

**Comparing accounts of formant normalization against US English listeners' vowel perception**

Anna Persson,<sup>1</sup> Santiago Barreda,<sup>2</sup> and T. Florian Jaeger<sup>3</sup>

<sup>1</sup>*Swedish Language and Multilingualism, Stockholm University, Stockholm, SE-106 91, Sweden*<sup>a</sup>

<sup>2</sup>*Linguistics, University of California, Davis*

<sup>3</sup>*Brain and Cognitive Sciences, Data Science, University of Rochester*

(Dated: 23 December 2024)

1 Human speech recognition tends to be robust, despite substantial cross-talker vari-  
2 ability. Believed to be critical to this ability are auditory normalization mechanisms  
3 whereby listeners adapt to individual differences in vocal tract physiology. This study  
4 investigates the computations involved in such normalization. Two 8-way alternative  
5 forced-choice experiments assessed L1 listeners' categorizations across the entire US  
6 English vowel space—both for unaltered and for synthesized stimuli. Listeners' re-  
7 sponds in these experiments were compared against the predictions of twenty influen-  
8 tial normalization accounts that differ starkly in the inference and memory capacities  
9 they imply for speech perception. This includes variants of *estimation-free* transfor-  
10 mations into psycho-acoustic spaces, *intrinsic* normalizations relative to concurrent  
11 acoustic properties, and *extrinsic* normalizations relative to talker-specific statistics.  
12 Listeners' responses were best explained by extrinsic normalization, suggesting that  
13 listeners learn and store distributional properties of talkers' speech. Specifically,  
14 *computationally simple* (single-parameter) extrinsic normalization best fit listeners'  
15 responses. This simple extrinsic normalization also clearly outperformed Lobanov  
16 normalization—a computationally more complex account that remains popular in  
17 research on phonetics and phonology, sociolinguistics, typology, and language acqui-  
18 sition.

---

<sup>a</sup>anna.persson@su.se

19 **I. INTRODUCTION**

20 One of the central challenges for speech perception originates in cross-talker variability:  
 21 depending on the talker, the same acoustic signal can encode different sound categories (Allen  
 22 *et al.*, 2003; Liberman *et al.*, 1967; Newman *et al.*, 2001). This results in ambiguity in the  
 23 mapping from acoustics to words and meanings. Research has identified several mechanisms  
 24 through which listeners resolve this ambiguity, ranging from early perceptual processes, to  
 25 adaptation of phonetic categories, all the way to adjustments in post-linguistic decision  
 26 processes (for review, see Xie *et al.*, 2023). The present study focuses on the first type of  
 27 mechanism, early auditory processes that transform and normalize the acoustic input into  
 28 the perceptual cues that constitute the input to linguistic processing (for reviews, Barreda,  
 29 2020; Johnson and Sjerps, 2021; McMurray and Jongman, 2011; Stilp, 2020; Weatherholtz  
 30 and Jaeger, 2016). We seek to respond, in particular, to recent calls to put theories of  
 31 adaptive speech perception to stronger tests (Baese-Berk *et al.*, 2018; Schertz and Clare,  
 32 2020; Xie *et al.*, 2023).

33 Evidence for the presence of early normalization mechanisms comes from neuroimaging  
 34 and neurophysiological studies (e.g., Oganian *et al.*, 2023; Skoe *et al.*, 2021), as well  
 35 as research on the peripheral auditory system suggesting automatic transformations of the  
 36 acoustic signal into scale-invariant spectral patterns (e.g., Patterson and Irino, 2014; Smith  
 37 *et al.*, 2005). Neurophysiological studies have further decoded effects of talker identity from  
 38 subcortical brain areas like the brain stem, and thus prior to the cortical regions believed to  
 39 encode linguistic categories (e.g., Sjerps *et al.*, 2019; Tang *et al.*, 2017). This includes brain

40 responses that lag the acoustic signal by as little as 20-50 msecs (Lee, 2009), suggesting very  
 41 fast and highly automatic processes. While this does not mean that *only* talker-normalized  
 42 auditory percepts are available to subsequent processing—there is now convincing evidence  
 43 that subcategorical information can enter listeners’ phonetic representations (e.g., Hay *et al.*,  
 44 2017, 2019; Johnson *et al.*, 1999; McGowan, 2015; Walker and Hay, 2011)—it does suggest  
 45 that normalized auditory percepts are available to subsequent processing. By removing  
 46 (some) cross-talker variability early during auditory processing, normalization offers an el-  
 47 egant and effective solution that can reduce the need for more complex adaptive processes  
 48 further upstream (Apfelbaum and McMurray, 2015; Xie *et al.*, 2023).

49 While it is relatively uncontroversial *that* normalization contributes to robust speech  
 50 perception, it is still unclear what types of computations this implicates. We address this  
 51 question for the perception of vowels, which cross-linguistically relies on peaks in the distri-  
 52 bution of spectral energy over acoustic frequencies (formants).<sup>1</sup> Vowel perception has long  
 53 been a focus in research on normalization (e.g., Bladon *et al.*, 1984; Fant, 1975; Gerstman,  
 54 1968; Johnson, 2020; Joos, 1948; Lobanov, 1971; Miller, 1989; Nearey, 1978; Nordström  
 55 and Lindblom, 1975; Syrdal and Gopal, 1986; Traunmüller, 1981; Watt and Fabricius, 2002;  
 56 Zahorian and Jagharghi, 1991; for review, see Barreda, 2020), with some reviews citing  
 57 over 100 competing proposals (Carpenter and Govindarajan, 1993). Importantly, these ac-  
 58 counts differ in the types and complexity of computations they assume to take place during  
 59 normalization.

60 On the lower end of computational complexity, *estimation-free* psycho-acoustic trans-  
 61 formations involve zero degrees of freedom that listeners would need to estimate from the

62 acoustic input. For example, there is evidence that a transformation of acoustic frequencies  
63 (measured in Hz) into the psycho-acoustic Bark-space better describes how listeners per-  
64 ceive differences along the frequency spectrum (in terms of critical bands, e.g., [Traunmüller](#),  
65 [1990](#); [Zwicker](#), [1961](#); [Zwicker et al.](#), [1957](#); [Zwicker and Terhardt](#), [1980](#)). It is thus possible  
66 that cross-talker variability in vowel pronunciations is reduced when formants are repre-  
67 sented in Bark, rather than Hz. Similar arguments have been made about other psycho-  
68 acoustic transformations (e.g., ERB, [Glasberg and Moore](#), [1990](#); Mel, [Stevens and Volkmann](#),  
69 [1940](#); or semitones, [Fant et al.](#), [2002](#)) most of which share that they log-transform acoustic  
70 frequencies—in line with neurophysiological evidence that the auditory representations in  
71 the brain seem to follow a roughly logarithmic organization, so that auditory perception is  
72 (up to a point) more sensitive to differences between lower frequencies than to the same  
73 difference between higher frequencies (e.g., [Merzenich et al.](#), [1975](#); for review, see [Saenz and](#)  
74 [Langers](#), [2014](#)). While each of these transformations was developed with different applica-  
75 tions in mind (e.g., ERB and Bark to explain frequency selectivity, [Glasberg and Moore](#),  
76 [1990](#); or semitones for the perception of musical pitch, [Balzano](#), [1982](#)), psycho-acoustic  
77 transformations might suffice for effective formant normalization. If so, this would offer a  
78 particularly parsimonious account of vowel perception as listeners would not have to infer  
79 talker-specific properties.

80 The parsimony of psycho-acoustic transformations contrasts with the majority of accounts  
81 for vowel normalization, which introduce additional computations. This includes accounts  
82 that normalize formants relative to other information that is available at the same point in  
83 the acoustic signal (intrinsic normalization, e.g., [Miller](#), [1989](#); [Peterson](#), [1961](#); [Syrdal and](#)

84 [Gopal, 1986](#)). For example, according to one proposal, listeners normalize vowel formants  
 85 by the vowel's fundamental frequency or other formants estimated at the same point in  
 86 time ([Syrdal and Gopal, 1986](#)). To the extent that the fundamental frequency is correlated  
 87 with the talkers' vocal tract size (for review, see [Vorperian and Kent, 2007](#)), this allows the  
 88 removal of physiologically-conditioned cross-talker variability in formant realizations. While  
 89 such intrinsic accounts arguably entail more computational complexity than estimation-  
 90 free transformations, they do not require that listeners *maintain* talker-specific estimates  
 91 over time. This distinguishes intrinsic from extrinsic accounts, which introduce additional  
 92 computational complexity.

93 According to extrinsic accounts, normalization mechanisms infer and store estimates  
 94 of talker-specific properties that then are used to normalize subsequent speech from that  
 95 talker ([Gerstman, 1968; Lobanov, 1971; Nearey, 1978; Nordström and Lindblom, 1975; Watt](#)  
 96 and Fabricius, 2002; for review, see [Weatherholtz and Jaeger, 2016](#)). At the upper end of  
 97 computational complexity, some accounts hold that listeners continuously infer and maintain  
 98 both talker-specific means for each formant and talker-specific estimates of each formant's  
 99 variability ([Gerstman, 1968; Lobanov, 1971](#)). These estimates are then used to normalize  
 100 formants, e.g., by centering and standardizing them (essentially z-scoring formants, [Lobanov,](#)  
 101 [1971](#)), removing cross-talker variability in the distribution of formant values. There are,  
 102 however, more parsimonious extrinsic accounts that require inference and maintenance of  
 103 fewer talker-specific properties. The most parsimonious of these is Nearey's *uniform scaling*  
 104 account, which assumes that listeners infer and maintain a single talker-specific parameter.  
 105 This parameter ( $\Psi$ ) can be thought of as capturing the effects of the talker's vocal tract

<sup>106</sup> length on the spectral scaling applied to the formant pattern produced by a talker (Nearey,  
<sup>107</sup> 1978).<sup>2</sup> Uniform scaling deserves particular mention here as it is arguably one of the most  
<sup>108</sup> developed normalization accounts, and rooted in principled considerations about the physics  
<sup>109</sup> of sound and the evolution of auditory systems (for review, see Barreda, 2020).

<sup>110</sup> In summary, hypotheses about the computations implied by formant normalization differ  
<sup>111</sup> in the flexibility they afford as well as the inference and memory complexity they entail.  
<sup>112</sup> Considerations about the complexity of inferences—essentially the number of parameters  
<sup>113</sup> that listeners are assumed to estimate at any given moment in time—arguably gain in  
<sup>114</sup> importance in light of the speed at which normalization seems to unfold. In the present  
<sup>115</sup> study, we thus ask whether computationally simple accounts are sufficient to explain human  
<sup>116</sup> vowel perception.

<sup>117</sup> While previous research has compared normalization accounts across languages, most of  
<sup>118</sup> this work has evaluated proposals in terms of how well the normalized phonetic space sup-  
<sup>119</sup> ports the separability of vowel categories (Adank *et al.*, 2004; Carpenter and Govindarajan,  
<sup>120</sup> 1993; Cole *et al.*, 2010; Escudero and Bion, 2007; Johnson and Sjerps, 2021; Syrdal, 1985).

<sup>121</sup> This approach is illustrated in Figure 1. These studies have found that computationally  
<sup>122</sup> more complex accounts—which also afford more flexibility—tend to achieve higher cate-  
<sup>123</sup> gory separability and higher categorization accuracy (for review, see Persson and Jaeger,  
<sup>124</sup> 2023). This includes Lobanov normalization, which continues to be highly influential in,  
<sup>125</sup> for example, variationist and sociolinguistic research because of its effectiveness in removing  
<sup>126</sup> cross-talker variability (for a critique, see Barreda, 2021). It is, however, by no means clear  
<sup>127</sup> that human speech perception employs the same computations that achieve the best cate-

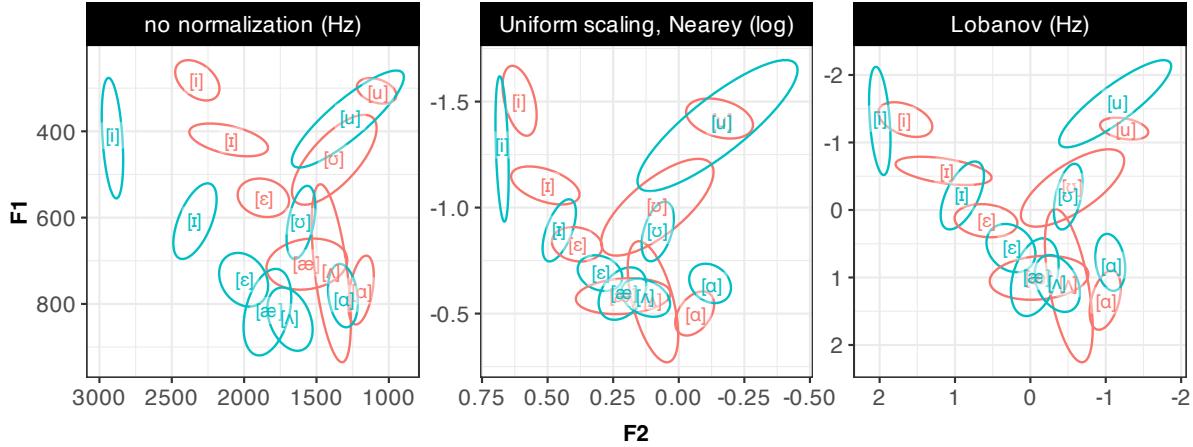


FIG. 1. Illustration of how height, which is positively correlated with vocal tract size, affects vowels' F1 and F2, and how normalization can partially remove this effect. Shown here are realizations of eight monophthong vowels of US English by a short (cyan) and a tall native talker (red). **Panel A:** In the acoustic space, prior to any normalization (Hz). **Panel B:** After uniform scaling (Nearey, 1978). **Panel C:** After Lobanov normalization (Lobanov, 1971). The present study compares these three accounts, along with 17 other normalization accounts. Here and throughout the paper, panel captions indicate the phonetic space in which normalization takes place in parenthesis. Note that this is not necessarily identical to the units of F1 and F2 *after* normalization (e.g., Lobanov normalization results in scale-free z-scores along the formant axes).

128 gory separability or accuracy (see also discussion in Barreda, 2021; Nearey and Assmann, 129 2007).

130 A substantially smaller body of research has addressed this question by comparing nor-  
 131 malization accounts against *listeners' perception* (Barreda and Nearey, 2012; Barreda, 2021;  
 132 Nearey, 1989; Richter *et al.*, 2017; for a review, see Whalen, 2016). Interestingly, these  
 133 works seem to suggest that computationally simpler accounts might provide a better fit  
 134 against human speech perception than the influential Lobanov model (Barreda, 2021; Richter  
 135 *et al.*, 2017). For example, Barreda (2021) compared the predictions of uniform scaling and  
 136 Lobanov normalization against listeners' categorization responses in a forced-choice cat-

137 egorization task over parts of the US English vowel space. In his experiment, listeners'  
 138 categorization responses were better predicted by uniform scaling than by Lobanov nor-  
 139 malization. Findings like these suggest that comparatively simple corrections for vocal tract  
 140 size—such as uniform scaling—might provide a better explanation of human perception than  
 141 more computationally complex accounts (see also [Johnson, 2020](#); [Richter \*et al.\*, 2017](#)).

142 This motivates the present work. We take a broad-coverage approach by comparing the  
 143 20 normalization accounts in Table 1 against the perception of eight monophthongs of US  
 144 English [i] as in *heed*, [ɪ] in *hid*, [ɛ] in *head*, [æ] in *had*, [ʌ] in *hut*, [ʊ] in *hood*, [ʊ] in *who'd*,  
 145 [ɑ] in *odd*).<sup>3</sup> We do so for the perception of both natural and synthesized speech. Our  
 146 broad-coverage approach complements previous studies, which have typically compared a  
 147 small number of accounts (up to 3) and focused on parts of the vowel inventory, and thus  
 148 parts of the formant space (typically 2-4 vowels, [Barreda, 2021](#); [Barreda and Nearey, 2012](#);  
 149 [Nearey, 1989](#); [Richter \*et al.\*, 2017](#)). The accounts we consider include the most influen-  
 150 tial examples of psycho-acoustic transformations ([Fant \*et al.\*, 2002](#); [Glasberg and Moore,](#)  
 151 [1990](#); [Stevens and Volkmann, 1940](#); [Traunmüller, 1981](#)), intrinsic ([Syrdal and Gopal, 1986](#)),  
 152 extrinsic ([Gerstman, 1968](#); [Johnson, 2020](#); [Lobanov, 1971](#); [McMurray and Jongman, 2011](#);  
 153 [Nearey, 1978](#); [Nordström and Lindblom, 1975](#)), and hybrid accounts that contain intrin-  
 154 sic and extrinsic components ([Miller, 1989](#)). This broad-coverage approach allows us to  
 155 assess, for example, whether the preference for computationally simple accounts observed  
 156 in [Barreda \(2021\)](#) replicates on new data that span the entire vowel space. It also allows  
 157 us to ask whether accounts even simpler than uniform scaling—such as psycho-acoustic  
 158 transformations—provide an even better fit to human perception.

TABLE I. Normalization accounts considered in the present study. Unless otherwise marked, formant variables ( $F$ s) in the right-hand side of normalization formulas are in Hz.

	Normalization procedure	Perceptual scale	Source	Formula
<b>extrinsic standardizing</b>		<b>No normalization</b>	<b>Hz</b>	<b>n/a</b>
<b>transformation</b>		log		$F_n^{\log} = \ln(F_n)$
Bark			Traunmüller (1990)	$F_n^{\text{Bark}} = \frac{26.81 \times F_n}{1960 + F_n} - 0.53$
ERB			Glasberg & Moore (1990)	$F_n^{\text{ERB}} = 21.4 \times \log_{10}(1 + F_n) \times 0.00437$
Mel			Stevens & Volkmann (1940)	$F_n^{\text{Mel}} = 2595 \times \log_{10}(1 + \frac{F_n}{700})$
Semitones conversion			Fant et al. (2002)	$F_n^{\text{ST}} = 12 \times \frac{\ln(\frac{F_n}{100})}{\ln}$
<b>intrinsic</b>				
Syrdal & Gopal 1 (Bark-distance model)		Bark	Syrdal & Gopal (1986)	$F1^{\text{SyrdalGopal1}} = F1^{\text{Bark}} - F0^{\text{Bark}}$
Syrdal & Gopal 2 (Bark-distance model)				$F2^{\text{SyrdalGopal1}} = F2^{\text{Bark}} - F1^{\text{Bark}}$
Miller (formant-ratio)		log	Miller (1989)	$F1^{\text{SyrdalGopal2}} = F1^{\text{Bark}} - F0^{\text{Bark}}$
				$F2^{\text{SyrdalGopal2}} = F3^{\text{Bark}} - F2^{\text{Bark}}$
				$SR = k(\frac{GMf0}{k})^{1/3}$
				$F1^{\text{Miller}} = \log(\frac{F1}{SR})$
				$F2^{\text{Miller}} = \log(\frac{F2}{F1})$
				$F3^{\text{Miller}} = \log(\frac{F3}{F2})$
<b>extrinsic centering</b>				
Nearey's uniform scaling		log	Nearey (1978)	$F_n^{\text{Nearey}} = \ln(F_n) - \text{mean}(\ln(F))$
Nordström & Lindblom (vocal tract scaling)		Hz	Nordström & Lindblom (1975)	$F_n^{\text{NordströmLindblom}} = \frac{F_n}{\text{mean}(\frac{F3 F1 \geq 600}{2.5})}$
Johnson (average formant spacing)		Hz	Johnson(2020)	$F_n^{\text{Johnson}} = \frac{F_n}{\text{mean}(\frac{F1 \cdot F2}{1.5 \cdot 2.5})}$
Nearey's formantwise log-mean		log	Nearey (1978)	$F_n^{\text{Nearey}} = \ln(F_n) - \text{mean}(\ln(F_n))$
C-CuRE		Hz	McMurray & Jongman (2011)	$F_n^{C-CuRE} = F_n - \text{mean}(F_n)$
		Bark		
		ERB		
		Mel		
Semitones conversion				
<b>extrinsic</b>				
<b>standardizing</b>				
Gerstman (range normalization)		Hz	Gerstman (1968)	$F_n^{\text{Gerstman}} = 999 \times \frac{F_n - F_n^{\text{min}}}{F_n^{\text{max}} - F_n^{\text{min}}}$
Lobanov (z-score)		Hz	Lobanov (1971)	$F_n^{\text{Lobanov}} = \frac{F_n - \text{mean}(F_n)}{\text{sd}(F_n)}$

<sub>159</sub> Next, we motivate and describe the two experiments we conducted. Then we compare  
<sub>160</sub> the normalization accounts in Table 1 against listeners' responses from these experiments.

<sub>161</sub> **A. Open Science Statement**

<sub>162</sub> All stimulus recordings, results, and the code for the experiment, data analysis, and com-  
<sub>163</sub> putational modeling for this article can be downloaded from the Open Science Framework  
<sub>164</sub> (OSF) at <https://osf.io/zemwn/>. The OSF repository also include extensive supple-  
<sub>165</sub> matory information (SI). Both the article and SI are written in R markdown, allowing readers  
<sub>166</sub> to replicate our analyses with the click of a button, using freely available software ([R Core](#)  
<sub>167</sub> [Team, 2024](#); [RStudio Team, 2020](#)). Readers can revisit the assumptions we committed to  
<sub>168</sub> for the present project—for example, by substituting alternative normalization accounts or  
<sub>169</sub> categorization models. Researchers can also substitute their own experiments on vowel nor-  
<sub>170</sub> malization for our Experiments 1a and 1b, to see whether our findings generalize to novel  
<sub>171</sub> data. We see this as an important contribution of the present work, as it should make it  
<sub>172</sub> substantially easier to consider additional normalization accounts—including variants to the  
<sub>173</sub> accounts we considered—and to assess the generalizability of the conclusions we reach based  
<sub>174</sub> on the present data.

<sub>175</sub> **II. EXPERIMENTS 1A AND 1B**

<sub>176</sub> To compare the performance of different normalization accounts against listeners' per-  
<sub>177</sub> ception, we conducted two small web-based experiments on US English listeners' perception  
<sub>178</sub> of US English vowels. Both experiments investigate listeners' perception of a single talker.

179 This choice was made so as to not confound questions about formant normalization with  
 180 questions about talker recognition, and inferences about talker switches (Magnuson and Nus-  
 181 baum, 2007). The two experiments employ the same eight-alternative forced-choice vowel  
 182 categorization task (Figure 2), and differ only in the whether they employed ‘natural’ (Ex-  
 183 periment 1a) or synthesized stimuli (Experiment 1b). To the best of our knowledge, these  
 184 two experiments are the first designed to compare normalization accounts against listeners’  
 185 perception over a larger portion of the monophthong inventory of a language.

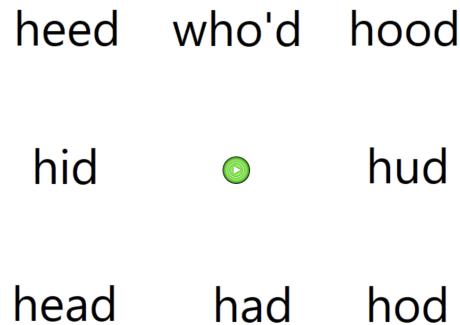


FIG. 2. Screen shot of the eight-alternative forced-choice (8-AFC) task used in both Experiment 1a and 1b.

186 Experiment 1a employs recordings of *hVd* word productions from a female talker of  
 187 US English, these recordings are ‘natural’ in the sense that they were not synthesized or  
 188 otherwise phonetically manipulated. One consequence of this is that the formant values  
 189 of these recordings are clustered around the talker’s category means, and thus span only  
 190 a comparatively small part of the phonetic space. This can limit the statistical power to  
 191 distinguish between competing accounts. Natural recordings furthermore vary not only  
 192 along the primary cues to vowel quality in US English (F1, F2) but also along secondary

193 cues (e.g., F0, F3, vowel duration, and vowel inherent spectral change—VISC) as well as  
194 other unknown properties, which can make it difficult to discern whether the performance  
195 of a normalization model is due to the normalization itself or other reasons, e.g., because  
196 a normalized cue happens to correlate with another cue that listeners are sensitive to but  
197 that is not included in the model.

198 Experiment 1b thus adopts an alternative approach and uses synthesized vowels. Unlike  
199 most previous work, which has used isolated vowels as stimuli ([Barreda, 2021](#); [Barreda and](#)  
200 [Nearey, 2012](#); [Nearey, 1989](#); [Richter \*et al.\*, 2017](#)), Experiment 1b uses synthesized *hVd* words  
201 to facilitate comparison to Experiment 1a. This allowed us to sample larger parts of the F1-  
202 F2 space, which has two advantages. First, it allowed us to collect responses over parts of the  
203 formant space for which we expect listeners to have more uncertainty, and thus exhibit more  
204 variable responses. This can increase the statistical power to distinguish between competing  
205 accounts. Second, differences in the predictions of competing normalization accounts will  
206 tend to become more pronounced with increasing distance from the category centers. By  
207 collecting responses at those locations, we can thus increase the contrast between competing  
208 accounts. Critically, an adequate model of formant normalization needs to capture human  
209 perception not only for prototypical vowel instances, but also instances of vowels that fall  
210 between category means.

211 The use of synthesized stimuli does, however, also come with potential disadvantages.  
212 Synthesized stimuli can suffer in ecological validity, lacking correlations between cues, and  
213 across the speech signal (e.g., due to co-articulation) that are characteristic of human speech.  
214 This raises questions about the extent to which processing of such stimuli engages the same

215 mechanisms as everyday speech perception. Additionally, it is possible that the use of robotic  
216 sounding synthesized speech affects listener engagement. This can lead to an increased rate  
217 of attentional lapses, and thus a decrease in the proportion of trials on which listeners'  
218 responses are based on the acoustics of the speech stimulus rather than random guessing  
219 (compare, e.g., [Kleinschmidt, 2020](#); [Tan and Jaeger, 2024](#)). By comparing normalization  
220 accounts against both natural and synthesized stimuli, we investigate the extent to which  
221 the accounts that best describe human perception depend on the type of stimuli used in the  
222 experiment.

223 **A. Methods**

224 **1. Participants**

225 We recruited 33 (Experiment 1a) and 33 (Experiment 1b) participants. The majority of  
226 these (24 for each experiment) were recruited from Amazon's Mechanical Turk. However,  
227 after exclusions we were left with a relatively low number of participants (for Experiment 1a,  
228 19, and for Experiment 1b, 22). We therefore decided to recruit an additional 18 participants  
229 from Prolific (9 for each experiment; October 2024). Exclusions described below left 28 and  
230 31 participants for analysis in Experiments 1a and 1b, respectively. Results did not change  
231 after inclusion of the new participants from Prolific.

232 Participants were paid \$6/hour (\$12/hour on Prolific) prorated by the duration of the  
233 experiments (15 minutes). Participants only saw the experiment advertised, and could only  
234 participate in it, if (i) they were located within the US, (ii) had an approval rating of

235 99% or higher, (iii) met the software requirements (a recent version of the Chrome browser  
236 engine), and (iv) had not previously completed any other experiments on vowel perception  
237 in our lab. Before the experiment could be accepted, participants had to confirm that they  
238 were (i) native speakers of US English (defined as having spent their childhood until the  
239 age of 10 speaking English and living in the United States), (ii) in a quiet room without  
240 distractions, (iii) wearing over-the-ear headphones. Participants' responses were collected via  
241 Javascript developed by the Human Language Processing Lab at the University of Rochester  
242 ([Kleinschmidt \*et al.\*, 2021](#)).

