**Authors’ response to Reviewer/Editor critique**

**Overview of changes**

We have uploaded the revisions for our paper “Comparing accounts of formant normalization against US English listeners’ vowel perception”. We thank both the associate editor and the reviewers for insightful comments on our work. We have revised the manuscript following the reviewers’ suggestions.

We are glad to see that both reviewers see that there is merit to this work. We briefly raise two main points here…

…

**Response to reviewers’ comments**

We respond in blue below.

***Associate editor***

Figures 1, 7, and 10. Lobanov used z scores. Is it still appropriate to use the unit “Hz” here?  
Around line 900: Should we expected the models to provide a better fit for later behavioral trials?  Did that happen?

This is a good point. Generally, we would expect models to provide a better fit for later behavioral trials. However, we expect relatively little signal here as each data point is quite noisy.

Thanks for catching these issues. We have now fixed all of them:  
Line 217: This instance of “ii” should be “iii”.

Line 386: Citation uses the last name only.

Figure 5, panel B: Numerals in yellow are very hard to see on a white background.

Line 537: “... to these data”?  
Line 654: I might put the text starting “i.e.,...” inside parentheses.   
Line 988: I think this parenthetical (??) should be omitted.

***Reviewer #1***

[summary omitted]

This is a classical problem in speech perception, with many accounts but insufficient empirical support to distinguish them. This study is an important step in the right direction. It uses an innovative approach based on Bayesian ideal observer models to formulate mappings of acoustic properties onto listener responses under different normalization schemes. The experiment and modeling are carefully designed and executed, and the overall conclusions seem reasonable. However, some caution is needed in framing the conclusions, as much more data is needed to assess the generality of the findings. There are discrepancies in the findings of the two experiments, one with natural speech produced by a single talker, the other with synthesized speech spanning a range of formant patterns and interpolating between the natural productions. The proposed explanation for the discrepancies seems plausible, but more data is needed to test the account using a wider range of stimuli to confirm the assumptions involved, together with a more careful matching of the dialects of the speakers and listeners. In addition to that general recommendation, the comments below include minor corrections and clarifications of the text.

It is assumed throughout that the information extracted from vowels in human perception can be adequately described by the formant pattern. There is a great deal of support for this position. It is not necessary to review this literature in detail, but it might be important to acknowledge that there are other perspectives that do not assume that listeners rely on formant frequencies (e.g., see Hillenbrand JM, Houde RA, Gayvert RT. Speech perception based on spectral peaks versus spectral shape. J Acoust Soc Am. 2006 Jun;119(6):4041-54. doi: 10.1121/1.2188369. PMID: 16838546). It should be noted that the problem of cross-talker variability is not resolved by adopting alternative “whole-spectrum” representations in place of formant frequencies; arguably, the problem is made even more difficult with spectral representations that do not separate formants from harmonics.

p. 1, lines 10-12: the terms *extrinsic* and *intrinsic* normalization need to be defined and introduced together, showing how they contrast (e.g., as in Nearey 1989).

p.1, bottom paragraph: the modeling work of Roy Patterson and colleagues deserves mention here, as an example of early / low-level auditory computations that may be engaged in talker normalization (e.g., Smith DR, Patterson RD, Turner R, Kawahara H, Irino T. The processing and perception of size information in speech sounds. J Acoust Soc Am. 2005 Jan;117(1):305-18. doi: 10.1121/1.1828637. PMID: 15704423; PMCID: PMC2346562).

p. 2, bottom, and elsewhere: the Mel scale is widely used by speech researchers (e.g., in Mel cepstral representations). However, as the authors correctly note, some studies have pointed out substantial problems with the original Mel scale proposal. My impression is that the Mel scale is no longer used (or is rarely used) in hearing science and psychoacoustics (for example, recent reviews of pitch perception in hearing do not cover this topic). It is also worth noting that the Mel scale was proposed as a model of human *pitch perception*, while the Bark and ERB scales were developed to model human auditory *frequency selectivity*; and the semitone scale is generally linked to musical pitch. It would be helpful for readers to point out these differences in modeling aims. These clarifications seem especially important in light of the conclusions reached in the Results section, top paragraph of p. 39.

