Overview of changes

We have uploaded the revisions for our paper “Evaluating normalization accounts against the dense vowel space of Central Swedish”. We thank both the editor and the reviewers for insightful comments on our work. We have revised the manuscript following reviewers’ suggestions.

This has led us to move one of the two studies (Study 1) into the SI. We still refer to it in the main text—both in the introduction and in the discussion—but only where relevant. As the reviewers had pointed out, this study mostly served as a base of comparison for the Study 2 (now the only study in the main text). Together with other cuts (footnotes, redundant passages identified by the reviewers, asides about the phonology of the Central Swedish vowel system), this shortened the main text from 41 pages to 28 pages (~33%).
Additionally, we have followed reviewers’ suggestion to a) anticipate our main take home points more clearly in the abstract, introduction, and discussion.

Finally, we decided to switch the evaluation of accounts from Luce’s choice to the criterion choice rule. This does not change any of the results but is more transparently interpretable (for those who are familiar with those choice rules).

Response to Reviewers' comments

We respond in blue. Sometimes we added **boldfacing to reviewers’ comments** in order to emphasize what points we are corresponding to. We hope this is helpful.

Reviewer #2

[summary skipped] The study's design and execution are commendable, demonstrating its robustness. However, the presentation and discussion of the findings leave room for improvement. Addressing these limitations could strengthen the paper and warrant its consideration for acceptance.

We thank the reviewer for the encouragement and helpful feedback below.

Strengths
1. This study compared two evaluation methods (separability index vs. ideal observers) with the same database which effectively illustrated the pros and cons of each method in a fairer way.
2. In addition to the commonly used F1 and F2 cues in vowel perception studies, this research incorporated more acoustic cues, especially duration, which had not been explored in previous works. This broadens our understanding of how normalization mechanisms operate in both frequency and time domains.
3. The study evaluated not only the contribution of individual cues but also the combined effects of different cues, providing a comprehensive analysis.
4. The authors have made their database and code publicly available, enabling other researchers to easily test their claims and further fostering collaboration and knowledge exchange within the field.

Limitations
1. The authors should consider reorganizing the article in a more concise and formal manner. For instance, there is no need to repeatedly mention data availability. The description of the SWEHVD DATABASE is redundant, and certain literature review content should be placed in the INTRODUCTION rather than the methods section. It would be beneficial to include a "the present study" subsection within the INTRODUCTION to outline the research plan and structure of the study. Please refer to the detailed comments below.

We have restructured the manuscript by moving Study 1 into the SI. This was in part motivated by comments from R3. The result is a much streamlined paper that still aims to err on the side of transparency. We have also followed many of the helpful suggestions by this reviewer.

2. Format: It is advisable to organize the manuscript in a conventional manner, with sections for introduction, materials and methods, results, and discussion.

Thank you. This is now the case.

3. The authors did not appear to specify the weight of each cue contributing to vowel normalization, assuming that each cue contributes equally to the results. It would be interesting to determine an optimal weight for each cue in the normalization models that incorporate all acoustic cues.

We intentionally chose to use *optimal* cue weighting. This was only briefly —and too indirectly— mentioned on p. 30. We have revised the method section to state it more clearly: by using multivariate Gaussian ideal observers, we provide our perceptual models with an estimate of the full joint distribution of all cues (that are included in a particular instance of the model). This *implicitly* weighs the cues optimally.

The reviewer might have been thinking of models that assume *independence* of cues, and then use *cue integration* with *cue weighting* to obtain the posterior of each category (e.g., Toscano & McMurray, 2010, using the model introduced in Jacobs, 2002). Independence of the different formants (F1, F2, …) is, however, *clearly* not warranted (see e.g., Fig 5 in the old manuscript). This, of course, does not mean that *listeners* could not rely on this (false) assumption. The present paper —much like the gross of research on normalization— cannot assess this point. We have thus chosen to continue to model predictions under optimal use of all cues (which is even 'more optimal' than optimal cue integration).

4. The variability of speech materials is limited, as only female talkers of one regional variety of Central Swedish were included in the database. As acknowledged by the authors, this does not provide enough variability, making it difficult to determine whether the findings remain valid when considering factors such as gender, age, and regional variability.

We agree (which is why we stated the caveat in the discussion), minus the word "enough". We would like to provide some context though. Normalization accounts were developed to address *physiologically* caused variability (incl. Talker variability). So, they were never intended to account for regional variability (for excellent reviews, see Barreda 2020; but also Nearey's "probabilistic sliding template model" paper). Age and gender are factors expected to affect physiology but, of these, only the latter (gender) is systematically varied in most research on normalization. So, yes, our study is based on data that exhibits *one* source of physiological variability less than most previous work.

Please see my detailed comments below for the evaluation of different parts of the manuscript including the validity of the methods, results, and data interpretation.

The detailed comments

ABSTRACT
1. line 9-11: it is not necessary to state the data variability in the abstract since this information is already included in the Data availability statement section. Agreed. We have removed this part.
2. It will be good to add a brief summary of key findings to the Abstract. We agree, it is now added.

