Trends in Cognitive Sciences

Understanding Voice Naturalness

--Manuscript Draft--

Manuscript Number:	TICS-D-24-00198R1
Article Type:	Review
Keywords:	Naturalness; Human-likeness; Voice perception; Authenticity; Voice synthesis
Corresponding Author:	Christine Nussbaum Friedrich Schiller University Jena Jena, GERMANY
First Author:	Christine Nussbaum
Order of Authors:	Christine Nussbaum
	Sascha Frühholz
	Stefan R. Schweinberger
Abstract:	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 suggest this is due to (a) conceptual underspecification, (b) heterogeneous 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.

Re: Your letter from September 11, 2024, per e-mail, regarding Manuscript ID TICS-D-24-00198 "Understanding Voice Naturalness”

Dear Lindsey,

we would like to thank you and the reviewers for your positive evaluation of the above manuscript, and for inviting us to submit a revised version of our Review Paper “Understanding Voice Naturalness”.

We included a response letter which gives details on how we addressed each individual point. We considered the feedback from you and the reviewers very helpful. In our opinion, the implementation of these suggestions improved our paper, by making it more accessible on the one hand, but also making it more specific to the topic of naturalness on the other hand.

Please note that one reviewer (Carolyn McGettigan; review signed) was concerned about the literature search and potential biases introduced by the search terms we chose. They made the valid suggestion that for a fully exhaustive literature search, one would need to incorporate more and different search terms. However, following your reminder that TiCS cannot report systematic reviews or results of unpublished systematic reviews, we kept our original literature search, but we made it more transparent to the reader, that this search is illustrative and not a fully systematic review. We nevertheless addressed the reviewers’ concern about potential biases by showing that other terms and keywords are sufficiently represented in our literature corpus. For our detailed response, please refer to points 9 and 11.

We hope you agree that the revised version is now suitable for publication in *TiCS* and we look forward to your response.

With kind regards,

Christine Nussbaum (on behalf of all authors).

Comments by the editor:

1. **First, they noted that some of the problems you identify and the solutions you offer are rather general and would apply to many research topics. To bring more specificity to the piece, please include a deeper discussion of the impacts of voice naturalness on listener perception and behavior. This will help increase the stakes of the piece. Additionally, please develop the future directions section to include examples specifically relevant to naturalness and understanding its impacts.**

Response: Thank you and the reviewers for pointing this out. To address the issue of specificity, we included a paragraph in the introduction, where we outlined the widespread impact of voice naturalness on listeners´ perception and behavior. This way, we now inform readers at the beginning of the review on the tremendous practical importance of voice naturalness research:

"*Importantly, variations in naturalness affect communicative quality [12,13]. Evidence from speech-language pathologies suggests that individuals with compromised speech naturalness are often perceived as withdrawn, cold, introverted or bored [14], potentially promoting social isolation and reduced quality of life [15–17] – even when speech intelligibility is preserved [18]. Accordingly, voice naturalness is a key target of speech therapy, across various voice alterations [18–20]. A recent survey on personalized speech synthesis for people who lost their biological voice further suggests that a majority prefers a more natural-sounding voice, even at the cost of some loss in intelligibility, both as users and listeners [21]. Thus, for human-to-human interaction, reduced voice naturalness consistently has negative implications. However, this is less clear for human-machine interaction (HMI). The Computers-Are-Social-Actors (CASA) framework proposed in the 1990s [22] assumed that we treat artificial agents like humans, fueling an (implicit) naturalness-is-better bias. In turn, this spurred efforts to create synthetic voices that resemble human vocal expression [23,24], even when the link between naturalness and success in HMI remains far from fully understood. While initial findings suggested that reduced naturalness in synthetic voices compromises likeability, trustworthiness, and pleasantness [11,25–28], contemporary synthetic voice design questions a "one size fits all" idea and instead advocates solutions tailored to specific applications [29]. Accordingly, maximum human-likeness of synthetic voices may not always be required or desirable. Instead, synthetic voice preferences may depend on the features of the listeners [27,30], the device [31–33], and its specific function [6,25,31]. Understanding and incorporating such preferences seems crucial for the success and acceptance of these devices [28].* " (page 3 and 4)

Further, addressing point 25 of the reviewers, we revised our recommendations in Box 2 and tailored them more to the specificities of voice naturalness research.

Finally, we revised the future directions section. For details, please refer to our responses of point 20 and 28.

2. **Second, the reviewers have indicated various points where you should provide more evidence that representations of voice naturalness differ from representations of other voice characteristics. If this evidence is not always available, this should be made clear and potentially highlighted as a future direction**.

Response: We added a paragraph in the future directions section that specifically addressed this point. For the specific changes, please refer to our responses to points 20 and 28.

3. **Third, Reviewer 1 has some questions about how the systematic search for articles was conducted and wonders whether different search terms might yield different results.**

Please note that TiCS articles cannot report the results of systematic reviews, nor can key points made in the piece rely on an unpublished systematic review. Given this, it will be important to find alternative ways to support the key premises of the piece, such as the lack of consistent definitions and the lack of exchange between different research domains. For example, to show the lack of consistent definitions, you could present representative examples of the definition-based, human-likeness-based, and combination definitions in a Table. I am open to including Figure 1C (with corresponding documentation included as a supplement), but you should provide additional evidence supporting the claim that there is a lack of exchange between the different research domains.

Response: We have addressed our intentions with the literature search in detail in response to points 9 and 11. Importantly, we clarified that we did not claim our literature work to be a fully systematic and exhaustive review of the literature (in line with the TiCS guidelines). Instead, because we make several claims about the field of naturalness research as a whole, we aimed to be transparent to readers about how we came to these conclusions. We have now outlined in Box 1 that this is an exemplary literature search. For the concerns about potential biases introduced by the search terms, please refer to point 11.

Following your suggestion, we have now included a Table with some definitions directly quoted from the original references (page 17 ff). The complete list of definitions is still available on OSF.

Finally, while we still think that Figure 1C provides the most compelling form of evidence for the lack of exchange between the different research domains, we have included an additional aspect:

"*In the case of voice naturalness, however, two recent systematic literature reviews on pathological [17] and synthetic voices [23] do not have a single reference in common.*" (page 6 ff)

4. **This is a Review (as opposed to an Opinion), so please avoid using highly opinionated language such as "we argue". "We discuss/suggest/show/etc." would be fine.**

Response: The rather opinionated language was originally intended, but in agreement with the guidelines for reviews, we have now applied the suggested changes throughout the manuscript (e.g. the abstract "argue" -> "suggest", page 4 "we argue" -> "our impression is", page 7 "argue" -> "we therefore conclude")

5. **Please do not number the different sections and subsections.**

Response: We removed the numbering of the sections (please note for the sake of readability, we did not mark this adjustment in the version with tracked changes).

6. **Many of the bolded terms do not appear in the Glossary. Please ensure that all bolded terms appear in the Glossary and that Glossary terms are in bold the first time they are used. Glossary terms should be listed in alphabetic order and should not include references. If a reference is needed, please include it when the Glossary term is first used.**

Response: Changed as suggested. (Please note that for the sake of readability, we did not mark this adjustment in the version with tracked changes.)

7. **Please include DOIs for any preprints, e.g. [101].**

Response: We adjusted the citation template to display DOIs for all references, if available. (Please note that for the sake of readability, we did not mark this adjustment in the version with tracked changes.)

Reviewer 1:

This is a very interesting paper on a timely topic, written by a team of authors that are exceedingly well placed to offer their expert perspective. Given the rapid rise in the sophistication of voice synthesis and its applications it's important that the various literatures concerned with the impact of human and synthetic voices strive for greater synthesis in their approach. This much-needed call to action paper offers suggestions for how to facilitate greater cross-disciplinary harmony, with the ultimate aim that research across fields can yield clearer and more applicable insights into how voices of different kinds might affect human behaviors. Although the paper focuses on voices, there are implications for perception of human vs "human-like" stimuli and entities across modalities and contexts. I have a few suggestions about ways to add depth and focus, which I think could greatly enhance the paper's impact.

Response: We were thankful for this positive evaluation, as well as the reviewer's impression of the timeliness of the topic. We appreciate your constructive feedback, which we address in detail below.

8. **The authors make a number of well-made observations about some of the insufficiencies of the existing literature on naturalness, including inconsistent terminology, missing/inconsistent definitions of terms for raters, lack of audio examples in published reports. However, these criticisms could be levelled at many topics of research in voice (and face) perception. For example, studies of voice/face/person trait perception often invoke low-dimensional social trait spaces to explain patterns of trait ratings. These dimensions may be conceptually equivalent or similar, but are labelled variably depending on the authors (e.g. Fiske describes these as "warmth and competence" while McAleer et al. uses "valence and dominance"; sociolinguists may use yet different approaches e.g. Bayard et al., 2001). Therefore, I think the current paper needs to make the specific case for *naturalness*. One way to address this would be to include more detailed motivation on the impacts of varying naturalness on listener perceptions and behavior. What are the implications of finding certain human or synthetic voices to be more or less natural-sounding? What is the published evidence that naturalness affects behaviors in different contexts, for example human-human communication vs. human-computer interaction? Can you use these examples to convince the reader of the importance/timeliness of studying naturalness, as a basis for some of the more specific methodological criticisms and suggestions?**

Response: These are several valid observations. In response, we decided to use the example of impression formation in voice and face perception, where we indeed observe competing two-factor models with different labels (e.g., warmth vs. competence, e.g., Fiske, 2018; or trustworthiness vs. dominance, e.g., Todorov et al., 2008). Crucially, however, unlike in the naturalness domain, there is substantial cross-talk in terms of cross-discussion and citations, as can be easily seen from several influential papers in that field. We included this line of argumentation in the manuscript as follows:

"Of course, poor interconnectivity is not unique to naturalness but affects many other research domains within voice or face perception. However, even when considering fields with highly divergent research traditions, such as impression formation from faces/voices for which two different two-factor models with different labels (e.g., warmth vs. competence, e.g. [70]; or trustworthiness vs. dominance, e.g. [71]) have been proposed, there is substantial research to link these distinct clusters and uncover both these specific taxonomies and their empirical relationships [72,73]. In the case of voice naturalness, however, two recent systematic literature reviews on pathological [17] and synthetic voices [23] do not have a single reference in common." (page 6/7)

Furthermore, we followed your suggestion and included a more detailed paragraph in the introduction where we outlined the impact of voice naturalness on perception and behavior, both in human-to-human and human-machine interaction (for details, see our response to point 1).

9. **Given the problems with terminology and cross-disciplinary awareness that are highlighted in the paper, it will be particularly important for the authors to make sure that they don't fall into the same trap of overlooking relevant research in other fields. The literature search yielded >300 papers, which is substantial, but based on only 2 search terms - if these are themselves informed by the authors' own preferred terminologies then the endeavour becomes circular. I wonder if it would be possible, for example, to use the ChatGPT analysis in a more task-driven way to generate more varied search terms for the literature search, rather than in the more illustrative way it is currently presented.**

Response: You raise an important issue here. In short, in point 3, the editor explicitly advised us not to base our key arguments on published or unpublished systematic reviews. Thus, we decided to keep the literature search in its current illustrative form.