243 An optional post-experiment survey recorded participant demographics using NIH pre-  
244 scribed categories, including participant sex (Male: 36, Female: 29), age (mean = 36.9 years;  
245 SD = 12.2; 95% quantiles = 22.6-66 years), race (White: 48, multiple: 3, Black: 10, Asian:  
246 3, declined to report: 1), and ethnicity (Non-Hispanic: 60, Hispanic: 4, declined to report:  
247 1). All but 1 participant completed the survey.

248 **2. Materials**

249 Experiment 1a employed *hVd* word recordings by one adult female talker of a Northeast-  
250 ern dialect (spoken in central Connecticut) from a phonetically annotated database of L1-US  
251 English vowel productions ([Xie and Jaeger, 2020](#)). Specifically, we used all nine recordings  
252 of each of the eight *hVd*-words—*heed*, *hid*, *head*, *had*, *hut*, *odd*, *hood*, *who'd* (the use of “hut”  
253 and “odd” rather than “hud” and “hod” follows [Assmann \*et al.\*, 2008](#); but see [Hillenbrand  
254 \*et al.\*, 1995](#)).

255 The stimuli for Experiment 1b were synthesized from a single *had* recording used in  
 256 Experiment 1a (see Figure 3 for example spectrograms). Specifically, we used a script  
 257 (based on descriptions in Wade *et al.*, 2007) in Praat (Boersma and Weenink, 2022) to  
 258 concatenate the original /h/ with a synthesized vowel and the original /d/ recording. Unlike  
 259 in Experiment 1a, all eight words thus had an *hVd* context (including “hud” and “hod”,  
 260 rather than “hut” and “odd”). The Praat script first segmented the original *had* token  
 261 into the three segments /h/, /æ/ and /d/, with the /d/ segment consisting of the voiced  
 262 closure and burst. The script then estimated the spectral envelope of the /h/ sound by  
 263 linear predictive coding (LPC; autocorrelation method), and used the resulting coefficients  
 264 to inversely filter the /h/. This resulted in an /h/ sound with effects of vocal tract removed,  
 265 leaving the source signal. Next, a glottal waveform was generated at each point in the pitch  
 266 contour from the original /æ/ sound using the point process to phonation functionality in  
 267 Praat. This waveform was multiplied with the intensity pattern from the same original  
 268 /æ/ sound. The resulting sound was concatenated with the neutral fricative /h/ sound,  
 269 to create a neutral hV-section that did not reflect any vocal tract resonances. The script  
 270 then created a formant grid that filtered the hV-section to create the intended vowel, and  
 271 finally concatenated this segment to the final /d/ to create an *hVd* word. For each *hVd*  
 272 word, the formant grid was populated with the F1, F2 and F3 values that we handed to  
 273 the script at five time-points transitioning from the /h/ to the steady-state vowel, to the  
 274 first portion of the voiced closure of the final /d/ segment through linear interpolation, thus  
 275 holding formants steady until transitioning into the final consonant. Formant bandwidths  
 276 were 500 Hz at the initial two time-points (the /h/ and beginning of transition to vowel),

277 and then decreased linearly during vowel onset and throughout the final three time-points  
 278 to 50 Hz (F1), 100 Hz (F2), 200 Hz (F3), 300 Hz (F4), and 400 Hz (F5-F8, following Wade  
 279 *et al.*, 2007). The bandwidth manipulation implied that the spectral peaks of the formants  
 280 became more defined and more separated as the vowel unfolded. We used this approach  
 281 to create synthesized vowels for arbitrary F1-F2 combinations. F3 was set based on those  
 282 F1-F2 values. Specifically, we ran a linear regression over the natural productions of the  
 283 talker from Experiment 1a, predicting F3 from F1, F2 and their interaction. We then used  
 284 that regression to predict F3 values for any F1-F2 combination in Experiment 1b. F4 to  
 285 F8, as well as vowel duration, were held identical across all tokens (using the automatically  
 286 extracted vowel duration and mean formant values across the vowel segment from the *had*  
 287 token used for resynthesis).

288 We generated 146 synthesized *hVd* recordings that spanned the F1 and F2 space. The  
 289 specific F1-F2 locations chosen were determined by a mix of modeling (using ideal observers  
 290 described in the next section to predict listeners' categorization responses) and intuition.  
 291 Specifically, we selected 64 recordings that we expected to fall within the bivariate 95%  
 292 confidence intervals (CIs) of the eight US English monophthongs, and 82 recordings that we  
 293 expected to fall between those CIs. Figure 4 under *Results* shows the distribution of stimuli  
 294 for both experiments. Of note, our procedure also generated formant combinations that are  
 295 physiologically unlikely to have all been produced by the same talker during 'normal' vowel  
 296 production (also known as "off-template" instances, Nearey, 1978).

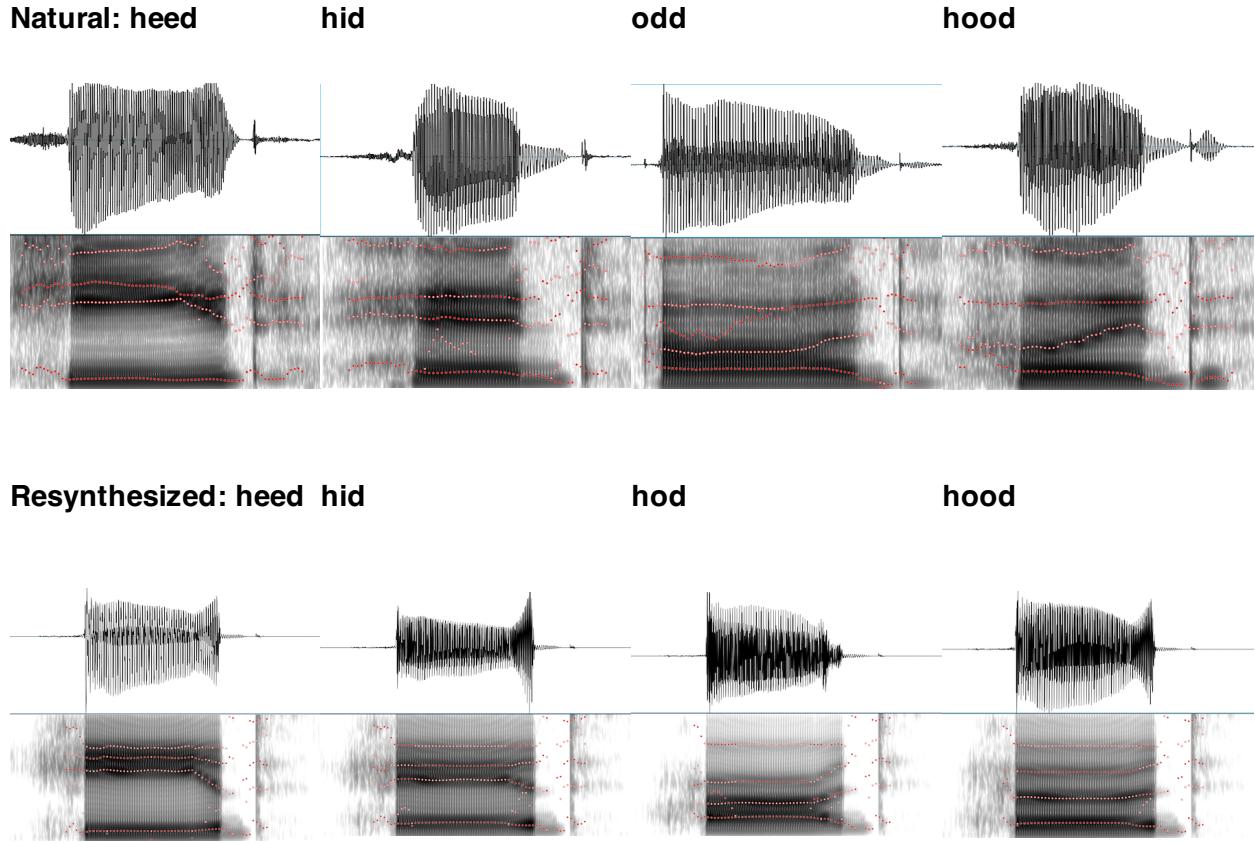


FIG. 3. **Top:** Spectrograms of four natural recordings from Experiment 1a. **Bottom:** Same for four synthesized tokens with similar formant values from Experiment 1b. Additional spectrograms are provided in the SI §2 C.

297      **3. Procedure**

298      The procedure for both experiments was identical. Live instances of each experiment  
 299      can be found at <https://www.hlp.rochester.edu/experiments/DLPL2S/experiment-A/experiments.html>. At the start of the experiment, participants acknowledged that they  
 300      met all requirements and provided consent, as per the Research Subjects Review Board of  
 301      the University of Rochester. Before starting the experiment, participants performed a sound  
 302      check. Participants were then instructed to listen to a female talker saying words, and click  
 303      on the word on screen to report what word they heard. On each trial, all eight *hVd*-words

305 were displayed on screen. Half of the participants in each experiment saw the response  
 306 options organized as in Figure 2 (resembling the IPA representation of a vowel space), half  
 307 saw the response options in the opposite order (flipping top and bottom and left and right  
 308 in Figure 2). Each trial started with the response grid on screen, together with a light green  
 309 dot centered on screen. After 1000 ms, an *hVd* recording played, and participants indicated  
 310 their response by a mouse-click. After a 1000 ms intertrial interval, the screen reset, and  
 311 the next trial started.

312 In both experiments, participants heard two blocks of the materials described in the  
 313 previous sections, for a total of 144 trials in Experiment 1a and 292 trials in Experiment  
 314 1b. Presentation within each block was randomized for each participant in order to reduce  
 315 confounds due to stimulus order (known to affect vowel perception, [Repp and Crowder, 1990](#), and references therein). Participants were not informed about the block structure of  
 317 the experiment.

318 After completing the experiment, participants filled out a language background question-  
 319 naire and the optional demographic survey. On average, participants took 9.3 minutes to  
 320 complete Experiment 1a ( $SD = 5.5$ ) and 17.9 minutes for Experiment 1b ( $SD = 6.5$ ).

321 **4. Exclusions**

322 We excluded participants who failed to follow instructions and did not wear over-the-ear  
 323 headphones (as indicated in the post-experiment survey). We also excluded participants  
 324 with mean (log-transformed) reaction times that were unusually slow or fast (absolute z-  
 325 score over by-participant means  $> 3$ ), or if they clearly did not do the task (e.g., by answering

326 randomly). This excluded 5 participants from Experiment 1a and 2 from Experiment 1b  
327 (for details, see SI §2A).

328 We further excluded all trials that were unusually fast or slow. Specifically, we first z-  
329 scored the log-transformed response times *within each participant* and then z-scored these  
330 z-scores *within each trial* across participants. Trials with absolute z-scores  $> 3$  were removed  
331 from analysis. This double-scaling approach was necessary as participants' response times  
332 decreased substantially over the first few trials and then continued to decrease less rapidly  
333 throughout the remainder of the experiment. The approach removes response times that are  
334 unusually fast or slow *for that participant at that trial*, while avoiding specific assumptions  
335 about the shape of the speed up in response times across trials. This excluded 1.3% of the  
336 trials in Experiment 1a and 0.9% in Experiment 1b. This left for analysis 3983 observa-  
337 tions from 28 participants in Experiment 1a, and 8970 observations from 31 participants in  
338 Experiment 1b.

339 **B. Results**

340 Participants' categorization responses in Experiments 1a and 1b are shown in Figure 4,  
341 with larger labels indicating recordings that participants agreed on more.<sup>4</sup> We make two  
342 observations. The first pertains to the degree of (dis)agreement between the two experiments.  
343 The second observation pertains to the degree of (dis)agreement across participants within  
344 each experiment.

345

### 1. Similarities and differences between Experiments 1a and 1b

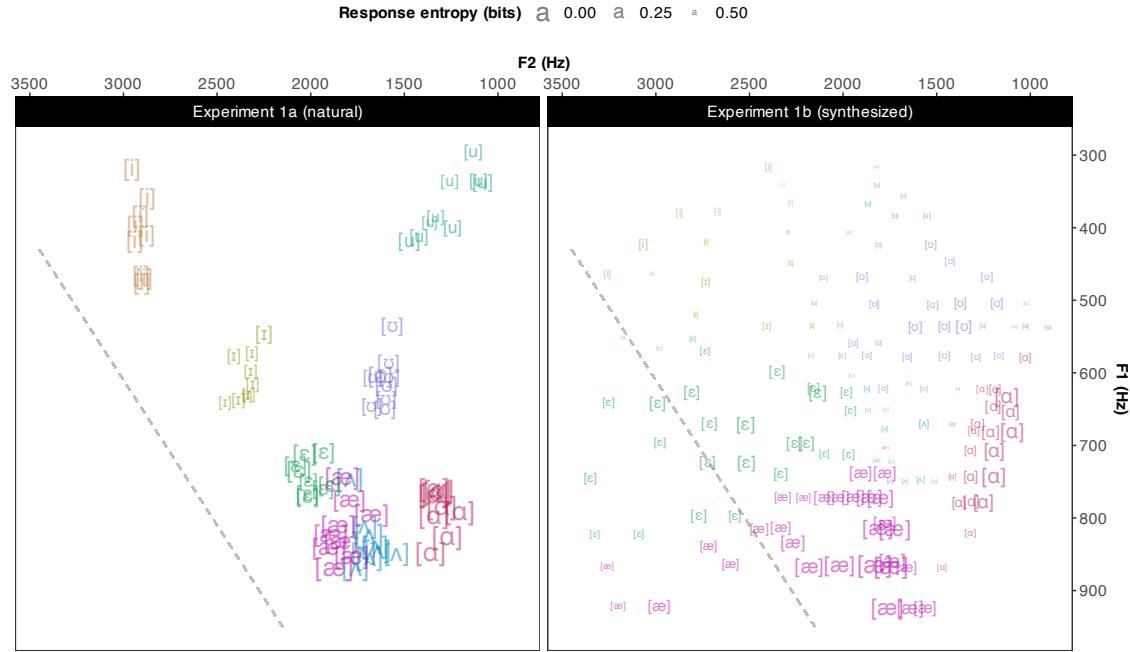


FIG. 4. Summary of listeners' categorization responses in Experiments 1a and 1b in F1-F2 space. The vowel label indicates the most frequent response provided across participants on each test location. Size indicates how consistent responses were across participants, which larger symbols indicating more consistent responses (lower entropy). F1-F2 combinations below the gray dashed line are unlikely to be articulated by the same talker.

346

Unsurprisingly, participants in both experiments divided the F1-F2 space into the eight

347 vowel categories in ways that qualitatively resembled each other (after taking into account

348 that Experiment 1b covers a larger range of F1-F2 values). Also unsurprisingly, there were

349 some differences between participants' responses across the two experiments, at least when

350 plotted in Hz. For example, [u] rarely was the most frequent response in Experiment 1b, even

351 for stimuli with similar F1-F2 values that were predominantly categorized as [u] in Experi-

352 ment 1a. There are at least two reasons to expect such differences. First, stimuli with similar

353 F1-F2 values across the two experiments still differed in other acoustic properties (e.g. vowel

duration or F3). These acoustic differences might have affected participants' responses. Second, it is possible that *formant normalization* affected participants' responses—i.e., the very mechanism we seek to investigate in the remainder of the paper. The two experiments differ in the means, variances, and other statistical properties that some normalization accounts predict to affect perception. As a consequence, Hz might not be the space in which we should expect identical responses across experiments.

Similarly, the two experiments differed in the extent to which participants agreed with each other. Participants in Experiment 1b exhibited overall less agreement in their responses (mean by-item response entropy = 0.45 bits, SE = 0.01) than participants in Experiment 1a (mean by-item response entropy = 0.19 bits, SE = 0.02). This was also confirmed by participants' responses during the post-experiment survey. Compared to participants in Experiment 1a, participants in Experiment 1b reported increased uncertainty about their responses, and that the stimuli were less distinguishable and more robotic-sounding (see SI [§2 B](#)).

This increased uncertainty in Experiment 1b was expected—and, indeed, intended by the design: Experiment 1b explored the entire F1-F2 space, including formant combinations located *between* the centers of the natural vowel categories. Experiment 1b therefore achieved its goal of eliciting less categorical response distributions, which is expected to facilitate comparison of competing normalization accounts.<sup>5</sup>

Auxiliary analyses presented in the SI ([§2 E](#)) suggest that *some but not all* of the differences in response entropy between the two experiments were caused by the placement of the stimuli in formant space: when comparing categorization responses for tokens from the

376 two experiments with similar acoustic properties (differences of  $\leq 30$  Hz along F1 and F2),  
 377 response entropies still differed substantially (for  $N = 40$  acoustically similar tokens, mean  
 378 by-item response entropy for Experiment 1a = 0.14 bits, SE = 0.02; Experiment 1b = 0.4  
 379 bits, SE = 0.03). The same section of the SI ([§2 E](#)) presents additional analyses grouping  
 380 acoustically similar tokens in the phonetic space defined by the normalization account we  
 381 find to best fit listeners' responses. These analyses support the same conclusion.

382 We see two mutually compatible explanations to this difference in listener agreement  
 383 between experiments. First, similar to the differences between experiments in the dominant  
 384 response pattern discussed above, differences in the degree of agreement between participants  
 385 might originate in *normalization*. Second, it is possible that the relation between formants  
 386 in the synthesized stimuli or some other unknown acoustic-phonetic differences between  
 387 the experiments explain the difference in response. For example, the absence of VISC  
 388 or differences in spectral tilt in the synthesized stimuli might have deprived listeners of  
 389 information that is actually crucial for establishing phonemic identity ([Hillenbrand and](#)  
 390 [Nearey, 1999](#)). This would result in increased uncertainty on each trial, leading to increased  
 391 entropy of listeners' responses. The computational study we present below shed some light  
 392 on these two mutually compatible possibilities.

393 ***2. Similarities and differences between participants***

394 Since the intended category was known for Experiment 1a, it was possible to calculate  
 395 participants' recognition accuracy. As also evident in the left panel of Figure 4, participants'  
 396 most frequent response *always* matched the intended vowel in Experiment 1a. Overall,

397 participants' responses matched the intended vowel on 84.7% (SE = 3.5%) of all trials  
 398 (Experiment 1b had no such ground truth). This is much higher than chance (12.5%). It  
 399 is, however, also quite a bit lower than 100%. To better understand the reasons for this,  
 400 Figure 5A plots the confusion matrix. This suggests that participants' performance was  
 401 largely affected by confusions between [i]-to-[ɛ] (*hid-to-head*), [ɛ]-to-[æ] (*head-to-had*), and  
 402 [u]-to-[ʊ] (*who'd-to-hood*).

403 One plausible explanation for this pattern of vowel confusions lies in the substantial  
 404 variation that exists across US English dialects (Labov *et al.*, 2006). Differences in the  
 405 realization of vowel categories, and associated representations, across dialects will directly  
 406 affect the expected classification for any given token. In addition, listeners might differ in  
 407 terms of experience with different dialects, or in the dialect they attribute to the talker who  
 408 produced the stimuli. To test this hypothesis, we calculated the [i]-to-[ɛ], [ɛ]-to-[æ], and  
 409 [u]-to-[ʊ] confusion rates for each participant in Experiment 1a. These data are summarized  
 410 in the left panel of Figure 5B. The data in the left panel suggest that most participants in  
 411 Experiment 1a either heard [i] tokens consistently as the intended [i] (clustering on the left  
 412 side of the panel) or as [ɛ] (clustering on the right side of the panel). Only a few participants  
 413 exhibited mixed responses for items intended to be [i]. Tellingly, many of the participants  
 414 who exhibited increased [i]-to-[ɛ] confusion *also* exhibited increased [ɛ]-to-[æ] confusion. This  
 415 is precisely what would be expected by listeners who assume a dialect in which these vowels  
 416 are articulated lower (with higher F1) than in the dialect of the talker in Experiment 1a.  
 417 A similar, but less pronounced, pattern was also found with regard to [u]-to-[ʊ] confusions.<sup>6</sup>  
 418 Finally, a qualitatively similar relation between [i]-to-[ɛ], [ɛ]-to-[æ], and [u]-to-[ʊ] confusions

A

Experiment 1a (natural)											Experiment 1b (synthesized)										
	[i]	[ɪ]	[ɛ]	[æ]	[ʌ]	[ɑ]	[ɔ]	[ʊ]			[i]	[ɪ]	[ɛ]	[æ]	[ʌ]	[ɑ]	[ɔ]	[ʊ]			
Response vowel	[u]	0	0.6	0.2	0	0.2	0	8.1	71.1		4.8	6.9	0.5	0.4	3.4	5.1	24.2	38.2			
[v]	0	0.4	0.4	0	0	0	0	78.7	27.3		2.1	7.7	2.9	0.9	12.6	7.7	45.9	32.5			
[ɑ]	0	0.2	0.2	0.8	0.4	98.4	0.2	0.2	0.2		0.9	1	0.2	3.1	14.3	59	5.8	3.3			
[ʌ]	0	0	1.2	0.2	0.2	96.2	0.2	3	0.4		2.5	1.8	3	4.8	39	18.2	13.1	13			
[æ]	0.4	0.4	12	93	0.8	0.6	2.6	0.4			2.1	4.7	22.5	67.3	19.4	9.1	2.6	2.4			
[ɛ]	1.8	28.3	83.6	5	0.8	0.2	5.1	0			9.7	20.9	55.6	21.4	10.7	0.4	3.8	2			
[i]	8.3	67.5	1.8	0.4	1.4	0.4	2	0.2			31.6	41.5	10.9	1.3	0.4	0.3	3	2.4			
[i]	89.5	2.6	0.6	0.6	0.2	0.2	0.2	0.4			46.2	15.6	4.4	0.6	0.1	0.2	1.6	6.1			

B

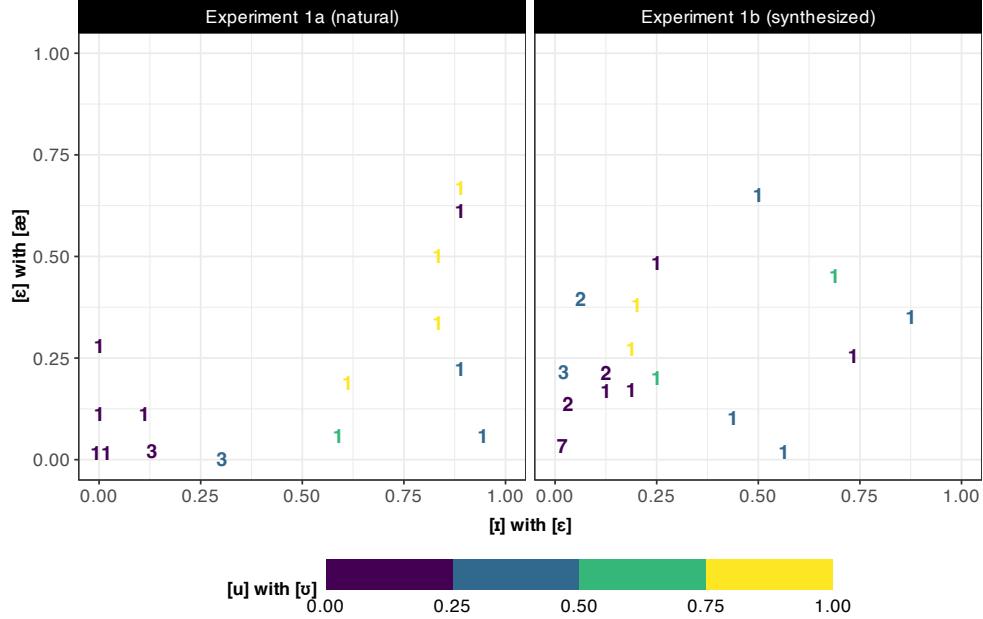


FIG. 5. Category confusability in Experiments 1a and 1b. **Panel A** summarizes the category confusability. Since correct responses were not defined for Experiment 1b, we grouped items along the x-axis based on most frequent response that listeners provided (for Experiment 1a, this was always identical to the intended response). Response percentages sum to 100 in each column, showing the response distribution depending on the most frequent response. **Panel B** summarizes individual differences across listeners, in terms of the listener-specific confusability of [i] with [ɛ] (x-axis), [ɛ] with [æ] (y-axis), and [u] with [ʊ] (color fill).

<sup>419</sup> was also observed in Experiment 1b (right panel of Figure 5B), though the pattern was

420 unsurprisingly less pronounced given that the stimuli in Experiment 1b by design often fell  
 421 into the ambiguous region *between* vowels. Taken together, vowel-to-vowel confusion rates  
 422 in Experiments 1a and 1b suggest that systematic dialectal differences contributed to the  
 423 relatively low categorization accuracy.

424 This highlights two important points. First, the data from Experiment 1a demonstrate  
 425 the perceptual challenges associated with an unfamiliar talker: in the absence of lexical or  
 426 other context to distinguish between the eight available response options, listeners can only  
 427 rely on the acoustic information in the input. In such a scenario, even listeners who are in  
 428 principle familiar with the dialect spoken by the talker have comparatively little informa-  
 429 tion to determine the talker’s dialect, making apparent what Winn (2018) aptly summarizes  
 430 as “speech [perception] is not as acoustic as [we] think”. Second, when dialect variability  
 431 is taken into account, listeners’ recognition accuracy improved substantially. After remov-  
 432 ing 8 listeners who heard more than 50% of the [i] items as [ɛ], *all* vowels were correctly  
 433 recognized at least 87.1% of the time (overall accuracy = 94.8%). This suggests that di-  
 434 alect differences affected the recognition of all vowels. This aspect of our results serves as  
 435 an important reminder that formant normalization is only expected to erase inter-talker  
 436 variability associated with *physiological* differences: variation in dialect, sociolect, or other  
 437 non-physiologically-conditioned variation pose separate challenges to human perception, and  
 438 require additional mechanisms (see discussion in Barreda, 2021; Weatherholtz and Jaeger,  
 439 2016). This introduces noise—variability in listeners’ responses that cannot be accounted  
 440 for by normalization—to any comparison of normalization accounts, potentially reducing  
 441 the power to detect differences between accounts.

442 **III. COMPARISON OF NORMALIZATION ACCOUNTS**

443 In order to evaluate normalization accounts against speech perception, it is necessary to  
 444 map the phonetic properties of stimuli—under different hypotheses about normalization—  
 445 onto listeners’ responses in Experiments 1a and 1b. Previous work has done so by directly  
 446 predicting listeners’ responses from the raw or normalized phonetic properties of stimuli  
 447 (Apfelbaum and McMurray, 2015; Barreda, 2021; Crinnion *et al.*, 2020; McMurray and  
 448 Jongman, 2011; Nearey, 1989). For example, McMurray and Jongman used multinomial  
 449 logistic regression to predict eight-way fricative categorization responses in US English (see  
 450 also Barreda, 2021).

451 Here we pursued an alternative approach by committing to a core assumption common to  
 452 contemporary theories of speech perception: that listeners acquire implicit knowledge about  
 453 the probabilistic mapping from acoustic inputs to linguistic categories, and draw on this  
 454 knowledge during speech recognition (e.g., TRACE, McClelland and Elman, 1986; exem-  
 455 plar theory, Johnson, 1997; Bayesian accounts, Luce and Pisoni, 1998; Nearey, 1990; Norris  
 456 and McQueen, 2008; ASR-inspired models like DIANA or EARSHOT, ten Bosch *et al.*,  
 457 2015; Magnuson *et al.*, 2020). Using a general computational framework for adaptive speech  
 458 perception (ASP, Xie *et al.*, 2023) we trained Bayesian ideal observers to capture the expec-  
 459 tations that a ‘typical’ L1 adult listener might have about the formant-to-vowel mappings of  
 460 US English. We approximated these expectations using a database of L1-US English vowel  
 461 productions (Xie and Jaeger, 2020)—transformed to reflect the different normalization ac-  
 462 counts. We then ask which of the different ideal observer models—corresponding to different

<sup>463</sup> hypotheses about formant normalization—best predicts listeners' responses in Experiments

<sup>464</sup> 1a and 1b.