p. 3, lines 79-81: “While such intrinsic accounts arguably entail more computational complexity than static transformations …” Confusing sentence – what does “static” mean here? This needs clarification.

p. 11, line 201: “random guessing”. Could there be an intermediate level between these two alternatives, not entirely random, but only based on incomplete or inaccurate assessment of the acoustic properties?

p. 14, bottom line: Were the parameters taken from Wade et al. 2007 similar to those of the talker in Experiment 1a?

These values were extracted from the \*had\* token used for resynthesis (and thus from the talker in Experiment 1a), and not taken from Wade et al. 2007. This is now clarified in the text.

p. 19, lines 337-338: "acoustically similar" may not be the best metric for this comparison; "auditorily similar” is what matters. A difference of 30 Hz in F2 may not be discriminable, while a 30 Hz difference in F1 likely would be (e.g., studies by Kewley-Port and colleagues). A quick check might be to make the same comparison in log Hz space.

p. 35: “For example, a model can exhibit high correlations with listeners’ responses even when its predictions are systematically ‘off’.” Can you give an example that might produce this outcome? Two lines below: “sufficiently much” - awkward (and vague) phrase.

p. 41, lines 699-700: “Figure 10 also shows how well accounts fit listeners’ responses for each test stimulus (opaqueness of the black points).” The shading is actually rather hard to see in the figure.

Thanks for catching these issues. We followed all these suggestions:

p. 1, line 17: Perhaps “suboptimal” might be a better wording than “inadequate”.

p. 6, bottom line: “all 8 monophthongs of US English”. The phrase “all 8” is potentially misleading. First, the set of US English monophthongs frequently includes /e/ and /o/, which generally exhibit substantial formant movement over their time course (vowel inherent spectral change, or VISC). But other monophthongs also exhibit VISC, and many studies have shown significant effects of VISC on perceptual judgments. Moreover, the number of monophthongs can vary as a function of dialect (e.g., /ɔ/ - /ɑ/ back vowel merger in Canadian English). Some studies also include /ə˜/ (“herd”) as a monophthong, so potentially there are 12 monophthongs in some dialects of US English. My recommendation is to drop the word “all” from this sentence.

p. 8, line 154: OSF, OSF repo, SI, R, etc.: be sure to define acronyms on first use. Not all readers will know what these refer to.

p. 10: “One consequence of this is that the formant values of these recordings are clustered around the category means, and thus span only a comparatively small part of the phonetic space”. Means across what observations? Stimulus sample or population? Is this assumed or based on actual measurements? In the next sentence, in the phrase “potential secondary cues”, the word "potential" seems ambiguous/unnecessary. It could mean "not well established in the literature", "varying in potency", "not always active or present (e.g., F0 in whispered vowels)". In the list of cues, consider including VISC.

p. 12, line 220: “optional post-experiment survey” How many participants opted not to complete the survey? (apologies if this is answered elsewhere in the paper, but I did not see it).

p. 13, section 2, Materials: I found it difficult to follow the description of the synthesis method. Since the study used a unique synthesis method, it is critical to provide sufficient information to readers to permit the study to be replicated. For example, “the /h/ sound [was filtered] inversely with its LPC, and concatenated … with a complex waveform generated from the pitch and intensity patterns of the original vowel”. This description is too general.

p. 14, middle of paragraph: “The bandwidth manipulation implied that formants became stronger as the vowel unfolded”. “Stronger” is unclear; did you mean that narrower formants produce more intense spectral peaks? This is hard to see in a spectrogram display.

p. 18, Fig. 4 caption: “F1-F2 combinations below the gray dashed line are articulatory unlikely to come from the same talker.” Not sure what this means.

p. 20, line 345: insert “spectral” in front of “tilt”

p. 24, line 422: “A welcome side effect of this is that far fewer degrees of freedom”. Expand what you mean by “this” (several things are discussed in the previous paragraph). “Far fewer” compared to what?

