Comments by the editor:

**To help you address the reviewers' comments, the revised manuscript may be up to 4000 words and may include up to 120 references and 6 elements.**

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:

* Make recommendations section more specific
* Impact of voice naturalness on listener perception and behavior: impact for pathological voices, maybe focus more on synthetic voices.
* Interplay with emotion
* Context-specificity?
* > wo im Manuscript einfügen?

1. **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: ToDo

* I would not argue that it does. Its just something that we have not studied sufficiently yet
* Ich glaube wir müssen nochmal diskutieren, was genau mit dem Punkt gemeint ist

1. **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that TiCS articles cannot report the results of systematic reviews, nor can key points made in the piece rely on an unpublished systematic review.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 one 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 16 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:

“*This is further corroborated by two recent systematic literature reviews on pathological [15] and synthetic voices [20] which do not have a single reference in common.*” (page 6)

1. **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 in changes throughout the manuscript (e.g. the abstract “argue” -> “suggest”, page 4 “we argue” -> “our impression is”, page 6 “argue” -> “we therefore conclude”)

1. **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).

1. **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 for the sake of readability, we did not mark this adjustment in the version with tracked changes).

1. **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 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 behaviours. 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: Thank you very much for this positive evaluation. We are particularly happy that you share our impression on the timeliness of the topic. We appreciate your constructive feedback, which we address in detail below.

1. **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 behaviour. 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 behaviours 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: ToDo

* Stefan: literature on impression formation and cross-references
* Todo: include papers in impact of naturalness on perception and behavior
* Maybe interplay with emotion?
* Add some more findings from papers
* Methodological criticism: make the point of ecological validity more prominent?

1. **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. We would like to address this by first clarifying 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 saw 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. Note that in point 3 the editor also explicitly advised us not to base our key arguments on published or unpublished systematic reviews. To make it more transparent to readers that this is NOT a fully systematic literature search, we reworded the first sentence in Box 1 as follows: “*For a more systematic overview on scientific insights into naturalness in voices, we conducted an exemplary literature search on Web of Science*” (page 16)

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 for several reasons: 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 were therefore not tied to the two search terms. This is corroborated by the keywords, showing that only ~20 of the papers included “naturalness” or “human-likeness”. In short, the literature we cover goes well beyond the two search terms. Since the literature we found was already very diverse and hard to integrate, adding more keywords may complicate the matter rather than complementing it. Terms like “voice quality”, “severity” or “clarity” for example could be dangerous because they are used to refer to several distinct concepts in the voice perception literature, potentially returning non-specific and chaotic search results.

1. **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: ToDo

* Point for discussion with Stefan and Sascha

1. **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 in an effort 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 were 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 that concern naturalness, research from all these keywords is somewhat represented in our literature overview. The only exception is “authenticity AND voice”, which picked up 0 of our naturalness papers. This is probably because we made an explicit effort to keep the concepts of naturalness and authenticity separate.

1. **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 wordcloud 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, the terms from the literature were manually extracted by us, 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, 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 promts 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 on 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.

1. **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. Changed as suggested (page 3).

1. **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)

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

Response: Thank you. Changed as suggested (page 7).

1. **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 8)

1. **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 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 9)

1. **"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 the complexity of the topic justice. 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)

1. **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. (pages 10 and 11)

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.

1. **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: ToDo Sascha

* Sascha, wäre das evtl ein Punkt, den du übernehmen könnest?
* Wichtig ist auch hier das Duville 2022 paper, das ist unser einziges mit EEG
* Evtl. Annetts Paper in “voiceness”

1. **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. Evaluations of naturalness influence food choice, environmental preferences, and 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.

1. **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 7)

Indeed, we discussed the definition “*By naturalness, we understand the voice stimulus to be perceived as a plausible outcome of the human speech production system*” (Nussbaum 2023), which is actually from our own previous paper again and we agree with you, that is it rather a combination of the two conceptualizations we introduce here. We therefore recoded is as such, and adjusted the respective information in Box 1 and the associated OSF repository. ToDo

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

1. **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 (for example, 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 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 … ToDo

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 the only way to approach an understanding of everyday experiences of evaluating voice naturalness.

1. **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].” 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].”

1. **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 many of our recommendations sound 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 targeting naturalness research.

1. **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 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 its an analog scale or a 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” [60] , in another one they rated voices on a 10-point-scale from “very natural, human-like” to “very mechanical, robot-like” [52], or made a binary classification of voices as either human or computer-generated [29] .*” (page 5)

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 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 even evidence from face perception that differences in rating scales may not even have a big impact after all [61], but we cannot conclude that this generalizes to naturalness ratings, and the insufficient report of empirical details impedes a meaningful comparison of findings.*” (page 5)

1. **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)

1. **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. And yes, it also implies that naturalness is assessed faster than authenticity – a hypothesis that must be put to the test in the future. So far, to the best of our knowledge, there is no evidence that directly compares the time course of naturalness to authenticity processing. In general, electrophysiological insights into these impressions are sparse. Duville et al. 2022 compared ERP responses to original and naturalness-reduced emotional voices and observed effects of the naturalness manipulation as early as the P200. For emotional authenticity, ERP insights suggest effects as early as the N100 (Conde et al. 2022, Kosilo et al 2021, Sarzedas 2024). On first sight, this seems to contradict our hypothesis. However, all publications on authenticity focused on non-verbal vocalizations, whereas Duville et al. 2022 used emotional utterances. Non-verbal vocalizations have been consistently shown to evoke faster electrophysiological responses than emotional speech prosody (c.f. Chapter 18 “The Electrophysiology and Time Course of Processing Vocal Emotion Expressions” by Paulmann & Kotz in “Understanding Voice Perception”), and thus a direct comparison between these studies is not informative and the question of the time-course is still outstanding.