<sup>465</sup> Training ideal observers on a database of vowel productions has the advantage that it

<sup>466</sup> reduces the degrees of freedom (DFs) used to predict listeners' responses. For example, using

<sup>467</sup> ordinary multinomial logistic regression trained on our perceptual data to predict eight-way

<sup>468</sup> vowel categorization as a function of F1, F2 and their interaction would require up to 28

<sup>469</sup> DFs. This problem increases with the number of cues considered. By instead training

<sup>470</sup> ideal observers on phonetic data that are independent of listeners' responses, the ASP-based

<sup>471</sup> approach we employ uses only two DFs to mediate the mapping from stimuli properties to

<sup>472</sup> listeners' responses, regardless of the number of cues considered. Over the next few sections,

<sup>473</sup> we describe how this parsimony is made possible through a commitment to strong linking

<sup>474</sup> hypotheses motivated by theories of speech perception.

<sup>475</sup> **A. Methods**

<sup>476</sup> **1. A general-purpose categorization model for  $J$ -AFC categorization tasks**

<sup>477</sup> Figure 6 summarises ASP's categorization model for a  $J$ -alternative forced-choice task

<sup>478</sup> (for an in-depth description, we refer to [Xie et al., 2023](#)). The model combines Bayesian ideal

<sup>479</sup> observers (as used in e.g., [Clayards et al., 2008](#); [Feldman et al., 2009](#); [Norris and McQueen,](#)

<sup>480</sup> [2008](#); [Xie et al., 2021](#); for a closely related approach, see also [Nearey and Hogan, 1986](#)) with

<sup>481</sup> psychometric lapsing models ([Wichmann and Hill, 2001](#)). To reduce researchers' degrees of

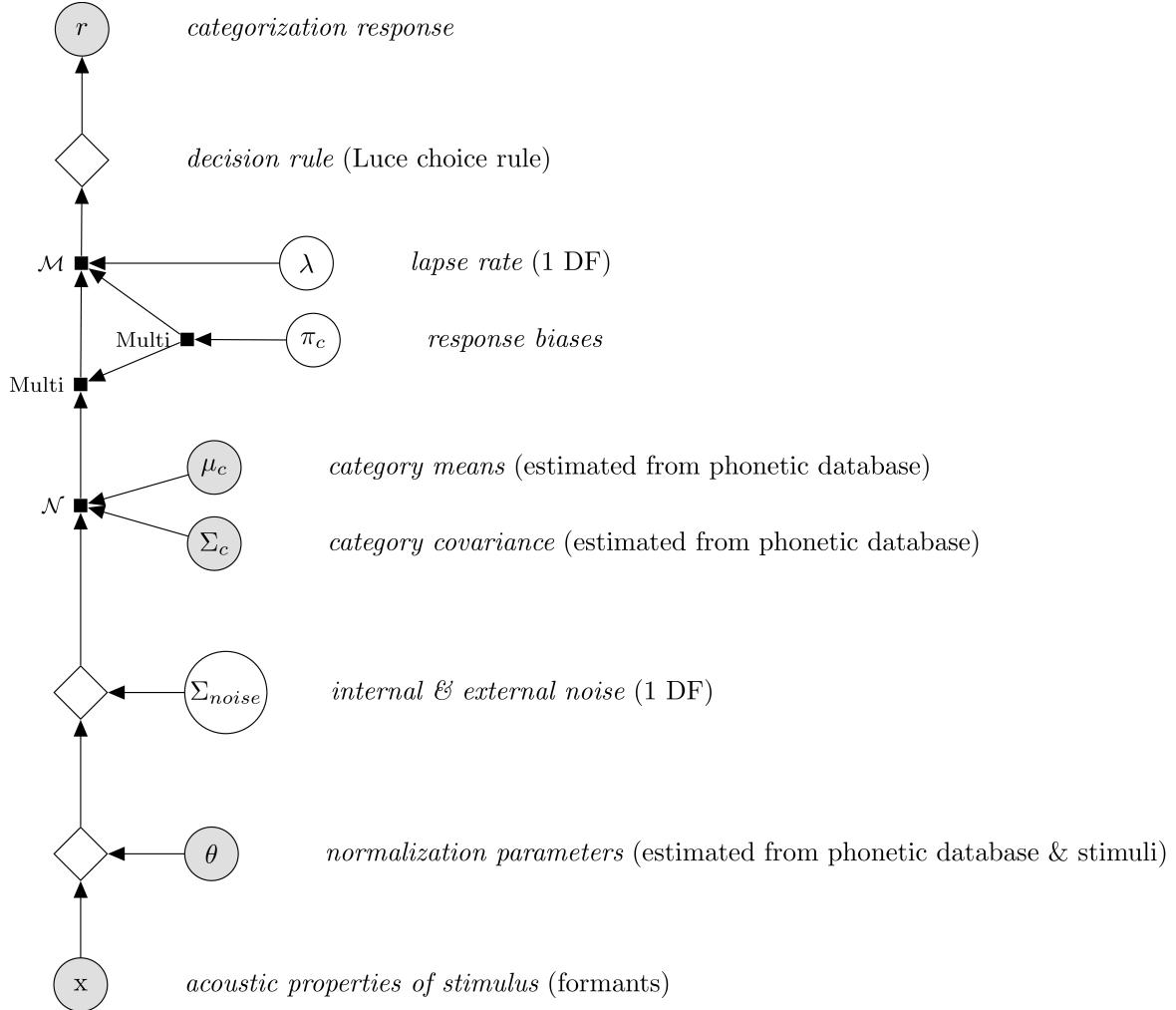


FIG. 6. Graphical model of ASP's general categorization framework (adapted for the current purpose from Xie et al., 2023, Figure 4). Here  $J = 8$  (the eight vowel response options in Experiments 1a and 1b). We use this framework to compare normalization accounts against listeners' categorization responses from Experiments 1a and 1b. Filled gray circles represent variables that are known to the researcher. Empty circles represent latent variables that are not observable. Diamonds represent variable-free processes, annotated with the distributions resulting at that level of the model:  $\mathcal{N}$ (ormal),  $\text{Multi}$ (nomial), and  $\mathcal{M}$ (ixture) distributions.

482 freedom, we adopt all assumptions made in Xie *et al.* (2023), and do not introduce additional  
 483 assumptions.

484 Starting at the bottom of the figure, the acoustic input  $x$  is normalized. Here, we follow  
 485 most previous evaluations of normalization accounts, and focus on the point estimates of  
 486 formants at the center of the vowel as the inputs to normalization. This leaves open the  
 487 question of how considerations of additional cues to vowel identity (e.g., VISC) or formant  
 488 dynamics might affect the findings we report below (a point to which we return in the general  
 489 discussion). Specifically, the main analysis we present here focus on  $x = F1$  and  $F2$ . As one  
 490 anonymous reviewer pointed out, this focus on  $F1-F2$  might underestimate the potential of  
 491 *intrinsic* normalization accounts, which might perform better when more acoustic-phonetic  
 492 features are considered. The SI, §3 E, thus reports additional analyses that instead employ  
 493  $F1-F3$ . These analyses indeed find that the fit of intrinsic normalization accounts improves  
 494 more than that of extrinsic accounts when  $F3$  is included in the analysis. However, the best-  
 495 fitting accounts were still the same extrinsic accounts we find to best fit listeners' responses  
 496 when only  $F1$  and  $F2$  is considered.

497 The specific computations applied to the input  $x$  depend on the normalization accounts  
 498 (see Table 1). We use  $\theta$  to refer to the parameters required by the normalization account.  
 499 For example, for Nearey's uniform scaling account (Nearey, 1978),  $\theta$  is the overall mean  
 500 of all log-transformed formants. For Lobanov normalization (Lobanov, 1971),  $\theta$  is a vector  
 501 of means and standard deviations for each formant (in Hz). The normalized input is then  
 502 perturbed by perceptual and environmental noise. Following Feldman *et al.* (2009), this  
 503 noise is assumed to be Gaussian distributed centered around the transformed stimulus with  
 504 noise variances that are independent and identical for all formants (i.e.,  $\Sigma_{noise}$  is a diagonal  
 505 matrix, and all diagonal entries have the same value).

506 Next, the likelihood of the normalized percept under each of the eight vowel categories  
 507 is calculated,  $p(F1, F2|vowel)$ . This requires specifying listeners' expectations about the  
 508 cue-to-category mapping (listeners' likelihood function). We followed [Xie et al. \(2023\)](#) and  
 509 previous work and assume that each vowel maps onto a multivariate Gaussian distribution  
 510 over the phonetic cues, here bivariate Gaussians over F1 and F2 (cf. [Clayards et al., 2008](#);  
 511 [Feldman et al., 2009](#); [Kleinschmidt and Jaeger, 2015](#); [Norris and McQueen, 2008](#); [Xie et al.,](#)  
 512 [2021](#)). We also followed previous models in assuming a single dialect template—i.e., a  
 513 single set of bivariate Gaussian vowel categories ([Nearey and Assmann, 2007](#)). The analyses  
 514 of participants' responses we provided above in the description of Experiments 1a and 1b  
 515 suggest that this assumption is wrong. However, more appropriate alternatives—such as  
 516 hierarchical or mixture models with multiple dialect templates—will require substantial  
 517 additional research as well as larger databases of vowel recordings that have high resolution  
 518 both within and across dialects. We return to this issue in the general discussion.

519 Once the likelihood function for each vowel is specified, the posterior probability of each  
 520 vowel is obtained by combining its likelihood with its prior probability or response bias  $\pi_c$ ,  
 521 according to Bayes theorem:<sup>7</sup>

$$p(vowel = c|F1, F2) = \frac{\mathcal{N}(F1, F2|\mu_c, \Sigma_c + \Sigma_{noise}) \times \pi_c}{\sum_{c_i} \mathcal{N}(F1, F2|\mu_{c_i}, \Sigma_{c_i} + \Sigma_{noise}) \times \pi_{c_i}} \quad (1)$$

522 Up to this point, the model is identical to a standard Bayesian ideal observer over noisy  
 523 input ([Feldman et al., 2009](#); [Kronrod et al., 2016](#)) for which the input has been transformed  
 524 based on the normalization account. ASP's categorization model adds to this the potential  
 525 that participants experience attentional lapses—or for other reasons do not respond based

526 on the input—on some proportion of all trials ( $\lambda$ , as in standard psychometric lapsing  
 527 models, [Wichmann and Hill, 2001](#)). On those trials, the posterior probability of a category  
 528 is determined solely by participants’ response bias, which we assume to be identical to the  
 529 response bias on non-lapsing trials (following [Xie et al., 2023](#)). This results in a posterior  
 530 that is described by weighted mixture of two components, describing participants’ posterior  
 531 on non-lapsing and lapsing trials, respectively:

$$p(vowel = v | F1, F2) = (1 - \lambda) \frac{\mathcal{N}(F1, F2 | \mu_c, \Sigma_c + \Sigma_{noise}) \times \pi_c}{\sum_{c_i} \mathcal{N}(F1, F2 | \mu_{c_i}, \Sigma_{c_i} + \Sigma_{noise}) \times \pi_{c_i}} + \lambda \frac{\pi_c}{\pi_{c_i}} \quad (2)$$

532 Finally, a decision rule is applied to the posterior to determine the response of the model,  
 533 conditional on the input (one of the eight vowels in Experiments 1a and 1b). We followed  
 534 the gross of research on speech perception and assume Luce’s choice rule ([Luce, 1959](#); for  
 535 discussion, see [Massaro and Friedman, 1990](#)). Under this choice rule, the model can be  
 536 seen as sampling from the posterior, responding with each category proportional to that  
 537 category’s posterior probability.

538 Next, we describe how we estimated the  $\theta$ s,  $\mu_c$ s and  $\Sigma_c$ s for each normalization account  
 539 from a phonetic database. We use this database as a—very coarse-grained—approximation  
 540 of a the speech input a ‘typical’ listener might have experienced previously. By fixing  $\theta$ ,  $\mu_c$   
 541 and  $\Sigma_c$  based on the distribution of phonetic cues in the database, we substantially reduce  
 542 the DFs that are allowed to mediate the mapping from stimulus properties to listeners’  
 543 responses (following [Xie et al., 2023](#)). In addition, this approach naturally penalizes overly  
 544 complex models by validating these against out-of-sample data. Finally, we describe how

545 we fit the remaining parameters as DFs to participants' responses from Experiments 1a and  
 546 1b.

547 **2. Modeling listeners' prior experience (and guarding against overfitting):  $\theta$ ,  $\mu_c$ ,**  
 548 **and  $\Sigma_c$**

549 By fixing  $\theta$ ,  $\mu_c$ , and  $\Sigma_c$  based on a database of vowel *productions*, we impose strong  
 550 constraints on the functional flexibility of the model in predicting listeners' responses. This  
 551 benefit is made possible by committing to a strong linking hypothesis—that listeners' cate-  
 552 gories are learned from, and reflect, the distributional mapping from formants to vowels in  
 553 previously experienced speech input (e.g., Abramson and Lisker, 1973; Massaro and Fried-  
 554 man, 1990; Nearey and Hogan, 1986). The database we use to approximate listeners' prior  
 555 experience was originally developed to compare the production of L1 and L2 speakers (Xie  
 556 and Jaeger, 2020). It contains 9-10 recordings of the eight *hVd* words from each of 17 (five  
 557 female) L1 talkers of a Northeastern dialect of US English (ages 18 to 35 years old). Since  
 558 Experiments 1a and 1b used recordings of one of these talkers, we excluded that talker prior  
 559 to fitting training ideal observers on the data. In total, this yields 5842 recordings that are  
 560 annotated for F0, F1-F3, and vowel duration. The SI (§3 A 1) summarizes the distribution  
 561 of these cues, and how the different normalization accounts affect those distributions.

562 To avoid over-fitting the ASP model to the database, we used 5-fold cross-validation:  
 563 we randomly split the Xie and Jaeger (2020) database into five approximately evenly-sized  
 564 folds (following Persson and Jaeger, 2023). This split was performed within each vowel to  
 565 guarantee that all five folds had the same relative amount of data for each vowel category.

566 These splits were combined into five training sets, each containing one of the folds (20% of  
 567 the data). This way, each training set was different from the others, increasing the variability  
 568 between sets.<sup>8</sup>

569 For each training set and for each normalization account, we then estimated the required  
 570 normalization parameters  $\theta$  for all talkers, and normalized all formants based on those  
 571 talker-specific parameters. This yielded 5 (training sets) \* 20 (accounts) = 100 normalized  
 572 training sets. For each of these normalized training sets, we fit the category means,  $\mu_c$ , and  
 573 covariance matrices,  $\Sigma_c$ , of all eight vowels, using the R package **MVBeliefUpdatr** (Jaeger,  
 574 2024).<sup>9</sup>

575 This yielded 100 ideal observer models, five for each of the 20 normalization accounts  
 576 in Table 1. Of note, the 20 ideal observers fit on each fold differ *only* in the assumptions  
 577 they make about the normalization that is applied to cues before they are mapped onto  
 578 the eight vowel categories. Figure 7 visualizes the resulting bivariate Gaussian categories  
 579 for four of the 20 normalization accounts. This illustrates one advantage of the cross-  
 580 validation approach: it takes a modest step towards simulating differences across listeners'  
 581 prior experience (represented by the five different folds).

582 **3. Transforming the stimuli from Experiments 1a and 1b into the normalized  
 583 phonetic spaces**

584 Next, we transformed the stimuli of Experiments 1a and 1b into the formant space de-  
 585 fined by the 20 normalization accounts in Table 1. This requires estimating the required  
 586 normalization parameters  $\theta$  for each experiment and normalization account. We calculated

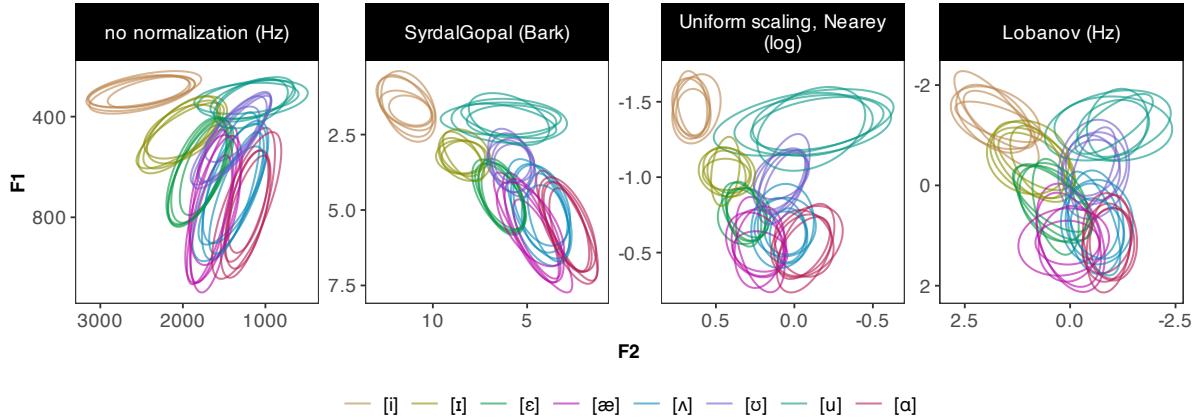


FIG. 7. Visualizing the bivariate Gaussian categories (prior to adding  $\Sigma_{noise}$ ) of four example normalization accounts in F1-F2 space. Separate ellipses are shown for each of the five training sets (each set corresponds to one set of eight ellipses). The relative stability of the category ellipses across training sets indicates that the database is sufficiently large for the present purpose.

587 these  $\theta$ s over all stimuli (of each experiment and normalization account). For example,  
 588 for Nearey's uniform scaling account (Nearey, 1978), we calculated the overall mean of all  
 589 log-transformed formants over all stimuli. For Lobanov normalization (Lobanov, 1971), we  
 590 calculated the mean and standard deviation of each formant (in Hz) over all stimuli. For  
 591 each combination of experiment and normalization account, we then normalized the stimuli  
 592 using those parameter estimates. The SI (§3 A 2) summarizes the  $\theta$  parameters of all nor-  
 593 malization accounts for each experiment and how they relate to the values obtained from the  
 594 training sets. For reasons outlined in that same section, we did not expect a clear relation  
 595 between an account's ability to predict listeners' responses for an experiment, and the degree  
 596 to which the account's normalization parameters differed between the experiment and the  
 597 training database (and, indeed, no such relation was found).

598 Combining the 100 normalized training sets described in the previous section with the  
 599 matching normalized stimuli from each of the two experiments yielded 200 data sets.

600 **4. Noise ( $\Sigma_{noise}$ ) and attentional lapses ( $\lambda$ )**

601 Finally, we describe the two parameters of the ASP model that we fit against listeners'  
 602 responses in Experiments 1a and 1b. These two parameters constitute the only DFs that  
 603 mediate the link from ideal observers' predictions to listeners' responses, and which are fit  
 604 to listeners' responses. The first DF ( $\Sigma_{noise}$ ) models the effects of internal (perceptual)  
 605 and external (environmental) noise on listeners' perception. While previous work provides  
 606 estimates of the internal noise in formant perception, these estimates were obtained under  
 607 *assumptions* about the relevant formant space. For example, [Feldman \*et al.\* \(2009\)](#) estimated  
 608 the internal noise variance to be about 15% of the average category variance along F1 and F2.  
 609 This estimate was based on the assumption that human speech perception transforms vowel  
 610 formants into Mel, without further normalization. Since we aim to *test* which normalization  
 611 account best explains speech perception, we cannot rely on this or other internal noise  
 612 estimates obtained under a single specific assumption. Additionally, internal noise can vary  
 613 across individuals and external noise can vary across environments (a point particularly  
 614 noteworthy, given that we conducted Experiments 1a and 1b over the web). We thus allowed  
 615 the noise variance  $\Sigma_{noise}$  to vary in fitting participants' responses. Following [Feldman \*et al.\*](#)  
 616 ([2009](#)), we assumed that perceptual noise had identical effects on all formants in the phonetic  
 617 space defined by the normalization account (see also [Kronrod \*et al.\*, 2016](#)). This reduces  
 618  $\Sigma_{noise}$  to a single DF, regardless of the normalization account (for details, see SI [§3 A 3](#)).

619 The magnitude of  $\Sigma_{noise}$  affects the slope of the categorization functions that predict  
 620 listeners' responses from stimulus properties (here, F1 and F2): higher  $\Sigma_{noise}$  imply more

shallow categorization slopes. To facilitate comparison of  $\Sigma_{noise}$  values across normalization accounts, we report results in terms of the best-fitting *noise ratios* ( $\tau^{-1}$ ), rather than  $\Sigma_{noise}$ s. Specifically,  $\Sigma_{noise}$  is best understood *relative* to the inherent variability of the vowel categories ( $\Sigma_c$ ). This variability in turn depends on the phonetic space defined by the normalization account. We thus divide  $\Sigma_{noise}$  by the mean of the diagonals of all  $\Sigma_c$ s to obtain the *noise ratio*  $\tau^{-1}$ . For example, noise ratio of 0 corresponds to the absence of any noise, and a noise ratio of 1 corresponds to noise variance of the same magnitude as the average category variance along F1 and F2 in the phonetic space defined by the normalization account.<sup>10</sup> Figure 8B illustrates the effects of this noise ratio for Nearey's uniform scaling account.

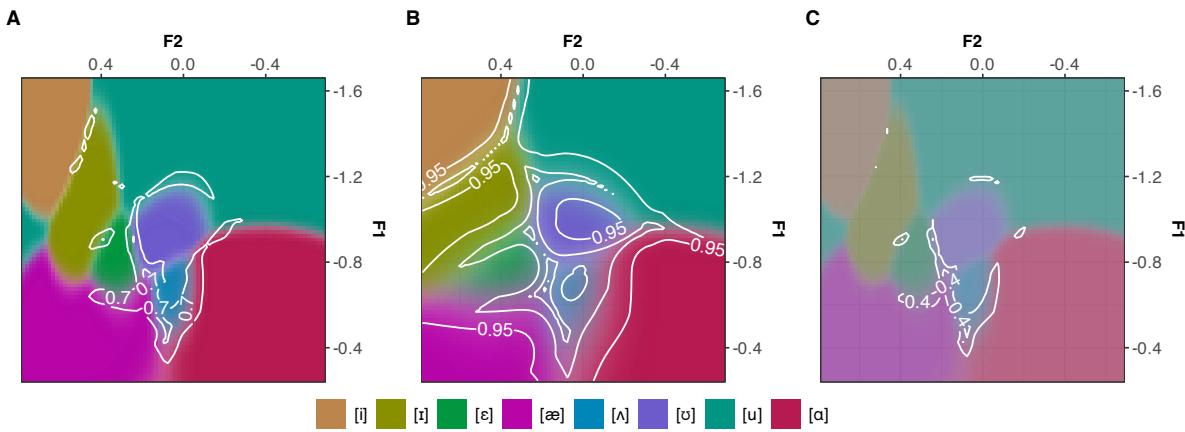


FIG. 8. Illustrating the consequences of perceptual and external noise ( $\Sigma_{noise}$ ) and attentional lapse rates ( $\lambda$ ) on the predicted posterior distribution of vowel categorizations. Shown are the average predicted posteriors across all five folds for Nearey's uniform scaling account. **Panel A:** Predicted posterior distribution for noise ratio  $\tau^{-1} = \lambda = 0$ . **Panel B:** Same for  $\tau^{-1} = 1$  and  $\lambda = 0$ . **Panel C:** Same for  $\tau^{-1} = 0$  and  $\lambda = 0.5$ . Transparency of a color is determined by that vowel's posterior probability. Contours indicate the highest posterior probability of any vowel (at .4, .5, .7, .95 probability level).

631 Second, participants can attentionally lapse or for other reasons reply without considering  
 632 the speech input. We thus allowed lapse rates ( $\lambda$ ) to vary while fitting human responses. This  
 633 introduces a second DF, which we fit against listeners' responses. Together, the inclusion  
 634 of freely varying lapse rates and a uniform response bias allows the ASP models to capture  
 635 that some unknown proportion of listeners' responses might be more or less random, rather  
 636 than reflecting properties of the vowel stimuli. This is illustrated in Figure 8C.

637 Finally, participants can have response biases that reflect their beliefs about the prior  
 638 probability of each category. However, to reduce the DFs fit to participants' responses, we  
 639 did *not* fit this response bias against listeners' responses (thus avoiding  $J - 1 = 7$  additional  
 640 DFs). Instead, we assumed uniform response biases—i.e., that listeners believed all eight  
 641 response options in the experiments to be equally likely ( $\forall c \pi_c = .125$ ). This decision implies  
 642 that our models would not be able to capture any potential non-uniformity in listeners'  
 643 response biases—including potential effects of additional acoustic differences (the absence  
 644 of [h] in *odd* or the coda [t], rather than [d] in *hut*) and orthographically particular response  
 645 options in Experiment 1a (“who’d”, “odd”, and “hut”). We do, however, see no reasons to  
 646 expect this decision to bias the comparison of normalization accounts.

647 **5. Fitting normalization accounts to listeners' responses**

648 For each of the 200 combinations of experiment, normalization account, training set,  
 649 we used constrained quasi-Newton optimization (Byrd *et al.*, 1995, as implemented in R's  
 650 `optim()` function) to find the  $\lambda$  and  $\tau^{-1}$  values that best described listener's responses.  
 651 Specifically, we used the 100 ideal observers described in the previous sections, applied them

652 to the normalized stimuli of the experiment, and determined which  $\lambda$  and  $\tau^{-1}$  maximized  
 653 the likelihood of listener's responses (for details, see SI [§3 A 3](#)). This procedure yielded five  
 654 maximum likelihood estimates for both  $\lambda$  and  $\tau^{-1}$  for each combination of experiment and  
 655 normalization account—one for each training set. All results presented below were validated  
 656 and confirmed by grid searches over the parameter spaces (SI, [§3 F](#)).

657 We compare normalization accounts in terms of the likelihood of listeners' responses under  
 658 these maximum likelihood estimates of  $\lambda$  and  $\tau^{-1}$ . Comparing accounts in terms of their  
 659 data likelihood follows more recent work (e.g., [Barreda, 2021](#); [McMurray and Jongman, 2011](#);  
 660 [Richter \*et al.\*, 2017](#); [Xie \*et al.\*, 2023](#)). Previous work has instead compared normalization  
 661 accounts in terms of their accuracy (e.g., [Johnson, 2020](#); [Nearey and Assmann, 2007](#); [Persson](#)  
 662 [and Jaeger, 2023](#)), or correlations with human response proportions (e.g., [Hillenbrand and](#)  
 663 [Nearey, 1999](#); [Nearey and Assmann, 1986](#)). Both of these approaches are problematic.  
 664 Correlations between the predictions of a model and human responses can be high even  
 665 when the model's predictions are systematically 'off'. Imagine three items for which listeners  
 666 respond [i] 10%, 30%, and 50% of the time. If a model predicts 30%, 50%, and 70%  
 667 [i] responses, respectively, for the same items, its predictions will perfectly correlate with  
 668 listeners' response proportions, and yet be systematically wrong. Similarly, a model can  
 669 achieve the highest possible accuracy in predicting listeners' responses simply because it  
 670 always predicts the most frequent response (see discussion of criterion choice rule in [Massaro](#)  
 671 [and Friedman, 1990](#)). In contrast, the likelihood of listeners' responses under a model is  
 672 a direct measure of how well the model captures the distribution of listeners' responses  
 673 conditional on the stimulus properties. In particular, data likelihood will be maximized if,

674 and only if, the model-predicted posterior probabilities of each vowel for each stimulus are  
 675 identical to the proportion with which those vowels occur in listeners' responses.