p. 25, Fig. 6: bottom line, the label “phonetic properties of stimulus (formants)”. Should the word “phonetic” be replaced with “acoustic” (or perhaps “acoustic-phonetic”)? Are you using the term “phonetic” to mean “perceptually relevant”?

p. 34, line 588 change “All result” to “All results”   
p. 36, Fig 9 caption: “Pointrange” -> “Point range”   
p. 43, lines 733-734: “researchers ought to adapt uniform scaling as our working hypothesis” Change “our” to “a”.   
p. 44, line 759: “the recognition of less categorically perceived consonants” Do you mean “less” or “more”?

p. 44, line 765: “Kronrod at al” => “Kronrod et al.” Replace “formants” with “formant frequencies”

p. 46, line 794: humans can hallucinate, but it is not clear that models can. Perhaps use another term here.

p. 46, line 796: listeners’ => listeners

p. 48, lines 845-847: Perhaps move this to a footnote?

p. 51, line 900: “us an” => “us as an”

p. 51, lines 910-912: “normalization accounts that best describe listeners’ perception share that they (1) learn and store talker-specific properties and (2) that they seem to be computationally very simple”. Drop the second occurrence of “that they” (it is provided before the first point).

***Reviewer #2***

[summary omitted]

In my opinion, the work, though very interesting and state-of-the-art in many regards, has some rather serious flaws that lead me to conclude that the authors should go back to the drawing board and produce a more definitive study. I think there should be new or expanded perception data against which to compare models, a better training set upon which to build the models, and a change in the way response bias is handled in the models.

1. Was the training data set adequate? I think that it may not have been. Based on the information on p. 29, line 496, the average number of vowel tokens that is used to estimate the extrinsic normalization parameters in the model fits is (10 \* 8) / 5 = 16 tokens per talker (about two tokens per vowel) for 5 female speakers and 12 male speakers. Thus, for each fold, the extrinsic normalization parameters were estimated using a relatively small number of vowel tokens. The authors should at a minimum report how variable the normalization parameters were, and should consider normalizing prior to splitting the set into folds. But aren’t there larger datasets that one could use, where this type of concern wouldn’t arise.

We appreciate this point but note that it is, in fact, already addressed in the paper. E.g., **Fig 7 continues to visualize the variability across the different folds**, and continues to include in its caption “*The relative stability of the category ellipses across training sets indicates that the database is sufficiently large for the present purpose.”* **Additionally, Figure S8 in the SI continues to visualize the cross-fold variability in the normalization parameters.**

The SI was shared on OSF, as indicated when we first mention the SI “*All stimulus recordings, results, and the code for the experiment, data analysis, and computational modeling for this article can be downloaded from OSF at https://osf.io/zemwn/. The OSF repo also include extensive supplementary information (SI).”*

Finally, we mention that this is the first paper that makes it *really* *easy* to test concerns like those by the reviewer: as we wrote in our Open Science statement in the manuscript, researchers can use a different database of vowel productions and rerun the entire code we have developed. This might require as little as a single line of code change.

2. Was the test set adequate? There were two problems with the listener data against which the models were tested. One is that both experiments used stimuli from a single female talker. Although this is fine as far as it goes, the conclusions would have been stronger if the models had been tested against listening results for several talkers. The other problem with the test set is that the test stimuli in experiment 1b include impossible vowels in the sense that vowels had formants outside of the range of those pronounceable by a single talker, and that the synthetic vowels did not exhibit duration or formant dynamic properties that should be associated with several of the possible response alternatives. Given that listeners were not very consistent in how they labeled these synthetic vowels, it isn’t clear what the models were capturing.

We agree with the reviewer that these are potential concerns. This is why we continue (as in the original manuscript) to transparently acknowledge them in the manuscript. These limitations are, of course, not unique to our work: as we continue to mention in the introduction, previous work has often only investigated small parts of the vowel space (often while offering only a small subset of response options). Previous work has also often focused on vowel-only stimuli (which are rarely observed in real life), or has used non-constant lexical context. In short, any test set comes with limitations. Going beyond previous work, the present work presents two test sets, each of which covers a larger part of the formant space and vowel inventory than most previous studies.