In what follows, we would like to explain this decision in more detail: First, we would like to clarify our objectives and scope of the literature search. In this manuscript, we present several claims about current shortcomings in the literature (e.g., lacking interconnectivity, inconsistent conceptualization, etc.), and we identified a strong need to substantiate these claims with a more objective approach. We intended Box 1 as a transparent roadmap allowing to reconstruct how we reached our conclusions. For this reason, we provided the reader with rather detailed information on how papers were searched and selected, in a manner that's reminiscent of systematic literature reviews or meta-analyses. However, we never aimed or claimed this to be a fully and exhaustive literature integration. To make it more transparent to readers that this is NOT a fully systematic literature search, we slightly reworded the first sentence in Box 1 as follows: "*For a more systematic overview on scientific insights into naturalness in voices, we conducted a focused literature search on Web of Science*" (page 18)

Nonetheless, your very valid concern about potential biases introduced by the search terms itself needs to be addressed. Unfortunately, it is very likely that there is still insightful work on naturalness out there that we were not able to find. In fact, there are some references where we contacted the authors several times to get access to the material but without success. We further suspect that there exist many papers where naturalness is not the main focus, but rather a small side note, resulting them to slip through any search we could have conducted. Thus, the key question here is not if we missed out on papers concerning naturalness - we most likely did - but if we missed out on something of crucial relevance, i.e. something that would reveal a blind spot in our current view on the literature or that would call for a critical expansion of the conceptual framework we proposed. While this possibility can never be fully excluded, we consider it rather unlikely: We went back to our search history again and found that only 38 of the 72 papers were identified directly from the Web of

Science search. The rest was found in the reference lists of identified papers and was therefore not tied to the two search terms. Thus, the literature we cover goes well beyond the two search terms (see also our response to point 11).

10. **With the suggestions for future research: I would again like to see more targeted examples that are specifically relevant to naturalness and understanding its impacts. With the birdsong example it's not clear that environment-dependent changes in vocal behavior would be specifically related to naturalness rather than "typicality", or specifically related to voice naturalness rather than syntactic/structural deviations.**

Response: This is a valid point. We used the questions raised from the reviewers under point 20, 23 and 28 and incorporated them into a new paragraph in the future directions section, which targets the time-course of naturalness perception and the role of context for the impression formation. For the specific changes, please refer to our response to these respective points.

11. **Some smaller points: Figures 1A and 1B: The large number of terms presented in these two word clouds might strengthen the argument that searching the literature for papers only on "naturalness" and "human-likeness" is not sufficient to capture all the relevant research (see my point above).**

Response: Please refer to our response to point 9. For a systematic review or a meta-analysis, you would be completely right, and a search would have needed to be much broader to catch all available literature on voice naturalness. As discussed above, this was not our primary goal since we did not aim for a systematic review. Nevertheless, it would be a pity if a relevant line of research was not represented, because we neglected important keywords. Therefore, we ran additional Web of Science searches on some keywords and checked how many of our included papers would come up in the results (all done in October 2024): “realism AND voice” (5 papers), “anthropomorphism AND voice” (6), “artificial* AND voice” (10), “normal* AND voice” (3), “accept* AND voice” (9), “clarity AND voice” (3), “ease* AND voice” (3), and “quality AND voice” (19). Thus, although we may not have found all papers concerning naturalness, research from all these keywords is somewhat represented in our literature overview. The only exception is “authent* AND voice,” which picked up none of our naturalness papers. This is probably because we made an explicit effort to keep the concepts of naturalness and authenticity separate.

We included a comment on this matter in the supplemental material on OSF and referred to it in the manuscript as follows:

“For a full documentation of all included papers and a reflection on potential biases in the literature search, please refer to OSF” *(p. 18)*

12. **What was the specific purpose of generating the ChatGPT wordcloud, and was ChatGPT prompted specifically for synonyms of voice naturalness? Perhaps this approach would be more motivating as a way to generate e.g. the top 10 words as search terms for the literature search.**

Response: The purpose of the ChatGPT word cloud was to complement the one created based on the literature in order to compensate for potential “blind spots” in Figure 1A. Despite our best efforts to provide transparency and reproducibility, we manually extracted the terms from the literature, which makes them prone to biases. We conducted the ChatGPT analysis to complement this with a more objective approach. We were interested if this would return some crucial terms we had completely overlooked. And indeed, it revealed a strong association to authenticity, which

contributed to motivating us to include a clear conceptual demarcation to naturalness in the manuscript.

The specific prompts and ChatGPT's original response are all available on the associated OSF repository (https://osf.io/asfqv/?view_only=62f8d88705bb4363903983c8bd08a2cf). The prompts were (1) "List synonyms for naturalness in pathological voices. Assign each synonym a frequency between 1 and 0, depending on how often it is used.", (2) "Now do the same for synthetic and manipulated voices.", (3) "Now do the same for healthy human voices.", (4) "Now combine all three lists and omit any repetitions." This gave us an overview of the association between naturalness and other terms in the non-scientific online literature. However, one needs to be aware that ChatGPT is also biased in several ways. This is why we presented both Figure 1A and 1B. Neither of them is flawless, but they complement each other. We agree that the resulting terms could be used as a starting point for a systematic literature search in the future. For the present paper, we considered its illustrative function sufficient.

We included a comment on this matter in the supplemental material on OSF and referred to it in the manuscript as follows:

"*The full prompt, the generated response, and a reflection on its strengths and limitations are accessible on* OSF" (page 15)

13. **Use of idioms and journalistic style: There are a few instances that I would recommend re-wording for clarity and precision. For example: P3, lines 50-53: "For synthetic voices, one can hardly keep up with the rapid developments which make indefatigable efforts to resemble human vocal expression" - I would suggest toning this down, e.g. "For synthetic voices, recent years have seen rapid developments in the effort to create stimuli that resemble human vocal expression".**

Response: Valid point. We substantially revised this paragraph and changed the specific sentence to:

"*In turn, this spurred efforts to create synthetic voices that resemble human vocal expression [23,24]*" (page 4).

14. **P4, lines 2-5: "we are currently looking at a rag rug rather than a research field" - this idiom is not so familiar for English speakers.**

Response: Good point. We changed it into "*patchwork*" (page 4)

15. **P6, line 21: ""Does this voice sound unusual" - missing question mark.**

Response: Thank you. Changed as suggested (page 8).

16. **P8, line 24: "They found that impressions of uncanniness resulted from "deviation from familiar categories" rather "categorical ambiguity"." - should this be "rather than"?**

Response: Yes, thank you. Changed as suggested (page 9)

17. **P9, line 22: "very prevalent danger" - I would tone this down a bit, especially as some deepfaking may be intentional and agreed (e.g. an actor allowing their voice to be cloned to make a documentary).**

Response: We personally consider deepfakes mostly as a danger (despite possible intentional and harmless applications), but we are happy to use a more balanced wording here and changed it into "*very prevalent challenge*". (page 10)

18. **"Likewise, voice gender cues can be rated for gender authenticity, which is closely related to judgement of gender conformity [71,72]." - I would personally not get into discussions of "authenticity" when it comes to gender perception, because this is a complex and emotive issue that goes well beyond the aims of the current paper. I would only include this if you feel that this adds something to your specific argument at this point in the paper, beyond the other examples given.**

Response: We do see your point, but as you say, the matter is complex. You are right that gender perception and gender identity goes beyond the aims of the current paper, and is an emotive topic. But there is substantial and important research on both naturalness and authenticity perception that specifically targets questions of gender (e.g. see the work by Baird at al. 2018 on non-binary synthetic voices, https://doi.org/10.17743/jaes.2018.0023). Therefore, to provide readers with a broad overview of the field, this research deserves mentioning. Having said that, we completely see your call for caution, because superficial or premature statements may come across as offensive if they do not do justice to the complexity of the topic. We therefore discussed our wording thoroughly and rephrased the sentence as follows:

"*In principle, authenticity can be assessed with regard to manifold social signals, including age, gender, or even personality [71,72].*" (page 9)

19. **Other word choices could be changed for better clarity: "processual" could be "processing" or "process-related" and "restitutes" could be "restores".**

Response: Changed as suggested. (page 11 and 12)

Reviewer 2:

It was a pleasure to read this piece. The writing is excellent, and the review offers a great summary of the existing evidence on voice naturalness, highlighting the current issues in this field and proposing ways to advance it. This review is much needed. It will bring conceptual clarity to research on this topic, making it more theory-driven and unifying different subfields coherently. Additionally, it provides numerous suggestions for improving future research. I find myself in the rare position of not having many suggestions to offer. There is just one topic that it would be interesting to speculate about:

Response: We are very grateful for this positive evaluation, and we are particularly happy that you share our opinion about the importance of the topic.

20. **In terms of processing time-course and underlying brain mechanisms, how are representations of naturalness different from representations of other voice characteristics (e.g., age, gender, trustworthiness)?**

Response: You raise a valid question here, and we decided to address this in the section on future directions:

"Our theoretical considerations on the processing of voice naturalness call for investigations of its time-course and underlying brain mechanisms – relative to authenticity assessment but also to other voice characteristics. Initial evidence suggests that voice naturalness affects the brain response as early as 200 ms after voice onset and interacts with the processing of vocal emotions [99–101]. Comparably early effects have been found for authenticity assessments [86,102,103]. Although the interpretability of these findings is limited due to the potential influence of acoustic confounds, they suggest that naturalness and authenticity assessments both are fast and fundamental parts of voice perception. However, electrophysiological insights directly comparing the time-course of naturalness and authenticity are elusive, as is their interplay with impressions of age, gender, or personality traits. A recent EEG study suggests that many first impressions formed from voices are highly intercorrelated [8], but for naturalness we are currently limited to behavioral data that point towards interactions with age, gender, and emotion perception [60,63,74]." (p 13)

21. **Minor point: the second sentence ('From a biological perspective, naturalness… evolutionary meaning.') is somewhat vague. I suggest the authors revise it for specificity and clarity.**

Response: We refined it as follows: "*From a biological perspective, naturalness may relate to an adaptive norm, with extreme deviations supposedly being rather "unnatural" instances. Perceptions of naturalness influence food choice, environmental preferences, as well as social trust and therefore carry evolutionary meaning [1–3].*" (page 3)

Reviewer 3:

The present manuscript presents an overview of the current developments in research on voice naturalness. Here the authors outline some of the problems with the current work and suggest a distinction between two types of naturalness (deviation-based and human-likeness-based naturalness) in order to further enhance and unite research in this area while also referring back to prominent voice perception models. I agree with the authors that this topic is very current and timely given the fast-paced improvements in AI technology making it more and more difficult to tell apart human and computer-generated material be it face images or voices. I do, however, question the two distinct types of naturalness proposed by the authors in the way they have been defined and whether this distinction is easily resolved by contextual cues. Please find my detailed comments below.

Response: Thank you for your overall evaluation and for proposing very helpful and valid suggestions. Below, we will address each of your points in detail.

22. **Asking participants to judge whether a particular voice is perceived as a plausible outcome of the human speech production system implies that some voices will not be produced by human speakers. It is therefore not clear how this is an example of deviation-based naturalness. The authors themselves acknowledge that human-likeness-based naturalness could be viewed as a type of deviation-based naturalness (with human voice as a reference point). They argue that the only difference between these two types of naturalness is the**

additional assumption of the existence of non-human voices. However, the earlier formulation seems to suggest the existence of non-human voices, so it is not clear why it is deemed as a more appropriate example of deviation- rather than human-likeness-based naturalness.

Response: This is a valid point. Our example instruction for participants is not ideal here because it blurs the conceptual distinction we introduce here. We therefore refined the sentence as follows: "*However, in many studies, raters are instructed to use an inner implicit reference which is based on their experience and expectations, e.g., judge whether "it conforms to the expected standard of unimpaired speech*". (page 8)

Indeed, we re-evaluated the definition "*By naturalness, we understand the voice stimulus to be perceived as a plausible outcome of the human speech production system*" (Nussbaum 2023), and we agree with you that it is rather a combination of the two conceptualizations we introduce here. We therefore recoded it as such and adjusted the respective information in Box 1 and the associated OSF repository.

Note that we followed the advice of the editor in point 3 and added Table 1 with some representative definitions, which hopefully helps readers to get a better understanding of our conceptual distinction as well.

23. **What is more, based on the authors' definitions, the distinction between the two types of naturalness could easily be resolved when everyday context is considered. For example, speaking with someone in person could address deviation-based naturalness only, whereas any online interaction leaves open the opportunity for voices to be non-human and therefore engages human-likeness-based naturalness. Perhaps this distinction is meant to be more helpful for experimental design rather than our everyday experiences of evaluating voice naturalness.**

Response: Here, we partly agree. The consideration of the interactional context is important of course and we agree that in fully human-to-human interactions, human-likeness definitions make little sense. But even the distinction between human-human and human-machine interactions could become smaller with the increasing adoption of technological innovations in daily life. A prominent example is the use of personalized synthesis for people who lost their original voice (see Hyppa-Martin 2024).

However, we conceptualize naturalness as a fully perceptual construct. Thus, the key point is not whether a voice is truly human or not (which is not an easy distinction anyway – is a recorded voice that went through some filtering, e.g. as in many normal telephone calls, or other manipulations still human or not?), or whether it deviates objectively on an acoustic spectrum. The key point is the *listeners' impression of this voice*. Hence, it can make sense to assess naturalness in a deviation-based manner even in fully synthesized voices, for instance in the case of the above-mentioned personalized voice synthesis approaches. Nevertheless, we agree that the context is always considered when forming these impressions. We therefore included the following:

"In a broad sense, naturalness impressions are always formed against a specific context, whether that context refers to the voice itself or the properties of the interaction. Accordingly, if the same voice is assessed in an all-human or HMI context could make a crucial difference." (page 13)

Finally, you are absolutely right that although our framework is primarily based on theoretical considerations, it was also one of our main concerns that it is of practical use, in the sense that it can

be easily implemented in experimental designs. This is because we believe that more systematic and conceptually well-defined research in this area is crucial to approach an understanding of everyday experiences of evaluating voice naturalness.

24. **At the beginning of their review, the authors refer to the first impressions literature but there is no empirical evidence provided that shows human listeners spontaneously form impressions of naturalness or how naturalness could affect the impressions we form. There are a couple of recent papers where listeners were asked to freely describe their first impressions from voices (Lavan, 2023 with a Western sample and Jiang et al., 2023 with a Chinese sample). Neither of these papers seems to mention evaluations of naturalness. Some additional references to previous first impressions work that evaluates naturalness are needed to support this point.**

Response: This is a good point. We added the two references that you mentioned here: *"When we hear voices, we form intuitive impressions about them within just a few hundred milliseconds [8–10]."* (page 3) Indeed, there is very little work in the first impressions literature on evaluations of naturalness. However, Kühne et al. 2020 included a qualitative analysis of participants' free descriptions on how they formed the naturalness impressions on the voices they heard. We added this reference now here: *"Unnatural voices may sound nasal or robotic, or may differ from the norm in pitch contour, temporal structure, or spectral composition; in short, there are many ways in which a voice can lack naturalness [11]."* (page 3)

25. **While generally helpful, the proposed practical recommendations for voice naturalness research could be applied to any other field and do not specifically target research in voice naturalness. Providing sufficient methodological details would improve work in any area.**

Response: Indeed, you are right that a subset of our recommendations sounds very general and may seem rather obvious. Still, they are not met by so many publications. There may be many reasons for that, one being that what seems obvious in one field may not be in another one. Thus, a rather simple and general "checklist" could still prove as very beneficial. Nevertheless, we reformulated some of our recommendations listed in Box 2 to make them more specific, and we added some examples that target naturalness research.

26. **In their discussion of the key problems in voice naturalness research, the authors mention the use of different rating scales to assess naturalness. Is there any evidence showing that using these scales leads to significantly different patterns of results? A recent paper by Kramer et al. (2024) compares the use of different types of rating scales for the evaluation of face attractiveness and they find very little evidence that scale use makes a considerable difference to the overall results reported.**

Response: Thank you for recommending this interesting work. It is reassuring to any experimental researcher that methodological subtleties such as the properties of a rating scale do not impact the results to a large degree. However, our concern about the inconsistencies in rating measures does not primarily target the number of levels or whether it's an analog scale or not. Instead, we saw a large variability in the denomination of endpoints (if they were reported). In some studies, responses ranged from "natural" to "unnatural", in others from "humanlike" to "robotlike", or from "natural" to "awkward". We made a small adjustment to the sentence:

"*For example, in one study participants were asked "How natural is the audio?" from "1 – natural" to "5 – unnatural" [65] , in another one they rated voices on a 10-point-scale from "very natural, human-like" to "very mechanical, robot-like" [58], or made a binary classification of voices as either human or computer-generated [37].*" (page 6)

Further, while the data presented by Kramer et al. (2024) show a robust pattern for attractiveness in faces this does not necessarily have to generalize to naturalness/human-likeness in voices. Diel et al. (2024) collected ratings on human-likeness on both a slider from 1-100 and later a binary classification as either human-like or not (to extract a measure of categorization certainty). From visual impression, some pathological voices were rated high on human-likeness (Figure 3), but in the binary response, participants showed considerable categorization uncertainty (Figure 6), meaning they struggled with the binary decision. However, there is no direct comparison of these two measures, which is why we didn't present it as hard evidence in the manuscript. Instead, we put it as follows:

"*There is recent evidence from face perception that differences in rating scales may not have a big impact on outcome [66], although we cannot conclude that this generalizes to naturalness ratings, and the insufficient report of empirical details impedes a meaningful comparison of findings.*" (page 6)

27. **The authors propose that human-based naturalness could be independent from distinctiveness - is there any empirical evidence to support this point?**

Response: To the best of our knowledge there is no evidence concerning the link between distinctiveness and naturalness yet. Thus, our elaborations on distinctiveness should be regarded as speculations, calling for empirical verification (or falsification) in the future. To highlight this, we rephrased some sentences as follows:

"*However, one may speculate that impressions of human-based naturalness could be quite independent from impressions of distinctiveness under certain conditions.*" (page 9)

And

"*In that vein, the link between distinctiveness and naturalness may not primarily be a conceptual but an empirical matter, requiring future inspection.*" (page 9)

28. **It is argued within the text that voice naturalness and authenticity are processed in different stages with naturalness based on voice properties or features whereas authenticity based on speech or social/affective analysis. Is it not possible to evaluate naturalness of speech content - e.g., how likely is it that this speech content is produced by a human speaker? Relatedly, this implies that naturalness is assessed faster than authenticity - is there any evidence to suggest this in the literature?**

Response: Indeed, the idea that naturalness and authenticity are linked to different processing stages is a prediction that can be derived from the model we propose. It may even imply that naturalness is assessed faster than authenticity- a hypothesis that must be put to the test in the future. However, they can also be assessed to some degree in parallel, which is why we refrained from very strong predictions in the manuscript at this stage. Instead, we included the following paragraph in the future directions section:

"Our theoretical considerations on the processing of voice naturalness call for investigations of its time-course and underlying brain mechanisms – relative to authenticity assessment but also to other voice characteristics. Initial evidence suggests that voice naturalness affects the brain response as

early as 200 ms after voice onset and interacts with the processing of vocal emotions [99–101]. Comparably early effects have been found for authenticity assessments [86,102,103]. Although the interpretability of these findings is limited due to the potential influence of acoustic confounds, they suggest that naturalness and authenticity assessments both are fast and fundamental parts of voice perception. However, electrophysiological insights directly comparing the time-course of naturalness and authenticity are elusive, as is their interplay with impressions of age, gender, or personality traits. A recent EEG study suggests that many first impressions formed from voices are highly intercorrelated [8], but for naturalness we are currently limited to behavioral data that point towards interactions with age, gender, and emotion perception [60,63,74]." (p 13)

Concerning your point that one might also evaluate naturalness or human-likeness of speech, we fully agree. There are anecdotes of smart devices using unusual word constructions, which makes the speech sound odd to the listener. One way to empirically test this would be to present participants with transcripts of synthesized and human speech and ask for their judgment of human likeness. However, a number of considerations made us decide to not include a discussion of this issue in the paper. The most important one is that our review is concerned with auditory voice processing irrespective of speech content, such that this issue seems to go beyond the scope of the current paper – especially when considering space limits. We hope this is deemed acceptable – but obviously would be happy to include a sentence or two on the topic if advised so by the editor.

Highlights:

- Voices elicit impressions about their naturalness, which affect interactions between humans as well as with artificial agents
- Despite its intuitive appeal and practical importance, a systematic understanding of voice naturalness is elusive – the concept is scientifically ill-defined
- We show that current voice naturalness research is situated within different research domains that resemble echo chambers within science – they neither cross-refer to one another nor to current voice perception theory
- We offer a concise conceptual framework by proposing a taxonomy with two distinct types: deviation-based naturalness and human-likeness-based naturalness
- We develop practical recommendations and perspectives for naturalness research. We argue that, in a world of digital agents, understanding the determinants for how humans perceive naturalness in social stimuli is a priority

Understanding Voice Naturalness

Christine Nussbaum[1,2,6], Sascha Frühholz[3,4,6], and Stefan R. Schweinberger[1,2,5,6,7]

[1]Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, 07743 Jena, Germany

[2]Voice Research Unit, Friedrich Schiller University, 07743 Jena, Germany

[3]Department of Psychology, University of Oslo, 0371 Oslo, Norway

[4]Cognitive and Affective Neuroscience Unit, University of Zurich, 8050 Zurich, Switzerland

[5]Swiss Center for Affective Sciences, University of Geneva, 1222 Geneva, Switzerland

[6]The Voice Communication Sciences (VoCS) MSCA Doctoral Network

[7]German Center for Mental Health (DZPG), Site Jena-Halle-Magdeburg, Germany

Correspondence should be addressed to Christine Nussbaum (https://www.allgpsy.uni-jena.de/christine-nussbaum/), Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Am Steiger 3/1, 07743 Jena, Germany. Tel: +49 (0) 3641 945934, E-Mail: christine.nussbaum@uni-jena.de. Supplemental materials to this work are accessible on the associated OSF-repository: https://osf.io/asfqv/?view_only=62f8d88705bb4363903983c8bd08a2cf

Abstract

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 suggest this is due to (a) conceptual underspecification, (b) heterogeneous 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.

Keywords: Naturalness, Human-likeness, Voice perception, Authenticity, Voice synthesis

Naturalness – a prominent aspect of voice perception

Naturalness is a prominent aspect of perception when we see, hear, smell, taste, or feel our environment. From a biological perspective, naturalness may relate to an adaptive norm, with extreme deviations supposedly being rather "unnatural" instances. Perceptions of naturalness influence food choice, environmental preferences, as well as social trust and therefore carry evolutionary meaning [1–3]. Beyond the biological context, the recent emergence of AI-generated digital and virtual contexts has brought human-machine interactions to everyday life, and therefore has brought questions of naturalness to the forefront of scientific research. One of the prime channels for communicative interactions is the voice [4], both in a purely human context and beyond – with current **voice synthesis** technology quickly invading everyday life, both in good use (e.g., in customer service calls, public transport, gaming, or support platforms [5,6]) and abuse (e.g., **deepfakes** [7]).

