Daliri, A.; Lester-Smith, R.A.; Lupiani, A.A.; Cilento, D.; Enos, N.M.; Weerathunge, H.R.; Tardif, M.C.; Stepp, C.E. Feedback and Feedforward Auditory-Motor Processes for Voice and Articulation in Parkinson's Disease. *J Speech Lang Hear Res* **2021**, *64*, 4682–4694, doi:10.1044/2021\_JSLHR-21-00153.

27. Klopfenstein, M. Relationship between acoustic measures and speech naturalness ratings in Parkinson's disease: A within-speaker approach. *Clin. Linguist. Phon.* **2015**, *29*, 938–954, doi:10.3109/02699206.2015.1081293.

28. Eadie, T.L.; Doyle, P.C. Direct Magnitude Estimation and Interval Scaling of Naturalness and Severity in Tracheoesophageal (TE) Speakers. *J Speech Lang Hear Res* **2002**, *45*, 1088–1096, doi:10.1044/1092-4388(2002/087).

29. Eadie, T.L.; Doyle, P.C.; Hansen, K.; Beaudin, P.G. Influence of speaker gender on listener judgments of tracheoesophageal speech. *J. Voice* **2008**, *22*, 43–57, doi:10.1016/j.jvoice.2006.08.008.

30. Yorkston, K.M.; Hammen, V.L.; Beukelman, D.R.; Traynor, C.D. The effect of rate control on the intelligibility and naturalness of dysarthric speech. *J. Speech Hear. Disord.* **1990**, *55*, 550–560, doi:10.1044/jshd.5503.550.

31. Euler, H.A.; Merkel, A.; Hente, K.; Neef, N.; Wolff von Gudenberg, A.; Neumann, K. Speech restructuring group treatment for 6-to-9-year-old children who stutter: A therapeutic trial. *J. Commun. Disord.* **2021**, *89*, 106073, doi:10.1016/j.jcomdis.2020.106073.

32. Assmann, P.F.; Dembling, S.; Nearey, T.M. Effects of frequency shifts on perceived naturalness and gender information in speech. In *INTERSPEECH*, 2006.

33. Venkatraman, A.; Sivasankar, M.P. Continuous Vocal Fry Simulated in Laboratory Subjects: A Preliminary Report on Voice Production and Listener Ratings. *Am. J. Speech Lang. Pathol.* **2018**, *27*, 1539–1545, doi:10.1044/2018\_AJSLP-17-0212.

34. Kapolowicz, M.R.; Guest, D.R.; Montazeri, V.; Baese-Berk, M.M.; Assmann, P.F. Effects of Spectral Envelope and Fundamental Frequency Shifts on the Perception of Foreign-Accented Speech. *Lang. Speech* **2022**, *65*, 418–443, doi:10.1177/00238309211029679.

35. Tamagawa, R.; Watson, C.I.; Kuo, I.H.; MacDonald, B.A.; Broadbent, E. The Effects of Synthesized Voice Accents on User Perceptions of Robots. *Int J of Soc Robotics* **2011**, *3*, 253–262, doi:10.1007/s12369-011-0100-4.

36. Mackey, L.S.; Finn, P.; Ingham, R.J. Effect of speech dialect on speech naturalness ratings: a systematic replication of Martin, Haroldson, and Triden (1984). *J. Speech Lang. Hear. Res.* **1997**, *40*, 349–360, doi:10.1044/jslhr.4002.349.

37. Goy, H.; Kathleen Pichora-Fuller, M.; van Lieshout, P. Effects of age on speech and voice quality ratings. *J. Acoust. Soc. Am.* **2016**, *139*, 1648, doi:10.1121/1.4945094.

38. Coughlin-Woods, S.; Lehman, M.E.; Cooke, P.A. Ratings of speech naturalness of children ages 8-16 years. *Percept Motor Skill* **2005**, *100*, 295–304, doi:10.2466/pms.100.2.295-304.

39. Baird, A.; Jørgensen, S.H.; Parada-Cabaleiro, E.; Hantke, S.; Cummins, N.; Schuller, B. Perception of Paralinguistic Traits in Synthesized Voices. In *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences.* AM '17: Audio Mostly 2017, London United Kingdom, 23 08 2017 26 08 2017; Fazekas, G., Barthet, M., Stockman, T., Eds.; ACM: New York, NY, USA, 2017; pp 1–5, ISBN 9781450353731.

40. Hardy, T.L.D.; Rieger, J.M.; Wells, K.; Boliek, C.A. Acoustic Predictors of Gender Attribution, Masculinity-Femininity, and Vocal Naturalness Ratings Amongst Transgender and Cisgender Speakers. *J. Voice* **2020**, *34*, 300.e11-300.e26, doi:10.1016/j.jvoice.2018.10.002.

41. Merritt, B.; Bent, T. Perceptual Evaluation of Speech Naturalness in Speakers of Varying Gender Identities. *J Speech Lang Hear Res* **2020**, *63*, 2054–2069, doi:10.1044/2020\_JSLHR-19-00337.

42. Baird, A.; Jørgensen, S.H.; Parada-Cabaleiro, E.; Cummings, N.; Hantke, S.; Schüller, B. The Perception of Vocal Traits in Synthesized Voices: Age, Gender, and Human Likeness. *J. Audio Eng. Soc.* **2018**, *66*, 277–285, doi:10.17743/jaes.2018.0023.