676 **B. Results**

677 We begin by comparing the fit of different accounts against listeners' responses in Ex-  
 678 periments 1a and 1b. Given the comparatively large number of accounts compared here, we  
 679 provide initial conclusions based on the best-fitting accounts along with the description of  
 680 the results (more in-depth discussion is provided in the general discussion). Following this  
 681 comparison, we visualize how different normalization accounts predict the formant space to  
 682 be divided into the eight vowel categories.

683 **1. Comparing normalization accounts in terms of fit against human behavior**

684 Figure 9 compares how well the different normalization accounts fit listeners' responses  
 685 in Experiments 1a and 1b. All accounts performed well above chance guessing (chance  
 686 per-token log-likelihood in both experiments:  $\ln(\frac{1}{8})=-2.08$ ) but also well below the highest  
 687 possible performance (in Experiment 1a, per-token log-likelihood = -0.46, in Experiment 1b:  
 688 -1.15).

689 Normalization significantly improved the fit to listeners' responses relative to no normal-  
 690 ization. This was confirmed by paired one-sided *t*-tests comparing the maximum likelihood  
 691 values for each normalization account against those in the absence of normalization (all *ps*  
 692  $< .05$  except for Gerstman normalization, log-transformation and semitones-transformation  
 693 and Experiment 1a; see SI §3B1). Not all normalization accounts achieved equally good fits,

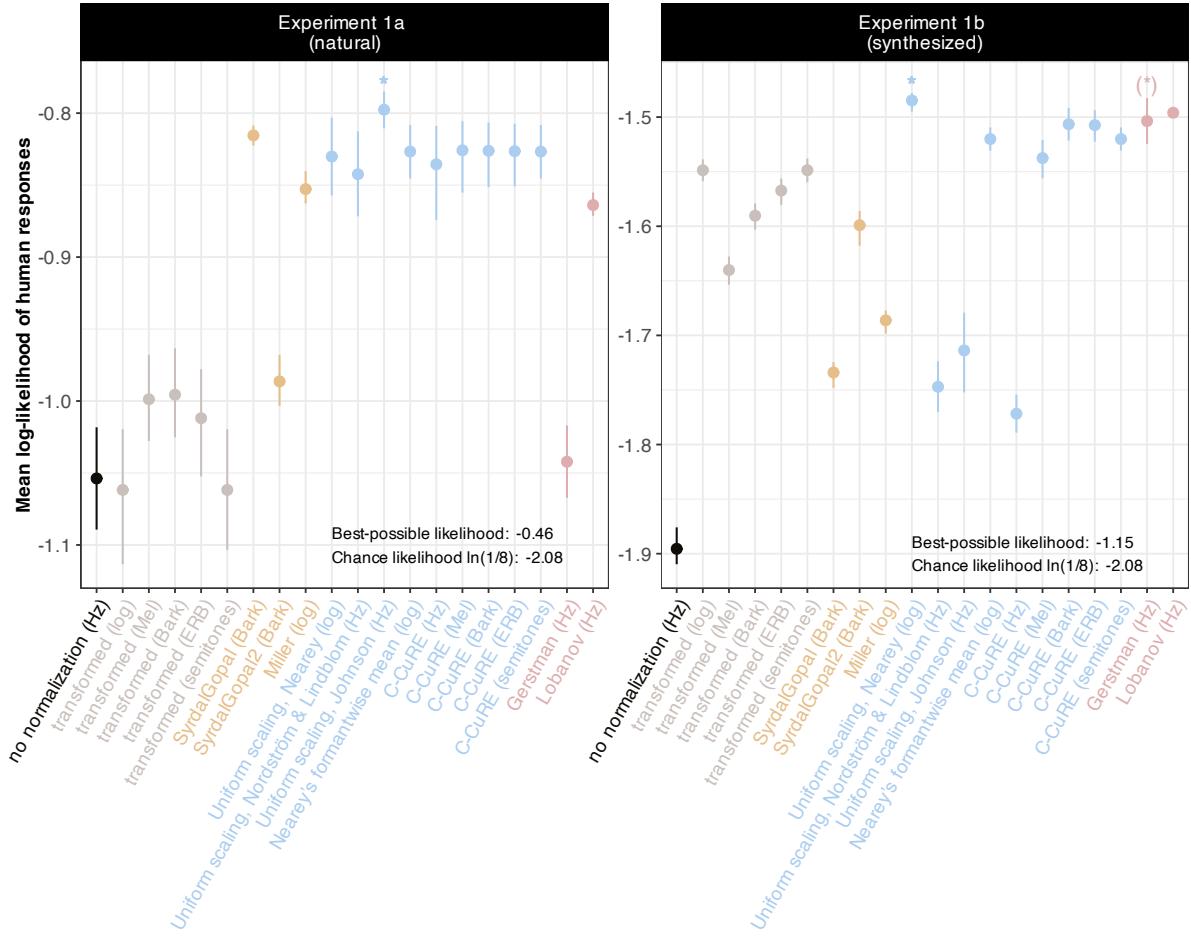


FIG. 9. Comparison of normalization accounts against listeners' responses. Point ranges indicate mean and 95% bootstrapped CIs of the per-token log-likelihoods summarized over the five training sets (higher is better), normalized by the number of listener responses in each experiment. Accounts that fit listeners' responses to an extent that is statistically indistinguishable from the best-fitting account are marked by (\*). Note that per-token likelihoods cannot be directly compared across experiments because the best-possible likelihoods differ across experiments (due to differences in stimulus placement and other factors).

694 however: only some extrinsic accounts fit listeners' behavior well across both experiments.

695 This supports two conclusions. First, it suggests that the normalization mechanisms oper-

696 ating during human speech perception involve computations that go beyond estimation-free

697 transformations into psycho-acoustic spaces. Second, it suggests that the input to these

computations is not limited to intrinsic information—i.e., that the computations draw on information beyond what is available in the acoustic signal *at that moment*. In particular, extrinsic normalization requires the estimation and memory maintenance of talker-specific properties from the speech signal.

While the accounts that achieved the best fit against listeners' responses differed between experiments, both were variants of uniform scaling. For Experiment 1a, Johnson normalization account provided the best fit (per-token log-likelihood = -0.8, SD = 0.02 across the five crossvalidation folds), while Nearey's uniform scaling account provided the best fit to Experiment 1b (per-token log-likelihood = -1.48, SD = 0.01). Both accounts essentially slide the representational 'template' of a dialect—here the eight bivariate Gaussian categories of an ideal observer—along a single line in the formant space. They differ only in *which* space this linear relation between formants is assumed. The same two accounts still fit listeners' responses best when F3 was included in the analysis in addition to F1 and F2 (SI, §3 E).<sup>11</sup>

This suggests that formant normalization might involve comparatively parsimonious maintenance of talker-specific properties: in its simplest form, uniform scaling employs a single formant statistic to normalize all formants. In contrast, computationally more complex accounts like Lobanov normalization might require the estimation and maintenance of two formant statistics (mean and standard deviation) for each formant that is normalized (e.g., a total of four formant statistics for F1 and F2, or six statistics for F1-F3).

Also of note is that accounts that were particularly stable across experiments operate in log space, whereas accounts that operate in Hz space seemed to display a more volatile performance (e.g., both standardizing accounts but also C-CuRE Hz, Nordström & Lindblom

720 and Johnson normalization). That accounts operating over log-transformed formants fit  
 721 human behavior better should not be surprising. While questions remain about the exact  
 722 organization of auditory formant representations, it is uncontroversial that the perceptual  
 723 sensitivity to acoustic frequency information is better approximated by a logarithmic scale  
 724 than by a linear scale (see [Moore, 2012](#)). As a result, a 30 Hz difference in an F1 of 300  
 725 Hz (a 10% change) is expected to be perceptually more salient than a 30 Hz change in an  
 726 F2 of 2500 Hz (a 1.2% change).<sup>12</sup> In summary, variability in how well different accounts  
 727 predict human behavior across the two experiments highlights the importance of psycho-  
 728 acoustic transformations for human speech perception. This also highlights the importance  
 729 of comparing normalization accounts against multiple types of data.

730 **2. *Visualizing the consequences of different normalization mechanisms***

731 Before we turn to the general discussion, we briefly visualize how different normalization  
 732 mechanisms affect vowel categorization. This sheds light on *why* the accounts differ in how  
 733 well they fit listeners' responses. Figure 10 visualizes the categorization functions predicted  
 734 by four different normalization accounts, using the best-fitting  $\lambda$  and  $\tau^{-1}$  values for each  
 735 account (i.e., the values that lead to the fit shown in Figure 9). Figure 10 highlights three  
 736 points. First, a comparison across panels A-C shows different normalization accounts can  
 737 result in very different predictions about how the acoustic space is carved into categories.

738 Second, the best-fitting parameters (shown at the top of each panel) were relatively com-  
 739 parable across accounts but differed more substantially across experiments. Specifically, the  
 740 best-fitting estimates of lapse rates  $\lambda$  were generally comparable across the two experiments

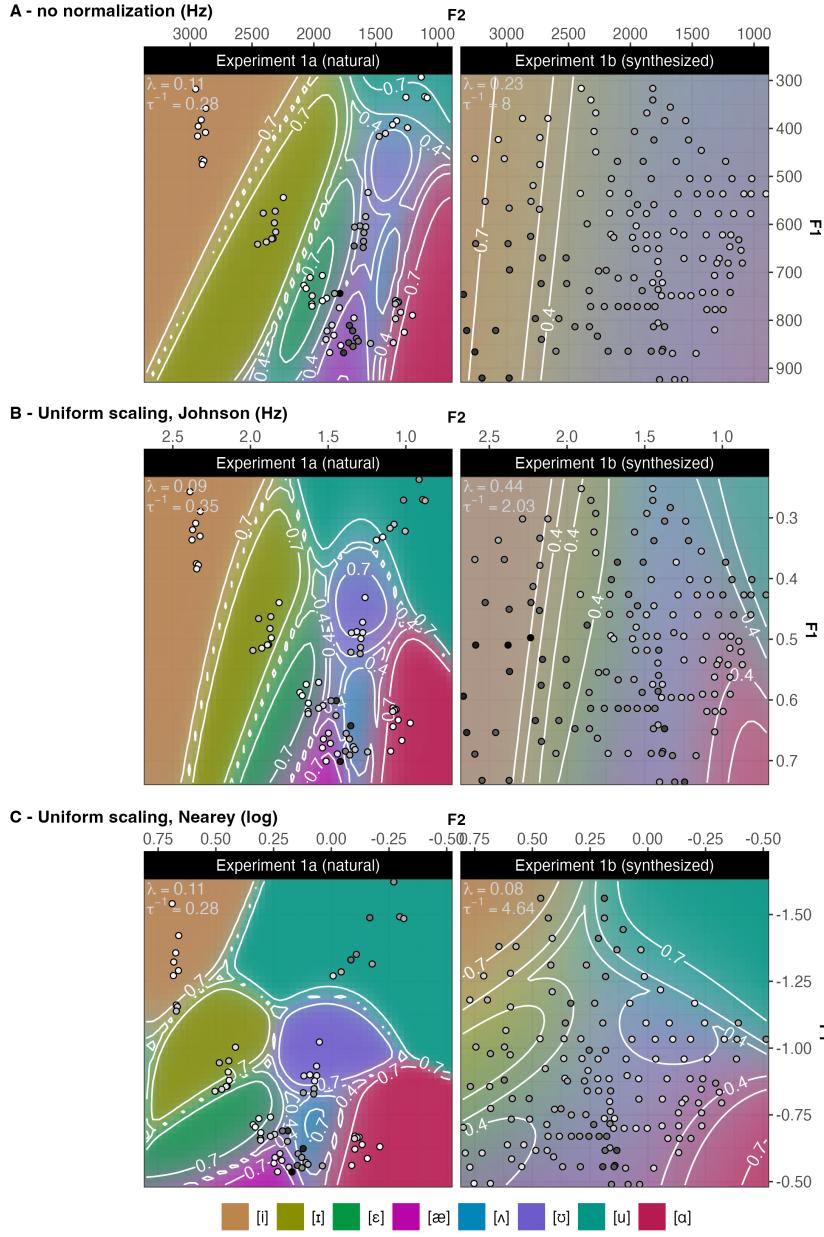


FIG. 10. Predicted categorization functions over the F1-F2 space under three different normalization accounts. For each account, we show the predicted posterior probabilities of all eight vowels obtained by averaging over the maximum likelihood parameterizations (of  $\lambda$  and  $\tau^{-1}$ ) for the five training sets (shown at top of each panel). **Panel A:** absence of normalization shown for reference. **Panel B:** the best-fitting account for Experiment 1a. **Panel C:** the best-fitting account for Experiment 1b. Contours indicate the highest posterior probability of any vowel. Points indicate location of test stimuli. The increasing brightness of points indicates a better match between the account's prediction and listeners' responses (higher log-likelihood; see text for detail).

741 (with the exception of Nordström & Lindblom and Johnson normalization, which exhibited  
 742 substantially higher lapse rates in Experiment 1b; SI [§3 B 2](#)). This suggests that partici-  
 743 pants in both experiments were about equally likely to pay attention to the stimulus. The  
 744 best-fitting noise ratios  $\tau^{-1}$ , however, differed substantially across experiments, and were  
 745 10 times larger for Experiment 1b (mean  $\tau^{-1} = 4.32$ , SD = 2.52 across normalization ac-  
 746 counts) than for Experiment 1a (mean  $\tau^{-1} = 0.42$ , SD = 0.46). This difference most likely  
 747 reflects the fact that the synthesized stimuli in Experiment 1b left listeners with substan-  
 748 tially more uncertainty about the intended category, as discussed during the description of  
 749 the experiments.

750 Since noise is assumed to be independent of category variability (see also [Feldman \*et al.\*, 2009](#);  
 751 [Kronrod \*et al.\*, 2016](#)), differences in noise ratios can substantially change the cate-  
 752 gorization function. This is particularly evident for the accounts that had more variable  
 753 performance across the two experiments. For example, Johnson normalization (Panel B)  
 754 resulted in very different best-fitting categorization functions for Experiments 1a and 1b.

755 Third and finally, Figure 10 also shows how well accounts fit listeners' responses for each  
 756 test stimulus (opaqueness of the points). This begins to explain *why* some accounts fit  
 757 listeners' responses in Experiment 1b less well. For example, the Johnson normalization  
 758 account (Panel B) predicts the responses to the test stimuli in Experiment 1a well, but fails  
 759 to predict the responses to the test stimuli in Experiment 1b. This drop in performance  
 760 seems to be primarily driven by stimuli that are unlikely to be articulated by the same talker  
 761 (lower left, cf. dashed line in Figure 4). This might suggest that this account was over-  
 762 engineered to explain naturally occurring productions—the type of data, it was originally

763 tested on (Johnson, 2020). A plausible account of normalization, however, should be able  
 764 to explain human perception to any type of stimulus, including synthesized stimuli. The SI  
 765 (§3B3) presents more detailed by-item comparisons of normalization accounts that might  
 766 be of interest to some readers.

767 **IV. GENERAL DISCUSSION**

768 Research on vowel normalization has an influential history. Cognitive scientists have  
 769 long aimed to understand the organization of frequency information in the human brain  
 770 (Siegel, 1965; Stevens and Volkmann, 1940), and how it helps listeners overcome cross-talker  
 771 variability in the formant-to-vowel mapping (e.g., Fant, 1975; Joos, 1948; Nordström and  
 772 Lindblom, 1975). Auditory processes that normalize speech inputs for differences in vocal  
 773 tract physiology are now recognized to be an integral part of speech perception (Johnson  
 774 and Sjerps, 2021; McMurray and Jongman, 2011; Xie *et al.*, 2023). Here, we set out to  
 775 investigate what types of computations are implicated in the normalization of the frequency  
 776 information that plays a critical role in the recognition of vowels.

777 Our results support three theoretical insights. First, human speech perception draws on  
 778 more than psycho-acoustic transformations or intrinsic information, in line with previous  
 779 research on normalization (Adank *et al.*, 2004; Ladefoged and Broadbent, 1957; Nearey,  
 780 1989). Rather, formant normalization seems to involve the estimation and storing of talker-  
 781 specific formant properties. Second, computationally simple uniform scaling accounts pro-  
 782 vide the best fit to listeners' responses, suggesting comparatively parsimonious maintenance  
 783 of talker-specific properties. This replicates and extends previous findings that uniform scal-

784 ing or similarly simple corrections for vocal tract size provide a better explanation for human  
785 perception than more complex extrinsic accounts (Barreda, 2021; Richter *et al.*, 2017). It is  
786 impossible to rule out more complex approaches to perceptual normalization given the large  
787 number of possible alternatives. However, given that uniform scaling provides a parsimo-  
788 nious explanation for human formant normalization, and the current absence of empirical  
789 evidence for more complex computations, we submit that researchers ought to adapt uni-  
790 form scaling as the working hypothesis. Third, the psycho-acoustic representation assumed  
791 by different normalization accounts matter, as indicated by the comparison of otherwise  
792 computationally similar accounts (e.g. Nearey's vs. Johnson's uniform scaling).

793 The results contribute to a still comparatively small body of work that has evaluated  
794 competing normalization accounts against listeners' perception, whereas most previous work  
795 evaluates accounts against intended productions. Complementing previous work, we took  
796 a broad-coverage approach: the present study compared 20 of the most influential normal-  
797 ization accounts against listeners' perception of *hVd* words with eight US English monoph-  
798 thongs in both natural and synthesized speech. This contrasts with previous work, which  
799 has typically focused on subsets of the vowel system, either using natural *or* synthesized  
800 speech, and considering a much smaller subset of accounts (typically 2-3 at a time). By  
801 considering a wider range of accounts, a wider range of formant values and vowel categories,  
802 and multiple types of speech, we aimed to contribute to a more comprehensive evaluation  
803 of competing accounts.

804 Next, we discuss the theoretical consequences of these findings for research beyond for-  
 805 mant normalization. Following that, we discuss limitations of the present work, and how  
 806 future research might overcome them.

807 **A. Consequences for theories of speech perception and beyond**

808 Understanding the perceptual space in which the human brain represents vowel categories—  
 809 i.e., the normalized formant space—has obvious consequences for research on speech percep-  
 810 tion. To illustrate how far reaching these consequences can be, we discuss a few examples.  
 811 For instance, research on *categorical perception* has found that vowels seem to be per-  
 812 ceived less categorically than some types of consonants. Recent work has offered an elegant  
 813 explanation for this finding: the perception of formants—relevant to the recognition of  
 814 vowels—might be more noisy than the perception of the acoustic cues that are critical to  
 815 the recognition of more categorically perceived consonants (Kronrod *et al.*, 2016). This is  
 816 a parsimonious explanation, potentially preempting the need for separate explanations for  
 817 the perception of different types of phonemic contrasts. Kronrod and colleagues based their  
 818 argument on estimates they obtained for the relative ratio of meaningful category variability  
 819 to perceptual noise ( $\tau$ , the inverse of our noise ratios,  $\tau^{-1}$ ). Critically, this ratio depends  
 820 both on (i) the perceptual space in which formants are assumed to be represented (Kronrod  
 821 *et al.* used Mel-transformed formant frequencies), and on (ii) whether the meaningful cate-  
 822 gory variability is calculated prior to, or following, normalization (Kronrod *et al.* assumed  
 823 the former, which increases estimates of category variability). Our point here is not to cast  
 824 doubt on the results of Kronrod *et al.* (2016) —the fact that the best-fitting noise ratios in

825 our study were relatively similar across accounts (while varying across experiments) suggests  
826 that the result of Kronrod and colleagues are likely to hold even under different assumptions  
827 about (i) and (ii)—but rather to highlight how research on the perception and recognition of  
828 vowels depends on assumptions about formant normalization. For example, similar points  
829 could be raised about experiments on statistical learning that manipulate formant or other  
830 frequency statistics (e.g., Chládková *et al.*, 2017; Colby *et al.*, 2018; Wade *et al.*, 2007; Xie  
831 *et al.*, 2021). Such experiments, too, need to make assumptions about the space in which  
832 formants are represented. If these assumptions are incorrect, this can affect whether the  
833 experimental manipulations have the intended effects, increasing the chance of null effects  
834 or misinterpretation of observed effects.

835 Understanding the perceptual space in which the human brain represents vowel cate-  
836 gories also has consequences for research beyond speech perception, perhaps more so than is  
837 sometimes recognized. For instance, in sociolinguistics and related fields, Lobanov remains  
838 the norm for representing vowels due to its efficiency in removing cross-talker variability (for  
839 review, see Adank *et al.*, 2004; Barreda, 2021). However, as shown in the present study, re-  
840 moving cross-talker variability is not the same as representing vowels in the perceptual space  
841 that listeners actually employ. Here, we do *not* find Lobanov to describe human perception  
842 particularly well. On the contrary, we find no support for the hypothesis that human speech  
843 perception employs these more complex computations that have been found to perform best  
844 at reducing category variability. This should worry sociolinguists. In order to understand  
845 how listeners infer a talker’s background or social identity, it is important to understand  
846 the perceptual space in which inferences are actually rooted. Critically, the representations

847 resulting from formant normalization presumably form an important part of the information  
848 that listeners use to draw social and linguistic inferences. It should thus be obvious that  
849 the use of normalization accounts that do not actually correspond to human perception can  
850 both mask real markers of social identity, and ‘hallucinate’ markers that are not actually  
851 present. For example, in order to determine how a talker’s social identity influences their  
852 vowel realizations, it is important to discount *all and only* effects that listeners will attribute  
853 to physiology, rather than social identity (Disner, 1980; Hindle, 1978).

854 Similar concerns apply to dialectology, research on language change, second language  
855 acquisition research, etc. For example, the perceptual space in which vowels are represented  
856 is critical to well-formed tests of hypotheses about the factors shaping the organization of  
857 vowel inventories across languages of the world (Lindblom, 1986; Stevens, 1972, 1989). It is  
858 essential in testing hypotheses about the extent to which the cross-linguistic realization of  
859 those systems is affected by perceptual processes (Flemming, 2010; Steriade, 2008), or by  
860 preferences for communicatively efficient linguistic systems (e.g., Hall *et al.*, 2018; Lindblom,  
861 1990; Moulin-Frier *et al.*, 2015). Similarly, tests of the hypothesis that vowel *articulation*  
862 during natural interactions is shaped by communicative efficiency do in obvious ways depend  
863 on assumptions about the perceptual space in which talkers—by hypothesis—aim to reduce  
864 perceptual confusion (cf. Buz and Jaeger, 2016; Gahl *et al.*, 2012; Scarborough, 2010; Wedel  
865 *et al.*, 2018). The same applies to any other line of research that aims to understand the  
866 perceptual consequences of formant variation across talkers, including research on infant- or  
867 child-directed speech (Eaves Jr *et al.*, 2016; Kuhl *et al.*, 1997), and research on whether non-  
868 native talkers are inherently more variable than native talkers (Smith *et al.*, 2019; Vaughn

869 *et al.*, 2019; Xie and Jaeger, 2020). In short, the perceptual space in which vowels are  
 870 represented is a critical component of understanding the structure of vowel systems, the  
 871 factors that shape them, and the ways in which they are used in natural language.

872 **B. Limitations and future directions**

873 As mentioned in the introduction, we take it as relatively uncontroversial *that* normalization  
 874 is part of human speech perception. Independent of any benefits that such normalization  
 875 conveys for speech perception, its existence is supported by evidence from cross-species  
 876 comparisons and neuro-physiological studies (for review, see Barreda, 2020). There are, how-  
 877 ever, important questions as to how decisions we made in comparing normalization accounts  
 878 against each other might have affected their fit against listeners' responses.

879 For instance, we followed previous work in focusing on formants, and specifically estimates  
 880 of the formants in the *center* of the vowel. There is, however, ample evidence that formant  
 881 dynamics throughout the vowel can strongly affect perception (Assmann and Katz, 2005;  
 882 Hillenbrand and Nearey, 1999; Nearey and Assmann, 1986). In addition, there are proposals  
 883 that entirely give up the assumption that formants are the primary cues to vowel identity  
 884 (e.g., whole-spectrum accounts, Hillenbrand *et al.*, 2006). While these proposals might  
 885 provide a more informative representation of vowels, we consider it unlikely that they would  
 886 entirely remove the problem of cross-talker variability. For instance, Richter *et al.* (2017) still  
 887 found benefits of normalization even when the entire frequency spectrum throughout vowels  
 888 was considered (in the form of Mel-Frequency Cepstral Coefficients and their derivatives).  
 889 For the present work, auxiliary analyses in the SI (§3E) replicated our core findings when

890 F3 was included in the model. Still, it remains unclear whether the inclusion of additional  
891 cues, such as VISC, or additional formant dynamics, would alter the results of the present  
892 study.

893 As is the case of any computational work, the present study committed to a number of  
894 assumptions that are not critical, but were necessary in order to deliver clear quantitative  
895 predictions. Quantitative tests of theories—as we have done here—require assumptions  
896 about *every* aspect of the model. Here, this included all the steps necessary to link properties  
897 of the stimuli to listeners' responses. For this purpose, we adopted the ASP framework (Xie  
898 *et al.*, 2023), and visualized the graphical model that links stimuli ( $x$ ) to responses ( $r$ ) in  
899 Figure 6.

900 Many of the assumptions we made should be relatively uncontroversial—e.g., the decision  
901 to include both external (environmental) and internal (perceptual) noise in our model. While  
902 these noise sources are often ignored in modeling human behavior, it is uncontroversial that  
903 they exist. Other assumptions we made were introduced as simplifying assumptions for  
904 the sake of feasibility—e.g., we expressed the effect of both types of noise through a single  
905 parameter that related the average within-category variability of formants to noise variability  
906 in the transformed and normalized formant space. In reality, however, environment noise  
907 can have effects that are independent of internal noise, and internal noise likely affects  
908 information processing at multiple (or all) of the steps shown in Figure 6. Such simplifying  
909 assumptions are both inevitable, and not necessarily problematic: as long as they do not  
910 introduce systematic bias to the evaluation of normalization accounts, they should not limit  
911 the generalizability of our results.

912 Some of our assumptions, however, might be more controversial. For example, we assumed  
 913 that category representations can be expressed as multivariate Gaussian distributions in the  
 914 formant space. This assumption, too, is a simplifying assumption—it simplified the com-  
 915 putation of likelihoods—rather than a critical feature of the ASP framework we employed.  
 916 While human category representations are unlikely to be Gaussians, the alternative, e.g.,  
 917 exemplar representations, would come with its own downsides, such as increased sensitiv-  
 918 ity to the limited size of phonetic databases and substantial increases in computation time  
 919 (exemplar representations afford researchers with much larger degrees of freedom). For re-  
 920 searchers curious how this and other assumptions we made affect our results, our data and  
 921 code are shared on OSF.

922 Like previous work, we further assumed that all listeners in our experiments use the  
 923 same underlying vowel representations—the same dialect template(s). However, as already  
 924 discussed, it is rather likely that not all of our listeners employed the same dialect tem-  
 925 plate(s). An additional analysis reported in the SI ([§3D](#)) thus compared normalization  
 926 accounts against only the subset of listeners who employed the dialect template used by  
 927 the majority of participants (see lower-left of Figure 5B). This left only 20 participants for  
 928 Experiment 1a (71.4%) and 23 for Experiment 1b (82.1%), substantially reducing statistical  
 929 power. Replicating the main analysis, uniform scaling accounts again fit listeners' behavior  
 930 well across both experiments. The best-performing account for Experiment 1a did, however,  
 931 differ from the one obtained for the superset of data (the intrinsic Syrdal & Gopal achieved  
 932 the best fit to listeners' responses in Experiment 1a for the shared dialect subset; see SI,  
 933 [§3D](#)).

934 A related assumption was introduced by the use of a phonetic database to approximate  
935 listeners' vowel representations. This deviates from most previous evaluations of normal-  
936 ization accounts (McMurray and Jongman, 2011; Barreda, 2021; but see Richter *et al.*,  
937 2017), and reflects our commitment to a strong assumption made by most theories of speech  
938 perception: that listeners' representations reflect the formant statistics previously experi-  
939 enced speech input. By using a phonetic database to estimate listeners' representations, we  
940 *substantially* reduced the degrees of freedom in the evaluation of normalization accounts,  
941 reducing the chance of over-fitting to the data from our experiments. Our approach does,  
942 however, also introduce two new assumptions.

943 First, our approach assumes that the mixture of dialect template(s) used by talkers in the  
944 database sufficiently closely approximates those of the listeners in our experiments. Some  
945 validation for this assumption comes from the additional analysis reported in the preceding  
946 paragraph: when we subset listeners to only those who used the majority dialect template,  
947 this improved the fit of all normalization accounts—as expected, if the category representa-  
948 tions we trained on the phonetic database primarily reflect those listeners' representations  
949 (see SI, §3 D). Future work could further address this assumption in a number of ways. On  
950 the one hand, dialect analyses like the ones we presented for our listeners (in Figure 5B)  
951 could compare listeners' templates against the templates used by talkers in the database.  
952 Alternatively or additionally, researchers could see whether our results replicate if ideal  
953 observers are instead trained on other databases that have been hypothesized to reflect a  
954 'typical' L1 listeners' experience with US English. Finally, it might be possible in future  
955 work to use larger databases of vowel recordings to train separate ideal observers for all ma-

956 jor dialects of US English, and to try to *estimate* for each listener which mixture of dialects  
 957 their responses are based on.

958 Second, we made the simplifying assumption that listeners' category representation—or  
 959 at least the representations listeners' drew on during the experiment—are talker-*independent*  
 960 (we trained a single set of multivariate Gaussian categories, rather than, e.g., hierarchically  
 961 organized set of multiple dialect templates). While this assumption is routinely made in  
 962 research on normalization and beyond, it might well be wrong (see e.g., [Xie \*et al.\*, 2021](#)).

963 Finally, the evaluation of normalization accounts in the present study shares with all  
 964 previous work (e.g., [Apfelbaum and McMurray, 2015](#); [Barreda, 2021](#); [Cole \*et al.\*, 2010](#); [Mc-  
 965 Murray and Jongman, 2011](#); [Nearey, 1989](#); [Richter \*et al.\*, 2017](#)) another simplifying assump-  
 966 tion that is clearly wrong: the assumption that listeners *know* the talker-specific formant  
 967 properties required for normalization. Specifically, we normalized the input for each ideal  
 968 observer using the maximum likelihood estimates of the normalization parameters over all  
 969 stimuli for the respective experiment. For example, for the evaluation of the ideal observer  
 970 trained on Lobanov normalized formants against listeners' responses in Experiment 1a, we  
 971 used the formant means and standard deviations of the stimuli used in Experiment 1a to  
 972 normalize F1 and F2. While this follows previous work, it constitutes a problematic as-  
 973 sumption for the evaluation of extrinsic normalization accounts. For extrinsic accounts,  
 974 the approach adopted here would seem to entail the ability to predict the future: even on  
 975 the first trial of the experiment, the input to the ideal observers were formants that were  
 976 normalized based on the normalization parameters estimated over the acoustic properties  
 977 of *all* stimuli. Listeners instead need to *incrementally infer* talker-specific properties from

978 the speech input (Barreda and Jaeger, [submitted](#); Nearey and Assmann, 2007; Xie *et al.*,  
 979 2023). An important avenue for future research is thus the development and evaluation of  
 980 incremental normalization accounts.

981 The present data only allow an initial, rather tentative, look at this question. For example,  
 982 for Experiment 1a, for which each trial had a known correct answer (the vowel intended by  
 983 the talker), we can assess whether participants' recognition accuracy improved across trials,  
 984 as would be expected if listeners need to incrementally infer the talker-specific normalization  
 985 parameters. Figure 11A suggests that this was indeed the case: the non-parametric listeners'  
 986 average recognition accuracy improved over the course of the experiment from about 65%  
 987 to 88%, with most of the improvements occurring during the first ten trials. To address  
 988 potential confounds due to differences in the distribution of stimuli across trials, we used a  
 989 generalized additive mixed-effect model to predict listeners' accuracy from log-transformed  
 990 trial order while accounting for random by-participant and by-item intercepts and slopes  
 991 for the log-transformed trial order (blue lines). Still, this result should be interpreted with  
 992 caution, as Experiment 1a was not designed to reliably address questions about incremental  
 993 changes across the experiment.

994 Figure 11B shows how the fit of the best-fitting normalization model changes across trials.  
 995 We used a generalized additive mixed-effect model to predict the log-likelihood of listeners'  
 996 responses from log-transformed trial order while accounting for random by-participant and  
 997 by-item intercepts and slopes for the log-transformed trial order (blue lines). Given that  
 998 our evaluation of normalization accounts assumed that the normalization parameters were  
 999 already known on the first trial of the experiment, we would expect that the likelihood of

1000 listeners' responses under a normalization model would improve the more input listeners  
 1001 have received (i.e., as the simplifying assumptions of our evaluation become increasingly  
 1002 more plausible). For Experiment 1a, this indeed appears to be the case. However, no clear  
 1003 evidence for such incremental improvements in the fit of the normalization model is observed  
 1004 for Experiment 1b. In short, the present data does not support decisive conclusions about  
 1005 the extent to which normalization proceeds incrementally.

1006 ## 'summarise()' has grouped output by 'Experiment'. You can override using the '.groups'  
 1007 ## 'summarise()' has grouped output by 'Experiment', 'Trial'. You can override using the

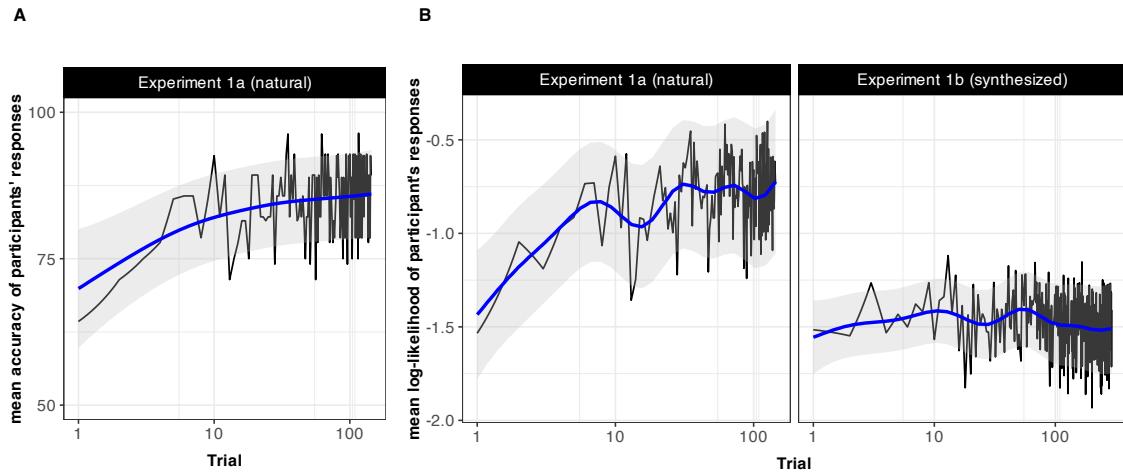


FIG. 11. **Panel A:** Changes across trials in listeners' average accuracy in recognizing the vowel intended by the talker in Experiment 1a, averaged across items and participants (black line). Blue line shows a generalized additive mixed-effects model predicting accuracy from log-transformed trial order, with 95% CIs. **Panel B:** Log-likelihood of listeners' responses under the best-fitting normalization account at each trial, averaged across items and participants (Johnson's uniform scaling for Experiment 1a and Nearey's uniform scaling for Experiment 1b). Blue lines show generalized additive mixed-effects models predicting log-likelihood from log-transformed trial order, with 95% CIs.

1008 **C. Concluding remarks**

1009 We set out to compare how well competing accounts of formant normalization explain  
 1010 listeners' perception of vowels. We developed a computational framework that makes it  
 1011 possible to compare a large number of different accounts against multiple data sets. The  
 1012 code we share on OSF makes it possible to 'plug in' different accounts of vowel normalization,  
 1013 different phonetic databases, and different perception experiments. This, we hope, will  
 1014 substantially reduce the effort necessary to conduct similar evaluations on other datasets,  
 1015 dialects, and languages.

1016 Comparing 20 of the most influential normalization accounts against L1 listeners' per-  
 1017 ception of US English monophthongs, we found that the normalization accounts that best  
 1018 describe listeners' perception share that they (1) learn and store talker-specific properties  
 1019 and (2) seem to be computationally very simple—taking advantage of the physics of sound  
 1020 generation to use as few as a single parameter to normalize inter-talker variability in vocal  
 1021 tract size. While the number of studies that have compared normalization accounts against  
 1022 *listeners'* behavior remains surprisingly small, these two results confirm the findings from  
 1023 more targeted comparisons that were focused on 2-3 accounts at a time (Barreda, 2021;  
 1024 Nearey, 1989; Richter *et al.*, 2017). Overall then, we submit that it is time for research in  
 1025 speech perception and beyond to consider simple uniform scaling the most-likely candidate  
 1026 for human formant normalization.

1027 **SUPPLEMENTARY MATERIAL**

1028 See supplementary material at <https://osf.io/zemwn/> for more details on participant  
1029 and experiment data, on the vowel database used, and on the computational models pre-  
1030 sented in the paper. The supplementary material also contains additional auxiliary analyses,  
1031 including models trained on different subsets of the data, and on additional cues besides F1-  
1032 F2.

1033 **ACKNOWLEDGMENTS**

1034 Earlier versions of this work were presented at 2023 ASA meeting, ExLing 2022, at the  
1035 Department of Computational Linguistics at the University of Zürich and at the Depart-  
1036 ment of Swedish language and multilingualism at Stockholm University. We are grateful to  
1037 Maryann Tan, Chigusa Kurumada, and Xin Xie for feedback on this work. We thank Travis  
1038 Wade for clarifications on the synthesis procedure used in his study. We thank Leslie Li  
1039 and Xin Xie for sharing their database of L1-US English \*hVd\* productions, and the JASA  
1040 copy editing staff for help with the Latex formatting. This work was partially funded by  
1041 grants to AP from Kungliga Vetenskapsakademien, Kungliga Vitterhetsakademien, and the  
1042 Department of Swedish Language and Multilingualism at Stockholm University, as well as  
1043 grants to TFJ by the Helge Ax:son Johnson foundation, the Stockholm University Board of  
1044 Human Science (Funding for Strategic Investments), and the Stockholm University Faculty  
1045 of Humanities' Research School (Kvalitetssäkrande medel grant).

1046 **AUTHOR CONTRIBUTIONS**

1047 AP designed the experiments and collected the data, with input from TFJ. TFJ pro-  
1048 grammed the experiments with input from AP. AP analyzed the experiments, with input  
1049 from TFJ. AP and TFJ wrote the code to implement and fit the normalization models, with  
1050 input from SB. AP developed the visualizations within input from SB and TFJ. AP wrote  
1051 the first draft of the manuscript with edits by SB and TFJ.

1052 **AUTHOR DECLARATIONS**

1053 **Conflict of Interest**

1054 The authors have no conflicts to disclose.

1055 **Ethics approval**

1056 This study was reviewed and approved Research Subjects Review Board (RSRB) of the  
1057 University of Rochester (STUDY00000417) under the OHSP and UR policies, and in ac-  
1058 cordance with Federal regulation 45 CFR 46 under the university's Federal-wide Assurance  
1059 (FWA00009386).

1060 **SUPPLEMENTARY INFORMATION FOR *PERSSON, BARREDA & JAEGER***  
 1061 **(2024). COMPARING ACCOUNTS OF FORMANT NORMALIZATION AGAINST**  
 1062 ***US ENGLISH LISTENERS' VOWEL PERCEPTION***

1063 **§1. REQUIRED SOFTWARE**

1064 Both the main text and these supplementary information (SI) are derived from the same  
 1065 R markdown document available via <https://osf.io/zemwn/>. It is best viewed using Acrobat  
 1066 Reader. The document was compiled using knitr in RStudio with R:

```
1067 ## -  

1068 ## platform      aarch64-apple-darwin20  

1069 ## arch          aarch64  

1070 ## os            darwin20  

1071 ## system        aarch64, darwin20  

1072 ## status  

1073 ## major         4  

1074 ## minor         4.2  

1075 ## year          2024  

1076 ## month         10  

1077 ## day           31  

1078 ## svn rev       87279  

1079 ## language      R  

1080 ## version.string R version 4.4.2 (2024-10-31)  

1081 ## nickname      Pile of Leaves
```

1082 Readers interested in working through the R markdown, and knitting it into a PDF will  
 1083 also need to download the IPA font **SIL Doulos** and a Latex environment like (e.g., **MacTex**  
 1084 or the R library **tinytex**).

1085 We used the following R packages to create this document: R (Version 4.4.2; **R Core**  
 1086 **Team, 2024**) and the R-packages *assertthat* (Version 0.2.1; **Wickham, 2019**), *brms* (Version  
 1087 2.22.0; **Bürkner, 2017, 2018, 2021**), *Cairo* (Version 1.6.2; **Urbanek and Horner, 2023**), *cmd-  
 1088 stanr* (Version 0.8.0; **Gabry et al., 2024**), *cowplot* (Version 1.1.3; **?**), *dplyr* (Version 1.1.4;

1089 `Wickham et al., 2023`), `ellipse` (Version 0.5.0; `Murdoch and Chow, 2023`), `english` (Ver-  
 1090 sion 1.2.6; `Fox et al., 2021`), `forcats` (Version 1.0.0; `Wickham, 2023a`), `furrr` (Version 0.3.1;  
 1091 `Vaughan and Dancho, 2022`), `fuzzyjoin` (Version 0.1.6; `Robinson, 2020`), `ggforce` (Version  
 1092 0.4.2; `Pedersen, 2024a`), `ggh4x` (Version 0.3.0; `van den Brand, 2024`), `ggnewscale` (Version  
 1093 0.5.0; `Campitelli, 2024`), `ggplot2` (Version 3.5.1; `Wickham, 2016`), `ggtext` (Version 0.1.2; `Wilke`  
 1094 and `Wiernik, 2022`), `itsadug` (Version 2.4.1; `van Rij et al., 2022`), `knitr` (Version 1.49; `Xie,`  
 1095 `2024`), `lubridate` (Version 1.9.4; `Grolemund and Wickham, 2011`), `magick` (Version 2.8.5;  
 1096 `?`), `magrittr` (Version 2.0.3; `Bache and Wickham, 2022`), `mgcv` (Version 1.9.1; `Wood et al.,`  
 1097 `2016`; `Wood, 2003, 2004, 2011`), `modelr` (Version 0.1.11; `Wickham, 2023b`), `MVBeliefUpdatr`  
 1098 (Version 0.0.1.10; `Jaeger, 2024`), `nlme` (Version 3.1.166; `Pinheiro et al., 2023`), `patchwork`  
 1099 (Version 1.3.0; `Pedersen, 2024b`), `phonR` (Version 1.0.7; `McCloy, 2016`), `phonTools` (Version  
 1100 0.2.2.2; `Barreda, 2023`), `plotfunctions` (Version 1.4; `van Rij, 2020`), `plotly` (Version 4.10.4;  
 1101 `Sievert, 2020`), `purrrr` (Version 1.0.2; `Wickham and Henry, 2023`), `Rcpp` (Version 1.0.13.1;  
 1102 `Eddelbuettel et al., 2024`), `readr` (Version 2.1.5; `Wickham et al., 2024a`), `remotes` (Version  
 1103 2.5.0; `Csárdi et al., 2024`), `RJ-2021-048` (`Bengtsson, 2021`), `rlang` (Version 1.1.4; `Henry and`  
 1104 `Wickham, 2024`), `stringr` (Version 1.5.1; `Wickham, 2023c`), `tibble` (Version 3.2.1; `Müller and`  
 1105 `Wickham, 2023`), `tidybayes` (Version 3.0.7; `Kay, 2024`), `tidyverse` (Version 1.3.1; `Wickham et al.,`  
 1106 `2024b`) and `tidyverse` (Version 2.0.0; `Wickham et al., 2019`).

1107 If opened in RStudio, the top of the R markdown document should alert you to any  
 1108 libraries you will need to download, if you have not already installed them. The full session  
 1109 information is provided at the end of this document.

1110 **A. Interested in using R markdown do create APA formatted documents that  
 1111 integrate your code with your writing?**

1112 A project template, including R markdown files that result in APA-formatted PDFs, is  
 1113 available at <https://github.com/hlplab/template-R-project>. Feedback is welcome by  
 1114 the contact author for this paper or [fjaeger@ur.rochester.edu](mailto:fjaeger@ur.rochester.edu). We aim to help oth-  
 1115 ers avoid the detours we made when first deciding to embrace literal coding to increase  
 1116 transparency in our projects.

1117 **§2. ADDITIONAL INFORMATION IN EXPERIMENTS 1A AND 1B**1118 **A. Exclusions**

1119 We first applied trial-level exclusion criteria to each participant, and then applied  
 1120 participant-level exclusion criteria.

1121 **1. *Trial exclusions***

1122 We excluded trials with RTs more than 3 standard deviations faster or slower than ex-  
 1123 pected. This was determined by first z-scoring the log-transformed RTs *within each partici-*  
 1124 *pant* (by subtracting the participants' mean from each observation and dividing through the  
 1125 participants standard deviation) and then z-scoring these z-scores *within each trial* across  
 1126 participants. This double-scaling approach was necessary as participants' RTs decreased  
 1127 substantially over the first few trials and then continued to decrease less rapidly until con-  
 1128 verging against a participant-specific minimum. This criterion did not remove just the first  
 1129 few trials but rather removed RTs that were unusually fast or slow *for that participant at*  
 1130 *that trial*. And, unlike more complicated methods (like developing a model of cross-trial  
 1131 decreases in RTs), the approach employed here does not make any assumptions about the  
 1132 shape of the speed up in RTs across trials. In total, 64 (1.4%) trials were excluded from  
 1133 Experiment 1a, and 89 (0.9%) from Experiment 1b.

1134 **2. *Participant exclusion***

1135 Participants were excluded if they (1) failed to pay attention to the instruction to wear  
 1136 over-the-ear-headphones, (2) had unusually slow or fast RT-means compared to other par-  
 1137 ticipants (more than 3 standard deviations faster or slower in their mean log-transformed  
 1138 RTs compared to other participants), (3) clearly did not do the task (e.g., randomly click-  
 1139 ing on different response options). Finally, participants were excluded if (4) the trial-level  
 1140 exclusions removed too many trials (>20% of all trials).

1141 For Experiment 1a, 1 participant was excluded based on the first criteria, as s/he used  
 1142 external speakers instead of head set (based on response in post-experiment questionnaire).  
 1143 One additional participant was excluded based on the second criterion. For Experiment 1b,  
 1144 no participants were excluded based on criteria 1 and 2.

1145 As the experiments did not contain catch trials, we visualized participants' individual re-  
 1146 sponses in order to identify participants that answered randomly, independent of the stimulus  
 1147 (third criterion). For Experiment 1a, Figure S1 suggests that 3 participants did not perform  
 1148 the task. This includes *had* responses to stimuli located across the entire vowel space, *heed*  
 1149 responses to stimuli located in the low back part of the space, or *odd* responses to stimuli  
 1150 located in the high front part of the space. For Experiment 1b, Figure S2 suggests that 2  
 1151 participants displayed unusual response patterns, responding *who'd* or *hod* for tokens in the  
 1152 high front part of the space, and *heed* for tokens in the high center and back parts.

1153 Finally, no participant was excluded based on the forth criteria (too many missing trials).  
 1154 In total, 5 participants were excluded from Experiment 1a (one of these were excluded based  
 1155 on multiple criteria), and 2 participants were excluded from Experiment 1b. This left 28  
 1156 participants for Experiment 1a, and 31 participants for Experiment 1b.

## 1157 B. Participant survey responses

1158 After completing the experiment, participants were asked to fill out a post-experiment  
 1159 survey. The survey contained questions about the type of audio equipment used, whether  
 1160 they experienced technical difficulties, alongside a series of questions on their perception  
 1161 of the talker and the stimuli used in the experiment. Participants' responses to three of  
 1162 these questions are summarized in Figure S3. It is clear from Figure S3 that participants  
 1163 experienced the natural and synthesized stimuli differently. Specifically, participants in Ex-  
 1164 periment 1b experienced more difficulties in distinguishing between words in the experiment  
 1165 – a larger proportion of participants reported that the words sounded more similar to each  
 1166 other than is the case for a typical native speaker of US English (for Experiment 1b, 68%  
 1167 reported word similarity, relative to 18 % of participants in Experiment 1a). Participants  
 1168 in Experiment 1b also experienced the stimuli as robotic-sounding to a larger extent (42%  
 1169 of participants characterized the speech as robotic in Experiment 1b, relative to 7% of par-  
 1170 ticipants in Experiment 1a). Finally, participants in Experiment 1b overall expressed more  
 1171 uncertainty about the stimuli (29% of participants stated that they were often unsure which  
 1172 word they heard, while no participant in Experiment 1a reported that they had difficulties  
 1173 identifying the word).

1174 The results of the survey confirm the results from the analysis of response entropy in  
 1175 the main text (II B), indicating substantial differences in participant uncertainty across

## Comparing normalization against perception

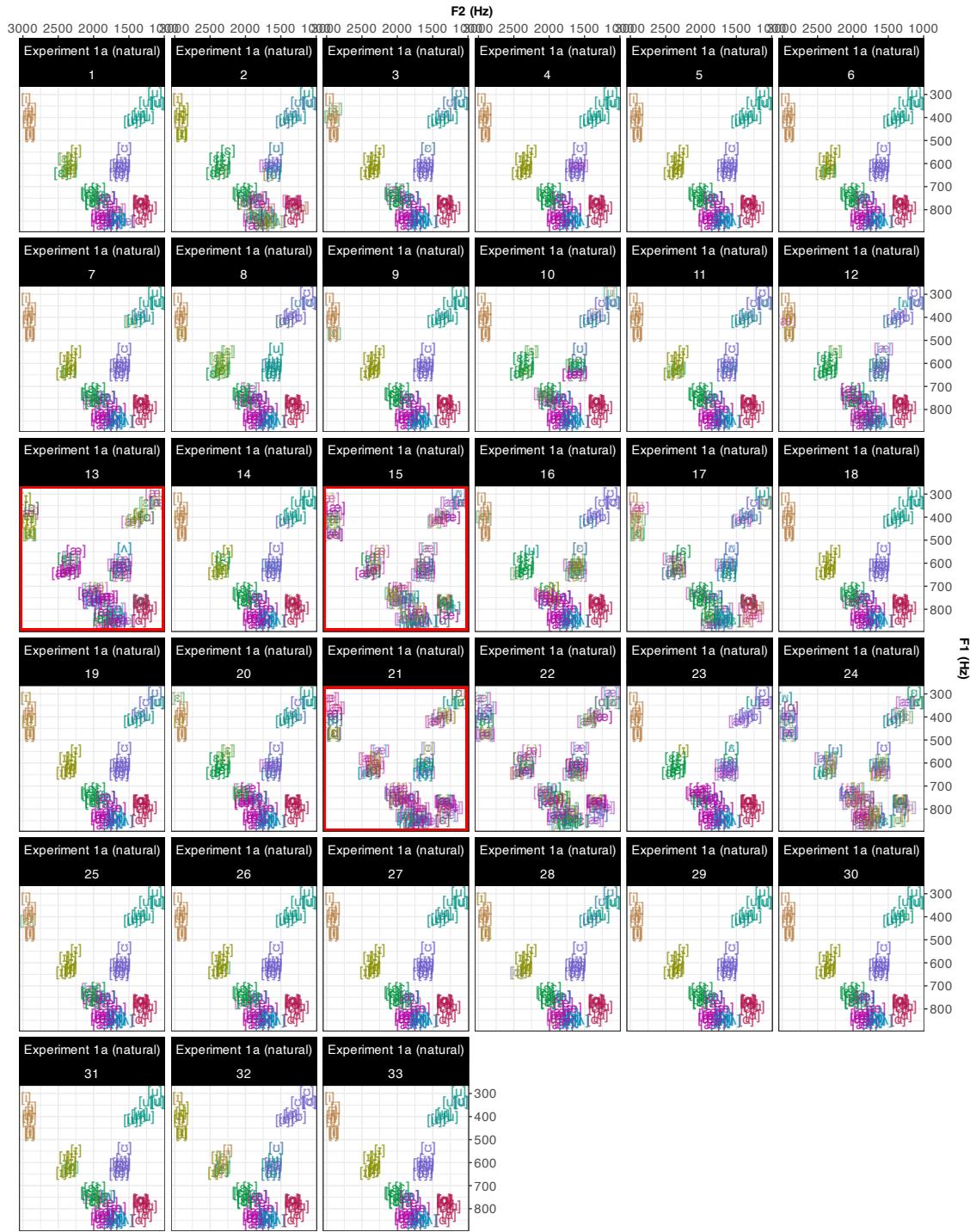


FIG. S1. Participants' response patterns in Experiment 1a. Color and vowel label indicate response provided by participants on each test location. Each vowel was repeated twice. Participants that displayed unusual response patterns and were therefore excluded, are highlighted.

## Comparing normalization against perception



FIG. S2. Participants' response patterns in Experiment 1b. Color and vowel label indicate response provided by participants on each test location. Each vowel was repeated twice. Participants that displayed unusual response patterns and were therefore excluded, are highlighted.

<sup>1176</sup> experiments. In addition, the survey responses also confirm that listeners' perception of the  
<sup>1177</sup> quality of the stimuli differed between experiments.

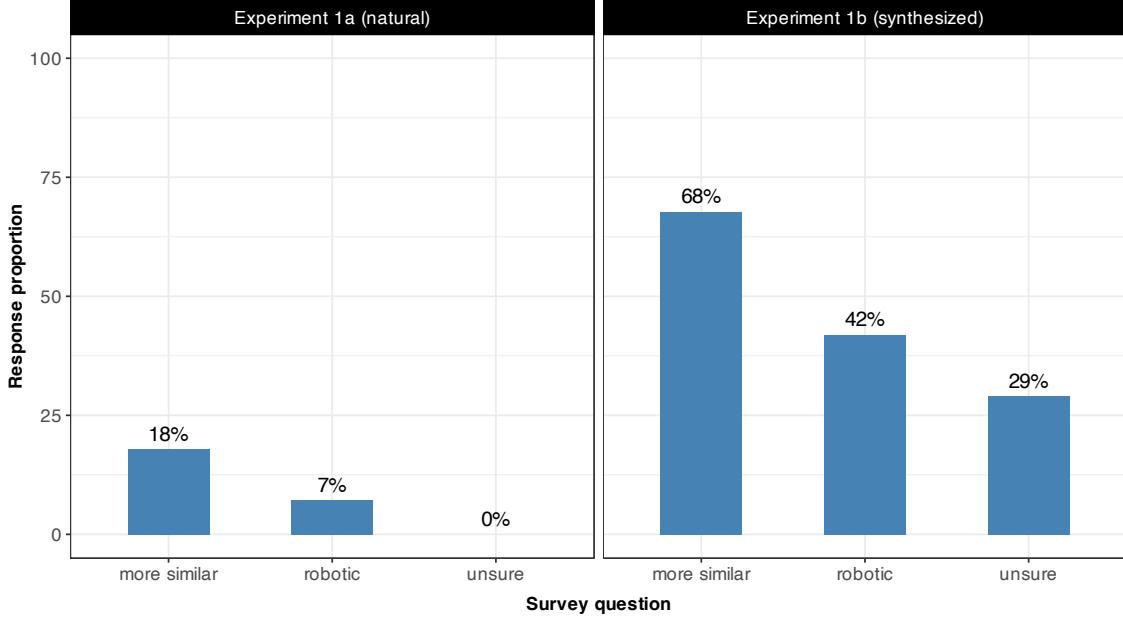


FIG. S3. Participants' responses to three of the post-experiment survey questions in Experiments 1a and 1b. The figure displays response proportions, normalized by the number of participants in each experiment. From left to right, proportion of participants that stated that: words sounded more similar to each other than what is usually the case; the stimuli sounded robotic; they were unsure about what word they heard.