Additionally, the SI reports additional subset analyses (references in the main text) that make sure that e.g., the results of Experiment 1b are not solely driven by the parts of the vowel space that are unlikely to come from the same talker as the rest of the vowel space (we respectfully point out to the reviewer, that these tokens are not “impossible”; they are just unlikely to come from the same talker as the rest of the tokens).

Finally, we note that decreased consistency is \*expected\* for Experiment 1 and any other experiment that does not solely present recordings of hyper-articulated prototypical vowel tokens. And critically, this is not a weakness but a strength: adequate models of normalization need to capture human perception not only for prototypical vowel instances but also instances of vowels that fall between the category means.

3. Were the models adequate? The main problem with the models is that the bias terms in the model were static and equal for all vowels. At first blush this seems like a fine modeling assumption - the task presents 8 response alternatives on the screen in each trial, leading to a demand characteristic for listeners that the 8 alternatives are equally likely. However, it is well-known in the vowel perception literature that there are strong order effects in vowel perception leading to dynamically changing response bias (references below). Given the observed magnitude of context effects in perception experiments, and the high degree of uncertainty engendered particularly by exp 1b, and low accuracy of even the best models of listeners in exp 1b, I think that the results for simulations of 1b are not reliable.

This is another great observation, and one that we had not discuss in the paper. Critically, **presentation order was randomized in the experiments** (as we continued to state on L281). This means that order effects are expected to average out across participants. **We now state this more clearly and cite Repp & Crowder (which contains references to other works) in that context.**

Of course, this means that some of the unexplained variance in participants’ responses might well be due to stimulus order (same as for a large number of other known effects on vowel categorization). Critically though this would constitute a form of statistical noise, not bias.

Repp, B., Healy, A. F. & Crowder, R. G. (1979), Categories and context in the perception of isolated steady-state vowels, Journal of Experimental Psychology: Human Perception and Performance, 5, 129-14

Cowan, N. & Morse, P. A. (1986), The use of auditory and phonetic memory in vowel discrimination, Journal of the Acoustical Society of America, 79, 500-507.

Repp, B. H. & Crowder, R. G. (1990), Stimulus order effects in vowel discrimination, Journal of the Acoustical Society of America, 88(5), 2080-2090.

4. Why do we reach different conclusions in experiments 1a and 1b? A couple of additional ways of looking at the data may help us understand this better.

a) To understand the variable performance of the extrinsic normalization methods, add a table showing their parameters (e.g. mean ln(F) for Nearey Uniform Scaling, …. mean(Fn) and sd(Fn) for z-score normalization) for the stimulus sets in exp1a and exp1b. Is it the case that extrinsic methods that had very little change in the normalization parameters from 1a to 1b were better at modeling listener behavior? Or were methods that were sensitive to the change in talker formant range thus better at modeling perception?

That’s an interesting idea! The suggested table is provided in the SI in the form of Figure S8. **We now discuss this point in the SI and in the results section (referring to Figure S8).**

b) Report the model fits using an interpretable parameter. Because log likelihood is a function of the number of observations in the dataset, the values being reported for exp 1a and 1b are on incomparable and unintuitive scales. Given that the ASP model (fig 6) produces a categorization response for each stimulus, it should be possible to measure model success in terms of the proportion of trials for which the model prediction matched the listener response. If my back of the envelope calculation is correct, the ASP model correctly predicts listener behavior on about 32%[[1]](#footnote-1) of trials in exp 1a with no normalization and about 41% of the time using the best normalization model. For exp 1b it looks like the model predictions are correct about 15% of the time with no normalization and 22% of the time with the best normalization model. One conclusion that I would draw from the model fits for experiment 1a and 1b is that the listener responses for exp 1b are not very modelable with this model architecture.

**While categorization accuracy is a more intuitive measure of fit, it is well known to be a more problematic measure of fit.** It makes no sense, for example, to compare the categorization accuracy across experiments: because the two experiments differ in where in the formant space they elicit responses, theories of speech perception *predict* that the stimuli in the two experiments are categorized with different accuracy (as is indeed the case). This weakness of accuracy as a measure of model performance remains under-appreciated in research in speech perception. So, we definitely would not want to contribute to this issue by committing the same mistake.

We note though that the categorization accuracies of models were much higher than the reviewer guestimated. **We now state so briefly at the end of the results section, where we summarize the categorization accuracy of the best performing normalization account for Experiments 1a and 1b (XXX% and XXX%, respectively), relative to the accuracy when no normalization is used.**

Specific comments: [1] means ‘line 1’

[121] - Earlier studies of the perception of synthetic steady-state vowels should be cited. While these don't explicitly address different normalization methods, they do help calibrate the level of success that we should expect for a perceptual model that only includes steady-state formants.

Lehiste, I., & Meltzer, D. (1973). Vowel and Speaker Identification in Natural and Synthetic Speech. Language and Speech, 16(4), 356-364. https://doi.org/10.1177/002383097301600406

Hillenbrand, J. & Gayvert, R.T. (1993) Identification of steady.state vowels synthesized from the Peterson and Barney measurements. J. Acoust. Soc. Am; 94 (2): 668–674. https://doi.org/10.1121/1.406884

[172] – At first blush it seems like such an odd choice to test the validity of techniques for dealing with between-talker variation using a within-talker experiment. It might be worth a comment regarding this choice.

Good idea. **We explicitly chose a single talker b/c it allowed us (to a first approximation) to disentangle two separate problems**: (1) the problem of how to normalize the input from a talker (which, at least for extrinsic accounts tends to be the context that is assumed to be the one over which extrinsic information is accumulated), and (2) the problem of recognizing when a talker switched (e.g., Magnuson & Nusbaum, 2007, among others). **We now state so clearly at the start of the section in Experiments 1a and 1b.**

[216] – also note which of the facts about participation are reported based on self-report?  
The text first describes objective criteria (not self-reported) and then uses the phrase “participants had to confirm that …” for any criteria that were self-reported.

[316] – “stimuli that were predominantly categorized as /u/” – same stimuli? or same measured formants at vowel midpoint?

Figure 4 – I believe that Lehiste and Meltzer also found that [ae] and [a] were relatively well perceived in steady-state synthetic vowel stimuli. Any thoughts about why this might be?

[361] – can you report the regional dialect of the speaker? Listeners?

Unfortunately, the information we have about the speaker is limited to whatever information is available in the original database (recording by Dr. Xin Xie; reported in Xie & Jaeger, 2020). For our web-based listeners, we refrained from having them report their dialect since such reports are notoriously unreliable if based on only self-reports.

Figure 5 – Does 0.5 on the x axis in figure 5B mean that the participant was equally likely to call the stimulus [I] or [E]?

Basically, yes. It means that this participant was about equally likely to respond [I] or [E] across all of the vowel recordings that the majority of participants heard as [I].

[417] – Having just said that dialect matters, this dialect-free formulation is inapt. Would it be possible to train on a dialect-matched database?

We share some of the sentiment evident in the reviewer’s point. That is why we---unlike all previous work---investigated and highlighted this issue. However, we wish it was as easy as “dialect-matching” the database. Match based on what? Self-reported dialect? Or some (non-trivial to obtain) objective estimate of the dialect template that the speaker/listener actually seems to use? And why stop at dialect? Different sociolects might well employ vowels, too. So, what we are pointing at here is a complex issue that will require careful thought to address. **We now briefly point to these complications when we first introduce the ASP architecture, and refer readers to the general discussion. We also expanded somewhat on this point in the general discussion.**

[441] – In vowel classification studies, if classification rates are greatly improved with a richer feature set (e.g. x = [F0, F1, F2, F3, dur]). This is a form of model-acquired intrinsic normalization, which is of great theoretical interest. The paper would be strengthened by the inclusion of at least one model with a richer feature set.