43. Martin, R.R.; Haroldson, S.K.; Triden, K.A. Stuttering and speech naturalness. *J. Speech Hear. Disord.* **1984**, *49*, 53–58, doi:10.1044/jshd.4901.53.

44. van Eck, N.J.; Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, *84*, 523–538, doi:10.1007/s11192-009-0146-3.

45. van der Linden, S. *Foolproof*: *Why we fall for misinformation and how to build immunity;* 4th Estate: London, 2023, ISBN 9780008466718.

46. Anand, S.; Stepp, C.E. Listener Perception of Monopitch, Naturalness, and Intelligibility for Speakers With Parkinson's Disease. *J. Speech Lang. Hear. Res.* **2015**, *58*, 1134–1144, doi:10.1044/2015\_JSLHR-S-14-0243.

47. Romportl, J. Speech Synthesis and Uncanny Valley. In *Text, speech and dialogue.* International Conference on Text, Speech, and Dialogue; Horák, A., Sojka, P., Kopeček, I., Pala, K., Eds.; Springer International Publishing: Cham (Alemania), 2014; pp 595–602, ISBN 978-3-319-10816-2.

48. Diel, A.; Lewis, M. Deviation from typical organic voices best explains a vocal uncanny valley. *Computers in Human Behavior Reports* **2024**, *14*, 100430, doi:10.1016/j.chbr.2024.100430.

49. van Prooije, T.; Knuijt, S.; Oostveen, J.; Kapteijns, K.; Vogel, A.P.; van de Warrenburg, B. Perceptual and Acoustic Analysis of Speech in Spinocerebellar ataxia Type 1. *Cerebellum* **2023**, doi:10.1007/s12311-023-01513-9.

50. Rao M V, A.; Victory J, S.; Ghosh, P.K. Effect of source filter interaction on isolated vowel-consonant-vowel perception. *J. Acoust. Soc. Am.* **2018**, *144*, EL95, doi:10.1121/1.5049510.

51. Ratcliff, A.; Coughlin, S.; Lehman, M. Factors influencing ratings of speech naturalness in augmentative and alternative communication. *Augmentative and Alternative Communication* **2002**, *18*, 11–19, doi:10.1080/aac.18.1.11.19.

52. Meltzner, G.S.; Hillman, R.E. Impact of Aberrant Acoustic Properties on the Perception of Sound Quality in Electrolarynx Speech. *J Speech Lang Hear Res* **2005**, *48*, 766–779, doi:10.1044/1092-4388(2005/053).

53. Diel, A.; Lewis, M. *The vocal uncanny valley: Deviation from typical organic voices best explains uncanniness*, 2023.

54. Valentine, T. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q J Exp Psychol A* **1991**, *43*, 161–204, doi:10.1080/14640749108400966.

55. Lima, C.F.; Arriaga, P.; Anikin, A.; Pires, A.R.; Frade, S.; Neves, L.; Scott, S.K. Authentic and posed emotional vocalizations trigger distinct facial responses. *Cortex; a journal devoted to the study of the nervous system and behavior* **2021**, *141*, 280–292, doi:10.1016/j.cortex.2021.04.015.

56. Sarzedas, J.; Lima, C.F.; Roberto, M.S.; Scott, S.K.; Pinheiro, A.P.; Conde, T. Blindness influences emotional authenticity perception in voices: Behavioral and ERP evidence. *Cortex; a journal devoted to the study of the nervous system and behavior* **2024**, *172*, 254–270, doi:10.1016/j.cortex.2023.11.005.

57. Anikin, A.; Lima, C.F. Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *Q J Exp Psychol (Hove)* **2017**, *71*, 622–641, doi:10.1080/17470218.2016.1270976.

58. Roswandowitz, C.; Kathiresan, T.; Pellegrino, E.; Dellwo, V.; Frühholz, S. Cortical-striatal brain network distinguishes deepfake from real speaker identity. *Commun. Biol.* **2024**, *7*, 711, doi:10.1038/s42003-024-06372-6.

59. Kachel, S.; Steffens, M.C.; Preuß, S.; Simpson, A.P. Gender (Conformity) Matters: Cross-Dimensional and Cross-Modal Associations in Sexual Orientation Perception. *Journal of Language and Social Psychology* **2020**, *39*, 40–66, doi:10.1177/0261927X19883902.

60. Mills, M.; Stoneham, G.; Georgiadou, I. Expanding the evidence: Developments and innovations in clinical practice, training and competency within voice and communication therapy for trans and gender diverse people. *International Journal of Transgenderism* **2017**, *18*, 328–342, doi:10.1080/15532739.2017.1329049.

61. Eiff, C.I. von; Frühholz, S.; Korth, D.; Guntinas-Lichius, O.; Schweinberger, S.R. Crossmodal benefits to vocal emotion perception in cochlear implant users. *iScience* **2022**, *25*, 105711, doi:10.1016/j.isci.2022.105711.

62. Schweinberger, S.R.; Eiff, C.I. von. Enhancing socio-emotional communication and quality of life in young cochlear implant recipients: Perspectives from parameter-specific morphing and caricaturing. *Front. Neurosci.* **2022**, *16*, 956917, doi:10.3389/fnins.2022.956917.

63. Belin, P.; Fecteau, S.; Bedard, C. Thinking the voice: neural correlates of voice perception. *Trends Cogn Sci* **2004**, *8*, 129–135, doi:10.1016/j.tics.2004.01.008.