<sup>1178</sup> **C. Spectrograms of stimuli used in Experiments 1a and 1b**

<sup>1179</sup> In the main text, we plot spectrograms of *heed*, *hid*, *odd* and *hood* tokens from Experiment  
<sup>1180</sup> 1a, together with four synthesized tokens with similar formant values from Experiment 1b  
<sup>1181</sup> (Figure 3). Figure S4 displays spectrograms of all eight categories from Experiment 1a  
<sup>1182</sup> together with eight synthesized tokens from Experiment 1b—the four categories from Figure  
<sup>1183</sup> 3 and the additional *head*, *had*, *hut* and *who'd* vowels.

Comparing normalization against perception

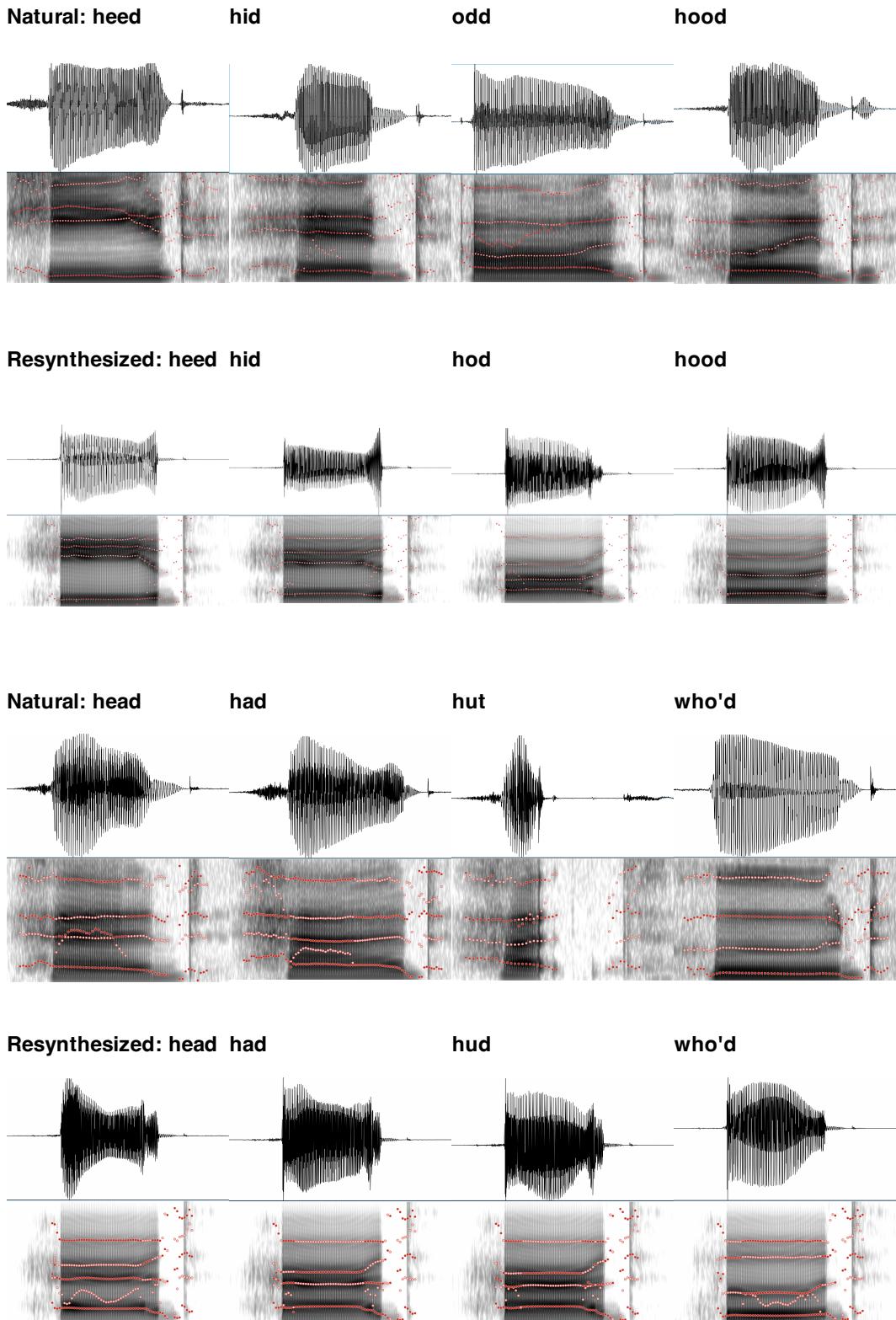


FIG. S4. Spectrograms of eight natural recordings from Experiment 1a and eight synthesized tokens with similar formant values from Experiment 1b.

1184 **D. Distribution of stimuli F1-F3 in Experiments 1a and 1b**

1185 The main text provides visualization of the stimuli's F1 and F2. Figure S5 shows the same  
 1186 stimuli in the 3D space defined by F1-F3. In addition, the F3 values in each synthesized token  
 1187 was set based on the F1-F2 values of the stimuli. Specifically, we used a linear regression  
 1188 based on the stimuli in Experiment 1a (predicting F3 from F1, F2 and their interaction)  
 1189 to predict F3 values for each F1-F2 combination in Experiment 1b. This difference in F3-  
 1190 distributions between experiment stimuli is clearly visible in Figure S5(b).

1191 Figure S5(b) further suggests that the talker used in Experiment 1a seems to produce  
 1192 some of the low (and high) back vowels with higher F3 values than might be expected. If  
 1193 this talker's F3-distribution is indeed unexpected or unusual (in comparison to listeners'  
 1194 expectations or to other talkers in the database), this could potentially partly explain the  
 1195 decrease in model fit for some of the normalization models when fit to F1-F3 compared to  
 1196 F1-F2 (see Section §3 E).

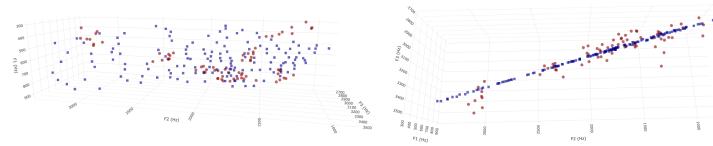


FIG. S5. Stimuli of Experiments 1a (red circles) and 1b (blue squares) in F1-F3 space. Shown are three different perspectives on the same data.

1197 **E. Auxiliary analysis of participant responses in Experiments 1a and 1b**

1198 Participants in Experiment 1b showed overall less agreement in their responses to the  
 1199 stimuli than participants in Experiment 1a, as indicated by the higher response entropy in  
 1200 Experiment 1b. As stated in the main text, response entropies differed even for tokens that  
 1201 were acoustically similar and only differed in  $\leq 30$  Hz along F1 and F2. Figure S6 visualizes  
 1202 differences in categorization behavior for these 40 acoustically similar tokens. For some of  
 1203 these tokens, participants selected the same category across experiments. However, even  
 1204 when selecting the same response option, participants still displayed higher disagreement  
 1205 for the tokens in Experiment 1b (mean by-item response entropy for the overlapping tokens

in Experiment 1a = 0.14 bits, SE = 0.02; and for the overlapping tokens in Experiment 1b = 0.4 bits, SE = 0.03).

We ran the same comparison in Nearey's uniform scaling space, for tokens with a difference of  $\leq .018$  log along F1 and F2. This comparison, however, returned highly similar results, with substantial differences in entropies for auditorily similar tokens (mean by-item response entropy for Experiment 1a = 0.15 bits, SE = 0.03; Experiment 1b = 0.35 bits, SE = 0.04).

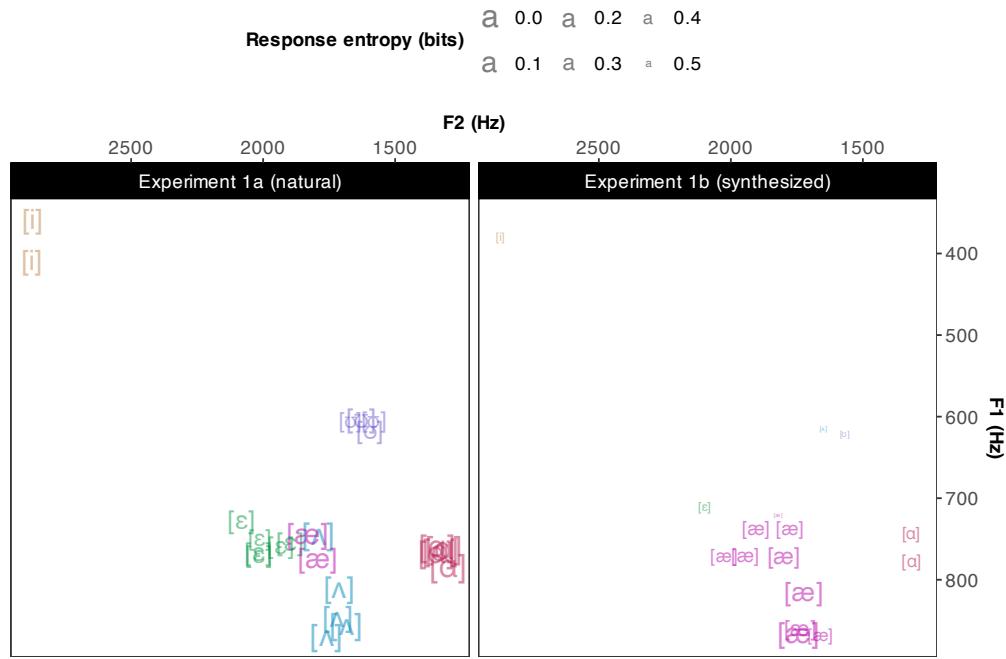


FIG. S6. Listeners' categorization responses in Experiments 1a and 1b, for comparable tokens in Hertz space. The vowel label indicates the most frequent response provided by participants on each test location. Size indicates how consistent responses were across participants, which larger symbols indicating more consistent responses (lower entropy).

To investigate the extent to which these differences in response entropy could be explained by differences in stimuli characteristics, we fit a series of general additive models (GAM). First, we confirmed that the difference in response entropy between experiments across all tokens was indeed statistically significant, by regressing response entropy against experiment (treatment coded;  $p < .0001$ , deviance explained = 57.6%). Next, we fit the same model to a subset of the data that excluded all *hut* and *odd* responses from Experiment 1a. This was done in order to assess whether differences in lexical context contributed to differences in

1220 entropy between experiments. This model confirmed that differences in entropy could not be  
 1221 reduced to differences in lexical context, since experiment remained a significant predictor  
 1222 of response entropy ( $p < .0001$ , deviance explained = 48.5%).

1223 In order to determine the extent to which this difference between experiments can be  
 1224 explained by differences in formants, we fit additional GAMs to the same subset. First, we  
 1225 regressed response entropy against both experiment and a tensor smooth of F1, F2, and  
 1226 F3. The tensor smooth allows for non-linear effects of, and interactions between, formants.  
 1227 To be maximally conservative, we normalized formants using one of the accounts that our  
 1228 computational studies find to be the best fit against human responses (Nearey's uniform  
 1229 scaling). The GAM with both formants and experiment explained 84.5% of the deviance in  
 1230 response entropy. Experiment still was a significant predictor of response entropy, suggesting  
 1231 that the difference in response entropy between experiments is not solely due to differences in  
 1232 formants. Formants did, however, explain a substantial amount of the difference in response  
 1233 entropy between experiments: a GAM with just the tensor smooth of F1-F3 explained 68.7%  
 1234 of the deviance.

1235 Experiment thus explained at least 15.8% of the deviance when it was added to this  
 1236 model—about 32.6% of the deviance explained if *only* experiment is included in the model  
 1237 (48.5%). These informal comparisons suggest that up to almost four fifths of the deviance  
 1238 that is explained by experiment is due to differences in formants between the experiments.

1239 Results using unnormalized formants confirmed these results (but, as expected if uniform  
 1240 scaling approximates the normalization employed by listeners, uniformly scaled formants  
 1241 explained about 4% more variance in response entropy).

1242 **§3. ADDITIONAL INFORMATION ON THE COMPUTATIONAL COMPAR-  
 1243 SON OF NORMALIZATION ACCOUNTS**

1244 **A. Methods**

1245 **1. Vowel data used to train ideal observers ([Xie and Jaeger, 2020](#))**

1246 In this subsection, we provide additional information on the formant data from the [Xie](#)  
 1247 [and Jaeger \(2020\)](#) database that we used to train the ideal observers, as described in the  
 1248 main text. The database consists of 1168 *hVd* word recordings from 16 (5 female) L1 talkers

1249 of a Northeastern dialect of US English (ages 18 to 35 years old). The talkers were recorded  
 1250 reading a list of 180 English monosyllabic words, a list of short sentences, and a list of ten  
 1251 *hVd* words—eight US English monophthongs (*heed*, *hid*, *head*, *had*, *hut*, *odd*, *hood*, *who'd*) as  
 1252 well as *aid* and *owed* (for further information, see [Xie and Jaeger, 2020](#)). For each talker, the  
 1253 database contains 9-10 recordings of each *hVd* word. An automatic aligner (Penn Phonetics  
 1254 Lab Forced Aligner, [Yuan and Liberman, 2008](#)) was used to obtain estimates for word and  
 1255 segment boundaries.

1256 The first author manually corrected the automatic alignments for all vowel segmentations.  
 1257 We then used the Burg algorithm in Praat ([Boersma and Weenink, 2022](#)) to extract estimates  
 1258 of the first three formants (F1-F3) at three time points of the vowel (35, 50, and 65 percent  
 1259 into the vowel). The following parameterization of the Burg algorithm was used:

- 1260     ● Time step (s): 0.01
- 1261     ● Max. number of formants: 5
- 1262     ● Formant ceiling (Hz): 5500 (5000 for the male talkers)
- 1263     ● Window length (s): 0.025
- 1264     ● Pre-emphasis from (Hz): 50

1265 In addition to F1-F3, we automatically extracted vowel duration and the fundamental  
 1266 frequency (F0) across the entire vowel. Figure S7, adapted from [Persson and Jaeger \(2023\)](#),  
 1267 visualizes the vowel data from the [Xie and Jaeger \(2020\)](#) database for all pairwise combina-  
 1268 tions of F0, F1, F2, F3 and vowel duration, in raw Hz. Unsurprisingly, the densities along  
 1269 the diagonal suggest that F1 and F2 carry most information for vowel category identity,  
 1270 as indicated by higher between-category separation than the other cues. We also note that  
 1271 there seems to be one female talker with substantially higher F0 and higher formants than  
 1272 the other female talkers.

1273 Figure S7, adapted from [Persson and Jaeger \(2023\)](#), visualizes the vowel data from the  
 1274 [Xie and Jaeger \(2020\)](#) database for all pairwise combinations of F0, F1, F2, F3 and vowel  
 1275 duration, in raw Hz. Unsurprisingly, the densities along the diagonal suggest that F1 and F2  
 1276 carry most information for vowel category identity, as indicated by higher between-category  
 1277 separation than the other cues. We also note that there seems to be one female talker with  
 1278 substantially higher F0 and higher formants than the other female talkers.

1279 Figure S8, adapted from [Persson and Jaeger \(2023\)](#), shows the distribution of F1 and F2  
 1280 in the different normalization spaces used in the main paper. We make a few observations.

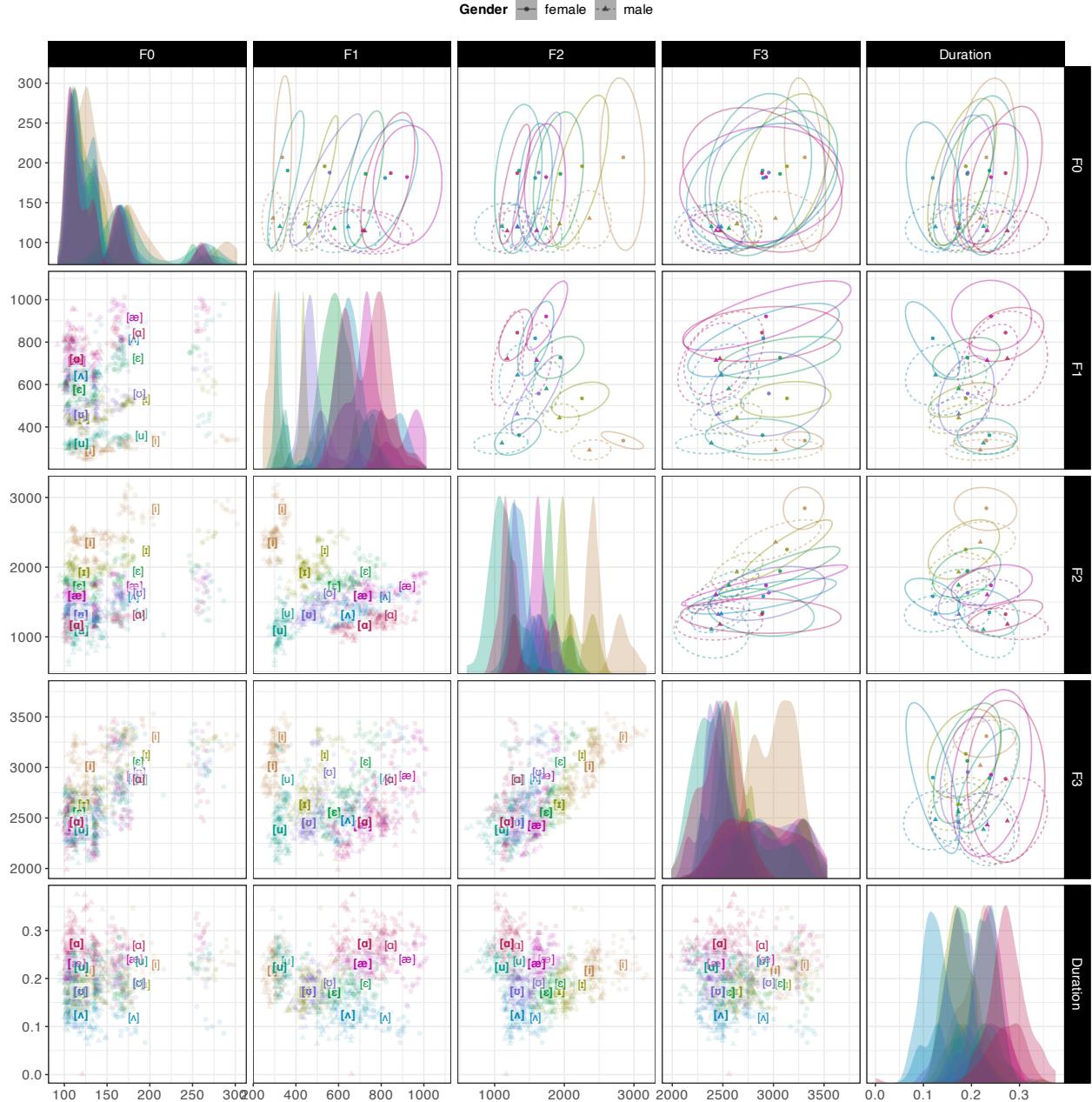


FIG. S7. The pairwise distributions of F0, F1, F2, F3, and duration for all 1168 recordings of eight monophthong  $hVd$  words in [Xie and Jaeger \(2020\)](#). Note that axis directions are not reversed.

**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.

1281 First, transforming the space into a perceptual scale (**grey**) does not seem to affect the vowel  
 1282 distributions much. Second, intrinsic normalization both increase category separability and  
 1283 more strikingly warp the space. Third, extrinsic centering and standardizing normalization  
 1284 seem to reduce category variability and increase separability, though it is not visually obvious  
 1285 which type of account achieves better separability.

1286 **2. Normalization parameters  $\theta$**

1287 Figure S9 relates the normalization parameters  $\theta$  obtained for each experiment to those  
 1288 obtained for the five training sets of the [Xie and Jaeger \(2020\)](#) database. This comparison  
 1289 serves two purposes. First, by comparing the  $\theta$ s of the stimuli in Experiment 1a, which  
 1290 was based on natural productions, to the  $\theta$ s obtained from [Xie and Jaeger \(2020\)](#), we can  
 1291 assess the extent to which the talker used for Experiment 1a is ‘typical’ relative to the other  
 1292 talkers of that database. Second, by comparing the range and variability of the  $\theta$ s across  
 1293 normalization accounts and experiments, we can assess the volatility of different types of  
 1294 parameters. In addition, we can also assess the difference between the beliefs the ideal  
 1295 observers have about the parameters and the parameters in the experiment stimuli.

1296 Figure S9 suggests that the talker used in Experiment 1a is overall aligned with the  
 1297 other talkers in the database, as indicated by the relative distance between the points to  
 1298 the dotted line. How closely the estimates obtained from the talker in Experiment 1a match  
 1299 the estimates obtained from the other talkers depend on the  $\theta$  assumed. For instance, for  
 1300 centering accounts assuming separate parameters for each formant,  $\theta$ s for F2 seem to be a  
 1301 closer match than  $\theta$ s for F1.

1302 Figure S9 further indicates that the reliability by which the formant statistics can be  
 1303 established for the same amount of data, seems to depend on the parameter. For instance,  
 1304 mean estimates display less variability than SDs and range values even within each experi-  
 1305 ment (as indicated by the larger CIs). Unsurprisingly, range values and SD also differ more  
 1306 between experiments than other estimates (distance between point and triangle for each ac-  
 1307 count; recall that the stimuli in Experiment 1b sampled larger parts of the phonetic space,  
 1308 Section II A 2).

1309 Finally, the differences between the  $\theta$ s for the database and those for the experiments  
 1310 ( $\delta_\theta$  in Figure S9, indicating the distance from horizontal dashed line) appear to be larger  
 1311 for standardizing accounts, especially for F2. This is likely due to the fact that the stan-

## Comparing normalization against perception

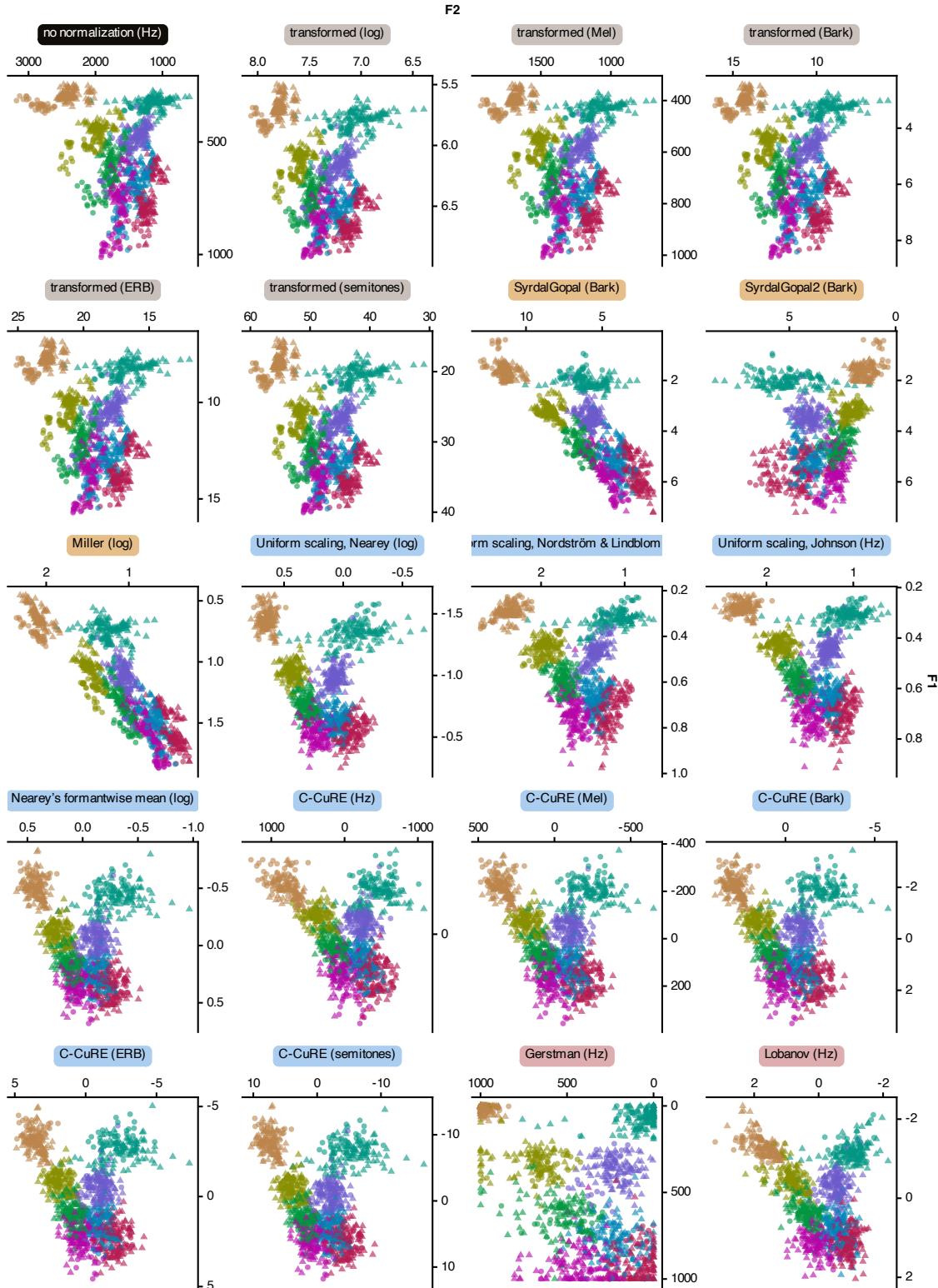


FIG. S8. Eight monophthong vowels of US English from the [Xie and Jaeger \(2020\)](#) database when F1 and F2 are transformed into a perceptual scale (grey), intrinsically normalized (yellow), or extrinsically normalized through centering (blue) or standardizing (purple). Each point corresponds to one recording, averaged across the three measurement points within each vowel segment. <sup>15</sup> Shape indicates gender (female talkers represented by points, male talkers by triangles).

