Comments by the editor:

1. **One big picture note: generally, Reviews should avoid using first-person language. Although this is not a hard rule, the idea is that Reviews should not be centering the perspective of the authors, as this would be more suitable for an Opinion. Please try to revise the highlights and the abstract to place less focus on your perspective. This will give a stronger impression of a Review. For example, “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” --> “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”. You need not contort the language awkwardly to avoid the first person, but please tilt the language away from this, throughout the piece. I’ve edited the introduction section to provide an example.**

Response:

Comments by Reviewer 1:

**Thank you to the authors for their thorough work on revising the paper in response to my previous comments. On this reading I found the paper to be more coherent and impactful. It will be a valuable and timely contribution to the literature, and I look forward to seeing it published!**

Response:

**I just have a few very, very minor comments on wording:**

1. **Page 6 line 20: "In principle, such empirical heterogeneity can be a powerful source of insight". I understood what is meant here, but somehow it didn't flow for me as it's immediately followed by a sentence suggesting that the impact of different scales might be minimal. So perhaps it requires a little more qualification about how heterogeneity might add to understanding rather than limiting it (or having no impact at all).**

Response:

1. **Page 6, line 35: "Finally, few studies only..." --> "Finally, only few studies.."**

Response:

1. **Page 6, line 48: "However, while the scientific findings are well-received within each domain" --> I'm not totally clear on the point that needs to be made here with the term "well-received". I usually take this term to mean that people like/appreciate something, but I'm not sure that can be said here without further qualification about something like the quality of the journals / citation rates in which the work appears. Is it that the work on naturalness is well-cited and/or theoretically impactful within individual research domains? Or more simply that it is appropriate to domain-specific audiences but doesn't translate across disciplinary boundaries?**

Response:

1. **Figure 2 caption: "....marked by the human voice border" --> I somehow just couldn't get this final sentence to scan easily. Are the samples defining the border, or rather (as I suspect) the human's perceptual evaluation of the samples? Would it be better to say something like "Human-likeness can be assessed from audio samples by judging whether or not they lie within the perimeter of an acceptable human voice space (or "human voice border").**

Response:

1. **Figure 3 caption: A suggestion for the penultimate sentence: "... or it could be real or fake in relation to person-related identity information."**

Response:

Comments by Reviewer 3:

**The authors have addressed most of my comments - I'd like to thank them for the thorough revision. There are just a few, minor issues that I'd like to push back on - just for the authors' consideration.**

Response:

1. **Spontaneous evaluation of voice naturalness - in my previous review I mentioned that I'm not aware of much evidence that listeners spontaneously evaluate voices on naturalness as some of the work on this topic does not seem to include judgements of naturalness when listeners are asked to freely describe their first impression of a particular voice. The authors offer a paper by Kuhne et al. (2020) in their response that does focus on naturalness, but it doesn't exactly address my point. In Kuhne et al. perceivers are explicitly asked to rate voices for naturalness or for how human they sound but they do not show any evidence that perceivers would evaluate voices on a dimension related to naturalness unprompted.**

Response:

1. **Face/voice impression models - when discussing the different labels used for the main dimensions of face/voice person perception, the authors distinguish between warmth and competence (Fiske et al., 2007) and the trustworthiness-dominance model of Oosterhof and Todorov (2008). I think it's important to acknowledge that these models are based on different types of cues - while the latter model is based on face-based impressions, the former one is a more general social cognition model, mostly based on text cues. It, therefore, might be factually inaccurate to suggest that the warmth-competence model has been proposed for the underlying structure of face or voice impressions.**

Response:

1. **Deviation-based vs human-likeness-based naturalness - I still struggle a little to differentiate between these two types as one does seem like a version of the other. It is not quite clear to me why the human-likeness-based naturalness has the additional assumption of the existence of non-human voices and that's not true for deviation-based naturalness. Why couldn't the deviation-based naturalness cross the boundaries between human and artificial voices? Perhaps some further elaboration on this point would help clarify the taxonomy proposed by the authors.**

Response:

1. **p.6, just before reference 67 - it is not clear what type of reliability the authors are referring to here**

Response:

1. **p. 6, last row - the authors refer to some highly divergent research traditions, but it is not clear what traditions they are referring to. It seems like they are referring to face vs voice impressions which I wouldn't necessarily call highly divergent, given that a lot of the voice impressions literature is strongly rooted and based on research on faces.**

Response: