# SUPPLEMENTARY INFORMATION FOR *PERSSON, BARREDA & JAEGER (2024). COMPARING ACCOUNTS OF FORMANT NORMALIZATION AGAINST US ENGLISH LISTENERS' VOWEL PERCEPTION*

## §1. REQUIRED SOFTWARE

Both the main text and these supplementary information (SI) are derived from the same R markdown document available via https://osf.io/zemwn/. It is best viewed using Acrobat Reader. The document was compiled using `knitr` in RStudio with R:

```
##                _
## platform       aarch64-apple-darwin20
## arch           aarch64
## os             darwin20
## system         aarch64, darwin20
## status
## major          4
## minor          4.0
## year           2024
## month          04
## day            24
## svn rev        86474
## language       R
## version.string R version 4.4.0 (2024-04-24)
## nickname       Puppy Cup
```

Readers interested in working through the R markdown, and knitting it into a PDF will also need to download the IPA font SIL Doulos and a Latex environment like (e.g., MacTex or the R library `tinytex`).

We used the following R packages to create this document: R (Version 4.4.0; R Core Team, 2024) and the R-packages *assertthat* (Version 0.2.1; Wickham, 2019), *brms* (Version 2.21.0; Bürkner, 2017, 2018, 2021), *Cairo* (Version 1.6.2; Urbanek and Horner, 2023), *cmdstanr*

(Version 0.7.1; Gabry *et al.*, 2024), *dplyr* (Version 1.1.4; Wickham *et al.*, 2023), *ellipse* (Version 0.5.0; Murdoch and Chow, 2023), *forcats* (Version 1.0.0; Wickham, 2023a), *furrr* (Version 0.3.1; Vaughan and Dancho, 2022), *fuzzyjoin* (Version 0.1.6; Robinson, 2020), *ggforce* (Version 0.4.2; Pedersen, 2024a), *ggh4x* (Version 0.2.8; van den Brand, 2024), *ggnewscale* (Version 0.4.10; Campitelli, 2024), *ggplot2* (Version 3.5.1; Wickham, 2016), *ggtext* (Version 0.1.2; Wilke and Wiernik, 2022), *knitr* (Version 1.47; Xie, 2024), *linguisticsdown* (Version 1.2.0; Liao, 2019), *lubridate* (Version 1.9.3; Grolemund and Wickham, 2011), *magrittr* (Version 2.0.3; Bache and Wickham, 2022), *mgcv* (Version 1.9.1; Wood *et al.*, 2016; Wood, 2003, 2004, 2011), *modelr* (Version 0.1.11; Wickham, 2023b), *MVBeliefUpdatr* (Version 0.0.1.10; Jaeger, 2024), *nlme* (Version 3.1.164; Pinheiro *et al.*, 2023), *patchwork* (Version 1.2.0; Pedersen, 2024b), *phonR* (Version 1.0.7; McCloy, 2016), *phonTools* (Version 0.2.2.2; Barreda, 2023), *plotly* (Version 4.10.4; Sievert, 2020), *purrr* (Version 1.0.2; Wickham and Henry, 2023), *Rcpp* (Version 1.0.12; Eddelbuettel *et al.*, 2024), *readr* (Version 2.1.5; Wickham *et al.*, 2024a), *remotes* (Version 2.5.0; Csárdi *et al.*, 2024), *RJ-2021-048* (Bengtsson, 2021), *rlang* (Version 1.1.4; Henry and Wickham, 2024), *stringr* (Version 1.5.1; Wickham, 2023c), *tibble* (Version 3.2.1; Müller and Wickham, 2023), *tidybayes* (Version 3.0.6; Kay, 2023), *tidyr* (Version 1.3.1; Wickham *et al.*, 2024b), and *tidyverse* (Version 2.0.0; Wickham *et al.*, 2019).

If opened in RStudio, the top of the R markdown document should alert you to any libraries you will need to download, if you have not already installed them. The full session information is provided at the end of this document.

## A. Interested in using R markdown do create APA formatted documents that integrate your code with your writing?

A project template, including R markdown files that result in APA-formatted PDFs, is available at `https://github.com/hlplab/template-R-project`. Feedback welcome. We aim to help others avoid the mistakes and detours we made when first deciding to embrace literal coding to increase transparency in our projects.

## §2. ADDITIONAL INFORMATION IN EXPERIMENTS 1A AND 1B

### A. Participant exclusion

We adopted the following exclusion criteria: participants would get excluded if the failed to pay attention to the instruction to wear over-the-ear-headphones, if they had unusually slow or fast RT-means compared to other participants, or if they clearly did not do the task (e.g., randomly clicking on different response options).

N=1 participant in Experiment 1a was excluded based on the first criteria, as s/he used external speakers instead of head set (based on response in post-experiment questionnaire). This participant was also more than 3 standard deviations faster in her/his mean log-RTs than other participants (second criteria), as were another participant in Experiment 1a. We decided to exclude participants who were more than 3 standard deviations faster or slower in their mean (log-transformed) RTs compared to other participants. We further excluded *all trials* with RTs more than 3 standard deviations faster or slower than expected. This was determined by first z-scoring the log-transformed RTs *within each participant* (by subtracting the participants' mean from each observation and dividing through the participants standard deviation) and then z-scoring these z-scores *within each trial* across participants. This double-scaling approach was necessary as participants' RTs decreased substantially over the first few trials and then continued to decrease less rapidly until converging against a participant-specific minimum. This criterion did not remove just the first few trials but rather removed RTs that were unusually fast or slow *for that participant at that trial.* And, unlike more complicated methods (like developing a model of cross-trial decreases in RTs), the approach employed here does not make any assumptions about the shape of the speed up in RTs across trials. In total, N=117 trials were excluded, however, no participant was excluded based on too high proportion of missing trials. Figure S1 summarizes participant exclusions due to reaction times and not wearing headphones.

The experiments did not contain independent catch trials. We therefore looked into participants' individual responses in order to identify participants that seem to have randomly answered, independent of stimulus. Figure S2 suggests that participants 13, 15, 21, 22, and 24, have not performed the task. Their responses are indicating that they have not payed attention to the stimuli but rather randomly selected responses irrespective of the vowel they heard, hence, e.g., *had* responses for all different locations in the vowel space (e.g., partici-

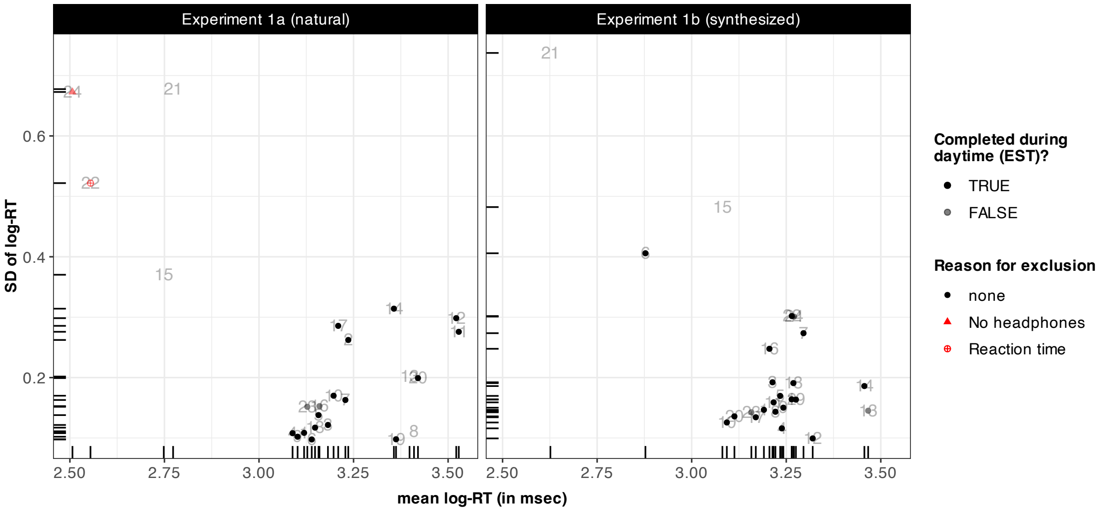

FIG. S1. Participant exclusions in Experiment 1a and Experiment 1b. Two participants in Experiment 1a (participants 22 and 24) were excluded based on their log-transformed RTs and/or for not wearing headphones.

pants 15, 21, 22, 24), or *heed* responses for the low back vowels (e.g., 24), or *odd* responses for high front vowels (e.g., participants 13, 21). Participant 8 seem to have performed the task but clearly used the phonetic space in ways different from everyone else, as s/he seems to have inverted the *who'd*-*hood* categories. This behaviour more likely indicates different dialect patterns, we still however, decided to exclude participant 8 as well. Participants 22 and 24 had already been excluded based on unusual RT-patterns (22) and not wearing over-the-ear headphones (24).

Excluding participants because of unusual vowel responses is more complicated for Experiment 1b as the stimuli is synthesized and more difficult to categorize in general. Figure S3 nevertheless indicates that most participants made use of the phonetic space in similar ways (and in line with where natural categories fall), besides participant 15 and 21. Participant 15 often responds *who'd* for tokens in the high front part of the space, and *heed* for tokens in the high center and back parts, while participant 21 is overall more random in responses, but often selects *heed* for high back tokens and several times selects *hod* for front tokens.

Unusual vowel responses is not an objective criterion. There are many participants that gave unexpected responses on some occasions, e.g., participants 10 and 12 in Experiment 1a, or participants 5 and 24 in Experiment 1b, however, we decided not to exclude them as they were not systematic in their response patterns, i.e., there were no indications of a definite dialect shift (as with participant 8 in Experiment 1a), or systematic randomness in selecting any kind of vowel for any kind of stimuli (as with participants 13, 15, 21, 22, 24, in Experiment 1a, and to some extent, participants 15 and 22 in Experiment 1b).

### B. Distribution of stimuli F1-F3 in Experiments 1a and 1b

Figure S4 visualizes the stimuli in Experiments 1a and 1b in F1-F3 space.

### C. Auxiliary analysis of participant responses in Experiments 1a and 1b

Participants in Experiment 1b showed overall less agreement in their responses to the stimuli than participants in Experiment 1a, as indicated by the higher response entropy in Experiment 1b. In order to assess the extent to which this was a result of the placement of the tokens in the F1-F3 space, we compared linear regression models that predicted response entropy from experiment, to models that employed residuals from a general additive model including response entropy and the tokens placement in the phonetic space (F1, F2) as response variable, and experiment as predictor. We furthermore compared against models based on the full data set to models that excluded all *hut* and *odd* responses from Experiment 1a in order to assess effects of lexical context.

When adding effects of lexical context to the model, the difference between experiments is reduced by 23.9%. Adding a nonlinear model with F1-F2 values of the tokens, the difference is reduced by an additional 35.6%, while adding F3-values reduces the difference by 50.9%. In sum, the result suggest that approximately two-thirds of the difference in response entropy between experiments can be attributed to the placement of stimuli in the formant space, while the remaining one-third is influenced by other factors, most likely the synthesized stimuli sounding highly unnatural.

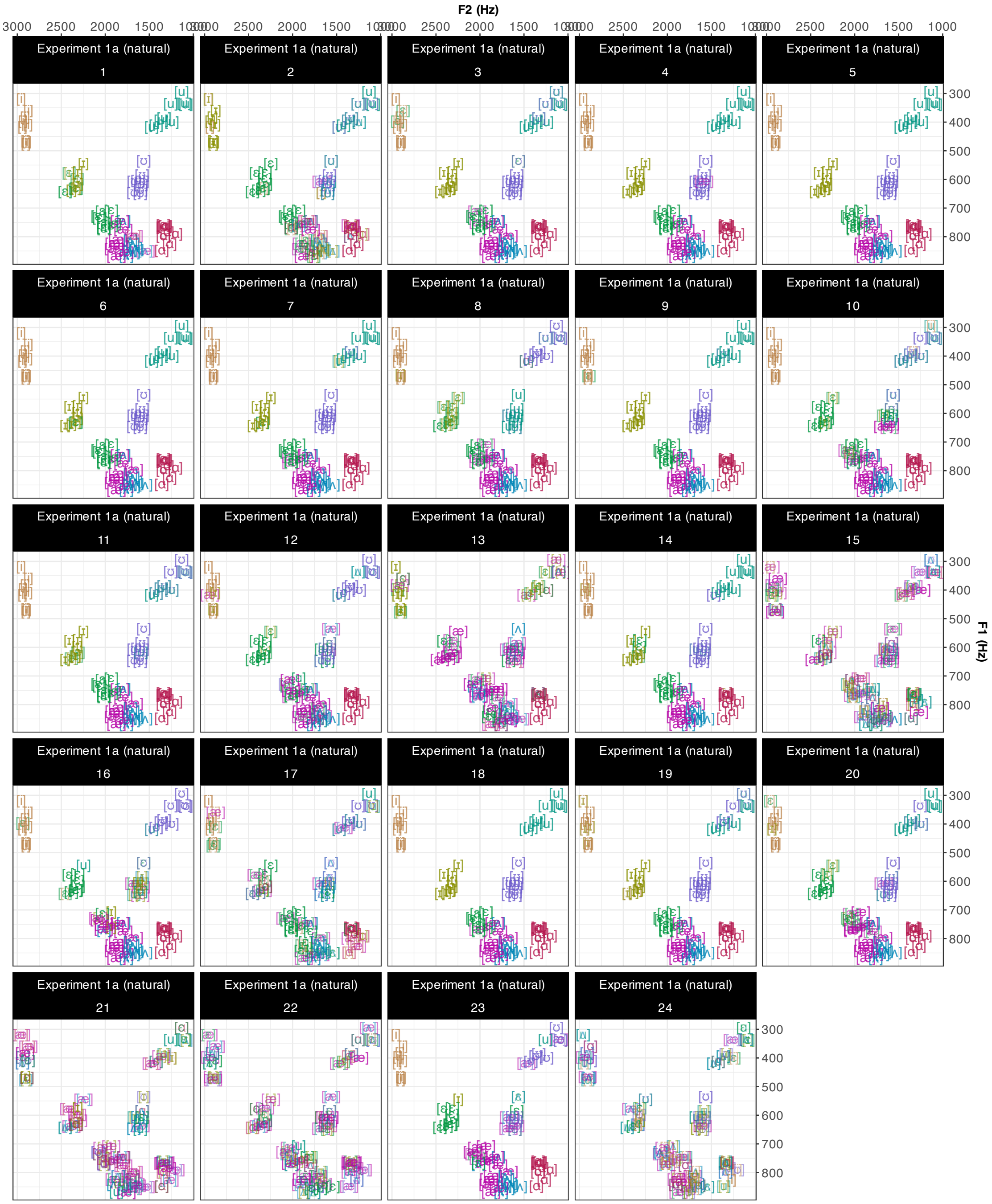

FIG. S2. Participants' categorization responses in Experiment 1a, shown in F1-F2 space. Color and vowel label indicate response provided by participants on each test location. Each vowel was repeated twice.

As stated in the main paper, response entropies differed even for tokens that overlap in Hertz space. Figure S5 visualizes differences in categorization behaviour for these tokens.

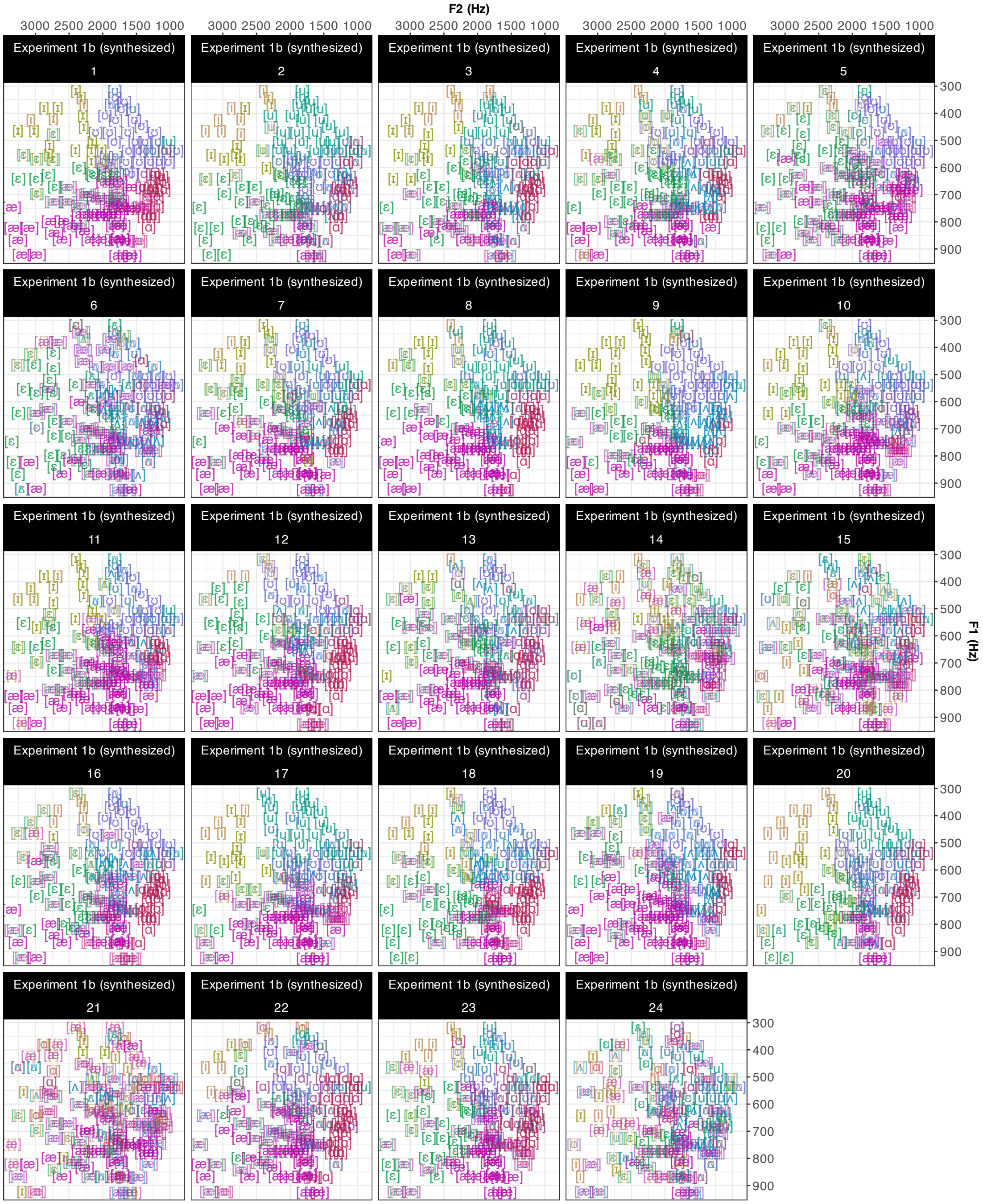

FIG. S3. Participants' categorization responses in Experiment 1b, shown in F1-F2 space. Color and vowel label indicate response provided by participants on each test location. Each vowel was repeated twice.

1086 For many of these tokens, the most frequent response is the same category across experiments, however, with substantially higher disagreement for tokens in Experiment 1b. In

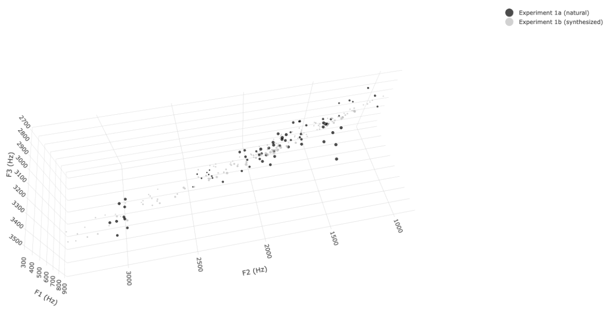

FIG. S4. Stimuli of Experiments 1a and 1b in F1-F3 space. Point size indicates response entropy: larger points represent higher listener agreement, and vice versa.

the bottom part of the acoustic space, participants in Experiment 1b seem to respond *had* disproportionally often.

## §3. ADDITIONAL INFORMATION ON THE COMPUTATIONAL COMPARISON OF NORMALIZATION ACCOUNTS

### A. Methods

#### *1. Vowel data used to train ideal observers (Xie and Jaeger, 2020)*

The Xie and Jaeger database consists of N=1168 *hVd* word recordings from 17 (5 female) L1 talkers of a Northeastern dialect of US English (ages 18 to 35 years old). The talkers were recorded reading a list of 180 English monosyllabic words, a list of short sentences, and a list of ten *hVd* words—the eight US English monophthongs as well as *aid* and *owed* (for further information, see Xie and Jaeger, 2020). For each talker, the database contains 9-10 recordings of each *hVd* word. An automatic aligner [Penn Phonetics Lab Forced Aligner; Yuan and Liberman (2008)] was used to obtain estimates for word and segment boundaries.[12]

The first author manually corrected the automatic alignments for all vowel segmentations. We then used the Burg algorithm in Praat (Boersma and Weenink, 2022) to extract estimates of the first three formants (F1-F3) at three points of the vowel (35, 50, and 65 percent into the vowel). The following parameterization of the Burg algorithm was used:

- Time step (s): 0.01

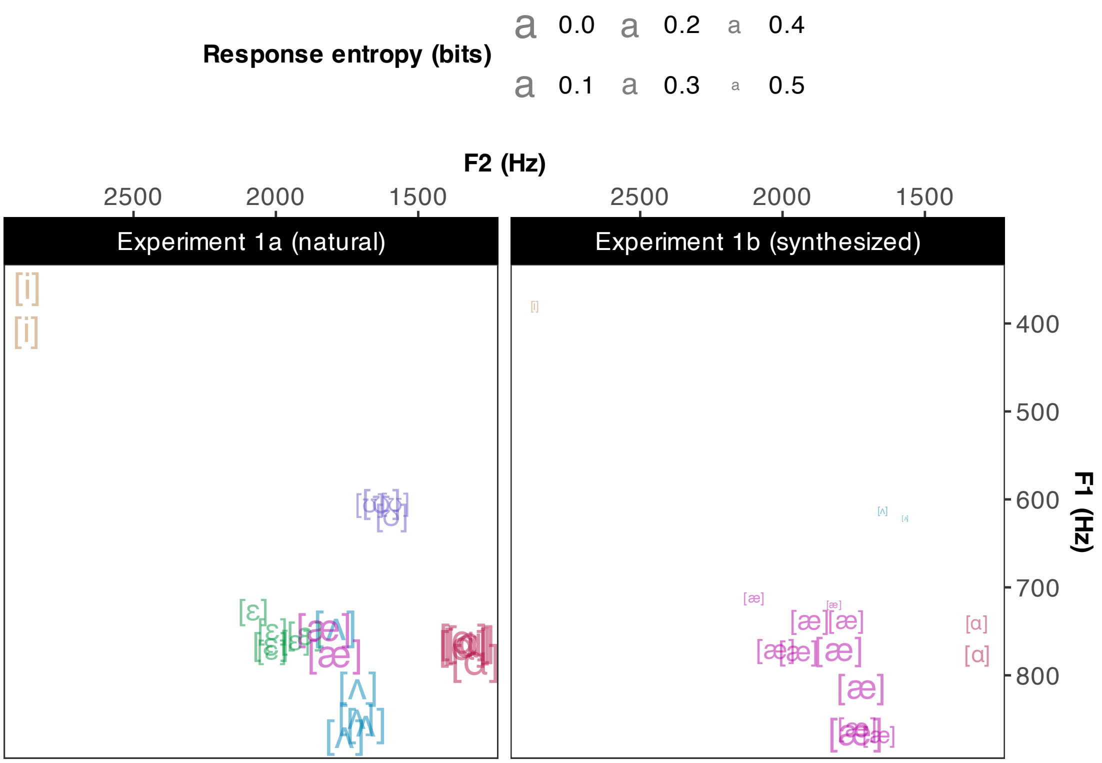

FIG. S5. Listeners' categorization responses in Experiments 1a and 1b, for comparable tokens in Hertz space. The vowel label indicates the most frequent response provided by participants on each test location. Size indicates how consistent responses were across participants, which larger symbols indicating more consistent responses (lower entropy).

- Max. number of formants: 5
- Formant ceiling (Hz): 5500 (5000 for the male talkers)
- Window length (s): 0.025
- Pre-emphasis from (Hz): 50

In addition to F1-F3, we automatically extracted vowel duration and the fundamental frequency (F0) across the entire vowel. These are the data that we used in the cross-validation procedure to train ideal observers, as described in the main text. Figure S6 visualizes the vowel data from the Xie and Jaeger (2020) for all pairwise combinations of F0, F1, F2, F3 and vowel duration, in raw Hertz. Figure S7 shows the distribution of F1 and F2 in the different normalization spaces used in the main study.

### 2. Normalization parameters $\theta$

Figure S8 relates the normalization parameters $\theta$ obtained for each experiment to those found for the five training sets of the Xie and Jaeger (2020) database. This servers two purposes. First, by comparing the $\theta$ of Experiment 1a, which was based on natural productions, to the $\theta$ obtained from Xie and Jaeger (2020), we can assess the extent to which the talker used for Experiment 1a is 'typical' relative to the other talkers of that database. Second, by comparing the range and variability of the $\theta$ across normalization accounts and experiments, we can assess the volatility of different types of parameters, and assess the difference between the beliefs the ideal observers have about the parameters and the parameters in the experiment. How reliably the statistics of the input is established for the same amount of data seems to depend on the space. For instance, parameters in Hertz space display more variability. Within a given scale, we also note that some parameters are more difficult to estimate than others, for instance, mean estimates display less variability than SD, and range values (min and max).

### 3. Optimization process to fit models to human responses

We used constrained quasi-Newton optimization (Byrd *et al.*, 1995) to determine the best-fitting values for the two degrees of freedom—lapse rate ($\lambda$) and noise ratio ($\tau^{-1}$). Optimization was performed separately for each of the 200 combinations of normalization account, experiment, and cross-validation fold. Specifically, we maximized the *likelihood* of the human categorization responses in each experiment under the categorization model conditional on the model's lapse rate and perceptual noise, $\Sigma_i^N \log p(response_i | F1_{i,\theta}, F2_{i,\theta}, M_{\theta,\lambda,\Sigma_{noise}})$, where $response_i$ is the $i$th categorization response, $F1_{i,\theta}, F2_{i,\theta}$ are the F1 and F2 values for the $i$th observation after normalization (with parameters $\theta$ being estimated based on the

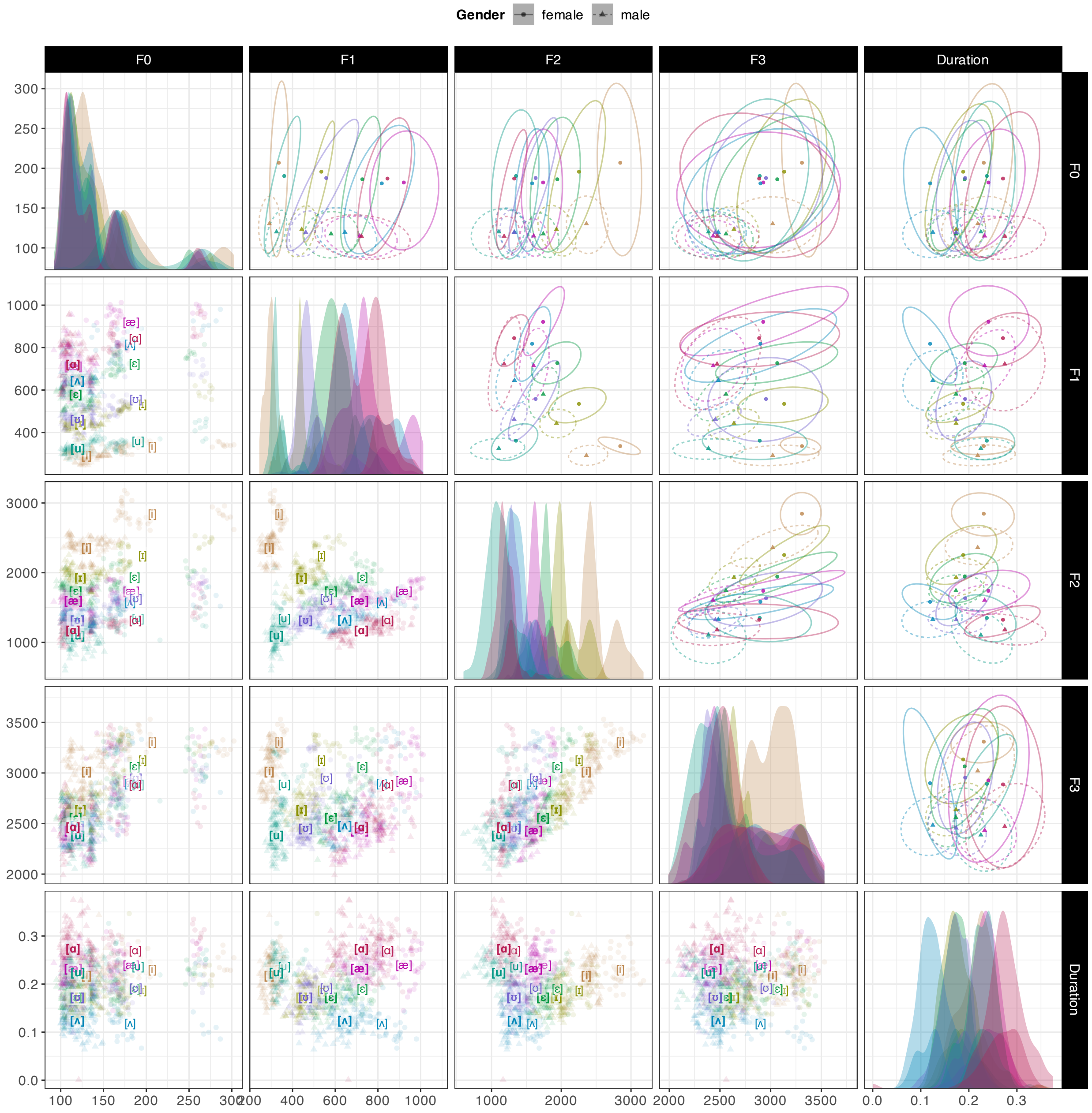

FIG. S6. The pairwise distributions of F0, F1, F2, F3, and duration for all 1168 recordings of the eight monophthong *hVd* words in Xie and Jaeger (2020). Note that axis directions are not reversed. **Panels on diagonal:** marginal cue densities of all five cues. **Lower off-diagonal panels:** each point corresponds to a recording, averaged across the three measurement points within each vowel segment. Vowel labels indicate category means across talkers. Male talkers' vowels are boldfaced. **Upper off-diagonal panels:** Same data as in the lower off-diagonal panels but showing bivariate Gaussian 95% probability mass ellipses around category means.

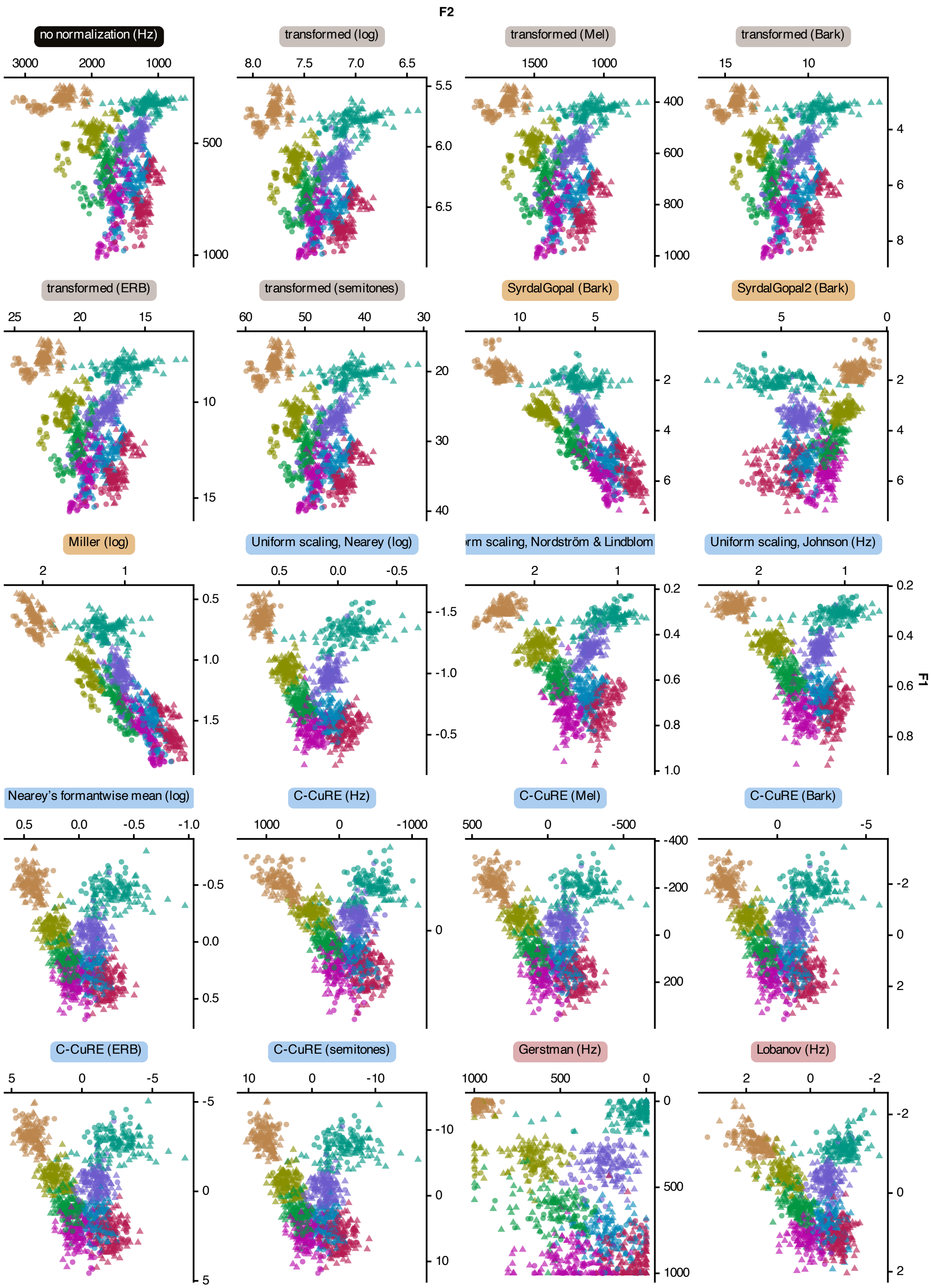

FIG. S7. The 8 monophthong vowels of US English from the Xie and Jaeger (2020) database when F1 and F2 are transformed into a perceptual scale (**grey**), intrinsically normalized (**yellow**), or extrinsically normalized through centering (**blue**) or standardizing (**purple**). Each point corresponds to one recording, averaged across the three measurement points within each vowel segment. Each panel combines the data from all five test folds. Shape indicates gender.

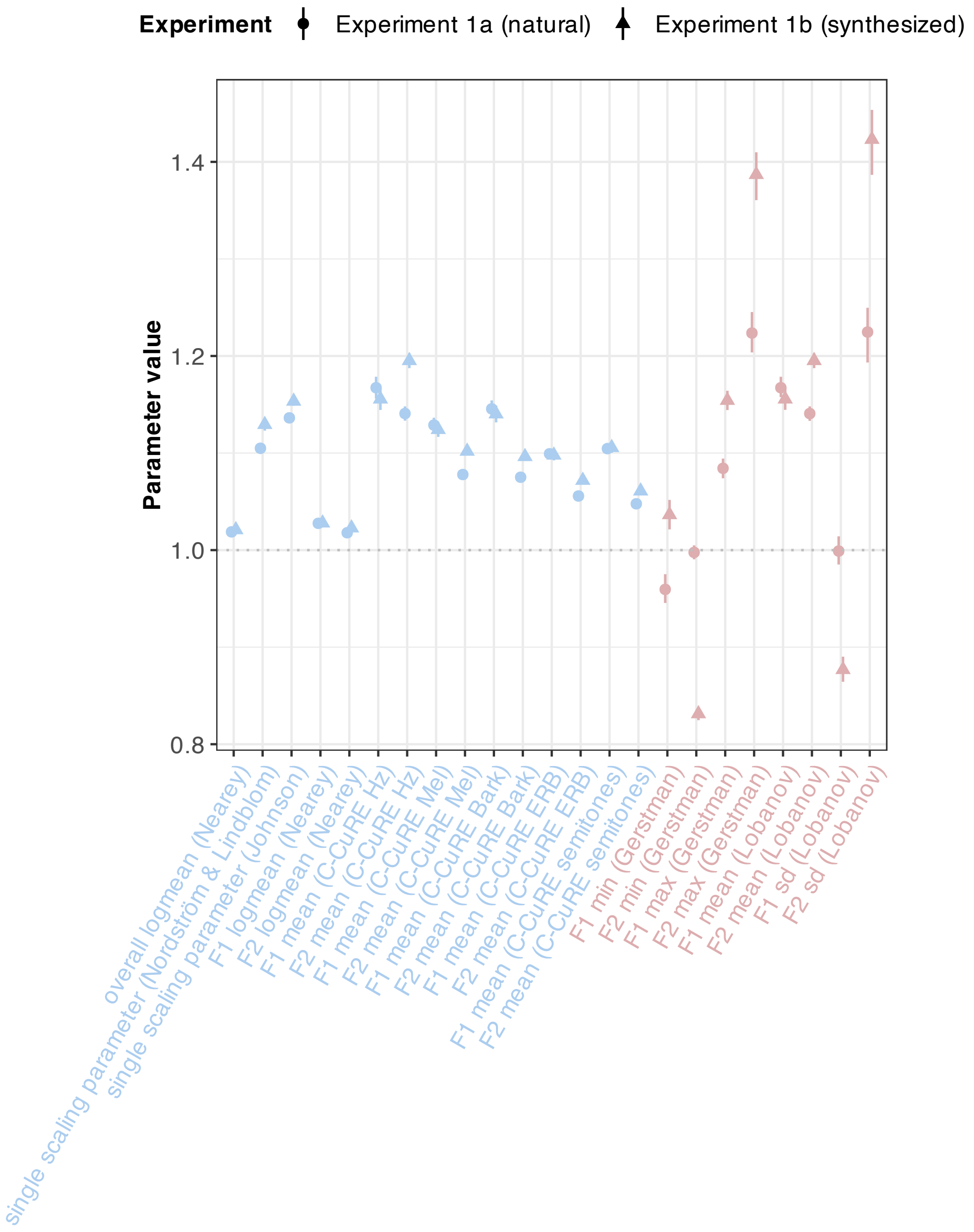

FIG. S8. Comparing normalization parameters $\theta$ across the phonetic database used to estimate listeners' prior experience (Xie and Jaeger, 2020) and Experiments 1a and 1b. Only accounts that assume talker-specific normalization parameters are shown.

distribution of phonetic cues across the stimuli in the experiment). $M_{\theta,\lambda,\Sigma_{noise}}$ is the categorization model in Figure 6, with normalization parameters $\theta$ fixed based from the prior cue distribution in the phonetic database (Xie and Jaeger, 2020), and $\lambda$ and $\Sigma_{noise}$ as the only free parameters to maximize the likelihood. The best-fitting parameterizations were determined by means of the `optim()` function in `R`'s `stats` package (R Core Team, 2024). The starting value of lapse rates and perceptual noise were set to 0.1 and 0.15, respectively. We set the lower and upper bounds to $10^{-10} \geq$ lapse rate $\geq 1$, and $10^{-10} \geq$ perceptual noise $\geq 10$ (well above previously observed estimates for perceptual noise in Kronrod *et al.*, 2016, p. 1698).

Section §3 F presents additional analyses that instead used a grid search over the parameter space. These analyses confirm the results presented in the main paper.

## B. Results for F1-F2

### *1. Significance test of model performance*

Tables S1 and S2 present the results from the paired one-sided t-tests conducted, predicting model log likelihood as a function of normalization account for Experiment 1a and 1b (dummy coded with no normalization model as reference model). The log likelihoods are averaged across the five cross-validation folds and ordered by best-fitting models.

### *2. Parameter estimates for best-fitting models*

In this section, we provide the estimates found for the two degrees of freedom—noise ($\Sigma_{noise}$) and attentional lapses ($\lambda$)—when fitting the models to human behaviour. This will provide insights into the relative contribution of these factors to explaining the variability found in the behavioral data between the two experiments, and to understanding the relative performance of the different normalization accounts as models of human behavior.

Figure S9 visualizes the parameter estimates for each account, averaged across the five training sets (see also Tables S3, S4 for summary of fitted values, and S10 for an illustration of how the fitted noise affects the bivariate Gaussian categories). The Figure indicates that fitted $\lambda$s are very similar across experiments. In Experiment 1a, the mean $\lambda$ across models was 0.1 (sd = 0.04), and for Experiment 1b, 0.12 (sd = 0.14). These estimates can in part

TABLE S1. T-test predicting the model log likelihood as a function of normalization account for Experiment 1a

| Normalization account | Statistic | Estimate mean | Diff. in means | p_value |
|---|---|---|---|---|
| Uniform scaling, Johnson (Hz) | -15.085 | -2523.406 | 611.644 | 0.000 |
| Uniform scaling, Nearey (log) | -9.100 | -2523.406 | 551.676 | 0.000 |
| C-CuRE (Bark) | -13.229 | -2523.406 | 549.192 | 0.000 |
| C-CuRE (Mel) | -12.722 | -2523.406 | 546.779 | 0.000 |
| C-CuRE (ERB) | -11.847 | -2523.406 | 546.472 | 0.000 |
| Nearey's formantwise mean (log) | -10.428 | -2523.406 | 543.718 | 0.000 |
| C-CuRE (semitones) | -10.428 | -2523.406 | 543.718 | 0.000 |
| Uniform scaling, Nordström & Lindblom (Hz) | -7.791 | -2523.406 | 517.075 | 0.001 |
| SyrdalGopal (Bark) | -9.381 | -2523.406 | 510.360 | 0.000 |
| C-CuRE (Hz) | -10.169 | -2523.406 | 498.443 | 0.000 |
| Miller (log) | -7.466 | -2523.406 | 464.305 | 0.001 |
| Lobanov (Hz) | -8.970 | -2523.406 | 461.405 | 0.000 |
| transformed (Bark) | -13.415 | -2523.406 | 228.190 | 0.000 |
| transformed (Mel) | -12.472 | -2523.406 | 214.317 | 0.000 |
| transformed (ERB) | -10.291 | -2523.406 | 192.156 | 0.000 |
| SyrdalGopal2 (Bark) | -3.673 | -2523.406 | 171.154 | 0.011 |
| Gerstman (Hz) | -2.104 | -2523.406 | 159.477 | 0.052 |
| transformed (log) | -2.617 | -2523.406 | 66.928 | 0.029 |
| transformed (semitones) | -2.617 | -2523.406 | 66.928 | 0.029 |

confirm what was hypothesized with regard to listeners' categorization accuracy—that the performance of listeners in inferring the category intended by the talker in part reflected attentional lapses (mean accuracy in Experiment 1a = 81.2% (SE = 4.8%); Experiment 1b had no such ground truth).

TABLE S2. T-test predicting the model log likelihood as a function of normalization account for Experiment 1b

| Normalization account | Statistic | Estimate mean | Diff. in means | p_value |
|---|---|---|---|---|
| Uniform scaling, Nearey (log) | -64.722 | -10372.71 | 2553.594 | 0.000 |
| Lobanov (Hz) | -82.707 | -10372.71 | 2521.647 | 0.000 |
| Gerstman (Hz) | -34.270 | -10372.71 | 2491.092 | 0.000 |
| C-CuRE (ERB) | -35.509 | -10372.71 | 2409.974 | 0.000 |
| C-CuRE (Bark) | -33.226 | -10372.71 | 2408.381 | 0.000 |
| Nearey's formantwise mean (log) | -35.495 | -10372.71 | 2305.669 | 0.000 |
| C-CuRE (semitones) | -35.495 | -10372.71 | 2305.669 | 0.000 |
| transformed (log) | -63.629 | -10372.71 | 2259.438 | 0.000 |
| transformed (semitones) | -63.629 | -10372.71 | 2259.438 | 0.000 |
| C-CuRE (Mel) | -28.933 | -10372.71 | 2221.903 | 0.000 |
| transformed (ERB) | -51.487 | -10372.71 | 2106.368 | 0.000 |
| transformed (Bark) | -46.651 | -10372.71 | 1942.912 | 0.000 |
| SyrdalGopal2 (Bark) | -29.866 | -10372.71 | 1816.140 | 0.000 |
| transformed (Mel) | -41.633 | -10372.71 | 1622.458 | 0.000 |
| Miller (log) | -26.334 | -10372.71 | 1244.978 | 0.000 |
| Uniform scaling, Johnson (Hz) | -8.574 | -10372.71 | 1069.757 | 0.001 |
| SyrdalGopal (Bark) | -20.657 | -10372.71 | 910.361 | 0.000 |
| Uniform scaling, Nordström & Lindblom (Hz) | -11.293 | -10372.71 | 878.721 | 0.000 |
| C-CuRE (Hz) | -10.441 | -10372.71 | 817.140 | 0.000 |

What is perhaps more obvious from Figure S9 is that $\Sigma_{noise}$ estimates clearly differ between experiments. In Experiment 1a, the best-fitting $\Sigma_{noise}$ estimates are comparable to what Kronrod *et al.* (2016) found (mean $\Sigma_{noise}$ in Experiment 1a = 0.52 (sd = 0.49). In Experiment 1b, this is not the case (mean $\Sigma_{noise}$ = 4.74 (sd=2.57). However, there is no a

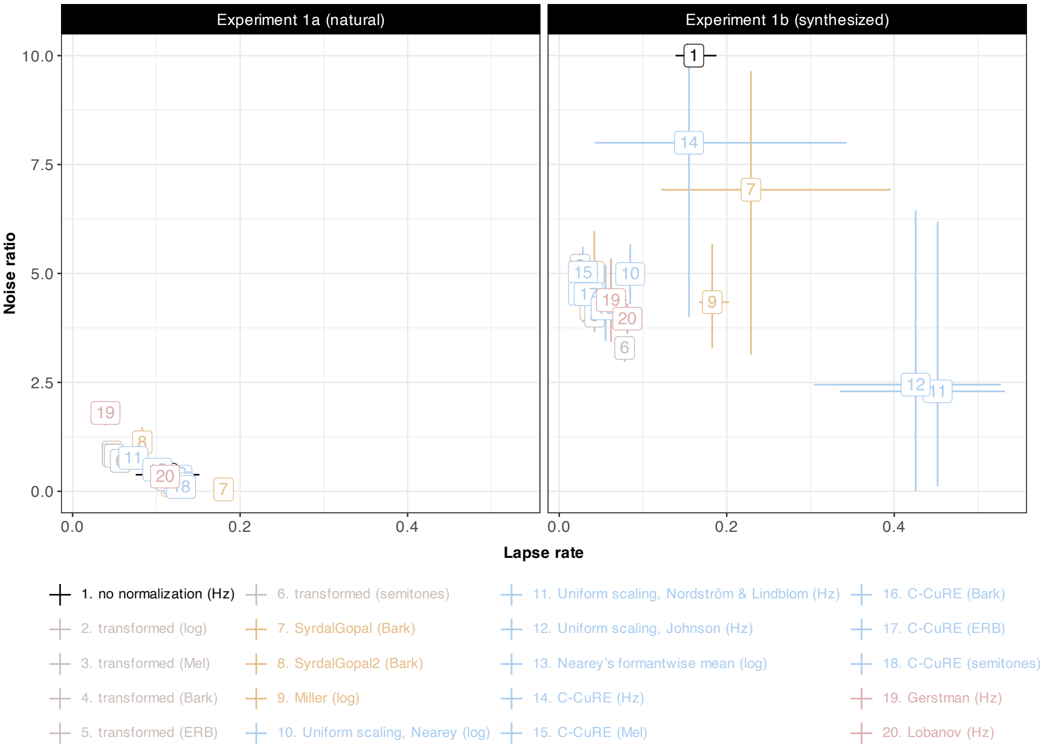

FIG. S9. Best-fitting estimates obtained for $\lambda$ and $\Sigma_{noise}$. Numeric label is placed at the mean across the five folds, line ranges represent the 95% CIs.

priori reason to expect internal perceptual noise to differ between experiments, which is why these high noise ratios likely reflect external noise. Given what was shown for the human data (cf. discussion on differences in response entropy between experiments, Section II B), this is perhaps not surprising. The stimuli in Experiment 1b were clearly more noisy and presumably left listeners with more uncertainty about the true value of the formants, and how to best make use of previous experience. Even if the task itself was identical across experiments, the nature of the stimuli in Experiment 1b likely contributed to making the experiment overall more demanding. In addition, Experiment 1b was longer (N=74 more trials, and took on average 8.1 more minutes to complete), both of these aspects might also affect the amount of attentional lapsing.

Finally, for the majority of accounts, there is little variability in parameter estimates—and likelihoods—across training sets (Tables S3 and S4). This suggests that models achieved their maximum likelihood fit to human data on similar estimates for the two degrees of freedom, which provides a sanity check of the modelling approach adopted.

### *3. By-item analysis*

To provide further insight into model performance, we visualize model fits against human behavior on a by-item level for three of the best-performing models across experiments, Nearey's uniform scaling, Johnson's uniform scaling and Lobanov. This allows us to assess whether normalization always improves model fit in absence of normalization, and if normalization models perform equally well in different parts of the acoustic-phonetic space.

Figure S11 indicates a general tendency for increased model performance as humans' predictions about human behavior become stronger, even though models' improvements are not limited to items for which humans have strong predictions. Normalization does not, however, improve model fit across the board. Relative to no normalization, all three accounts both increase and decrease in performance on a by-item level. The advantage of Nearey's uniform scaling relative to no normalization seems to be driven by smaller improvements (<35% change) on many items in Experiment 1a (proportion of items with increase in performance = 76.4%, mean improvement in likelihood by item = 13.92 (sd = 11.87), mean likelihood by item for items where there is *no* improvement = -12.59 (sd = 10.81)), whereas for Experiment 1b, Nearey's uniform scaling improves substantially on many items (proportion = 93.2, mean improvement in likelihood by item = 19.32 (sd = 15.38), mean likelihood by item for items where there is *no* improvement = -7.37 (sd = 4.62)). Johnson follows the same pattern for Experiment 1a only (proportion = 80.6%, mean improvement in likelihood by item = 13.16 (sd = 11.36), mean likelihood by item for items where there is *no* improvement = -10.84 (sd = 9.91)), while for Experiment 1b, improvements are less pronounced (proportion = 75.3%, mean improvement in likelihood by item = 11.9 (sd = 7.74), mean likelihood by item for items where there is *no* improvement = -6.65 (sd = 4.47)). Lobanov seems to follow the same pattern as Nearey (for Experiment 1a, proportion = 73.6%, mean improvement in likelihood by item = 12.67 (sd = 12.08), mean likelihood by item for items where there is *no* improvement = -11.06 (sd = 9.82); for Experiment 1b, proportion = 87%, mean improvement in likelihood by item = 20.59 (sd = 17.07), mean likelihood by item for items where there is *no* improvement = -4.92 (sd = 4.62)).

TABLE S3. The best-fitting estimates obtained for noise ratios and lapse rates in Experiment 1a (averaged across the five cross-validation folds and ordered by best-performing models)

| Normalization account | mean log likelihood | noise percentage | lapse rate |
|---|---|---|---|
| Uniform scaling, Johnson (Hz) | -2214.93 | mean=0.5 (SD=0.28) | mean=0.1 (SD=0.01) |
| Uniform scaling, Nordström & Lindblom (Hz) | -2219.16 | mean=0.77 (SD=0.39) | mean=0.07 (SD=0) |
| C-CuRE (Bark) | -2261.27 | mean=0.28 (SD=0.26) | mean=0.12 (SD=0.02) |
| C-CuRE (Mel) | -2261.61 | mean=0.29 (SD=0.27) | mean=0.12 (SD=0.02) |
| C-CuRE (ERB) | -2262.53 | mean=0.16 (SD=0.23) | mean=0.13 (SD=0.02) |
| C-CuRE (semitones) | -2264.28 | mean=0.1 (SD=0.17) | mean=0.13 (SD=0.01) |
| Nearey's formantwise mean (log) | -2264.28 | mean=0.1 (SD=0.17) | mean=0.13 (SD=0.01) |
| Uniform scaling, Nearey (log) | -2273.68 | mean=0.34 (SD=0.36) | mean=0.13 (SD=0.02) |
| C-CuRE (Hz) | -2305.16 | mean=0.19 (SD=0.15) | mean=0.13 (SD=0.02) |
| SyrdalGopal (Bark) | -2357.42 | mean=0.05 (SD=0.04) | mean=0.18 (SD=0.01) |
| Miller (log) | -2374.96 | mean=0.15 (SD=0.06) | mean=0.12 (SD=0.01) |
| Lobanov (Hz) | -2390.23 | mean=0.34 (SD=0.27) | mean=0.11 (SD=0.02) |
| transformed (Bark) | -2569.87 | mean=0.83 (SD=0.2) | mean=0.05 (SD=0) |
| transformed (Mel) | -2587.65 | mean=0.89 (SD=0.18) | mean=0.05 (SD=0) |
| transformed (ERB) | -2598.16 | mean=0.81 (SD=0.24) | mean=0.05 (SD=0) |
| Gerstman (Hz) | -2647.38 | mean=1.8 (SD=0.33) | mean=0.04 (SD=0) |
| SyrdalGopal2 (Bark) | -2661.45 | mean=1.12 (SD=0.39) | mean=0.08 (SD=0.01) |
| transformed (semitones) | -2682.60 | mean=0.7 (SD=0.34) | mean=0.06 (SD=0.01) |
| transformed (log) | -2682.60 | mean=0.7 (SD=0.34) | mean=0.06 (SD=0.01) |
| no normalization (Hz) | -2779.13 | mean=0.38 (SD=0.26) | mean=0.11 (SD=0.05) |

TABLE S4. The best-fitting estimates obtained for noise ratios and lapse rates in Experiment 1b (averaged across the five cross-validation folds and ordered by best-performing models)

| Normalization account | mean log likelihood | noise percentage | lapse rate |
|---|---|---|---|
| Gerstman (Hz) | -9474.08 | mean=4.39 (SD=1.24) | mean=0.06 (SD=0.03) |
| Uniform scaling, Nearey (log) | -9551.73 | mean=4.99 (SD=0.87) | mean=0.08 (SD=0) |
| C-CuRE (Bark) | -9595.30 | mean=4.53 (SD=0.9) | mean=0.03 (SD=0.02) |
| C-CuRE (ERB) | -9601.26 | mean=4.52 (SD=0.93) | mean=0.04 (SD=0.02) |
| Lobanov (Hz) | -9608.45 | mean=3.96 (SD=0.47) | mean=0.08 (SD=0.01) |
| Nearey’s formantwise mean (log) | -9702.20 | mean=4.2 (SD=1.08) | mean=0.06 (SD=0.03) |
| C-CuRE (semitones) | -9702.20 | mean=4.2 (SD=1.08) | mean=0.06 (SD=0.03) |
| C-CuRE (Mel) | -9772.00 | mean=5.02 (SD=0.71) | mean=0.03 (SD=0.01) |
| transformed (semitones) | -9815.97 | mean=3.29 (SD=0.37) | mean=0.08 (SD=0.01) |
| transformed (log) | -9815.97 | mean=3.29 (SD=0.37) | mean=0.08 (SD=0.01) |
| transformed (ERB) | -9956.82 | mean=4.03 (SD=0.42) | mean=0.04 (SD=0.01) |
| transformed (Bark) | -10123.34 | mean=4.18 (SD=0.42) | mean=0.04 (SD=0.01) |
| SyrdalGopal2 (Bark) | -10231.19 | mean=5.02 (SD=1.14) | mean=0.04 (SD=0.01) |
| transformed (Mel) | -10431.06 | mean=5.17 (SD=0.39) | mean=0.02 (SD=0.01) |
| Uniform scaling, Johnson (Hz) | -10791.39 | mean=2.45 (SD=4.33) | mean=0.43 (SD=0.14) |
| Miller (log) | -10848.88 | mean=4.34 (SD=1.54) | mean=0.18 (SD=0.02) |
| Uniform scaling, Nordström & Lindblom (Hz) | -11085.94 | mean=2.29 (SD=4.32) | mean=0.45 (SD=0.13) |
| C-CuRE (Hz) | -11098.93 | mean=8 (SD=4.47) | mean=0.15 (SD=0.21) |
| SyrdalGopal (Bark) | -11187.23 | mean=6.92 (SD=4.17) | mean=0.23 (SD=0.19) |
| no normalization (Hz) | -12118.49 | mean=10 (SD=0) | mean=0.16 (SD=0.03) |

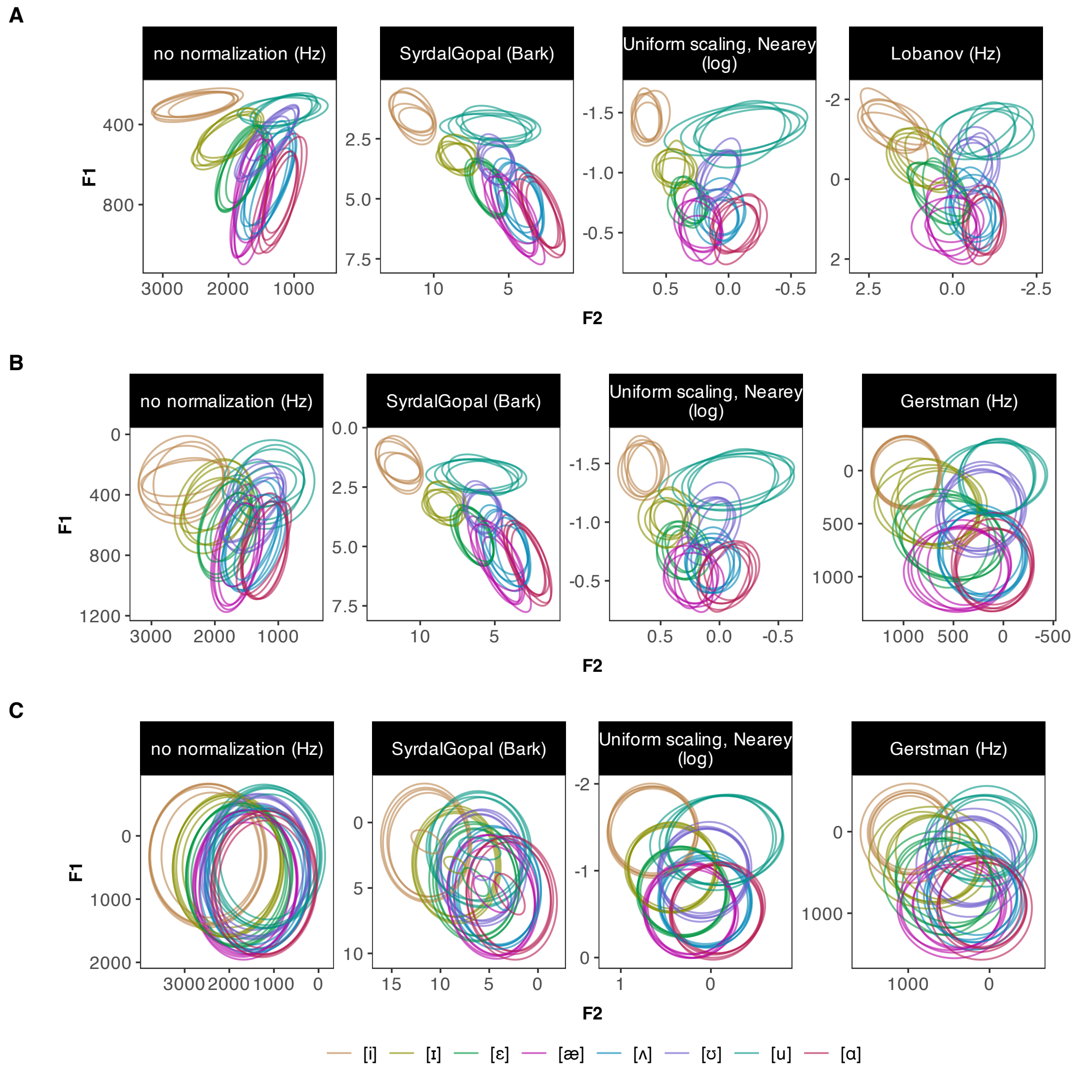

FIG. S10. Visualizing the bivariate Gaussian categories of four example normalization accounts for each of the five cross-validation folds (each fold corresponds to one set of eight ellipses). **Panel A** prior to adding $\Sigma_{noise}$, **Panel B** with added noise from best-fitting models in Experiment 1a, **Panel C** with added noise from best-fitting models in Experiment 1b. For most of the accounts in Panel B and C, noise ratios adds so much category variability that models could presumably only make correct predictions at the outer range of the ellipses. If allowing for separate noise estimates for F1 and F2, this might however not be the case.

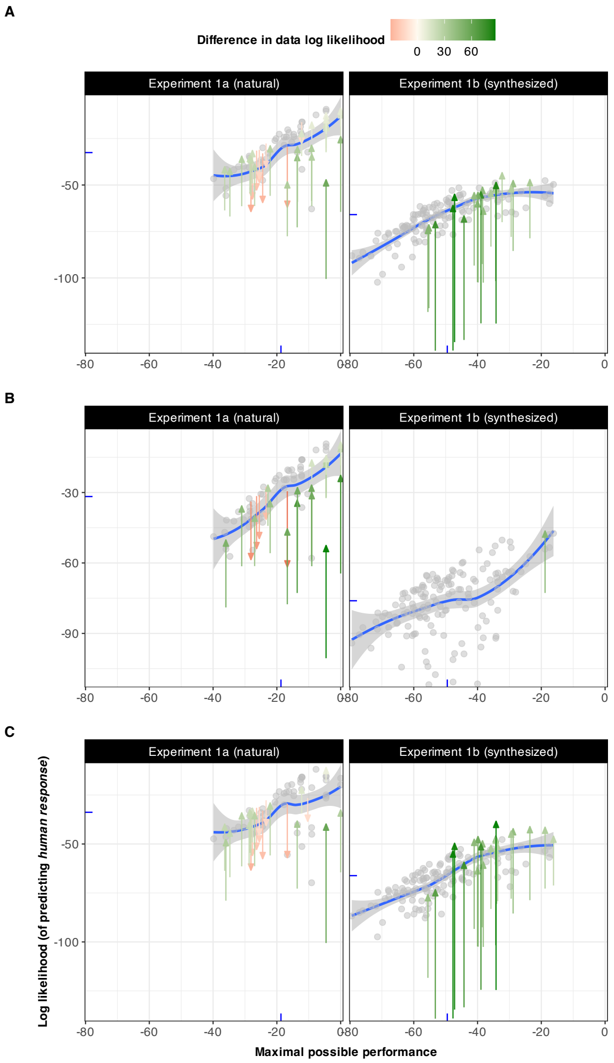

FIG. S11. By-item model improvement from no normalization, relative to the maximum possible performance (predicting human responses from human responses). Maximum log likelihood across items indicated by ticks on axis. Arrows indicate change from no normalization to Nearey's uniform scaling (**panel A**), Johnson (**panel B**), and Lobanov (**panel C**), for items with a change of more than 35%. Points represent items for which change is less than 35%. Color and arrow head indicate decrease or increase in log likelihood.

To explore whether differences in model performance are related to where in the acoustic-phonetic space items are located, we plot the likelihood of the unnormalized model in the acoustic-phonetic space, along with likelihood differences between the best-performing models (see Figure S12).

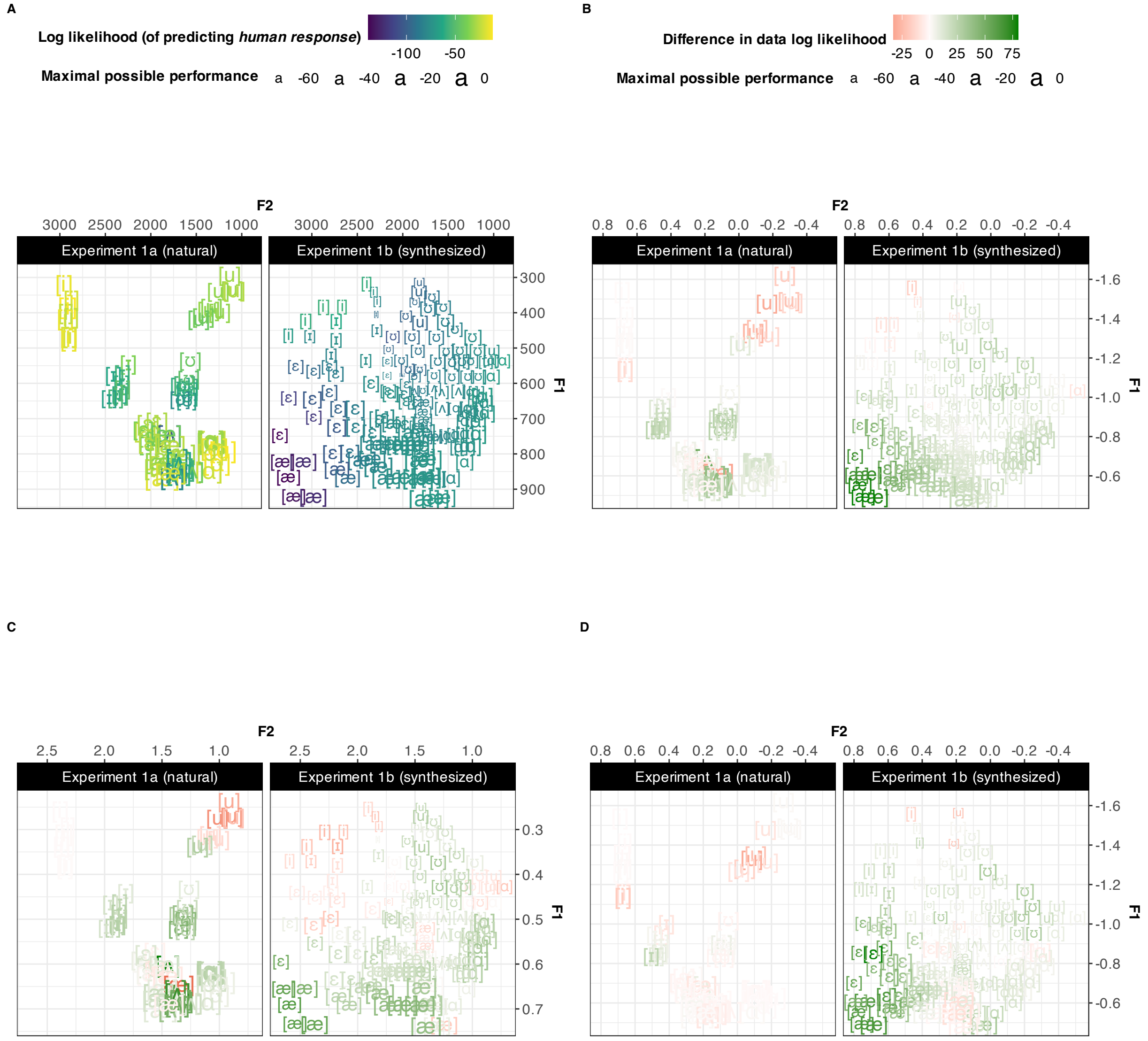

FIG. S12. In which part of the acoustic-phonetic space does normalization fail to improve fit against human responses? For each test location, the vowel label indicates the most frequent response provided by participants. Size of vowel label relates model performance to maximum performance (predicting human responses from human responses). **Panel A** shows the likelihood of the unnormalized model in predicting human responses to both experiments. **Panels B-D** shows difference in likelihood between models, Nearey's uniform scaling vs. no normalization (**panel B**), Johnson vs. no normalization (**panel C**), Nearey's uniform scaling vs. Johnson (**panel D**).

Figure S12 suggests that normalization does not improve things universally across the acoustic-phonetic space. Overall, model performance is better for items for which human predictions are stronger, that is, models perform better in parts of the acoustic space where humans can easier predict human behavior (Figure S12, *Panel A*). To the extent that this is not the case, it seems that normalization in general can adjust for this, improving model performance on many tokens where the maximum performance is high but the unnormalized model's predictions are low, e.g., in the left bottom and center part of the acoustic space (*Panels B-C*). There is overall less improvement in Experiment 1a, presumably because models are already performing well predicting human behavior in the first experiment. Both Nearey's uniform scaling and Johnson clearly perform worse relative to the unnormalized model in the upper right part of the space, more specifically for the [u] category (*Panels B-C*; left), which could indicate that models are overly categorical in a part of the space where humans are less categorical. Possible reasons to this, could be 1) the stimuli sounding more like a neighbouring category to many listeners, or 2) potential effects of orthography, making humans less inclined to select the [u] category. The potential effect of the infrequent non-word response option *who'd* could have been checked against the synthesized stimuli in Experiment 1b. If there was indeed an effect of orthography, we should have observed a better model fit and larger between-account differences in predictions in this part of the acoustic space. Unfortunately, we under-sampled that part, which is an important caveat for Experiment 1b. For the items closest to the area in question, participants however often responded *hood*, which might indicate that items in this part of the space for this talker overall sounded more like *hood* and not *who'd* for many listeners (c.f., discussion on listeners' dialect templates in Section II B).

Comparing the two best-performing models across experiments (*Panel D*), there are no evident patterns of improvement in one model relative to the other. In Experiment 1a, Johnson provided the best fit to listeners' responses and appears to improve the fit relative to Nearey across the entire space (with the exception of one [ɪ] token). For Experiment 1b, Nearey overall improves the fit relative to Johnson, with the exception of some locations in the mid part of the phonetic space (including high, center and low vowels).

### C. Results for F1-F2 (subsets of Experiments 1a and 1b)

To evaluate two potential concerns with our stimuli, we decided to compare the 20 normalization accounts against a subset of the data from Experiment 1a and 1b. For Experiment 1a, we excluded listeners' responses to the two *hVd* stimuli that differed in phonological context from all other words: *odd* and *hut*. For Experiment 1b, we excluded responses to stimuli that were presumed physiologically implausible under the assumption of a single talker (all stimuli below the diagonal dashed line in Figure 4).

This analysis overall replicates the results from the main analysis: uniform scaling accounts again provide the best fit against listeners' responses in both experiments (Figure S13). For Experiment 1a, Nordström & Lindblom achieved the best fit (log likelihood = -1804, SD = ), while Nearey's uniform scaling again provided the best fit to Experiment 1b (log likelihood = -8382, SD = 71).

### D. Results for F1-F2 (subset of listeners sharing dialect template)

Analyses in the main paper suggested that not all listeners in Experiment 1a and 1b shared dialect template (Section II B). To investigate the effect of excluding listeners that likely did not use the same underlying vowel representations for categorization, we compared the 20 normalization accounts against a subset of listeners who employed the dialect template used by the majority of participants (see lower-left of both panels in Figure 5B). This left 11 participants for Experiment 1a (61.1%) and 14 for Experiment 1b (77.8%). Under the assumptions that 1) our model of listeners is adequate, that 2) the subset group of listeners now share dialect template and that 3) the priors, the phonetic database, can approximate this template, we would expect all model to increase their likelihood fit to listeners' responses (c.f., Section IV).

As expected, Figure S14 suggests that all models overall provide higher likelihood fits against human responses in both experiments compared to the main model. To increase comparability to the results of the main model, we scaled the log likelihood of models in the subset data to those of the main analysis by multiplying the model log likelihoods with the ratio of the number of observations in the main model over the number of observations in the subset model. This suggested that the improvement in likelihood for the dialect subset model to the original dataset was 41.1%.

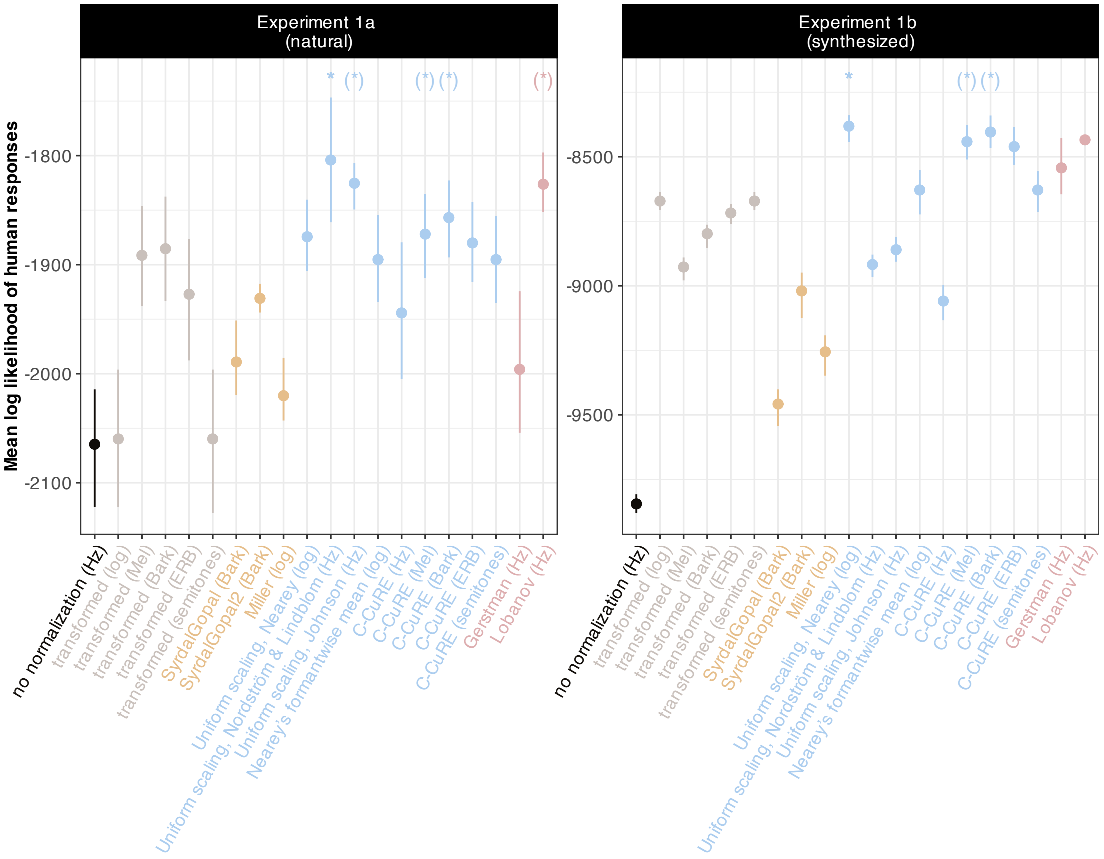

FIG. S13. Results of model fit to subset data. Pointranges indicate mean and 95% bootstrapped CIs of the log-likelihood summarized over the five training sets (higher is better). Accounts that fit listeners' responses to an extent that is statistically indistinguishable from the best-fitting account are marked by (*).

Replicating the results from the main analysis, uniform scaling accounts again fit listeners' behavior well across both experiments. While Nearey's uniform scaling provided the best-fit in Experiment 1b (log likelihood = -5591, SD = 47), Syrdal & Gopal now achieved the best fit to Experiment 1a (log likelihood = -618, SD = 19). Only one additional account performed within the range of the best-performing accounts: for Experiment 1a, all *ps* < .0341; for

Experiment 1b, Lobanov achieved likelihood fits statistically indistinguishable from Nearey's uniform scaling ($p > .15$, log likelihood = -5612, SD = 30).

While Nearey's uniform scaling displayed relatively stable performance across experiments, Syrdal & Gopal varied drastically, achieving one of the worst fits to listeners' responses in Experiment 1b (log likelihood = -6684, SD = 75). As mentioned in Section III B 2, a possible explanation to large fluctuations in model fits between experiments, is that this account has been over-engineered on specific types of natural vowel productions. Given that formant normalization is a pre-linguistic mechanism, it ought to be able to explain listeners' responses to any type of data, including data that does not follow correlations in natural data. This would suggest that Syrdal & Gopal might not be a plausible account of normalization.

### E. Results for F1-F3

To investigate whether the inclusion of F3, a cue known to be important for vowel category distinctions, would improve the model fit to human behavior, we trained ideal observers on multivariate (F1-F2-F3) categories from the same database as in the main study. Here, we first report the results of the F1-F3 model and qualitatively compare them to the results in the main text for F1-F2. This will highlight that the results are largely similar and support the same conceptual conclusions, but there are some differences in model fit. To understand these differences better, we then also directly compare the results quantitatively to see for which accounts the inclusion of F3 improved the fit against listeners' responses and for which accounts it decreased the fit.

Figure S15 summarizes how well the different accounts fit listeners' responses in Experiments 1a and 1b when assuming F1-F2-F3 multivariate category representations. Many aspects replicate the F1-F2 results reported in the main text. First, normalization significantly improved the fit relative to no normalization. Second, the same uniform scaling accounts again achieved the best fit against listeners' responses: for Experiment 1a, Johnson normalization account provided the best fit (log likelihood = -2345, SD = 23), while Nearey's uniform scaling account provided the best fit to Experiment 1b (log likelihood = -9610, SD = 76). However, we note that the inclusion of F3 does not improve the fit to listeners' responses for several accounts (compare *squares* and *circles* in Figure S15). In fact, with the exception of the raw Hertz, scale transformations, and intrinsic accounts, most extrinsic

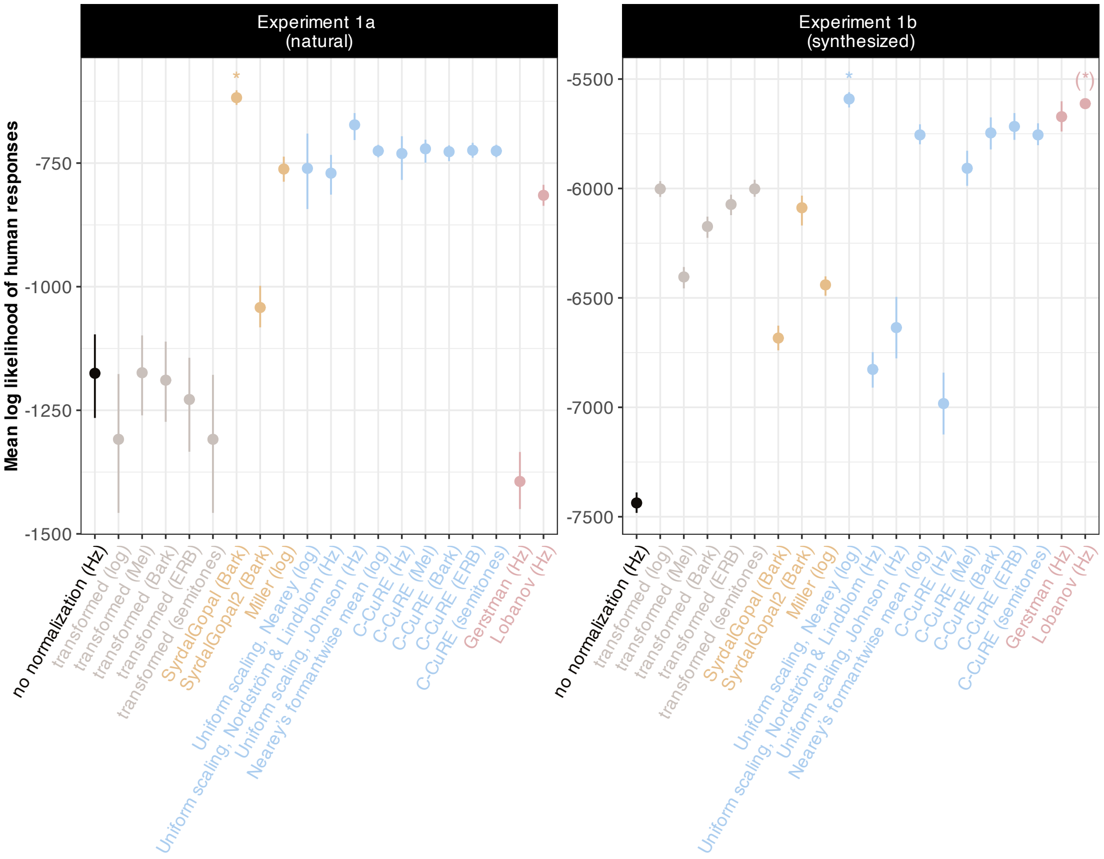

FIG. S14. Results of model fit to data excluding listeners that do not seem to share dialect template. Pointranges indicate mean and 95% bootstrapped CIs of the log-likelihood summarized over the five training sets (higher is better). Accounts that fit listeners' responses to an extent that is statistically indistinguishable from the best-fitting account are marked by (*).

accounts seem to decrease their fit, more so in Experiment 1a than 1b. This includes the overall best-performing account in the main text, Nearey's uniform scaling, that no longer achieves a statistically indistinguishable fit from Johnson in Experiment 1a. At first blush, this is puzzling given that the model now has access to more information of a type that is broadly believed to be informative for US English vowel recognition (Hillenbrand *et al.*, 1995; Nearey, 1989; Peterson and Barney, 1952). What might be underlying the apparent lack of improvement, and why does it appear as if some accounts actually achieve worse fits?

One possible explanation is that listeners were only exposed to one talker in Experiment 1a. According to some theories, F3 is expected to contribute to vowel recognition when there are multiple talkers, acting as a sort of normalizer for vocal tract length (Nearey, 1989). In the absence of other talkers, this advantage might instead introduce noise to the models—an additional source of information that are not of use for listeners in this context. It is also possible that this particular talker has a pattern of F3 use across categories that is atypical given the other talkers in the database. This might explain why the raw Hertz model improves the fit with F3-inclusion. We checked for additional outliers along F3 for this talker, and also inspected the talker's categories in 3D-space (S4), but we could not find that outliers would be a likely explanation. To gain further knowledge into this talker's use of F3 compared to other talkers in the database, we used the same models to predict the ground truth, i.e., the category the talker actually intended to produce. These models patterned with the other prediction results, again indicating that F3-inclusion did not improve model performance. We take this to suggest that the F1-F3 results is not about how our model uses F3, but rather about how this specific talker uses F3. The results might thus link back to the potential dialect differences between talkers in the database, reported in Section II B.

### F. Grid search over parameter space for F1-F2 and F1-F3

As an alternative to the quasi-Newton optimization presented in the main text, we also conducted a grid search over the space defined by the two parameters lapse rate and noise ratio. Figure S16 summarizes the results for a grid of lapse rates ∈ 0, .02, .06, .18, .36, .72 and noise ratios ∈ 0, .3, .6, 1.25, 2.5, 5 for Experiment 1a. For Experiment 1b (Figure S17), the range of noise ratios explored was ∈ 0, 1.5, 3, 6, 12.5, 25.

This search confirmed the pattern described in the main text. Additional grid searches confirmed this pattern held beyond the values shown here. For all normalization accounts, all combinations of cues, and both experiments, the goodness of fit of the ideal observers initially improved with increasing lapse rate and increasing noise ratios, and then decreased once lapse rates or noise ratios reached the best-fitting values (which depended on the combination of normalization account, cues, and experiment). It further indicated that Nearey's uniform scaling, together with the other uniform scaling accounts and some of the

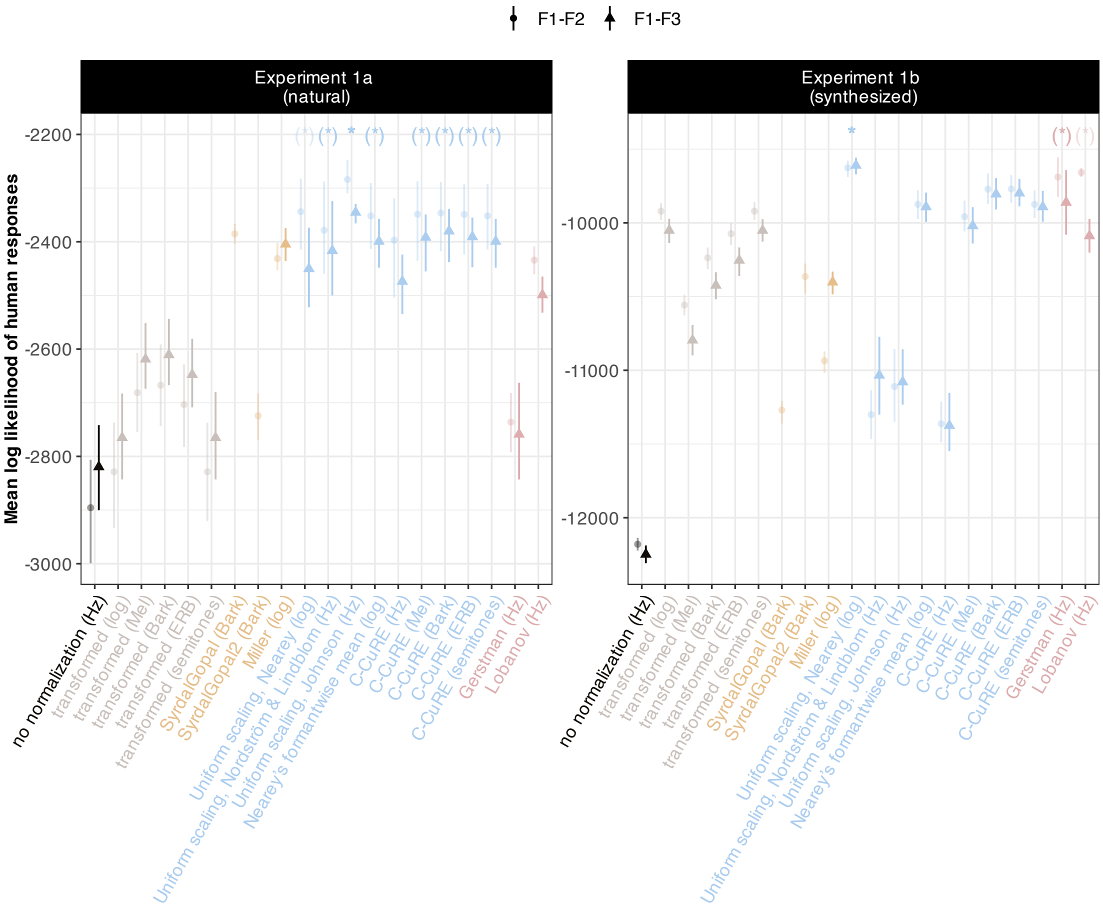

FIG. S15. Results of ideal observer models trained on F1, F2 and F3 as cues to vowel identity. As in Figure 9 in the main text, pointranges indicate mean and 95% bootstrapped CIs of the log-likelihood summarized over the five training sets (higher is better). For comparison, results from the F1-F2 models are included (more transparent circles).

1353 C-CuRE accounts (Experiment 1a) and Gerstman (Experiment 1b), improved faster and
1354 performed consistently well for a good range of parameters, even for high $\tau^{-1}$. Many of the
1355 other models were less consistent and only performed well for a smaller range of estimates.

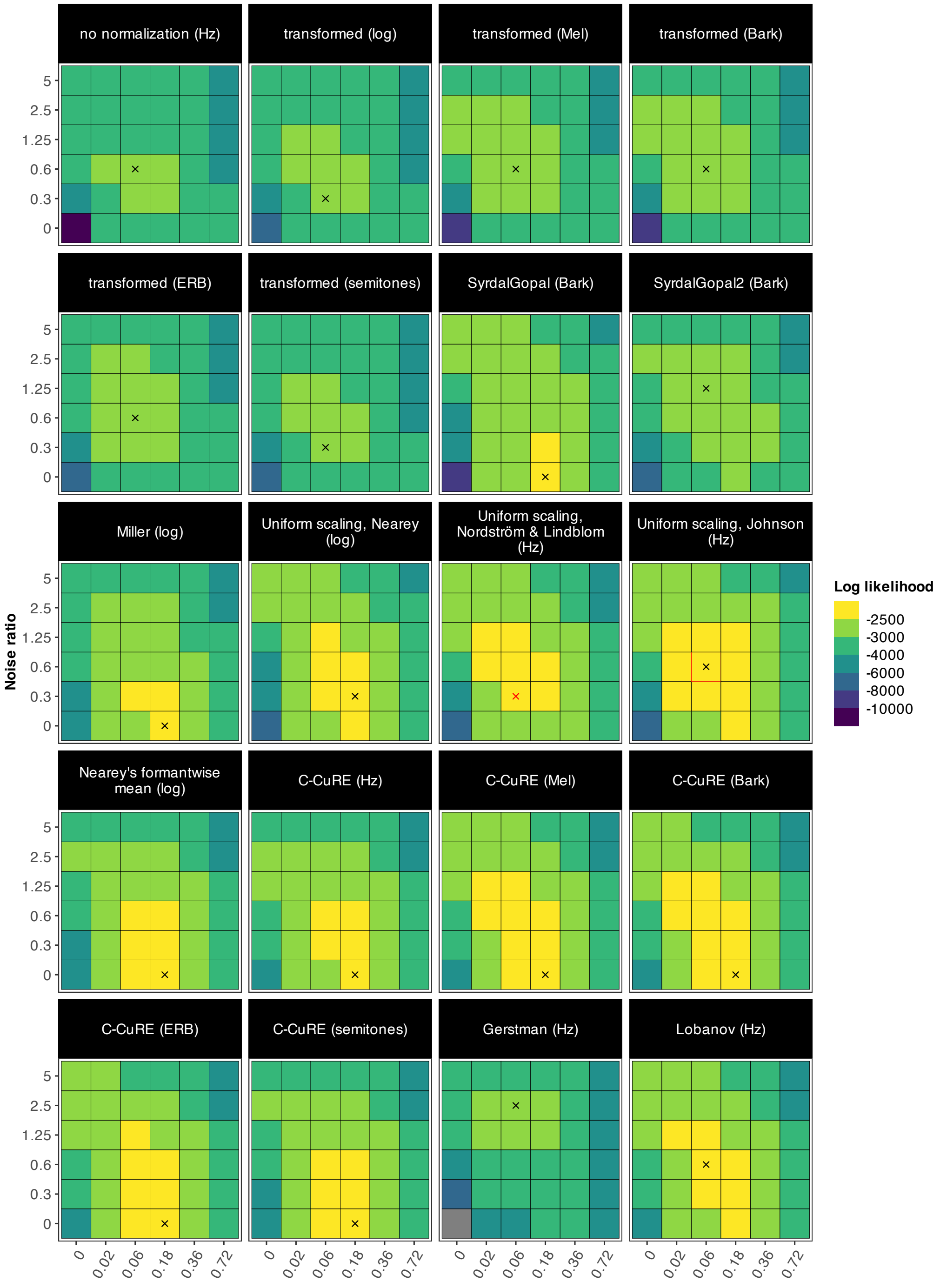

FIG. S16. Predicted likelihoods of ideal observer for human vowel responses in Experiment 1a, under different normalization accounts, different $\lambda$s and different $\tau^{-1}$s. Likelihood is aggregated across vowels. Crosses are placed at the combination of parameters for which the maximum likelihood for an account and a cross-validation fold was found. The red cross indicates the maximum likelihood across all accounts and folds.

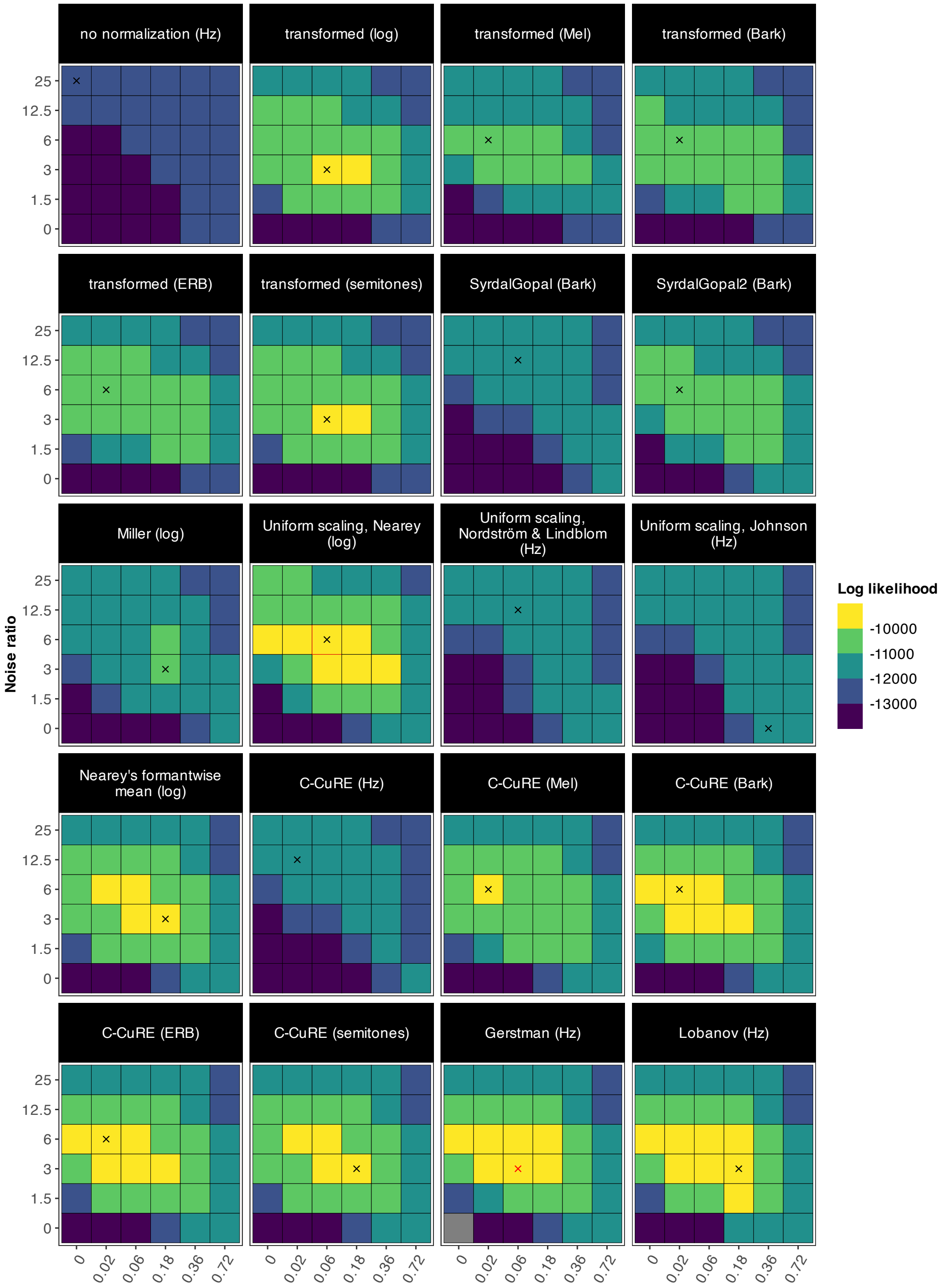

FIG. S17. Predicted likelihoods of ideal observer for human vowel responses in Experiment 1b, under different normalization accounts, different $\lambda$s and different $\tau^{-1}$s. Likelihood is aggregated across vowels. Crosses are placed at the combination of parameters for which the maximum likelihood for an account and a cross-validation fold was found. The red cross indicates the maximum likelihood across all accounts and folds.