INTRODUCTION
1. line 29: Sjerps et al. (2019) found the normalized vowel representations in the auditory cortex. That does not mean the talker-normalized speech can be decoded from areas as early as the brain stem. Thanks for catching that. We meant to (and now do) cite Skoe et al. 2021.
2. line 31-34. Here the authors identified a research gap: the specific nature of the operations involved in normalization and stated that this study will contribute to this line of research. However, in the Discuss part, the authors did not address this question in a clear way. We agree and have made additions to discussion.

3. line 95-116. Move this part after line 140 and integrate it with line 141-161. Line 117-line 140 are all about how to evaluate different methods in table 1 and it will be better to present this part close to Table 1. Line 95-116 and 141-161 are about Swedish vowels. Done.
4. line 162-167. Move this part to the end of the introduction or discussion where the authors addressed the limitations. Now moved to introduction.

The SWEHVD DATABASE
General Comment: The authors are advised to simplify this section, as the primary focus of the study is to evaluate various normalization methods rather than to present a database. A concise description of the Swedish vowel inventory, along with information about the included vowels, recording process, and preprocessing of the recordings, should suffice. We have now moved some subsections of the SwehVd description into the SI, and cut out other parts of the characterization of the vowel space in SwehVd.

STUDY 1
1. line 379-393. This part should be moved to INTRODUCTION where the authors first mentioned the evaluation method of reducing inter-talker variability. Done.
2. line 406-409. Integrate this part with line 396-402. Done.
3. line 421-424. The authors should remove this section, as similar information has already been addressed in the INTRODUCTION. It would be more appropriate to emphasize these points in the DISCUSSION section. Moved to discussion.
4. line 426-432. This should be included in INTRODUCTION. This is now moved to the SI (where this study is now presented as an auxiliary study).
5. Figure 8 caption: change "… and long and short vowels together (columns)…" to "…and all vowels (columns)..." Done.
6. line 586-592. Separability index conceptually is to reduce cue variability around the category mean. Therefore, if the primary interest of this study is to assess the expected consequences of normalization for perception, as stated here, there is no need to carry out Study 1. The authors may consider rephrasing this part. Study I is moved to SI, and we have adjusted text accordingly.

STUDY 2.
1. page 28 is empty. Footnote 13 should be on page 29.
2. line 602-638. The review of previous studies should be presented in INTRODUCTION. Done.
3. line 653-656. This should be presented in DISCUSSION or CONCLUSION. Moved to a footnote in Discussion.
4. Figure 12 caption. change "… and long and short vowels together (columns)…" to "…and all vowels (columns)..." Done.
5. Figure 13. Why choose these vowels? Although the authors stated in line 784 that …five vowels that illustrate some of the vowel-specific effects across different types of normalization, they only explained vowel [i:] in this part. Excellent point. We have decided to remove this plot from the main paper. We refer the reader to the complete vowel specific plots in SI but keep a summary of main findings in the main paper.

GENERAL DISCUSSION
1. General Comments: This study incorporates several unique design elements, such as using two evaluation methods to assess different normalization accounts and employing Swedish, a language with a dense vowel inventory of 21 vowels varying in both quality and quantity. However, these distinctive aspects have not been thoroughly discussed. The authors may want to consider elaborating on how these specific design choices contribute to the study's goals or the conclusions derived from them. For example:
(a). Compare and contrast the two different evaluation methods (Separability index vs. predicted recognition accuracy). Although this has been mentioned elsewhere in the manuscript, it would be beneficial to include it in the GENERAL DISCUSSION section.
(b). Explain how the Swedish vowel dataset has helped reveal findings that were not reported in previous studies.
(c). Discuss the role of duration (a cue that has never been included in comparisons of normalization accounts) in vowel normalization within this study.

2. line 809-810. It would be good to move the discussions about Study 1 and the comparison of two methods of evaluating normalization to GENERAL DISCUSSION. Besides, the sentence “we focus here on the discussion of Study 2” is misleading. It indicated that the discussion below was about the findings in Study 2. Actually, the authors also included the findings in Study 1.

Excellent points. We have restructured and edited the general discussion to include these aspects.

3. line 823-824. “languages with dense vowel spaces ……complex normalization mechanisms” is a little bit overstated. As the authors suggested that they limited evaluation to a single level of normalization, if more levels or other stages of normalization were included, languages with dense vowel spaces may require a more complex normalization process.

Here, the point we intended is that simple mechanisms are sufficient to achieve substantial increases in normalization accuracy. If a single level of simple mechanisms can achieve high accuracy, so can multiple levels of simple mechanisms. Since we already state the caveat—shared with preceding work—that we only consider individual accounts, we feel that readers can draw their own conclusions as to whether it is likely that consideration of combined accounts would show a benefit of more complex accounts for languages with dense vowel spaces.

4. line 846-854. It seems two advantages are derived from previous studies. Therefore, they are more like the reasons why the authors chose this method and should be put in INTRODUCTION but not DISCUSSION. It will be good if the authors can illustrate that how this study showed the advantages of this method in DISCUSSION.

We have moved the motivation of choice of method to the introduction.

5. line 856-858. No need to repeat this point. This point has now been removed.
6. line 874-875. Delete “Next, we close by discussing additional limitations of the present work and future directions.” Done.