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:

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:

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that TiCS articles cannot report the results of systematic reviews, nor can key points made in the piece rely on an unpublished systematic review.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

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.**
2. **Please do not number the different sections and subsections.**
3. **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.**
4. **Please include DOIs for any preprints, e.g. [101].**

Reviewer 1:

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.
2. 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?
3. 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.
4. 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 behaviour would be specifically related to naturalness rather than "typicality", or specifically related to voice naturalness rather than syntactic/structural deviations.
5. 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).
6. 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.
7. 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"
8. 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.
9. P6, line 21: ""Does this voice sound unusual" - missing question mark.
10. P8, line 24: "They found that impressions of uncanniness resulted from "deviation from familiar categories" rather "categorical ambiguity"." - should this be "rather than"?
11. 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).
12. "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.
13. Other word choices could be changed for better clarity: "processual" could be "processing" or "process-related" and "restitutes" could be "restores".

Reviewer 2:

It was a pleasure to read this piece. The writing is excellent, and the review offers an 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:

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)?
2. 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.

Reviewer 3:

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. The authors propose that human-based naturalness could be independent from distinctiveness - is there any empirical evidence to support this point?
8. 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?