When we hear voices, we form intuitive impressions about them within just a few hundred milliseconds [8–10]. Crucially, listeners seem to be very sensitive to impressions of voice (un-)naturalness. Unnatural voices may sound nasal or robotic, or may differ from the norm in pitch contour, temporal structure, or spectral composition; in short, there are many ways in which a voice can lack naturalness [11]. Importantly, variations in naturalness affect communicative quality [12,13]. Evidence from speech-language pathologies suggests that individuals with compromised speech naturalness are often perceived as withdrawn, cold, introverted or bored [14], potentially promoting social isolation and reduced quality of life [15–17] – even when speech intelligibility is preserved [18]. Accordingly, voice naturalness is a key target of speech therapy, across various voice alterations [18–20]. A recent survey on personalized speech synthesis for people who lost their biological voice further suggests that a majority prefers a more natural-sounding voice, even at the cost of some loss in intelligibility, both as users and listeners [21]. Thus, for human-to-human

interaction, reduced voice naturalness consistently has negative implications. However, this is less clear for human-machine interaction (HMI). The Computers-Are-Social-Actors (CASA) framework proposed in the 1990s [22] assumed that we treat artificial agents like humans, fueling an (implicit) naturalness-is-better bias. In turn, this spurred efforts to create synthetic voices that resemble human vocal expression [23,24], even when the link between naturalness and success in HMI remains far from fully understood. While initial findings suggested that reduced naturalness in synthetic voices compromises likeability, trustworthiness, and pleasantness [11,25–28], contemporary synthetic voice design questions a "one size fits all" idea and instead advocates solutions tailored to specific applications [29]. Accordingly, maximum human-likeness of synthetic voices may not always be required or desirable. Instead, synthetic voice preferences may depend on the features of the listeners [27,30], the device [31–33], and its specific function [6,25,31]. Understanding and incorporating such preferences seems crucial for the success and acceptance of these devices [28].

Given its widespread practical importance, the role of voice naturalness deserves scientific scrutiny. But although many recent studies provide useful empirical insights, we are currently looking at a patchwork 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) heterogeneous operationalization, (c) lack of exchange between research domains and (d) insufficient anchoring in voice perception theory. Our impression is that these problems have so far precluded a systematic understanding of vocal naturalness, impeded visibility to a wider readership, concealed 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.

Current problems in voice naturalness research

Conceptual underspecification

Voice naturalness lacks a consistent definition and terminology in the literature (see **Figure 1A-B**). In fact, the majority of papers 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 often vary tremendously across research contexts (see **Table 1** for examples). In speech-language pathology, several researchers refer to the definition provided by Yorkston and colleagues (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*" [17,34]. In contrast, research on synthetic and non-human voices usually defines naturalness as "*speech most closely perceived as a human voice*" [35] or "*the degree to which a user feels a certain technology or system is human-like*" [36]. 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 [37]. Second, listeners' naturalness perception is the result of a complex multifactorial impression formation, presumably based on the integration and weighting of many **acoustic cues** [38]. Beyond that, however, 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. We therefore see a strong need to unite them under a concise conceptual framework, which we provide in Section 3.

[Insert Figure 1 and Table 1 about here, please]

Heterogeneous operationalization

A common consequence of inconsistent conceptualization is heterogeneous operationalization. Primarily, this concerns the studied vocal categories and features, which include human vs. synthetic

voices [30,39–42]; cartoon voices [43]; pathological voices such as in individuals with Parkinson's disease [44–47], **tracheoesophageal speech** [48,49], **dysarthria** [50–53], Down syndrome [54], or stuttering [19]; acoustically manipulated human voices [55]; vocal fry [56]; as well as different accents [57,58], dialects [59], age groups [60–62], and gender identities [20,63,64]. In addition, it concerns the experimental designs and measurements, especially rating scales which differ in the number of levels and denominations of endpoints. For example, in one study participants were asked "How natural is the audio?" from "1 – natural" to "5 – unnatural" [65] , in another one they rated voices on a 10-point-scale from "very natural, human-like" to "very mechanical, robot-like" [58], or made a binary classification of voices as either human or computer-generated [37] . In principle, such empirical heterogeneity can be a powerful source of insight. There is recent evidence from face perception that differences in rating scales may not have a big impact on outcome [66], although we cannot conclude that this generalizes to naturalness ratings, and the insufficient report of empirical details impedes a meaningful comparison of findings. Specifically, it is often not stated how naturalness and the related experimental task were explained to the listeners – but instructions can be crucial determinants of study outcome. Further, the precise acoustic properties of voice material often remain elusive, bearing a risk for potential undetected confounds. Finally, few studies only provide measurements on reliability [67]. To help address these issues, we compiled some practical recommendations as guidance for future research in **Box 2**.

Lack of exchange between different research domains

Research on voice naturalness is inherently interdisciplinary, with two main domains: speech-language pathology and synthetic voices. However, while the scientific findings are well-received within each domain, these domains are remarkably poorly interconnected. **Figure 1C** illustrates this via a cross-citation analysis using VOSViewer [68], showing several distinct clusters of studies reminiscent of echo chambers which are frequently discussed in social media [69]. Of course, poor interconnectivity is not unique to naturalness but affects many other research domains within voice or face perception. However, even when considering fields with highly divergent research traditions,

such as impression formation from faces/voices for which two different two-factor models with different labels (e.g., warmth vs. competence, e.g. [70]; or trustworthiness vs. dominance, e.g. [71]) have been proposed, there is substantial research to link these distinct clusters and uncover both these specific taxonomies and their empirical relationships [72,73]. In the case of voice naturalness, however, two recent systematic literature reviews on pathological [17] and synthetic voices [23] do not have a single reference in common. One might 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 [14,26,74]. Further, while several studies failed to find an **uncanny valley** [75] effect for synthetic voices [11,76], a recent study suggests it might exist for pathological ones [77]. We therefore conclude that the lack of 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

The majority of naturalness research comes from applied fields, aiming to optimize artificial agents or to improve 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, and a simultaneous lack of understanding of basic mechanisms on the other hand. To fully understand how naturalness affects our perception and response to voices, this void needs to be filled.

Towards a concise framework for voice naturalness

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

We propose a taxonomy with two distinct types: Deviation-based naturalness and human-likeness-based naturalness (**Figure 2**). 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 [78]). However, in many studies, raters are instructed to use an inner implicit reference that is based on their experience and expectations, e.g., judge whether "*it conforms to the expected standard of unimpaired speech*" [52]. 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 [79–81]) to complex multivariate vocal patterns (e.g., in distorted or pathological voices [82]).

Human-likeness-based naturalness defines naturalness by its resemblance to a real human voice. Instructions 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 may 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 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 (see **Table 1**), 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 [77] studied the uncanny valley effect in different types of unnatural voices. They found that impressions of uncanniness resulted from "deviation from familiar categories" rather than "categorical ambiguity". This could reflect initial empirical observations in line with our proposed conceptual distinction.

[Insert Figure 2 about here, please]

Delimiting distinctiveness and authenticity

In the following, we briefly discuss the demarcation of the proposed definitions of naturalness from two established concepts in perception research, starting with distinctiveness. *Distinctiveness*, as opposed to typicality**,** has been defined as the degree to which faces or voices stick out due to rare or unusual features, and this concept is commonly used to refer to identity [83,84]. 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. 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 different concept that can capture multiple forms of deviations beyond naturalness. Accordingly, while unnatural voices would commonly be perceived as somewhat distinctive, natural voices can be distinct or typical. However, one may speculate that impressions of

human-based naturalness could be quite independent from impressions of distinctiveness under certain conditions. For instance, a person who is very accustomed to 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, requiring future inspection.

A second concept that deserves particular consideration is *authenticity*. In the scientific literature, authenticity is an established term with meaning that may refer to vocal emotion, identity or gender – rather than the holistic impression of a voice. Emotional authenticity, for example, refers to the distinction between a posed and a "real"/spontaneous emotional expression, which leads to differential behavioral and neural outcomes [85–87]. In the context of voice cloning and the now very prevalent challenge of deepfakes [7], identity authenticity is assessed with regard to a specific speaker. In principle, authenticity can be assessed with regard to manifold social signals, including age, gender, or even personality [88,89]. In fact, when prompted for synonyms of naturalness, authenticity was **ChatGPT**´s first reply (**Figure 1B**), suggesting semantic relatedness between these two terms in openly accessible online sources. At first sight, it might 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, if considered against the backdrop of voice perception theory, it becomes apparent that assessments of naturalness and authenticity appear at different stages of voice processing (see Section 5 and **Figure 3**). Thus, we tend to keep the concepts of naturalness and authenticity rather separate.

Converging evidence

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 about 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 propose that conceptual progress for disintegrated – but also highly interdisciplinary – naturalness research can be achieved by two measures: (a) converting empirical heterogeneity from an impediment into an advantage and (b) fostering mutually beneficial exchange between fields. Awareness of the interdisciplinary nature of the field is crucial for implementing both measures: First, publications need to be findable and accessible, preferably through the establishment of common terminology that feeds into common keywords. Second, findings need to be communicated inclusively for readerships from diverse backgrounds. 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 hope 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 restores hearing in people with sensorineural deafness by resynthesizing auditory signals for direct electrical stimulation of the cochlea [90], and real-time synthesis in CI sound processors could be modified to achieve better perceptual outcomes, ultimately benefitting quality of life [91]. For people who are predicted to lose their personal voice due to progressive disorders such as ALS or due to planned **laryngectomy**, current voice banking technology already allows for personalized speech synthesis with the patient´s former individual voice, often with remarkably high ratings of both naturalness and authenticity [21,92].

Naturalness research rooted in voice perception theory

Several authors have pointed out that research on voice naturalness is rather insufficiently rooted in theoretical perspectives on voice perception and voice analysis [17,23]. As discussed in Section 2.4, the topic of voice naturalness is highly influenced by research perspectives from applied sciences and

seemingly less by basic voice research and its theoretical approaches. However, neurocognitive models of voice perception can provide process-related perspectives on multi-level voice perception and voice information analysis. This allows rooting the mechanisms and types of voice naturalness assessments at relevant levels of voice analysis. Influential theories of voice perception propose sequential and partly hierarchical stages of voice processing, including a major distinction between mechanisms for voice object analysis as initial stages that are followed by the analysis of communicative and social content carried by the voice signal [4,93–95].

This processing distinction between voice object analysis and voice content analysis is relevant as it pertains to the necessary conceptual distinction between voice naturalness assessments on the one hand and the assessment of the authenticity of expressed voice content on the other hand (**Figure 3**). Assessing the naturalness of voices is conceptually associated with the initial levels of voice object analysis, including the stages of low-level auditory analysis and the analysis of structural voice patterns. Humans presumably assess acoustic feature deviations and acoustic feature likeness as low-level naturalness assessments [96], whereas assessing pattern deviations and pattern likeness concerns the assessments of natural or unnatural spectrotemporal voice profiles [97].

Unlike the rooting of naturalness assessments at the processing levels of voice feature and object analysis, authenticity assessments most likely appear at the level of voice information analysis. Voices are used as carriers to express communicative and social content. For example, voices are used for speech communication, emotional expressions, and to produce individual voice characteristics that are detected by cognitive and neural recognition mechanisms. Such voice content could be either spontaneous and authentic, or it could be acted and thus rather nonauthentic [98]. This authentic/non-authentic distinction specifically also concerns person-specific identity information in voices, which could be real or fake [7]. Such authenticity assessments might be independent of naturalness assessments, although we consider the possibility of mutual influences.

For instance, perceiving a voice as unnatural might bias non-authenticity judgments of voice content, and vice versa.