We agree. This is why SI continues to include analyses that use F1-F3 (Section 3E). We also considered models of F1-F3, plus duration. These analyses continue to be referenced on p.26 in the main text. **We have, however, revised the text to be clearer that (1) these additional analyses indeed find that some intrinsic accounts improve when F3 is also considered, but (2) the best-fitting accounts are the same types of extrinsic accounts as for F1-F2.**

[495] – a very male dominated database (12 men, 5 women). Why not use something like the Hillenbrand database?  
Please see main part of letter above.

[509] – So the average number of vowel tokens that is used to estimate the extrinsic normalization parameters is 10/5 \* 8 = 16 tokens per talker? I would like to know how variable the theta are for the different training sets.  
See above. The SI contains this information. And Figure 7 also shows how little variability there is across folds.

[524] – Much larger number of stimuli going into the calculation of the theta for this talker (natural 9\*8 = 72; 146 stimuli for synthetic stimuli). How much different are the theta distributions for natural versus synthetic (with not humanly possible vowels)?  
The SI contains this information (Figure S8).

[573] – Given that in Exp 1a each vowel category was presented equally often, could you estimate response bias for Exp 1b from the responses in 1a?

[631] – Is there some more intuitive way to represent the degree of fit between model prediction and human response? How are we supposed to intuitively get a grasp on what -2284 means relative to -9626. These seem to be very different.

Unfortunately, sometimes the only available intuitive measures are misleading. Consider, for example, the long-standing discussion of ‘intuitive’ data transforms for reaction time analyses that has been used to justify the use of intuitive but inadequate models of analysis (reviewed in Burchill & Jaeger, 2024); or the long-standing discussion that linear regression or ANOVA is more intuitive for analyses of proportions---alas it is a bad choice, compared to less intuitive but well-formed approaches like logistic regression (reviewed in Jaeger, 2008).

We note though that **the text states the both the likelihoods for chance-guessing (floor baseline) and the best-possible strategy of exactly mirroring the probability distribution of listeners (ceiling baseline).** The likelihood numbers in the figure can be meaningfully interpreted relative to those baselines. We have now also added these numbers into the figures.

[664] -- I wonder if the impact of log transformation is similar to the increased reliance on higher formants that is found in Lammert & Narayanan (2014) estimation of vocal tract length? Per Johnson (2021), this is relevant in the case of a model that requires extrinsic VT length estimation from a small number of tokens, as may be the case here.

Thank you for raising this point. **This is indeed what we assume**--following reasoning initially developed in T. Nearey’s work (and echoed in Johnson, 2021). **We now briefly clarify this in the general discussion.**

Thanks for catching these issues. We followed all these suggestions:

[208] -- note here how many people were excluded.  
[227] – Since dialect figures in the interpretation of the data later, you should report as much as you can about the dialect spoken by the speaker for this experiment. Please note that “NE US English” isn’t very limiting. Was she from Boston, New York, Maine, Buffalo? It matters.  
[241] – Stimulus construction - when it is said that the final /d/ was concatenated onto the vowel, do you mean the /d/ burst? /d/ voiced closure + burst, or /d/ transition, closure and burst?  
[241] – Fig 3 indicates that vowel formants were held steady until the final consonant transition. Please confirm that this is so.

Figure 3. Since confusions between words is an issue in discussing the perception results, please show all eight test words in natural and synthetic versions.  
[247] -- Regarding the narrowing of bandwidth over time in the stimulus, remind us why Wade et al. did this? It is a bit vague to describe the effect of bandwidth narrowing as making formants ‘stronger’. Perhaps point out that narrowing bandwidth results in higher amplitude spectral peaks, and greater separation of peaks,

1. P(c) = exp(ln(L)/n). Thus for ln(L) = -2900 for the no normalization case, and n = 2565, the estimated proportion of correct model predictions is exp(-2900/2565) = 0.32 [↑](#footnote-ref-1)