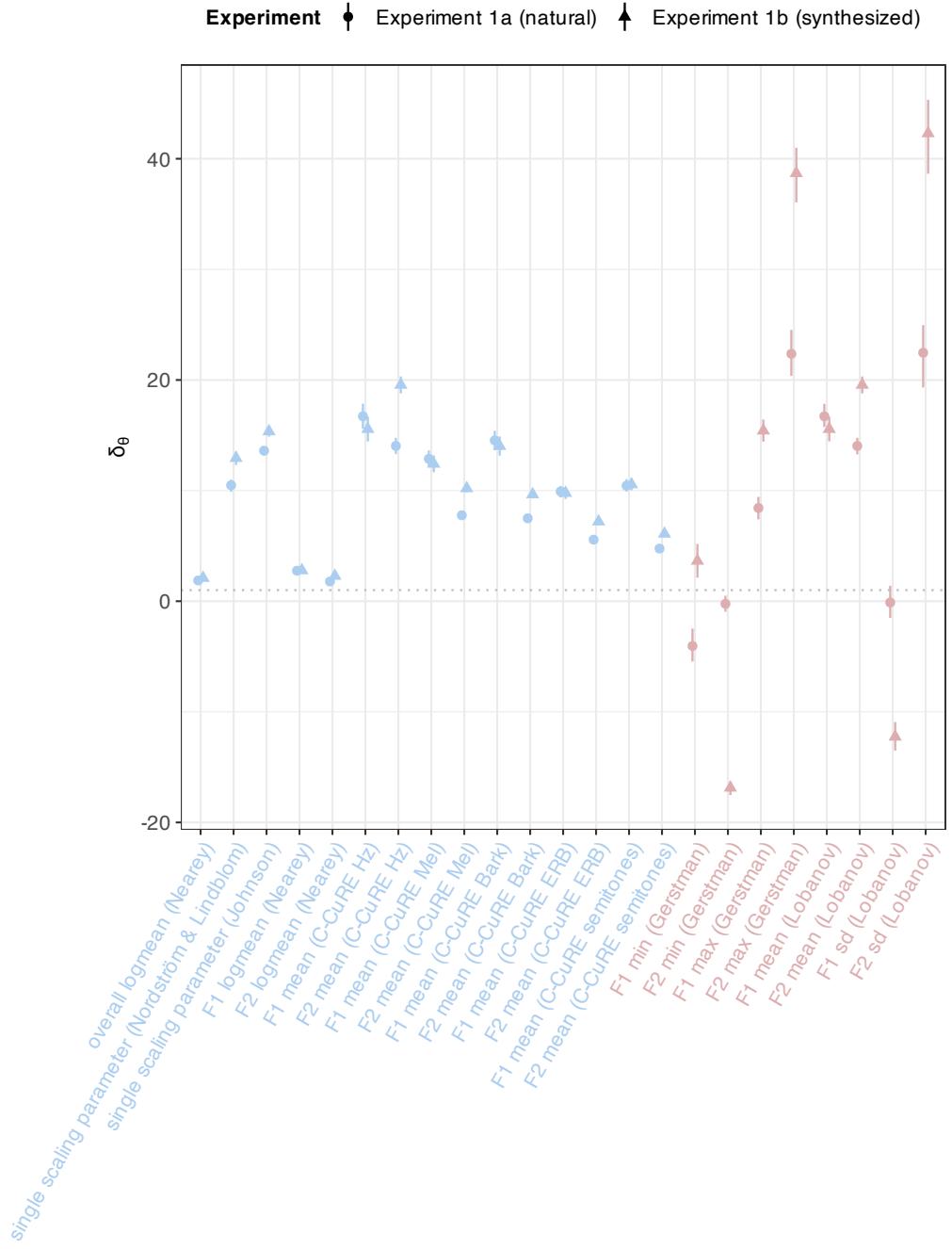


FIG. S9. Comparing normalization parameters  $\theta$  across the phonetic database used to estimate listeners' prior experience (Xie and Jaeger, 2020) and the stimuli from Experiments 1a and 1b. Only accounts that assume talker-specific normalization parameters are shown. To put all parameters on a similar scale, we divided each parameter by its average value for the five folds of the phonetic database, multiplied by 100, and subtracted 100. The resulting  $\delta_\theta$  indicates the difference in percent between the  $\theta$  estimated for the experiment relative to the  $\theta$  for the phonetic database. The horizontal dotted line ( $\delta_\theta = 1$ ) indicates perfect alignment between  $\theta$  value in database and experiment stimuli.

1312 standardizing accounts operate over Hertz space—note, for example, a similar pattern for F2 for  
 1313 the centering accounts operating in Hertz space. Larger differences between the  $\theta$ s for the  
 1314 database and those from an experiment imply that there is theoretically more room for that  
 1315 account to improve perception in the experiment. However, larger differences might also  
 1316 entail that listeners require more input before they have successfully converged against the  
 1317 correct normalization parameters for the experiment (as we mention in the general discus-  
 1318 sion, the present study shares with all previous work the limitation that it does *not* account  
 1319 for this incremental inference process). Without a model of how listeners infer normalization  
 1320 parameters from the input, it is thus not obvious whether larger differences between the  $\theta$ s  
 1321 for the database and those for an experiment should result in better or worse fit against  
 1322 listeners' responses. In any case, however, there was no systematic relationship between the  
 1323 degree of difference in the parameters and how well accounts fit listeners' responses: when  
 1324  $\delta_\theta$ s differed between experiments, they were almost always larger for Experiment 1b; yet,  
 1325 some accounts performed better in predicting listeners' responses for Experiment 1a, and  
 1326 others performed better in predicting listeners' responses for Experiment 1b.

1327 **3. Optimization process to fit models to human responses**

1328 To determine the best-fitting values for the two degrees of freedom—lapse rate ( $\lambda$ ) and  
 1329 noise ratio ( $\tau^{-1}$ )—when fitting the ideal observers to listeners' responses in Experiment 1a  
 1330 and 1b, we used constrained quasi-Newton optimization (Byrd *et al.*, 1995). Optimiza-  
 1331 tion was performed separately for each of the 200 combinations of normalization account,  
 1332 experiment, and training set. Specifically, we maximized the *likelihood* of the human cate-  
 1333 gorization responses in each experiment under the categorization model conditional on the  
 1334 model's lapse rate and noise,  $\sum_i^N \log p(response_i | F1_{i,\theta}, F2_{i,\theta}, M_{\theta,\lambda,\Sigma_{noise}})$ , where  $response_i$  is  
 1335 the  $i$ th categorization response,  $F1_{i,\theta}, F2_{i,\theta}$  are the F1 and F2 values for the  $i$ th observa-  
 1336 tion after normalization (with parameters  $\theta$  being estimated based on the distribution of  
 1337 phonetic cues across the stimuli in the experiment).  $M_{\theta,\lambda,\Sigma_{noise}}$  is the categorization model  
 1338 in Figure 6, with normalization parameters  $\theta$  fixed based from the prior cue distribution in  
 1339 the phonetic database (Xie and Jaeger, 2020), and  $\lambda$  and  $\tau^{-1}$  as the only free parameters  
 1340 to maximize the likelihood. The best-fitting parameterizations were determined by means  
 1341 of the `optim()` function in R's `stats` package (R Core Team, 2024). The starting value of  
 1342 lapse rates and noise were set to 0.1 and 0.15, respectively. We set the lower and upper

1343 bounds to  $10^{-10} \geq$  lapse rate  $\geq 1$ , and  $10^{-10} \geq$  noise  $\geq 10$  (i.e., including values well beyond  
1344 previously observed estimates for perceptual noise in [Kronrod et al., 2016](#), 1698).

1345 Section [§3 F](#) presents additional analyses that instead used a grid search over the parameter  
1346 space. These analyses overall confirm the results presented in the main text.

1347 **B. Results for F1-F2**

1348 **1. Significance test of model performance**

1349 To assess whether normalization significantly improved the fit to listeners' responses,  
1350 we conducted paired one-sided *t*-tests comparing the maximum likelihood values (averaged  
1351 across the five training sets) for each normalization account against those in the absence  
1352 of normalization (dummy coded with the unnormalized model as reference model). The  
1353 results of these tests indicated that normalization accounts achieved a better fit to listeners'  
1354 responses, compared to no normalization (all  $p < .05$ ), except for Gerstman normalization,  
1355 log-transformation and semitones-transformation and Experiment 1a, see Tables [S1](#) and [S2](#).

1356 **2. Parameter estimates for best-fitting models**

1357 In this section, we present the best-fitting estimates for the noise ratio ( $\tau^{-1}$ ) and attentional lapses ( $\lambda$ ). This provides insights into the relative contributions of these parameters in  
1358 explaining the variability found in the behavioral data between the two experiments. Figure  
1359 [S10](#) visualizes the best-fitting parameter estimates for each account (see also Tables [S3](#), [S4](#)  
1360 for summary of fitted values, and [S11](#) for an illustration of how the fitted noise affects the  
1361 bivariate Gaussian categories).

1363 The mean  $\lambda$  was overall similar across experiments (for Experiment 1a, overall mean  
1364 across accounts = 0.09, SD = 0.03; for Experiment 1b, overall mean = 0.12, SD = 0.15). This  
1365 suggests that listeners' categorization behavior was indeed affected by attentional lapses,  
1366 and to similar extents across the two experiments. Figure [S10](#) also indicates that some  
1367 accounts were fitted with clearly higher  $\lambda$  in Experiment 1b, e.g., Nordström & Lindblom,  
1368 and Johnson normalization.

1369 In contrast to the lapse rates, the two experiments clearly differed in the best-fitting  
1370 noise ratios,  $\tau^{-1}$ . For Experiment 1a, the best-fitting  $\tau^{-1}$  estimates are comparable to what

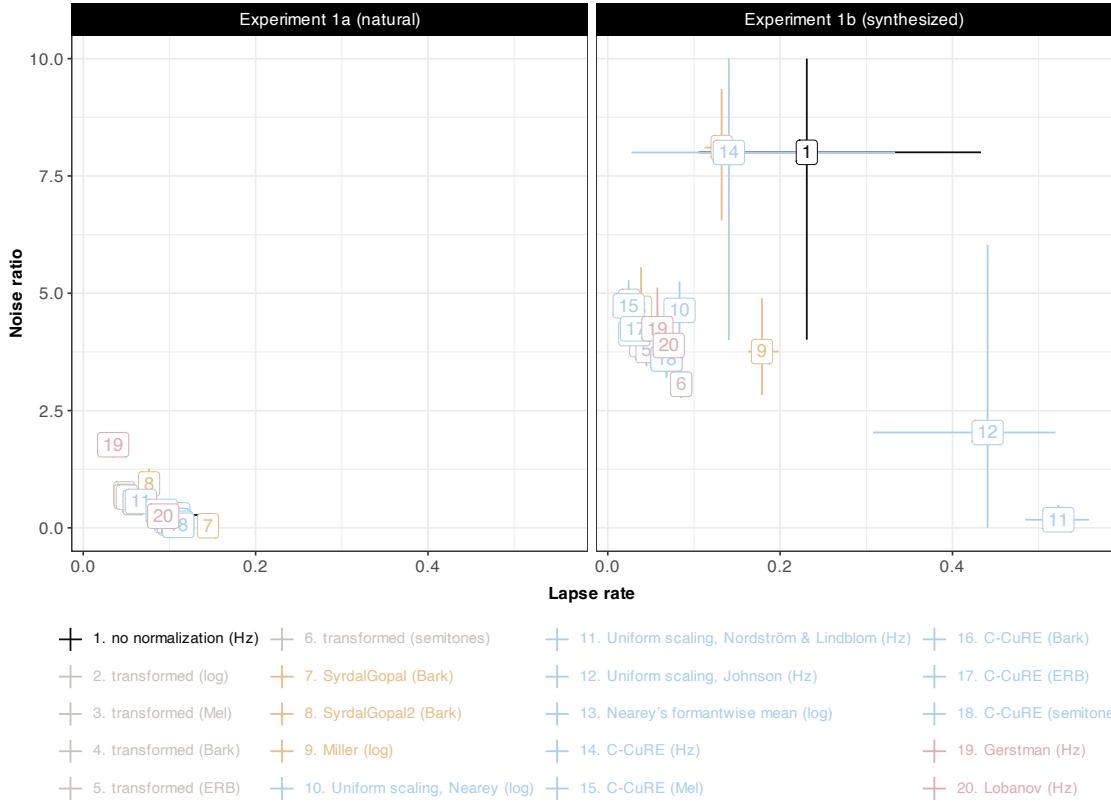


FIG. S10. Best-fitting estimates obtained for  $\lambda$  and  $\tau^{-1}$ . Numeric label is placed at the mean across the five training sets, line ranges represent the 95% CIs.

1371 Kronrod *et al.* (2016) found (mean  $\tau^{-1} = 0.42$ , SD = 0.46). For Experiment 1b, however,  
 1372 much larger noise ratios were observed (mean  $\tau^{-1} = 4.32$ , SD = 2.52). Of note, differences  
 1373 in noise ratios between experiments were observed even for the best-fitting accounts. This  
 1374 suggests that the difference is due to properties of the experiments, rather than universally  
 1375 bad predictions of the normalization accounts.

1376 Since there is no obvious reason to expect *internal* (perceptual) noise to differ between  
 1377 experiments, the observed large noise ratios for Experiment 1b must have other sources. We  
 1378 suspect that the speech synthesis procedure used for Experiment 1b removed useful infor-  
 1379 mation from those stimuli, which caused listeners to be more uncertain about the category  
 1380 identity. This is in line with the analyses presented in §2E, showing that listeners in Ex-  
 1381 periment 1b exhibited higher response entropy even when differences in the distribution of  
 1382 formant values across across the two experiments were taken into account. Of the two DFs  
 1383 we fit to listeners' responses, this increased uncertainty is relatively naturally accounted for

TABLE S1. *t*-test predicting the model log likelihood as a function of normalization account for Experiment 1a (ordered in descending order by best-fitting models). Estimate mean represents the estimate of the reference account (no normalization), and Diff. in means indicates the increase in estimate mean for that account relative to the reference account.

Normalization account	Statistic	Estimate mean	Diff. in means	<i>p</i> -value
Uniform scaling, Johnson (Hz)	-16.870	-3620.93	1020.284	0.000
SyrdalGopal (Bark)	-11.087	-3620.93	949.264	0.000
C-CuRE (Mel)	-14.652	-3620.93	907.904	0.000
C-CuRE (Bark)	-15.019	-3620.93	906.590	0.000
C-CuRE (ERB)	-13.711	-3620.93	905.508	0.000
Nearey's formantwise mean (log)	-12.217	-3620.93	904.476	0.000
C-CuRE (semitones)	-12.217	-3620.93	904.476	0.000
Uniform scaling, Nearey (log)	-9.354	-3620.93	890.753	0.000
C-CuRE (Hz)	-12.906	-3620.93	869.304	0.000
Uniform scaling, Nordström & Lindblom (Hz)	-9.112	-3620.93	841.884	0.000
Miller (log)	-8.074	-3620.93	800.627	0.001
Lobanov (Hz)	-8.253	-3620.93	756.465	0.001
SyrdalGopal2 (Bark)	-3.427	-3620.93	268.549	0.013
transformed (Bark)	-9.315	-3620.93	231.641	0.000
transformed (Mel)	-8.245	-3620.93	219.092	0.001
transformed (ERB)	-5.953	-3620.93	166.781	0.002
Gerstman (Hz)	-0.355	-3620.93	46.230	0.370
transformed (log)	0.688	-3620.93	-31.765	0.735
transformed (semitones)	0.688	-3620.93	-31.765	0.735

TABLE S2. *t*-tests predicting the model log likelihood as a function of normalization account for Experiment 1b (ordered in descending order by best-fitting models). Estimate mean represents the estimate of the reference account (no normalization), and Diff. in means indicates the increase in estimate mean for that account relative to the reference account.

Normalization account	Statistic	Estimate mean	Diff. in means	p_value
Uniform scaling, Nearey (log)	-41.250	-14405.18	3684.082	0.000
Lobanov (Hz)	-47.328	-14405.18	3583.837	0.000
Gerstman (Hz)	-42.344	-14405.18	3515.437	0.000
C-CuRE (Bark)	-30.829	-14405.18	3489.366	0.000
C-CuRE (ERB)	-31.757	-14405.18	3481.959	0.000
Nearey's formantwise mean (log)	-33.301	-14405.18	3367.455	0.000
C-CuRE (semitones)	-33.301	-14405.18	3367.455	0.000
C-CuRE (Mel)	-26.506	-14405.18	3211.180	0.000
transformed (log)	-37.128	-14405.18	3111.364	0.000
transformed (semitones)	-37.128	-14405.18	3111.364	0.000
transformed (ERB)	-33.480	-14405.18	2943.964	0.000
transformed (Bark)	-31.040	-14405.18	2737.680	0.000
SyrdalGopal2 (Bark)	-25.364	-14405.18	2659.434	0.000
transformed (Mel)	-27.293	-14405.18	2290.697	0.000
Miller (log)	-19.932	-14405.18	1877.593	0.000
Uniform scaling, Johnson (Hz)	-8.235	-14405.18	1630.218	0.001
SyrdalGopal (Bark)	-15.194	-14405.18	1448.857	0.000
Uniform scaling, Nordström & Lindblom (Hz)	-20.021	-14405.18	1331.748	0.000
C-CuRE (Hz)	-13.117	-14405.18	1110.344	0.000

1384 through increased noise ratios (which reduce the signal contained in a stimulus, rather than  
 1385 increasing the rate of random guessing).

1386 Finally, for the majority of accounts, there is little variability in estimates of  $\tau^{-1}$  and  
 1387  $\lambda$ —and likelihoods—across training sets (Tables S3 and S4, Figure S10, Figure S11). This  
 1388 suggests that models achieved their maximum likelihood fit to human data on similar es-  
 1389 timates for the two degrees of freedom. Important exceptions are parameter estimates for  
 1390 Syrdal & Gopal, Johnson, and C-CuRE (Hz) in Experiment 1b, that all display considerable  
 1391 variability across training sets.

1392 **3. By-item analysis**

1393 To provide further insight into model performance across the phonetic space, we visualize  
 1394 model fits against human behavior on a by-item level for three of the best-performing mod-  
 1395 els across experiments, Nearey’s uniform scaling, Johnson’s uniform scaling and Lobanov.  
 1396 This allows us to assess whether normalization always improves model fit in absence of  
 1397 normalization, and if normalization models perform equally well in different parts of the  
 1398 acoustic-phonetic space.

1399 While Figure S12 indicates a general tendency for increased model performance as hu-  
 1400 mans’ predictions about human behavior become stronger, models also improve relative to  
 1401 no normalization on items for which humans have weaker predictions. Normalization does  
 1402 not, however, improve model fit across the board. Relative to no normalization, all three  
 1403 accounts both increase and decrease in performance on a by-item level. The advantage of  
 1404 Nearey’s uniform scaling relative to no normalization seems to be driven by smaller improve-  
 1405 ments (<35% change) on many items in Experiment 1a (proportion of items with increase  
 1406 in performance = 75%; mean improvement in likelihood by item = 22.21, SD = 20.3; mean  
 1407 likelihood by item for items where there is *no* improvement = -17.13, SD = 15.54), whereas  
 1408 for Experiment 1b, the improvements are numerically larger and more frequent (proportion  
 1409 = 95.9%; mean improvement in likelihood by item = 26.89, SD = 21.66; mean likelihood by  
 1410 item for items where there is *no* improvement = -13.31, SD = 7.25). Johnson normalization  
 1411 follows the same pattern for Experiment 1a only (proportion = 75%; mean improvement  
 1412 in likelihood by item = 23.3, SD = 20.01; mean likelihood by item for items where there  
 1413 is *no* improvement = -13.21, SD = 15.85), while for Experiment 1b, improvements are less  
 1414 pronounced (proportion = 74%; mean improvement in likelihood by item = 18.29, SD =

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)

model	mean log likelihood	noise percentage	lapse rate
Uniform scaling, Johnson (Hz)	-3176.96	mean=0.35 (SD=0.25) mean=0.09 (SD=0.02)	
SyrdalGopal (Bark)	-3247.98	mean=0.05 (SD=0.05) mean=0.14 (SD=0.01)	
C-CuRE (Mel)	-3289.34	mean=0.15 (SD=0.18) mean=0.11 (SD=0.02)	
C-CuRE (Bark)	-3290.65	mean=0.16 (SD=0.19) mean=0.1 (SD=0.02)	
C-CuRE (ERB)	-3291.74	mean=0.1 (SD=0.18) mean=0.11 (SD=0.02)	
Nearey's formantwise mean (log)	-3292.77	mean=0.05 (SD=0.11) mean=0.11 (SD=0.01)	
C-CuRE (semitones)	-3292.77	mean=0.05 (SD=0.11) mean=0.11 (SD=0.01)	
Uniform scaling, Nearey (log)	-3306.49	mean=0.28 (SD=0.29) mean=0.11 (SD=0.01)	
C-CuRE (Hz)	-3327.94	mean=0.16 (SD=0.1) mean=0.1 (SD=0.01)	
Uniform scaling, Nordström & Lindblom (Hz)	-3355.36	mean=0.56 (SD=0.27) mean=0.07 (SD=0)	
Miller (log)	-3396.62	mean=0.1 (SD=0.04) mean=0.1 (SD=0.01)	
Lobanov (Hz)	-3440.78	mean=0.25 (SD=0.2) mean=0.09 (SD=0.02)	
SyrdalGopal2 (Bark)	-3928.70	mean=0.94 (SD=0.4) mean=0.08 (SD=0.02)	
transformed (Bark)	-3965.60	mean=0.67 (SD=0.22) mean=0.05 (SD=0.01)	
transformed (Mel)	-3978.15	mean=0.72 (SD=0.21) mean=0.05 (SD=0.01)	
transformed (ERB)	-4030.46	mean=0.65 (SD=0.25) mean=0.05 (SD=0)	
Gerstman (Hz)	-4151.01	mean=1.77 (SD=0.32) mean=0.03 (SD=0)	
no normalization (Hz)	-4197.24	mean=0.28 (SD=0.26) mean=0.11 (SD=0.04)	
transformed (semitones)	-4229.01	mean=0.54 (SD=0.32) mean=0.06 (SD=0.01)	
transformed (log)	-4229.01	mean=0.54 (SD=0.32) mean=0.06 (SD=0.01)	

<sup>1415</sup> 11.96; mean likelihood by item for items where there is *no* improvement = -9.09, SD =  
<sup>1416</sup> 5.68). Lobanov seems to pattern with Nearey for both Experiment 1a (proportion = 72.2%;  
<sup>1417</sup> mean improvement in likelihood by item = 20.6, SD = 20.59; mean likelihood by item for  
<sup>1418</sup> items where there is *no* improvement = -15.74, SD = 13.5), and Experiment 1b (proportion  
<sup>1419</sup> = 91.8%; mean improvement in likelihood by item = 27.52, SD = 23.14; mean likelihood by  
<sup>1420</sup> item for items where there is *no* improvement = -8.69, SD = 9.47).

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)

model	mean log likelihood	noise percentage	lapse rate
Uniform scaling, Nearey (log)	-13318.80	mean=4.64 (SD=0.81) mean=0.08 (SD=0)	
Lobanov (Hz)	-13419.04	mean=3.89 (SD=0.51) mean=0.07 (SD=0.01)	
Gerstman (Hz)	-13487.44	mean=4.24 (SD=1.12) mean=0.06 (SD=0.03)	
C-CuRE (Bark)	-13513.51	mean=4.14 (SD=0.6) mean=0.03 (SD=0.01)	
C-CuRE (ERB)	-13520.92	mean=4.24 (SD=0.88) mean=0.03 (SD=0.02)	
Nearey's formantwise mean (log)	-13635.42	mean=3.6 (SD=0.44) mean=0.07 (SD=0.01)	
C-CuRE (semitones)	-13635.42	mean=3.6 (SD=0.44) mean=0.07 (SD=0.01)	
C-CuRE (Mel)	-13791.70	mean=4.73 (SD=0.68) mean=0.02 (SD=0.01)	
transformed (log)	-13891.52	mean=3.06 (SD=0.36) mean=0.09 (SD=0.01)	
transformed (semitones)	-13891.52	mean=3.06 (SD=0.36) mean=0.09 (SD=0.01)	
transformed (ERB)	-14058.92	mean=3.78 (SD=0.39) mean=0.04 (SD=0.01)	
transformed (Bark)	-14265.20	mean=3.91 (SD=0.38) mean=0.04 (SD=0.01)	
SyrdalGopal2 (Bark)	-14343.45	mean=4.68 (SD=1.06) mean=0.04 (SD=0.01)	
transformed (Mel)	-14712.18	mean=4.83 (SD=0.35) mean=0.02 (SD=0.01)	
Miller (log)	-15125.29	mean=3.76 (SD=1.31) mean=0.18 (SD=0.02)	
Uniform scaling, Johnson (Hz)	-15372.66	mean=2.03 (SD=4.45) mean=0.44 (SD=0.15)	
SyrdalGopal (Bark)	-15554.02	mean=8.1 (SD=1.74) mean=0.13 (SD=0.03)	
Uniform scaling, Nordström & Lindblom (Hz)	-15671.13	mean=0.17 (SD=0.35) mean=0.52 (SD=0.04)	
C-CuRE (Hz)	-15892.54	mean=8 (SD=4.47) mean=0.14 (SD=0.22)	
no normalization (Hz)	-17002.88	mean=8 (SD=4.46) mean=0.23 (SD=0.23)	

1421 To explore whether differences in model performance are related to where in the acoustic-  
 1422 phonetic space items are located, we plot the by-item likelihood of the unnormalized model in  
 1423 the acoustic-phonetic space, along with likelihood differences between the best-performing  
 1424 models in each experiment, Johnson's and Nearey's uniform scaling accounts (see Figure  
 1425 S13).

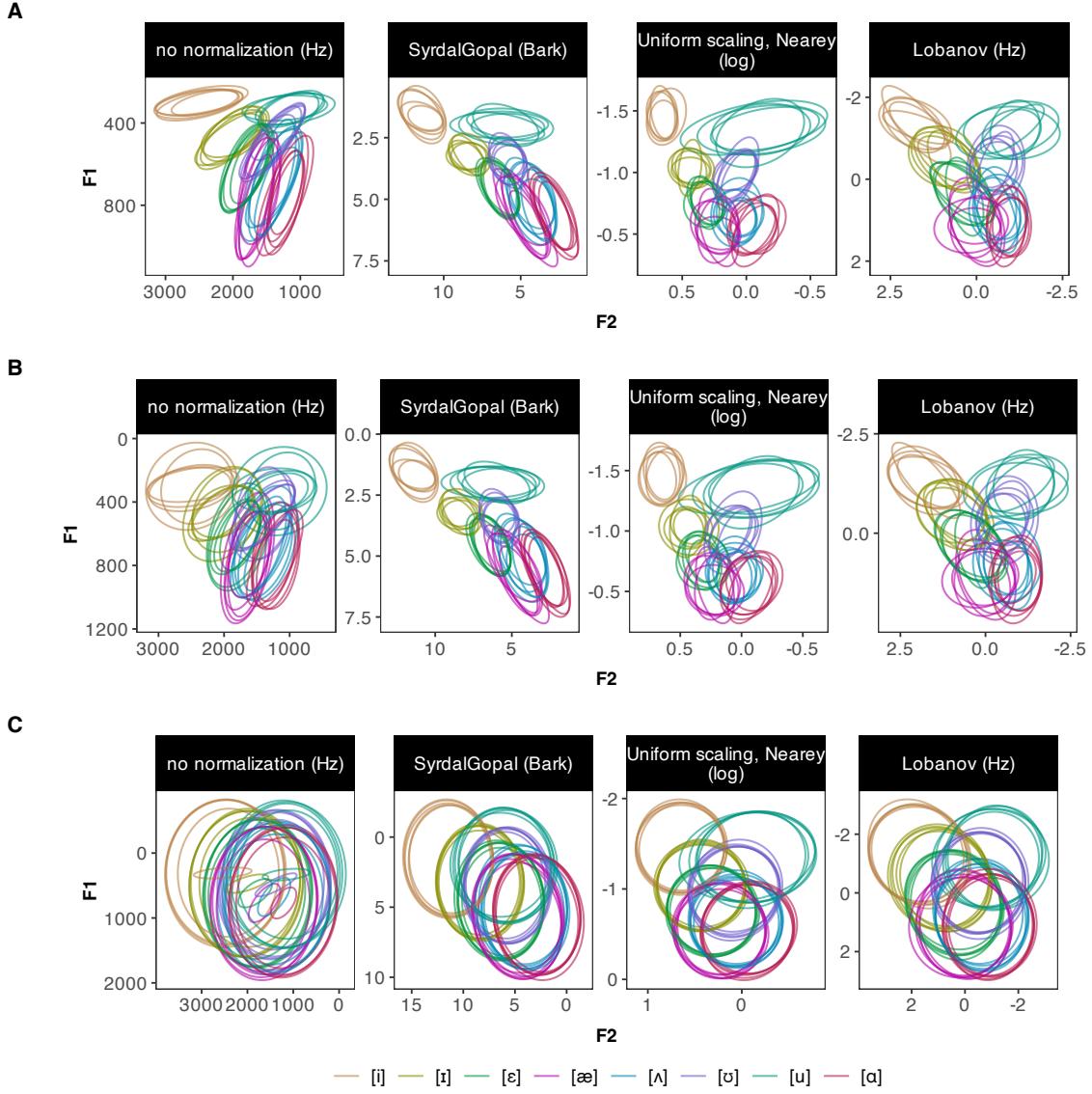


FIG. S11. Visualizing the bivariate Gaussian categories of four example normalization accounts for each of the five training sets (each training set corresponds to one set of eight ellipses). **Panel A** prior to adding  $\tau^{-1}$ , **Panel B** with added noise from best-fitting parameter estimates in Experiment 1a, **Panel C** with added noise from best-fitting parameter estimates 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.