[Insert Figure 3 about here, please]

Perspectives for future research

Our theoretical considerations on the processing of voice naturalness call for investigations of its time-course and underlying brain mechanisms – relative to authenticity assessment but also to other voice characteristics. Initial evidence suggests that voice naturalness affects the brain response as early as 200 ms after voice onset and interacts with the processing of vocal emotions [99–101]. Comparably early effects have been found for authenticity assessments [86,102,103]. Although the interpretability of these findings is limited due to the potential influence of acoustic confounds, they suggest that naturalness and authenticity assessments both are fast and fundamental parts of voice perception. However, electrophysiological insights directly comparing the time-course of naturalness and authenticity are elusive, as is their interplay with impressions of age, gender, or personality traits. A recent EEG study suggests that many first impressions formed from voices are highly intercorrelated [8], but for naturalness we are currently limited to behavioral data that point towards interactions with age, gender, and emotion perception [60,63,74]. In a broad sense, naturalness impressions are always formed against a specific context, whether that context refers to the voice itself or the properties of the interaction. Accordingly, if the same voice is assessed in an all-human or HMI context could make a crucial difference.

In that vein, while this article focuses on understanding naturalness in voices from an interdisciplinary perspective, we wish to emphasize the multisensory perspective of naturalness research. In fact, substantial research in the domain of faces has compared the perceived naturalness or realism of synthesized versus real faces (for a systematic review and meta-analysis, see [104]). Recent research even demonstrated conditions in which synthesized faces can be perceived as more human than genuine human faces. Moreover, an attempt to identify the visual features that trigger

such a paradoxical facial “hyperrealism” effect suggested contributions of typicality, familiarity, attractiveness and low memorability [105]. Although this interpretation was based on qualitative reports and requires converging evidence, it seems clear how such research can inspire systematic search for commonalities or differences between mechanisms that trigger voice or face naturalness. Ultimately, we believe that naturalness research should also systematically consider interactions between vocal and visual aspects of naturalness in combination. Indeed, accumulating evidence suggests a complex interplay of visual appearance, vocal features, behavior and the interactional context for the acceptance of virtual agents [28,31–33,106–113].

Beyond humans, vocalizations are abundant in the animal kingdom. Many animals can manipulate and adapt their vocal calls to specific situations or needs. For instance, birds living in urban environments modify their song in frequency or amplitude, to avoid masking by constant anthropogenic noise [114]. While this reduces risk of not being heard by conspecifics, the degree to which such urban-induced changes to natural patterns of vocalization may have other consequences to communication seems unclear at present. We imagine that, with appropriate adaptations, the present taxonomy could be useful to promote an understanding of animal voice naturalness as well.

Finally, very recent fMRI research has uncovered a cortical-striatal brain network that is involved when listeners try to distinguish deepfake from real speaker identities [7]. Such research is relevant also because the accelerating spread of misinformation via social media is now considered a major problem which compromises societal cohesion [69,115]. While large-scale misinformation is still mostly text-based as of today, next-generation deepfakes likely will be even more efficient vehicles of misinformation. This is because they efficiently instrumentalize person-related trust via high-level perceptual deception. On that perspective, better understanding of characteristics of “successful” vocal deepfakes and their processing in the brain may be one important component for strengthening human resilience to fake information of the future.

Concluding remarks

Naturalness in voices is a highly intuitive concept, but one that is scientifically underspecified and far from systematically understood, despite considerable research efforts. To address this, we propose a conceptual framework for voice naturalness. Our taxonomy, comprised of deviation-based naturalness and human-likeness-based naturalness, is rooted in voice perception theory, and is inspired by interdisciplinary empirical findings. The new framework offers the flexibility that is necessary to be applicable across diverse empirical designs, while at the same time promoting comparability across research domains. We complement this conceptual groundwork with several practical recommendations to bridge previously unconnected approaches and better integrate this highly interdisciplinary field. We hope to provide a foundation for conjoined efforts towards more systematic future research on numerous **outstanding questions** on voice naturalness. While we here focus on voices, we ultimately opt for a multisensory perspective on naturalness research. In a world that is increasingly dominated by digitally synthesized agents, it seems important to identify the multifaceted determinants for human perception of naturalness in social stimuli.

Figure Legends

Figure 1

Terminology and interconnectivity of voice naturalness research

Note. ***A)*** *Word cloud depicting synonyms and closely related concepts from 72 publications that target naturalness in voices (for details, see* ***Box 1****). Word size represents number of occurences.* ***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 occurrence frequency. The full prompt, the generated response, and a reflection on its strengths and limitations are accessible on OSF.* ***C)*** *A bibliographic network*

visualization using VOSviewer [68], 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. 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. A full documentation and an interactive version of the bibliographic network can be found on OSF.

Figure 2

A conceptual framework for the definition of voice naturalness

Note. Assessing the naturalness of voices requires a reference frame (left panel), which is most commonly represented by the voice production system of humans. This human production system sets reference either as individual voice samples (explicit target voice) or as prototype voice representations (implicit prototype voice), against which test voice samples (right panel) are assessed for naturalness. Two types of naturalness assessments are proposed (mid panel). The deviation-based approach assesses naturalness in terms of distance away from the reference, while the human-likeness-based approach assesses naturalness according to its similarity towards the reference. Deviation in voice naturalness can occur, for example, due to clinical conditions, voice manipulations, and acoustic artifacts. Human-likeness-based naturalness defines naturalness by its resemblance to a real human voice. Human likeness can be assessed based on audio samples within (human samples) and outside the human voice space (synthetic samples) marked by the human voice border.

Figure 3

Rooting voice naturalness in voice processing theory

Note. Theories of voice perception propose a multi-level processing of voice samples (left panel) and analyzing these samples according to their feature and auditory object patterns (mid panel), followed by the analysis of information carried by the voice signals (right panel). Assessing the naturalness of voices appears at the level of voice features (low-level auditory analysis) and voice object analysis (voice structural analysis) and includes the assessment of acoustic deviations and acoustic likeness, as well as the assessment of pattern deviations and pattern likeness to reference voice samples. Unlike naturalness assessments, authenticity judgments mainly concern the assessment of communicative and social content carried by the voice signal at the level of voice information analysis. Such voice content can be expressed either spontaneously (authentic) or can be enacted (non-authentic), or it could be of a real or fake nature when it specifically concerns person-related identity information. Naturalness and authenticity assessments may have mutual influences.

Table 1

Examples for definitions of deviation-based and human-likeness-based naturalness of voices in the literature

Conceptualization	Definition	Reference
Deviation-based naturalness	"Naturalness was 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." (p. 4687)	Abur et al. (2021) [44]
	"Speech naturalness can be described as how the speech of a person with a speech disorder compares with that of typical speech or, in the case of an acquired disorder, how an individual's speech compares to its premorbid state" (p. 1134)	Anand & Stepp (2015) [14]
	"Speech naturalness refers to a rather broad perceptual impression representing the overall quality of a person's speech output in relation to what is conceptualized as normal or natural" (p. 1633/1634)	Schölderle et al. (2023) [51]
	"[…] degree to which individuals sound 'different' from healthy peers" (p. 1265)	Vogel et al. (2019) [53]

Human-likeness-based naturalness	"Human likeness has been used […] to describe how accurately the machine is able to imitate a human." (p. 2864)	Baird et al. (2018) [26]
	"Naturalness refers to whether synthetic speech is perceived as uniquely human, despite being computer-generated." (p. 5)	Hyppa-Martin et al. (2024) [21]
	"Natural speech is the speech most closely perceived as a human voice." (p. 10)	Mawalim et al. (2022) [35]
	"Naturalness refers to how closely the output sounds like human speech." (p. 389.e1)	Yamasaki et al. (2017) [42]
Combination of both	"By naturalness, we understand the voice stimulus to be perceived as a plausible outcome of the human speech production system" (p. 1)	Nussbaum et al. (2023) [74]
	"[…] voices which sound like they could come from an actual human being (which should be rated as more natural) and voices that sound more fictitious, such as a cartoon character or a monster (which should be rated as less natural)." (p.429)	Kapolowicz et al. (2022) [57]

Note. Definitions are all original quotes from the respective references. The full compilation of extracted definitions can be accessed on OSF. Note that the mapping of definitions to the conceptualization of naturalness was carried out by us and not the authors of the original publications.

Box 1: A field in numbers

For a more systematic overview on scientific insights into naturalness in voices, we conducted a focused literature search on Web of Science on 26 April 2023 using the search terms "naturalness AND voice" or "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 measured 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 all included papers and a reflection on potential biases in the literature search, please refer to *OSF*.

In total, we identified 72 articles, covering a time range from 1984 to 2024. Thirty-eight (53%) were published in the last 5 years. Sixty-seven report behavioral empirical data, of which 48 are predominantly ratings. Two are literature reviews, and three used neurophysiological measures. Regarding voice category, 33 used synthetic, 18 human-pathological, 6 human-manipulated and 5 healthy human voices. Ten used more than one of these voice categories. In only 32 papers, we could identify an explicit definition of naturalness (see Table 1 for examples and *OSF* for a full list). 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 1A**. Subsequently, we compared these to the articles' keywords: 58 papers provided keywords, but only 32 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 26 employed a deviation-based conceptualization, 35 used human-likeness, and 11 used a combination of both.

Box 2: Practical recommendations for voice naturalness research

Research on voice naturalness is highly interdisciplinary. To make future research accessible to a wider readership across disciplines, and allow comparability and integration of findings, 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 participants and readers. With the 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 under study. If used consistently, this taxonomy offers quick orientation for readers and fosters comparability across findings.
- Use consistent keywords to make relevant research findable across disciplines. We recommend “naturalness”, “human-likeness” or, in cases discussed in Section 3.2, “authenticity”.
- Include full reports on methodological details. Specifically, this concerns acoustic manipulations that target voice naturalness, measurements (i.e. rating scales used to assess naturalness impressions), instructions to raters, and reports on reliability. For synthetic voices, be as specific as possible on synthesis methods, toolboxes and their settings, as well as any additional processing you applied.
- Wherever possible, provide stimulus examples. This is important because readers may have a clear idea how a male vs. female voice sounds or how an angry voice differs from a happy one, but their imagination of an (un)-natural or synthetic voice could be quite vague and differ tremendously from the actual audio material. Often, direct auditory impression can be complementary to, and more insightful than, a list of acoustic measures and descriptions. In some cases (i.e. when very different synthesis methods were used), differences in audio material may offer a straightforward explanation for different empirical outcomes.
- Communicate findings inclusively enough for readerships from diverse backgrounds. Provide explicit definitions (e.g. for terms like “prosody”, “dysarthria”, or “anthropomorphism”), avoid technical jargon including abbreviations unfamiliar to other fields (e.g. synthesis algorithms, machine learning approaches, or acoustic measures), adopt scientific standards from other fields where appropriate, and discuss findings against the wider interdisciplinary literature (i.e. linking insights into pathological voices to synthetic ones and vice versa).

- Quantify naturalness, whenever it could have important implications for ecological validity of the stimulus material, even when naturalness is not the primary focus of the study. This is especially important when using acoustic manipulations which could have unintended side effects on perceived naturalness [74,116].

