Reviewer #3 (Andrey Anikin)

[summary omitted] Together with the theoretical discussion of vowel normalization in the Introduction and General Discussion, there is a lot of good material here on speaker normalization. I found the paper well-written and clearly anchored in the relevant literature. I do have some comments and suggestions, but I would like to emphasize that these are merely my recommendations, which the authors may or may not wish to follow – I don't think any of these concerns should preclude publication.

We appreciate the encouragement, and are glad to see that the reviewer finds merit in this work.

My first question is about why the authors have chosen to present and publish the work before data collection is complete. They repeatedly point out that the corpus should include 24 male + 24 female speakers, but only 23 female speakers have been recorded and analyzed so far. This leaves some gaps in the argument – e.g., when the authors discuss possible limitations of comparing normalization methods for female speakers only (line 902): "it is theoretically possible that the high performance of general normalization accounts... might not replicate when talkers of different genders are included. Future releases of the SwehVd database will contain data from male talkers, which will allow us or other researchers to revisit these questions." It would be strange to publish another paper once all speakers have been recorded, and at the very least, this unusual approach needs to be explicitly justified.

This is a great question. The answer is simply a trade-off between feasibility and informativity. Annotating 210 productions from 48 talkers (for a total of 10 080 productions) is very time-consuming. We therefore decided to publish an initial result with half of the database finalized. We submit that this practice is more common than meets the eye—we are just more transparent about it in the manuscript. Additionally, anyone doubting that the results generalize will need to change *one* line in the code that generates our paper (the line that loads the data), and all plots, numbers, etc. will update accordingly.

We have now finalized annotation for the 24 female talkers, and updated the results. We emphasize that we will complete the annotation of the database. While we do not plan another paper on normalization (unless the results contradict the once published here) but we are working on a paper on the phonetics of Central Swedish vowel productions (beyond the effects of normalization) that will employ the entire database.

Second, in my opinion, the lasting contribution of this work is likely to be the corpus itself and its carefully verified phonetic measurements. The authors say that the audio is available on OSF, but I don't see any audio files there. As for the measures themselves, I'm not entirely convinced that the reported formant values are sufficiently noise-free. The chosen method of analysis (automatic LPC in Praat followed by manual verification of outliers) is common in the field, but in my opinion it is bound to leave a lot of noise in the data. A seemingly reasonable formant configuration returned by Praat may not be an outlier, yet entirely incorrect (e.g., if LPC locks to harmonics in relatively high-pitched vowels). If at all possible, I would suggest checking all measurements manually (in most cases a quick check should do – just a few seconds of work per vowel), or at least the two authors could check a proportion of all recordings and report the agreement between automatic and manual measurements (and inter-rater reliability). A good reference is Whalen, D., Chen, W.-R., Shadle, C. H., & Fulop, S. A. (2022). Formants are easy to measure; resonances, not so much: Lessons from Klatt (1986). The Journal of the Acoustical Society of America, 152(2), 933–941. I also discuss this issue and possible solutions in this preprint, which you absolutely don't need to cite (it's not even published yet):
https://cogsci.se/publications/2023_guideVTL/vtn_2023.01.26_preprint.pdf

Please note that the SwehVd corpus is published on an OSF (https://osf.io/ruxnb/) different from the paper OSF (https://osf.io/zb8gx/). We apologize that this was not clearly stated in the paper, we have now made this clearer.---- (incl. all files for the 24 male talkers, which are recorded but as of yet not annotated).

We agree that formant measures can be noisy, and now alert readers to that possibility. We plan to progressively correct those values on the OSF data (manual corrections will be noted separately, ensuring backward replicability of the published results). Critically, (1) we have no reason to believe that the *random* noise introduced by

automatic annotation would explain our results, and (2) the practice we follow is—as the reviewer also notes—common in the field.

Third, I would recommend streamlining and shortening the paper, particularly with respect to Results and Figures, which strike me as rather "raw". For example, Figures 6-10 are very detailed and descriptive, with a separate F-ratio shown for each normalization method and separately for f0/F1-F3/duration. The main findings, especially of Study 1, could be summarized better – for instance, with a measure of multidimensional clustering quality on all acoustic descriptives simultaneously.

We have restructured the paper and moved Study 1 into the SI. Instead, we now briefly discuss issues with measures like the separability index as part of what used to be Study 2 (now the only study in the main text).

The authors themselves discuss the many limitations of their chosen measure of separability (e.g., line 570) - so why not try another? The classifier and cross-validation in Study 2 are described and extolled at great length, when this is a standard simple problem in machine learning, and any reasonable approach would arguably do just as well (SVM, Random Forest, multi-logistic regression, ...).

We completely agree that cross-validation (CV) is a standard method. However, many researchers in the language, psychological, and cognitive sciences remain sufficiently unfamiliar with it to warrant description.

SVM, RF, and MLR are, however, not substitutes for CV. As we stated on p. 30, these methods *would* be substitutes for the ideal observers we use. But, as stated in the same place, these alternatives would introduce additional degrees of freedom. This makes these methods also less likely to be plausible models for a cognitive mechanism that is meant to *quickly and efficiently* remove talker- and context-specific variability from the speech signal (more parameters require more data to be reliably estimated). **We have slightly revised the wording of this section to make this clearer.** We hope that these changes, together with moving Study 1 into the SI, will address the reviewer's point.

Finally, a minor comment: while it is interesting to find the method of normalization that achieves the best accuracy of vowel recognition, improved performance does not guarantee that the method captures the cognitive mechanisms used by human listeners, so I would be more careful when talking of the normalization procedures and models in Study 2 as "perceptual".

We fully agree, and are now clearer about this. But that is not the same as studying fit against perception. We made (and still make) that very point in the conclusion (former p. 41):

944 Fourth and finally, we followed the majority of previous work and evaluated normalization
945 accounts against production data. This is potentially problematic, especially when measures like
946 category separability or reduced cross-talker variability in category means are used to evaluated
947 normalization accounts (as in our Study 1 and many previous studies). These evaluations essentially
948 assume that the goal of speech perception is to make the perceptual realizations of the same category
949 by different talkers as similar as possible in the normalized space (for an in-depth critique, see
950 Barreda, 2021). However, the goal of speech perception is presumably to reliably infer the category
951 intended by the talker,19 and this aim does not necessarily entail perfect removal of cross-talker
952 variability (as evidenced, for example, by the different findings of Studies 1 and 2).

In the revised manuscript (former) Study 2 is no longer described as "perceptual". We do, however, continue to refer to "models of perception".