Glossary:

- Acoustic cues: physical and measurable features of sounds (such as voices); these may include fundamental frequency, intensity, a range of timbre cues, or temporal characteristics. Used by listeners to inform manifold impressions about voices, such as emotion, identity, age, gender or naturalness.
- Anthropomorphism: the attribution of human characteristics, emotions, or behaviors to non-human entities
- ChatGPT: a chatbot developed by OpenAI, based on a large language model, that generates text based on input-prompts (GPT stands for generative pre-trained transformer)
- Deepfakes: digitally manipulated media, such as images, videos, or voice recordings, created using deep learning techniques with the goal to convincingly display the appearance of a specific individual.
- Dysarthria: impairments of speech motor subsystems due to various neurological conditions such as Parkinson's disease, amyotrophic lateral sclerosis (ALS), developmental conditions, strokes, or traumatic brain injury.
- Laryngectomy: surgical removal of the larynx, typically in the context of larynx cancer treatment
- Synthetic/artificial voices: computer generated voices. Common methods are articulatory synthesis, concatenative synthesis, and statistical parametric synthesis, including deep learning algorithms

- Tracheoesophageal speech: a method of vocalization following total laryngectomy via a tracheoesophageal prosthesis that enables speech through esophageal vibrations.
- Uncanny valley: a sudden feeling of eeriness evoked by humanoid robots that almost approach, but do not entirely reach a human-like appearance

Acknowledgements and Funding

We thank Simone Dahmen and Fatma Bilem for their support with the literature analysis, and the members of the Jena Voice Research Unit (https://www.voice.uni-jena.de/) for helpful suggestions on this project.

The authors gratefully acknowledge the award of funding through an EU-MSCA doctoral network "Voice Communication Sciences" (action 101168998, https://www.vocs.eu.com/).

CN: I dedicate this work to our stillborn son. Thanks for changing our lives.

References

1. Román, S. et al. (2017) The importance of food naturalness for consumers: Results of a systematic review. *Trends in Food Science & Technology* 67, 44–57. DOI: 10.1016/j.tifs.2017.06.010
2. Meier, B.P. et al. (2019) Naturally better? A review of the natural-is-better bias. *Social & Personality Psych* 13. DOI: 10.1111/spc3.12494
3. Ode, A. et al. (2009) Indicators of perceived naturalness as drivers of landscape preference. *Journal of environmental management* 90, 375–383. DOI: 10.1016/j.jenvman.2007.10.013
4. Young, A.W. et al. (2020) Face and voice perception: Understanding commonalities and differences. *Trends Cogn Sci* 24, 398–410. DOI: 10.1016/j.tics.2020.02.001
5. Rodero, E. and Lucas, I. (2023) Synthetic versus human voices in audiobooks: The human emotional intimacy effect. *New Media & Society* 25, 1746–1764. DOI: 10.1177/14614448211024142
6. Rodero, E. (2017) Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices. *Computers in Human Behavior* 77, 336–346. DOI: 10.1016/j.chb.2017.08.044

7. Roswandowitz, C. et al. (2024) Cortical-striatal brain network distinguishes deepfake from real speaker identity. *Communications biology* 7, 711. DOI: 10.1038/s42003-024-06372-6
8. Lavan, N. et al. (2024) The time course of person perception from voices in the brain. *Proc Natl Acad Sci U S A* 121, e2318361121. DOI: 10.1073/pnas.2318361121
9. Lavan, N. (2023) How do we describe other people from voices and faces? *Cognition* 230, 105253. DOI: 10.1016/j.cognition.2022.105253
10. Jiang, Z. et al. (2024) Comparison of face-based and voice-based first impressions in a Chinese sample. *Br. J. Psychol.* 115, 20–39. DOI: 10.1111/bjop.12675
11. Kühne, K. et al. (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, 1–16. DOI: 10.3389/fnbot.2020.593732
12. Ilves, M. and Surakka, V. (2013) Subjective responses to synthesised speech with lexical emotional content: the effect of the naturalness of the synthetic voice. *Behaviour & Information Technology* 32, 117–131. DOI: 10.1080/0144929X.2012.702285
13. Ilves, M. et al. (2011) The Effects of Emotionally Worded Synthesized Speech on the Ratings of Emotions and Voice Quality. In , pp. 588–598, Springer, Berlin, Heidelberg
14. Anand, S. and Stepp, C.E. (2015) Listener Perception of Monopitch, Naturalness, and Intelligibility for Speakers With Parkinson's Disease. *J Speech Lang Hear Res* 58, 1134–1144. DOI: 10.1044/2015_JSLHR-S-14-0243
15. Moya-Galé, G. and Levy, E.S. (2019) Parkinson's disease-associated dysarthria: prevalence, impact and management strategies. *JPRLS* Volume 9, 9–16. DOI: 10.2147/JPRLS.S168090
16. Damico, J.S. and Ball, M.J., eds (2019) *The SAGE Encyclopedia of Human Communication Sciences and Disorders,* SAGE Publications, Inc
17. Klopfenstein, M. et al. (2020) The study of speech naturalness in communication disorders: A systematic review of the literature. *Clinical Linguistics & Phonetics* 34, 327–338. DOI: 10.1080/02699206.2019.1652692
18. Frankford, S.A. et al. (2024) Contributions of Speech Timing and Articulatory Precision to Listener Perceptions of Intelligibility and Naturalness in Parkinson's Disease. *J Speech Lang Hear Res* 67, 2951–2963. DOI: 10.1044/2024_JSLHR-23-00802
19. Euler, H.A. et al. (2021) Speech restructuring group treatment for 6-to-9-year-old children who stutter: A therapeutic trial. *Journal of communication disorders* 89, 106073. DOI: 10.1016/j.jcomdis.2020.106073
20. Hardy, T.L.D. et al. (2020) Acoustic Predictors of Gender Attribution, Masculinity-Femininity, and Vocal Naturalness Ratings Amongst Transgender and Cisgender Speakers. *Journal of Voice* 34, 300.e11-300.e26. DOI: 10.1016/j.jvoice.2018.10.002
21. Hyppa-Martin, J. et al. (2024) A large-scale comparison of two voice synthesis techniques on intelligibility, naturalness, preferences, and attitudes toward voices banked by individuals with amyotrophic lateral sclerosis. *Augmentative and Alternative Communication* 40, 31–45. DOI: 10.1080/07434618.2023.2262032
22. Nass, C. et al. (1994) Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94,* ACM Press
23. Seaborn, K. et al. (2021) Voice in Human–Agent Interaction. *ACM Comput. Surv.* 54, 1–43. DOI: 10.1145/3386867
24. Triantafyllopoulos, A. et al. (2023) An overview of affective speech synthesis and conversion in the deep learning era. *Proceedings of the IEEE*
25. Schreibelmayr, S. and Mara, M. (2022) Robot Voices in Daily Life: Vocal Human-Likeness and Application Context as Determinants of User Acceptance. *Frontiers in Psychology* 13, 1–17. DOI: 10.3389/fpsyg.2022.787499

26. Baird, A. et al. (2018) The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech. In *Interspeech 2018,* pp. 2863–2867, ISCA
27. Lee, E.-J. (2010) The more humanlike, the better? How speech type and users' cognitive style affect social responses to computers. *Computers in Human Behavior* 26, 665–672. DOI: 10.1016/j.chb.2010.01.003
28. Lu, L. et al. (2021) Leveraging "human-likeness" of robotic service at restaurants. *International Journal of Hospitality Management* 94, 1–9. DOI: 10.1016/j.ijhm.2020.102823
29. Cambre, J. and Kulkarni, C. (2019) One Voice Fits All? *Proc. ACM Hum.-Comput. Interact.* 3, 1–19. DOI: 10.1145/3359325
30. Eyssel, F. et al. (2012) 'If you sound like me, you must be more human'. In *HRI' 12. Proceedings of the seventh annual ACM/IEEE Conference on Human-Robot Interaction : March 5-8, 2012 Boston, Massachusetts, USA* (Yanco, H. et al., eds), pp. 125–126, Association for Computing Machinery
31. Im, H. et al. (2023) Let voice assistants sound like a machine: Voice and task type effects on perceived fluency, competence, and consumer attitude. *Computers in Human Behavior* 145, 107791. DOI: 10.1016/j.chb.2023.107791
32. McGinn, C. and Torre, I. (2019 - 2019) Can you Tell the Robot by the Voice? An Exploratory Study on the Role of Voice in the Perception of Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI),* pp. 211–221, IEEE
33. Mitchell, W.J. et al. (2011) A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception* 2, 10–12. DOI: 10.1068/i0415
34. Yorkston, K.M. et al. (1999) *Management of motor speech disorders in children and adults,* Pro-ed Austin, TX
35. Mawalim, C.O. et al. (2022) Speaker anonymization by modifying fundamental frequency and x-vector singular value. *Computer Speech & Language* 73, 1–17. DOI: 10.1016/j.csl.2021.101326
36. Hu, P. et al. (2021) Dual humanness and trust in conversational AI: A person-centered approach. *Computers in Human Behavior* 119, 106727. DOI: 10.1016/j.chb.2021.106727
37. Nusbaum, H.C. et al. (1997) Measuring the naturalness of synthetic speech. *International Journal of Speech Technology* 2, 7–19
38. Mayo, C. et al. (2011) Listeners' weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis. *Speech Commun* 53, 311–326. DOI: 10.1016/j.specom.2010.10.003
39. Abdulrahman, A. and Richards, D. (2022) Is Natural Necessary? Human Voice versus Synthetic Voice for Intelligent Virtual Agents. *MTI* 6, 51. DOI: 10.3390/mti6070051
40. Urakami, J. et al. (2020) 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), pp. 244–259, Springer International Publishing
41. Velner, E. et al. (2020) Intonation in Robot Speech. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Belpaeme, T. et al., eds), pp. 569–578, ACM
42. Yamasaki, R. et al. (2017) Perturbation Measurements on the Degree of Naturalness of Synthesized Vowels. *Journal of Voice* 31, 389.e1-389.e8. DOI: 10.1016/j.jvoice.2016.09.020
43. Ko, S. et al. (2023) The Effects of Robot Voices and Appearances on Users' Emotion Recognition and Subjective Perception. *Int. J. Human. Robot.* 20. DOI: 10.1142/S0219843623500019
44. Abur, D. et al. (2021) Feedback and Feedforward Auditory-Motor Processes for Voice and Articulation in Parkinson's Disease. *J Speech Lang Hear Res* 64, 4682–4694. DOI: 10.1044/2021_JSLHR-21-00153

45. Klopfenstein, M. (2015) Relationship between acoustic measures and speech naturalness ratings in Parkinson's disease: A within-speaker approach. *Clinical Linguistics & Phonetics* 29, 938–954. DOI: 10.3109/02699206.2015.1081293
46. Klopfenstein, M. (2016) Speech naturalness ratings and perceptual correlates of highly natural and unnatural speech in hypokinetic dysarthria secondary to Parkinson's disease. *JIRCD* 7, 123–146. DOI: 10.1558/jircd.v7i1.27932
47. Moya-Galé, G. et al. (2024) Perceptual consequences of online group speech treatment for individuals with Parkinson's disease: A pilot study case series. *International Journal of Speech-Language Pathology,* 1–16. DOI: 10.1080/17549507.2024.2330538
48. Eadie, T.L. and Doyle, P.C. (2002) Direct Magnitude Estimation and Interval Scaling of Naturalness and Severity in Tracheoesophageal (TE) Speakers. *J Speech Lang Hear Res* 45, 1088–1096. DOI: 10.1044/1092-4388(2002/087)
49. Eadie, T.L. et al. (2008) Influence of speaker gender on listener judgments of tracheoesophageal speech. *Journal of Voice* 22, 43–57. DOI: 10.1016/j.jvoice.2006.08.008
50. Yorkston, K.M. et al. (1990) The effect of rate control on the intelligibility and naturalness of dysarthric speech. *The Journal of speech and hearing disorders* 55, 550–560. DOI: 10.1044/jshd.5503.550
51. Schölderle, T. et al. (2023) Speech Naturalness in the Assessment of Childhood Dysarthria. *American Journal of Speech-language Pathology* 32, 1633–1643. DOI: 10.1044/2023_AJSLP-23-00023
52. Lehner, K. and Ziegler, W. (2022) Clinical measures of communication limitations in dysarthria assessed through crowdsourcing: specificity, sensitivity, and retest-reliability. *Clinical Linguistics & Phonetics* 36, 988–1009. DOI: 10.1080/02699206.2021.1979658
53. Vogel, A.P. et al. (2019) Speech treatment improves dysarthria in multisystemic ataxia: a rater-blinded, controlled pilot-study in ARSACS. *Journal of neurology* 266, 1260–1266. DOI: 10.1007/s00415-019-09258-4
54. Jones, H.N. et al. (2019) Auditory-Perceptual Speech Features in Children With Down Syndrome. *American journal on intellectual and developmental disabilities* 124, 324–338. DOI: 10.1352/1944-7558-124.4.324
55. Assmann, P.F. et al. (2006) Effects of frequency shifts on perceived naturalness and gender information in speech. In *INTERSPEECH*
56. Venkatraman, A. and Sivasankar, M.P. (2018) Continuous Vocal Fry Simulated in Laboratory Subjects: A Preliminary Report on Voice Production and Listener Ratings. *American Journal of Speech-language Pathology* 27, 1539–1545. DOI: 10.1044/2018_AJSLP-17-0212
57. Kapolowicz, M.R. et al. (2022) Effects of Spectral Envelope and Fundamental Frequency Shifts on the Perception of Foreign-Accented Speech. *Language and speech* 65, 418–443. DOI: 10.1177/00238309211029679
58. Tamagawa, R. et al. (2011) The Effects of Synthesized Voice Accents on User Perceptions of Robots. *Int J of Soc Robotics* 3, 253–262. DOI: 10.1007/s12369-011-0100-4
59. Mackey, L.S. et al. (1997) Effect of speech dialect on speech naturalness ratings: a systematic replication of Martin, Haroldson, and Triden (1984). *J Speech Lang Hear Res* 40, 349–360. DOI: 10.1044/jslhr.4002.349
60. Goy, H. et al. (2016) Effects of age on speech and voice quality ratings. *The Journal of the Acoustical Society of America* 139, 1648. DOI: 10.1121/1.4945094
61. Coughlin-Woods, S. et al. (2005) Ratings of speech naturalness of children ages 8-16 years. *Percept Motor Skill* 100, 295–304. DOI: 10.2466/pms.100.2.295-304

62. Baird, A. et al. (2017) Perception of Paralinguistic Traits in Synthesized Voices. In *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences* (Fazekas, G. et al., eds), pp. 1–5, ACM
63. Merritt, B. and Bent, T. (2020) Perceptual Evaluation of Speech Naturalness in Speakers of Varying Gender Identities. *J Speech Lang Hear Res* 63, 2054–2069. DOI: 10.1044/2020_JSLHR-19-00337
64. Baird, A. et al. (2018) The Perception of Vocal Traits in Synthesized Voices: Age, Gender, and Human Likeness. *J. Audio Eng. Soc.* 66, 277–285. DOI: 10.17743/jaes.2018.0023
65. Aylett, M.P. et al. (2020) Speech Synthesis for the Generation of Artificial Personality. *IEEE Trans. Affective Comput.* 11, 361–372. DOI: 10.1109/TAFFC.2017.2763134
66. Kramer, R.S.S. et al. (2024) The psychometrics of rating facial attractiveness using different response scales. *Perception* 53, 645–660. DOI: 10.1177/03010066241256221
67. Martin, R.R. et al. (1984) Stuttering and speech naturalness. *The Journal of speech and hearing disorders* 49, 53–58. DOI: 10.1044/jshd.4901.53
68. van Eck, N.J. and Waltman, L. (2010) Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 84, 523–538. DOI: 10.1007/s11192-009-0146-3
69. van der Linden, S. (2023) *Foolproof: Why we fall for misinformation and how to build immunity,* WW Norton & Company.
70. Fiske, S.T. (2018) Stereotype Content: Warmth and Competence Endure. *Curr Dir Psychol Sci* 27, 67–73. DOI: 10.1177/0963721417738825
71. Todorov, A. et al. (2008) Understanding evaluation of faces on social dimensions. *Trends Cogn Sci* 12, 455–460. DOI: 10.1016/j.tics.2008.10.001
72. Sutherland, C.A.M. et al. (2013) Social inferences from faces: ambient images generate a three-dimensional model. *Cognition* 127, 105–118. DOI: 10.1016/j.cognition.2012.12.001
73. Sutherland, C.A.M. et al. (2016) Integrating social and facial models of person perception: Converging and diverging dimensions. *Cognition* 157, 257–267. DOI: 10.1016/j.cognition.2016.09.006
74. Nussbaum, C. et al. (2023) Perceived naturalness of emotional voice morphs. *Cognition & Emotion,* 1–17. DOI: 10.1080/02699931.2023.2200920
75. Mori, M. et al. (2012) The Uncanny Valley. *IEEE Robot. Automat. Mag.* 19, 98–100. DOI: 10.1109/mra.2012.2192811
76. Romportl, J. (2014) Speech Synthesis and Uncanny Valley. In *Text, speech and dialogue* (Horák, A. et al., eds), pp. 595–602, Springer International Publishing
77. Diel, A. and Lewis, M. (2024) Deviation from typical organic voices best explains a vocal uncanny valley. *Computers in Human Behavior Reports* 14, 100430. DOI: 10.1016/j.chbr.2024.100430
78. van Prooije, T. et al. (2024) Perceptual and Acoustic Analysis of Speech in Spinocerebellar ataxia Type 1. *Cerebellum,* 112–120. DOI: 10.1007/s12311-023-01513-9
79. Moore, B.C.J. and Tan, C.-T. (2003) Perceived naturalness of spectrally distorted speech and music. *The Journal of the Acoustical Society of America* 114, 408–419. DOI: 10.1121/1.1577552
80. Rao M V, A. et al. (2018) Effect of source filter interaction on isolated vowel-consonant-vowel perception. *The Journal of the Acoustical Society of America* 144, EL95. DOI: 10.1121/1.5049510
81. Ratcliff, A. et al. (2002) Factors influencing ratings of speech naturalness in augmentative and alternative communication. *Augmentative and Alternative Communication* 18, 11–19. DOI: 10.1080/aac.18.1.11.19
82. Meltzner, G.S. and Hillman, R.E. (2005) Impact of Aberrant Acoustic Properties on the Perception of Sound Quality in Electrolarynx Speech. *J Speech Lang Hear Res* 48, 766–779. DOI: 10.1044/1092-4388(2005/053)

83. Andics, A. et al. (2010) Neural mechanisms for voice recognition. *Neuroimage* 52, 1528–1540. DOI: 10.1016/j.neuroimage.2010.05.048
84. Valentine, T. et al. (2016) Face-space: A unifying concept in face recognition research. *Q J Exp Psychol (Hove)* 69, 1996–2019. DOI: 10.1080/17470218.2014.990392
85. Lima, C.F. et al. (2021) Authentic and posed emotional vocalizations trigger distinct facial responses. *Cortex* 141, 280–292. DOI: 10.1016/j.cortex.2021.04.015
86. Sarzedas, J. et al. (2024) Blindness influences emotional authenticity perception in voices: Behavioral and ERP evidence. *Cortex* 172, 254–270. DOI: 10.1016/j.cortex.2023.11.005
87. Anikin, A. and Lima, C.F. (2017) Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *Q J Exp Psychol (Hove)* 71, 622–641. DOI: 10.1080/17470218.2016.1270976
88. Kachel, S. et al. (2020) Gender (Conformity) Matters: Cross-Dimensional and Cross-Modal Associations in Sexual Orientation Perception. *Journal of Language and Social Psychology* 39, 40–66. DOI: 10.1177/0261927X19883902
89. Mills, M. et al. (2017) 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* 18, 328–342. DOI: 10.1080/15532739.2017.1329049
90. Eiff, C.I. von et al. (2022) Crossmodal benefits to vocal emotion perception in cochlear implant users. *iScience* 25, 105711. DOI: 10.1016/j.isci.2022.105711
91. Schweinberger, S.R. and Eiff, C.I. von (2022) Enhancing socio-emotional communication and quality of life in young cochlear implant recipients: Perspectives from parameter-specific morphing and caricaturing. *Frontiers in Neuroscience* 16, 956917. DOI: 10.3389/fnins.2022.956917
92. Yamagishi, J. et al. (2012) Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoust. Sci. & Tech.* 33, 1–5. DOI: 10.1250/ast.33.1
93. Belin, P. et al. (2004) Thinking the voice: neural correlates of voice perception. *Trends Cogn Sci* 8, 129–135. DOI: 10.1016/j.tics.2004.01.008
94. Belin, P. et al. (2011) Understanding voice perception. *Br. J. Psychol.* 102, 711–725. DOI: 10.1111/j.2044-8295.2011.02041.x
95. Lavan, N. and McGettigan, C. (2023) A model for person perception from familiar and unfamiliar voices. *Commun Psychol* 1, 1–11. DOI: 10.1038/s44271-023-00001-4
96. Staib, M. and Frühholz, S. (2023) Distinct functional levels of human voice processing in the auditory cortex. *Cerebral Cortex* 33, 1170–1185. DOI: 10.1093/cercor/bhac128
97. Staib, M. and Frühholz, S. (2021) Cortical voice processing is grounded in elementary sound analyses for vocalization relevant sound patterns. *Progress in neurobiology* 200, 101982. DOI: 10.1016/j.pneurobio.2020.101982
98. Pinheiro, A.P. et al. (2021) Emotional authenticity modulates affective and social trait inferences from voices. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 376, 20200402. DOI: 10.1098/rstb.2020.0402
99. Duville, M.M. et al. (2022) Neuronal and behavioral affective perceptions of human and naturalness-reduced emotional prosodies. *Frontiers in computational neuroscience* 16, 1022787. DOI: 10.3389/fncom.2022.1022787
100. Duville, M.M. et al. (2024) Improved emotion differentiation under reduced acoustic variability of speech in autism. *BMC medicine* 22, 121. DOI: 10.1186/s12916-024-03341-y
101. Nussbaum, C. et al. (2022) Contributions of fundamental frequency and timbre to vocal emotion perception and their electrophysiological correlates. *Social Cognitive and Affective Neuroscience* 17, 1145–1154. DOI: 10.1093/scan/nsac033

102. Kosilo, M. et al. (2021) The neural basis of authenticity recognition in laughter and crying. *Scientific reports* 11, 23750. DOI: 10.1038/s41598-021-03131-z
103. Conde, T. et al. (2022) The time course of emotional authenticity detection in nonverbal vocalizations. *Cortex; a journal devoted to the study of the nervous system and behavior* 151, 116–132. DOI: 10.1016/j.cortex.2022.02.016
104. Miller, E.J. et al. (2023) How do people respond to computer-generated versus human faces? A systematic review and meta-analyses. *Computers in Human Behavior Reports,* 100283. DOI: 10.1016/j.chbr.2023.100283
105. Miller, E.J. et al. (2023) AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychol Sci* 34, 1390–1403. DOI: 10.1177/09567976231207095
106. Cabral, J.P. et al. (2017) The Influence of Synthetic Voice on the Evaluation of a Virtual Character. In *Interspeech 2017,* pp. 229–233, ISCA
107. Ehret, J. et al. (2021) Do Prosody and Embodiment Influence the Perceived Naturalness of Conversational Agents' Speech? *ACM Trans. Appl. Percept.* 18, 1–15. DOI: 10.1145/3486580
108. Ferstl, Y. et al. (2021) Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents,* pp. 76–83, ACM
109. Gong, L. and Nass, C. (2007) When a Talking-Face Computer Agent is Half-Human and Half-Humanoid: Human Identity and Consistency Preference. *Human Comm Res* 33, 163–193. DOI: 10.1111/j.1468-2958.2007.00295.x
110. Higgins, D. et al. (2022) Sympathy for the digital: Influence of synthetic voice on affinity, social presence and empathy for photorealistic virtual humans. *Computers & Graphics* 104, 116–128. DOI: 10.1016/j.cag.2022.03.009
111. Li, M. et al. (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. DOI: 10.1016/j.chb.2022.107645
112. Parmar, D. et al. (2022) Designing Empathic Virtual Agents: Manipulating Animation, Voice, Rendering, and Empathy to Create Persuasive Agents. *Autonomous agents and multi-agent systems* 36. DOI: 10.1007/s10458-021-09539-1
113. Sarigul, B. and Urgen, B.A. (2023) Audio–Visual Predictive Processing in the Perception of Humans and Robots. *Int J of Soc Robotics* 15, 855–865. DOI: 10.1007/s12369-023-00990-6
114. Lowry, H. et al. (2013) Behavioural responses of wildlife to urban environments. *Biological reviews of the Cambridge Philosophical Society* 88, 537–549. DOI: 10.1111/brv.12012
115. Kauk, J. et al. (2024) The adaptive community-response (ACR) method for collecting misinformation on social media. *J Big Data* 11. DOI: 10.1186/s40537-024-00894-w
116. Malisz, Z. et al. (2020) Modern speech synthesis for phonetic sciences: a discussion and an evaluation. DOI: 10.31234/osf.io/dxvhc

Outstanding questions:

- Vocal communication is abundant in the animal kingdom, and many animals manipulate their vocal behavior in an adaptive manner – is there demand for a comparative perspective on voice naturalness?
- How is a listener´s perception of naturalness shaped through experience (e.g., with voice assistants, smart home devices, or patients with voice disorders)?
- With respect to the present conceptual framework, (how) are human-likeness based naturalness and deviation-based naturalness dissociable in the brain?
- In the trade-off between precise experimental control and open field recordings, can we identify converging evidence for how and when reduced naturalness in voices critically affects the ecological validity of research? In depth, will we need a dynamic definition of ecological validity in view of an ever more digital world of social interaction?
- Are natural voices always preferred, or is naturalness preference context dependent? Can natural voices impede rather than promote communication success in some situations?
- Many domains of social perception are characterized by individual variability, but it is unclear whether there are substantial individual differences in the tolerance of or preference for unnatural voice features. If so, can these be related to other domains of auditory cognition, or to other person traits?
- To what extent is naturalness perception affected by factors such as age, gender, or cultural background?

Figure 1

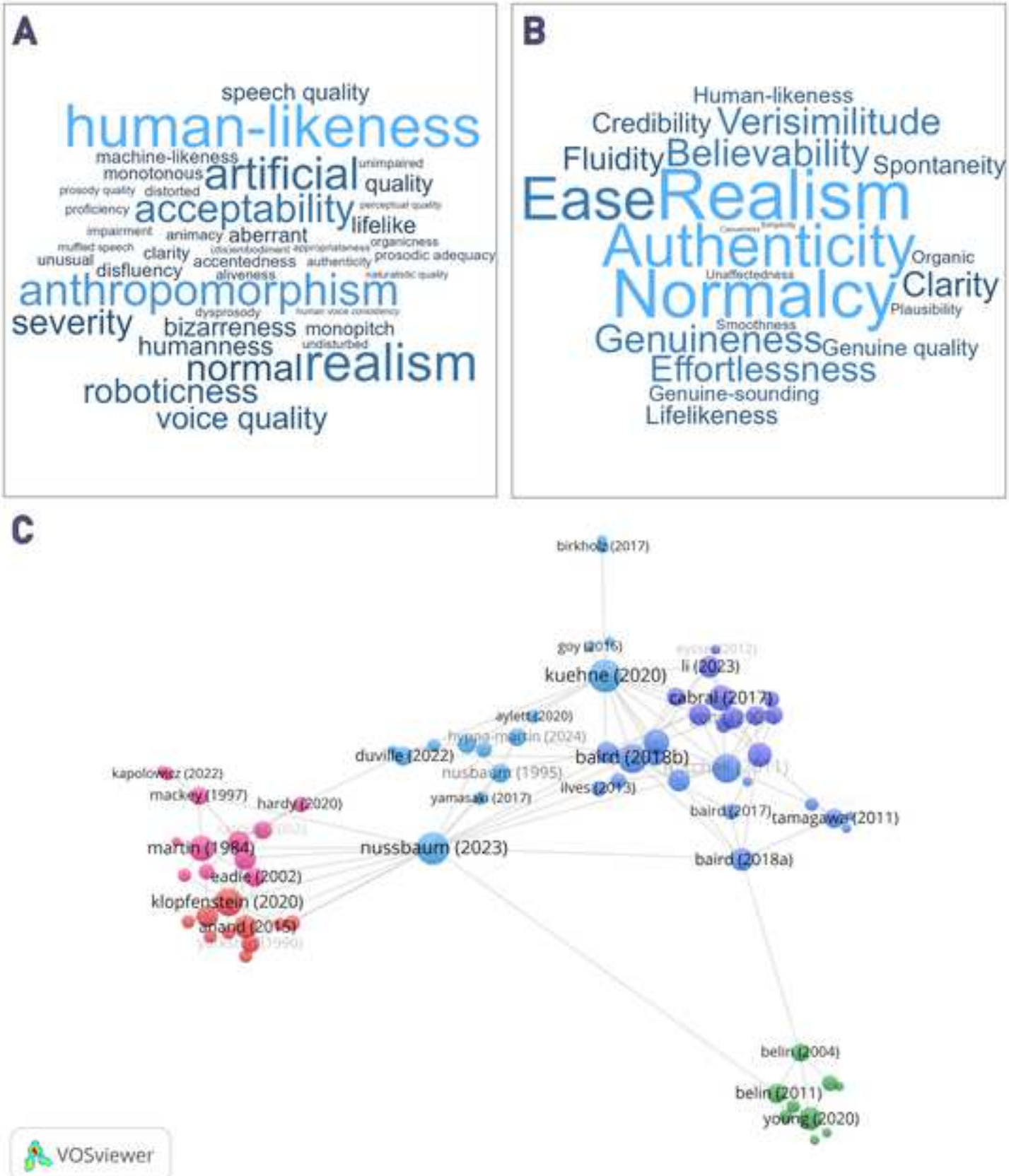


Figure 2 (Key Figure)

Click here to access/download;Figure;Figure2.tif

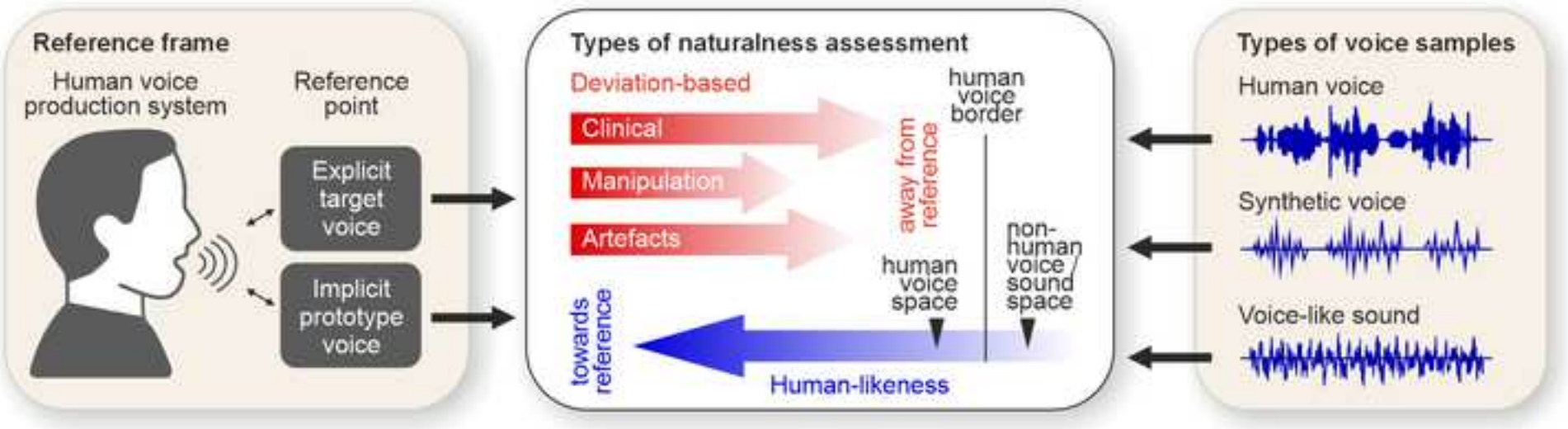
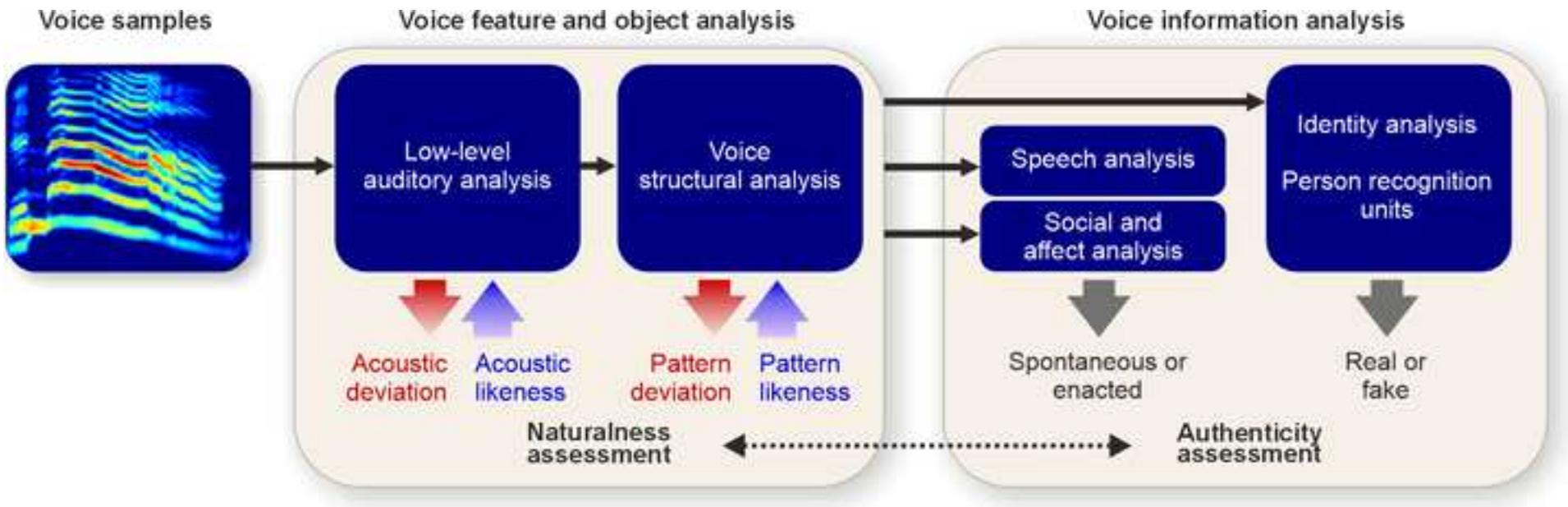


Figure 3



Declaration of interest:

The authors declare no competing interests.

Manuscript - with tracked changes

Click here to view linked References