1426 Figure S13 suggests that normalization does not improve the likelihood fit universally  
 1427 across the acoustic-phonetic space. Mirroring Figure S12, models achieve better fits in parts  
 1428 of the acoustic space where humans can more easily predict human behavior (Figure S13,

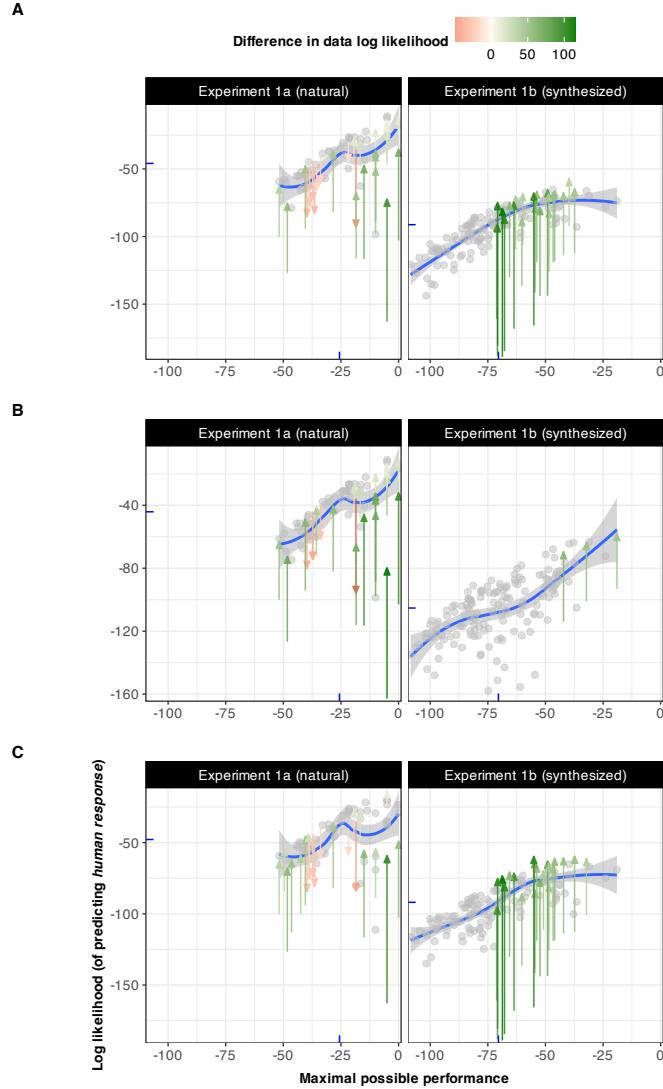


FIG. S12. By-item model improvement from no normalization, relative to the maximum possible performance (predicting human responses from human responses). Maximum log likelihood across items indicated by ticks on axis. Arrows indicate change from no normalization to Nearey's uniform scaling (**panel 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.

1429 *Panel A*). To the extent that this is not the case, it seems that normalization in general  
 1430 can adjust for this, improving model performance on many tokens where the maximum  
 1431 performance is high but the unnormalized model's predictions are low, e.g., in the left  
 1432 bottom and center part of the acoustic space (*Panels B-C*).

## Comparing normalization against perception

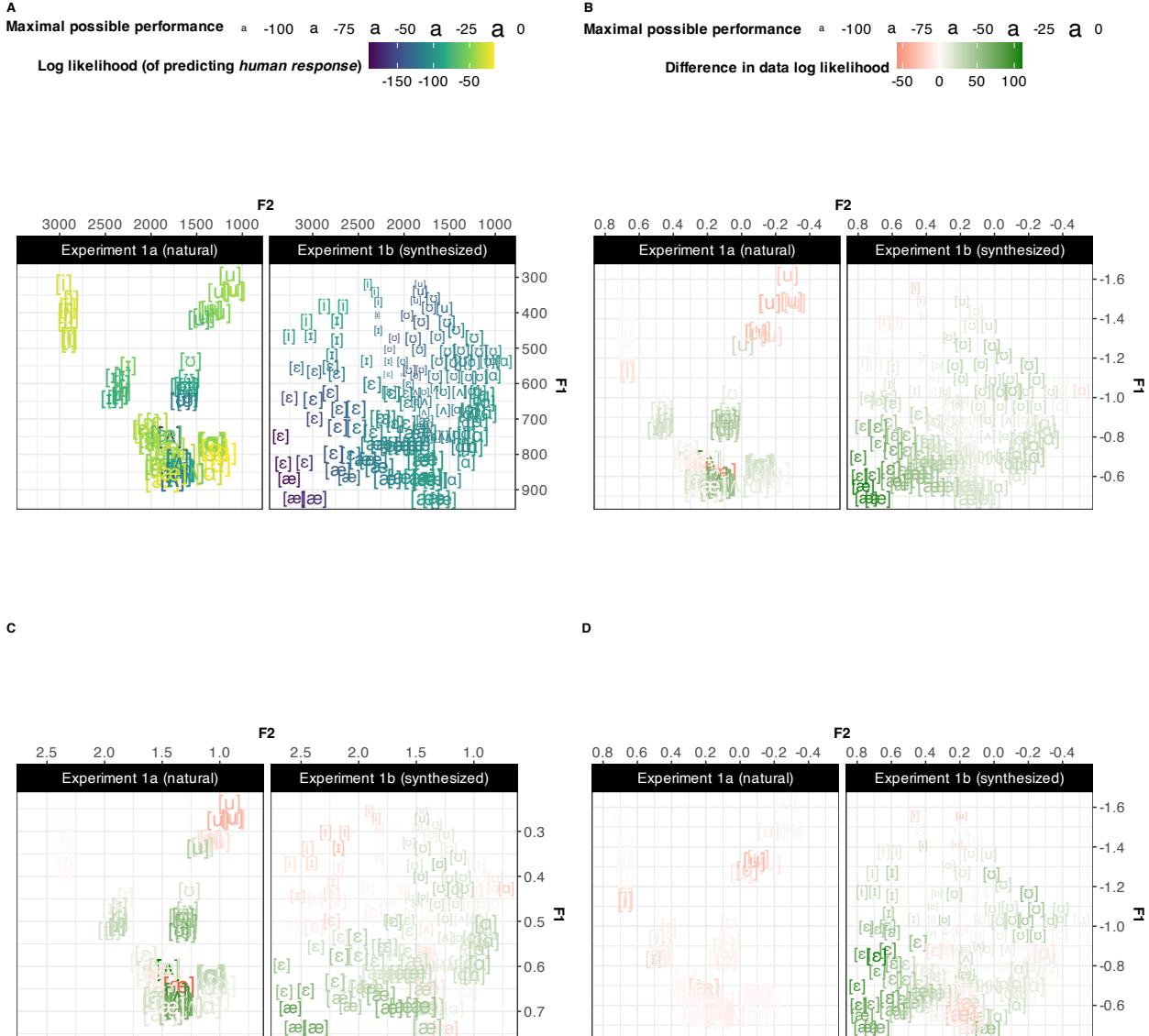


FIG. S13. In which part of the acoustic-phonetic space does normalization fail to improve fit against human responses? For each test location, the vowel label indicates the most frequent response provided by participants. Size of vowel label relates model performance to maximum performance (predicting 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**).

<sup>1433</sup> In Experiment 1a, both Nearey's uniform scaling and Johnson clearly perform worse  
<sup>1434</sup> relative to the unnormalized model in the upper right part of the space, more specifically

1435 for the [u] category, which could indicate that models are overly categorical in a part of the  
 1436 space where humans are less categorical (*Panels B-C*; left). Possible reasons could be 1)  
 1437 the stimuli sounding more like a neighbouring category to many listeners, or 2) potential  
 1438 effects of orthography, making humans less inclined to select the [u] category. The potential  
 1439 effect of the infrequent non-word response option *who'd* could have been evaluated against  
 1440 the synthesized stimuli in Experiment 1b. If there was indeed an effect of orthography, we  
 1441 should have observed a better model fit and larger between-account differences in predictions  
 1442 in this part of the acoustic space. Unfortunately, we under-sampled that part, which is  
 1443 an important caveat for Experiment 1b. For the items closest to the area in question,  
 1444 participants, however, often responded *hood*, which might indicate that items in this part  
 1445 of the space for this talker overall sounded more like *hood* and not *who'd* for many listeners  
 1446 (c.f., discussion on listeners' dialect templates in Section [II B](#)).

1447 Comparing the two best-performing models across experiments (*Panel D*), no evident  
 1448 pattern of improvement in one model relative to the other seems to emerge. In Experiment  
 1449 1a, Johnson provided the best fit to listeners' responses and appears to improve the fit  
 1450 relative to Nearey's uniform scaling across almost the entire space. For Experiment 1b,  
 1451 Nearey's uniform scaling overall improves the fit relative to Johnson, with the exception of  
 1452 some locations in the mid part of the phonetic space (including high, center and low vowels).

1453 **4. Comparing accounts in terms of accuracy in predicting listeners' responses**

1454 An alternative approach to maximum likelihood fitting is to compare the accuracies of  
 1455 accounts in predicting listeners' responses. Tables [S5](#) and [S6](#) report the mean accuracies of  
 1456 the accounts averaged across the five folds for each experiment. The account with the highest  
 1457 accuracy in each experiment is boldfaced. For Experiment 1a, an intrinsic account, Syrdal  
 1458 & Gopal, achieved the highest accuracy in predicting listeners' responses, while Nearey's  
 1459 uniform scaling achieved the highest accuracy for Experiment 1b (see Section [III A 5](#) in the  
 1460 main text for a discussion on the use of accuracy as evaluation method).

1461 **C. Results for F1-F2 (subset of Experiments 1a and 1b)**

1462 To address two potential concerns with our stimuli, we decided to compare the 20 normal-  
 1463 ization accounts against a subset of the data from Experiment 1a and 1b. For Experiment

TABLE S5. Accuracy predicting listeners' responses in Experiment 1a (under Luce choice rule)

Normalization account	accuracy
no normalization (Hz)	mean=52.3% (SD=2.1)
transformed (log)	mean=49.3% (SD=4.2)
transformed (Mel)	mean=51% (SD=2.1)
transformed (Bark)	mean=50.6% (SD=2.3)
transformed (ERB)	mean=49.7% (SD=2.5)
transformed (semitones)	mean=49.3% (SD=4.2)
<b>SyrdalGopal (Bark)</b>	<b>mean=65.1% (SD=0.5)</b>
SyrdalGopal2 (Bark)	mean=52.9% (SD=1.4)
Miller (log)	mean=60.2% (SD=1.4)
Uniform scaling, Nearey (log)	mean=59.7% (SD=3.4)
Uniform scaling, Nordström & Lindblom (Hz)	mean=58.9% (SD=2.2)
Uniform scaling, Johnson (Hz)	mean=61.8% (SD=1.9)
Nearey's formantwise mean (log)	mean=61.5% (SD=1.1)
C-CuRE (Hz)	mean=60.8% (SD=2.1)
C-CuRE (Mel)	mean=61.5% (SD=1.2)
C-CuRE (Bark)	mean=61.3% (SD=1)
C-CuRE (ERB)	mean=61.5% (SD=1.1)
C-CuRE (semitones)	mean=61.5% (SD=1.1)
Gerstman (Hz)	mean=43.6% (SD=1.9)
Lobanov (Hz)	mean=58.5% (SD=1.9)

<sup>1464</sup> 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 <sup>1465</sup> stimuli that could be considered physiologically implausible under the assumption of a single <sup>1466</sup> talker (all stimuli below the diagonal dashed line in Figure 4). <sup>1467</sup>

TABLE S6. Accuracy predicting listeners' responses in Experiment 1b (under Luce choice rule)

Normalization account	accuracy
no normalization (Hz)	mean=16.9% (SD=1.2)
transformed (log)	mean=26.5% (SD=0.8)
transformed (Mel)	mean=23% (SD=0.9)
transformed (Bark)	mean=24.8% (SD=0.9)
transformed (ERB)	mean=25.4% (SD=0.9)
transformed (semitones)	mean=26.5% (SD=0.8)
SyrdalGopal (Bark)	mean=21.6% (SD=0.3)
SyrdalGopal2 (Bark)	mean=24.8% (SD=0.5)
Miller (log)	mean=24.7% (SD=1.3)
<b>Uniform scaling, Nearey (log)</b>	<b>mean=29.2% (SD=0.3)</b>
Uniform scaling, Nordström & Lindblom (Hz)	mean=25% (SD=2.6)
Uniform scaling, Johnson (Hz)	mean=23.7% (SD=3.8)
Nearey's formantwise mean (log)	mean=27.7% (SD=0.6)
C-CuRE (Hz)	mean=20.7% (SD=2.6)
C-CuRE (Mel)	mean=26% (SD=1)
C-CuRE (Bark)	mean=27.3% (SD=0.9)
C-CuRE (ERB)	mean=27.4% (SD=0.9)
C-CuRE (semitones)	mean=27.7% (SD=0.6)
Gerstman (Hz)	mean=27.8% (SD=1.4)
Lobanov (Hz)	mean=27.7% (SD=1.1)

<sup>1468</sup> This subset analysis overall replicates the results from the main analysis: uniform scaling  
<sup>1469</sup> accounts again provide the best fit against listeners' responses in both experiments (Figure  
<sup>1470</sup> S14). For Experiment 1a, Nordström & Lindblom provides the best fit (per-token log-  
<sup>1471</sup> likelihood = -0.85, SD = 0.04), whereas Nearey's uniform scaling provides the best fit to

1472 listeners' responses in Experiment 1b; for Experiment 1b, per-token log-likelihood = -1.51,  
 1473 SD = 0.01). While the relative ordering of accounts is similar compared to the main analysis,  
 1474 more accounts perform within the range of the best-fits in the subset analysis.

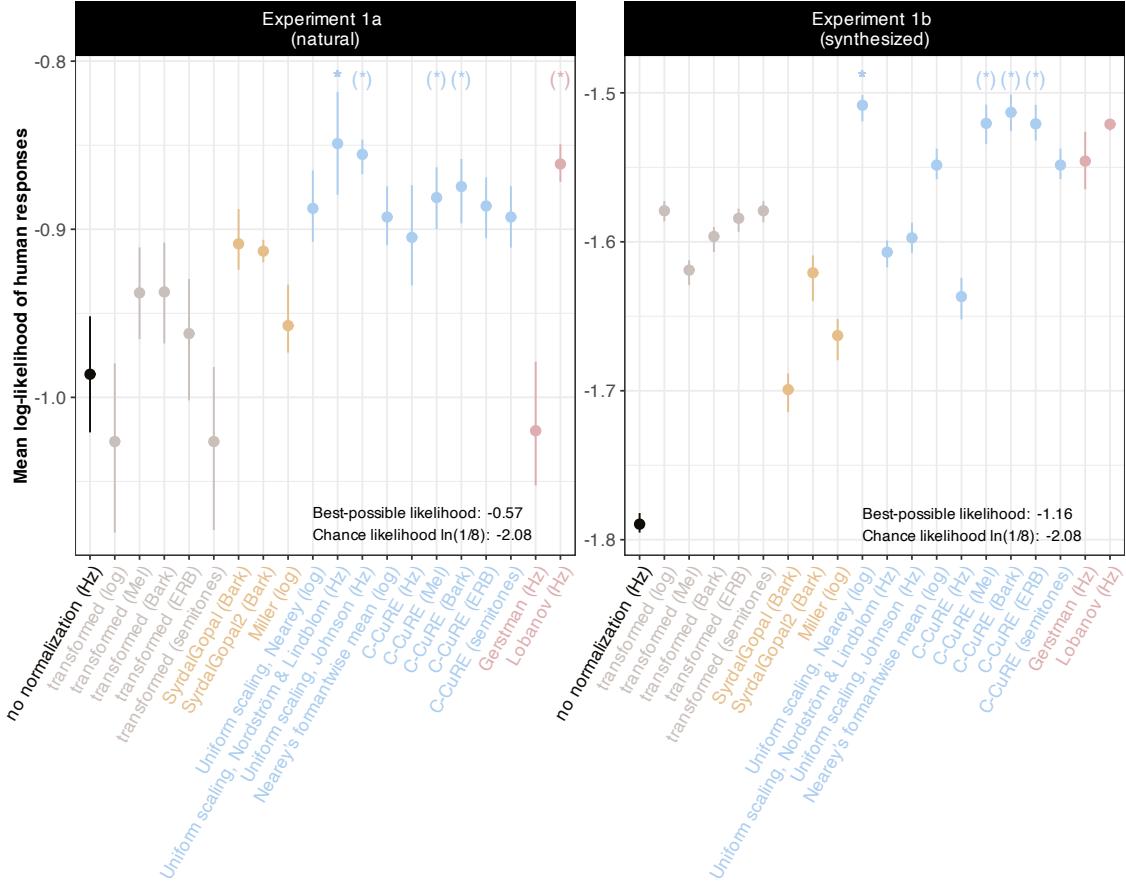


FIG. S14. Results of model fit to subset data. Point ranges indicate mean and 95% bootstrapped CIs of the per-token log-likelihood summarized over the five training sets (higher is better). Accounts that fit listeners' responses to an extent that is statistically indistinguishable from the best-fitting account are marked by (\*).

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

1475 Analyses in the main paper suggested that not all listeners in Experiments 1a and 1b  
 1476 shared dialect template (Section II B). To investigate the effect of excluding listeners that  
 1477 likely did not use the same underlying vowel representations for categorization, we compared  
 1478 the 20 normalization accounts against a subset of listeners who employed the dialect template  
 1479

1480 used by the majority of participants (see lower-left of both panels in Figure 5B). This left  
 1481 20 participants for Experiment 1a (71.4%) and 23 for Experiment 1b (82.1%). Under the  
 1482 assumptions that 1) our model of listeners is adequate, 2) the subset group of listeners now  
 1483 share dialect template, and that 3) the phonetic database can approximate this template, we  
 1484 would expect all models to increase their likelihood fit to listeners' responses (c.f., Section  
 1485 IV).

1486 Replicating the results from the main analysis, Figure S15 indicates that uniform scaling  
 1487 accounts again fit listeners' behavior well across both experiments. Nearey's uniform scaling  
 1488 again provides the best-fit in Experiment 1b (per-token log-likelihood = -1.39, SD = 0.01).  
 1489 However, an intrinsic account, Syrdal & Gopal, now achieves the best fit to Experiment  
 1490 1a (per-token log-likelihood = -0.44, SD = 0.01). While Nearey's uniform scaling displays  
 1491 relatively stable performance across experiments, Syrdal & Gopal does not, achieving one  
 1492 of the worst fits to listeners' responses in Experiment 1b (per-token log-likelihood = -1.65,  
 1493 SD = 0.02). As mentioned in Section III B 2, a potential explanation to large fluctuations in  
 1494 model fits between experiments, is the possibility of over-engineered normalization accounts.  
 1495 Given that formant normalization is a pre-linguistic mechanism, it ought to be able to explain  
 1496 listeners' responses to any type of data, including data that does not follow correlations in  
 1497 natural data.

1498 Finally, as expected, all models overall provide higher likelihood fits against human re-  
 1499 sponds in both experiments compared to the main model. When scaling the log likelihood  
 1500 of models in the subset data to those of the main analysis, the results suggest that the over-  
 1501 all improvement in likelihood across accounts for the dialect subset model to the original  
 1502 dataset was 63.7%.

1503 **E. Results for F1-F3**

1504 The models evaluated in the main text were trained and evaluated on bivariate (F1-F2)  
 1505 categories. Here, we investigate whether the inclusion of F3, a cue known to be important for  
 1506 vowel category distinctions, would improve the model fit to human behavior. We therefore  
 1507 trained ideal observers on multivariate (F1-F2-F3) categories from the same database as in  
 1508 the main study. We first report the results of the F1-F3 model and qualitatively compare  
 1509 them to the results in the main text for F1-F2. This will highlight that the results are  
 1510 largely similar and support the same conceptual conclusions, but there are some differences

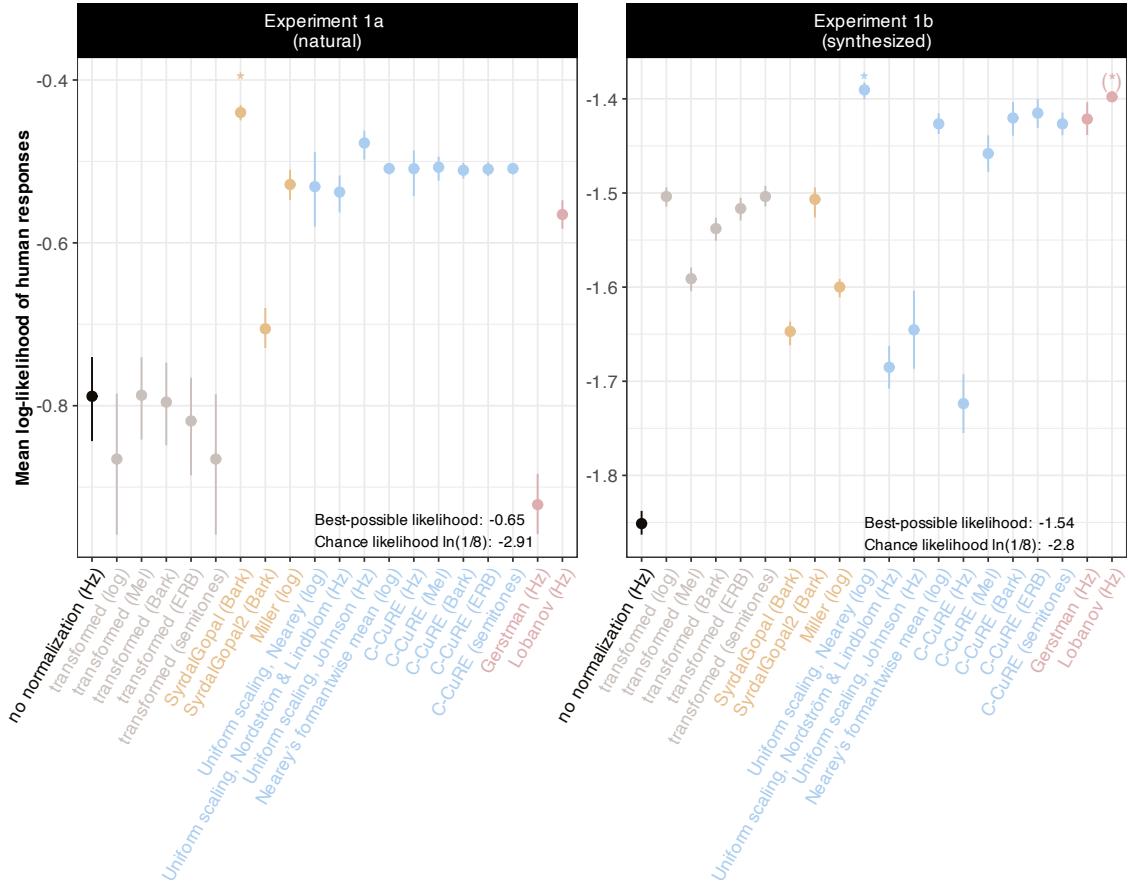


FIG. S15. Results of model fit to data excluding listeners that do not seem to share dialect template. Point ranges indicate mean and 95% bootstrapped CIs of the per-token log-likelihood summarized over the five training sets (higher is better). Accounts that fit listeners' responses to an extent that is statistically indistinguishable from the best-fitting account are marked by (\*).

1511 in model fit. To understand these differences better, we then also directly compare the  
 1512 results quantitatively to see for which accounts the inclusion of F3 improved the fit against  
 1513 listeners' responses and for which accounts there were no improvements.

1514 Figure S16 summarizes how well the different accounts fit listeners' responses in Experiments  
 1515 1a and 1b when assuming F1-F2-F3 multivariate category representations. Many  
 1516 aspects replicate the F1-F2 results reported in the main text. First, normalization signif-  
 1517 icantly improves the fit relative to no normalization. Second, the same uniform scaling  
 1518 accounts again achieve the best fit against listeners' responses: for Experiment 1a, Johnson  
 1519 normalization account provides the best fit (per-token log-likelihood = -0.83, SD = 0.01),  
 1520 while Nearey's uniform scaling account provides the best fit to Experiment 1b (per-token

1521 log-likelihood = -1.48, SD = 0.01). However, we also note that the inclusion of F3 does  
 1522 not improve the fit to listeners' responses for several accounts (compare *squares* and *circles*  
 1523 in Figure S16). In fact, most extrinsic accounts seem to decrease their fit, more so in Ex-  
 1524 periment 1a than in Experiment 1b. At first blush, this is puzzling given that the model  
 1525 now has access to more information of a type that is broadly believed to be informative for  
 1526 US English vowel recognition (Hillenbrand *et al.*, 1995; Nearey, 1989; Peterson and Barney,  
 1527 1952). What might be underlying the lack of improvement, and why does it appear as if  
 1528 some accounts actually achieve worse fits?

1529 One possible explanation is that listeners were only exposed to one talker in Experiment  
 1530 1a. According to some theories, F3 is expected to contribute to vowel recognition when there  
 1531 are multiple talkers, acting as a sort of normalizer for vocal tract length (Nearey, 1989). In  
 1532 the absence of other talkers, this advantage might instead introduce noise to the models—  
 1533 an additional source of information that is not useful for listeners in this context. It is also  
 1534 possible that the F3-distribution across categories for this particular talker is atypical given  
 1535 the other talkers in the database. This might explain why the raw Hz model improves the fit  
 1536 with F3-inclusion. We checked for additional outliers along F3 for this talker, but we could  
 1537 not find that outliers would be a likely explanation. To gain further knowledge into this  
 1538 talker's use of F3 compared to other talkers in the database, we used the same models to  
 1539 predict the ground truth, i.e., the category the talker actually intended to produce. These  
 1540 models patterned with the other prediction results, again indicating that F3-inclusion did  
 1541 not improve model performance. We take this to suggest that the F1-F3 results is not about  
 1542 how our model uses F3, but rather about how this specific talker uses F3 (c.f., the potential  
 1543 dialect differences between talkers in the database, reported in Section II B, as well as the  
 1544 talker's formant distributions in 3D-space, Figure S5).

## 1545 F. Grid search over parameter space for F1-F2 and F1-F3

1546 As an alternative to the quasi-Newton optimization presented in the main text, we also  
 1547 conducted a grid search over the space defined by the two parameters lapse rate and noise  
 1548 ratio. Figure S17 summarizes the results for a grid of lapse rates  $\in 0, .02, .06, .18, .36, .72$   
 1549 and noise ratios  $\in 0, .3, .6, 1.25, 2.5, 5$  for Experiment 1a and models trained on F1-F2. For  
 1550 Experiment 1b and F1-F2 (Figure S18), the range of noise ratios explored was  $\in 0, 1.5, 3,$   
 1551 6, 12.5, 25. Figures S19 and S20 visualize results for models trained on F1-F3.

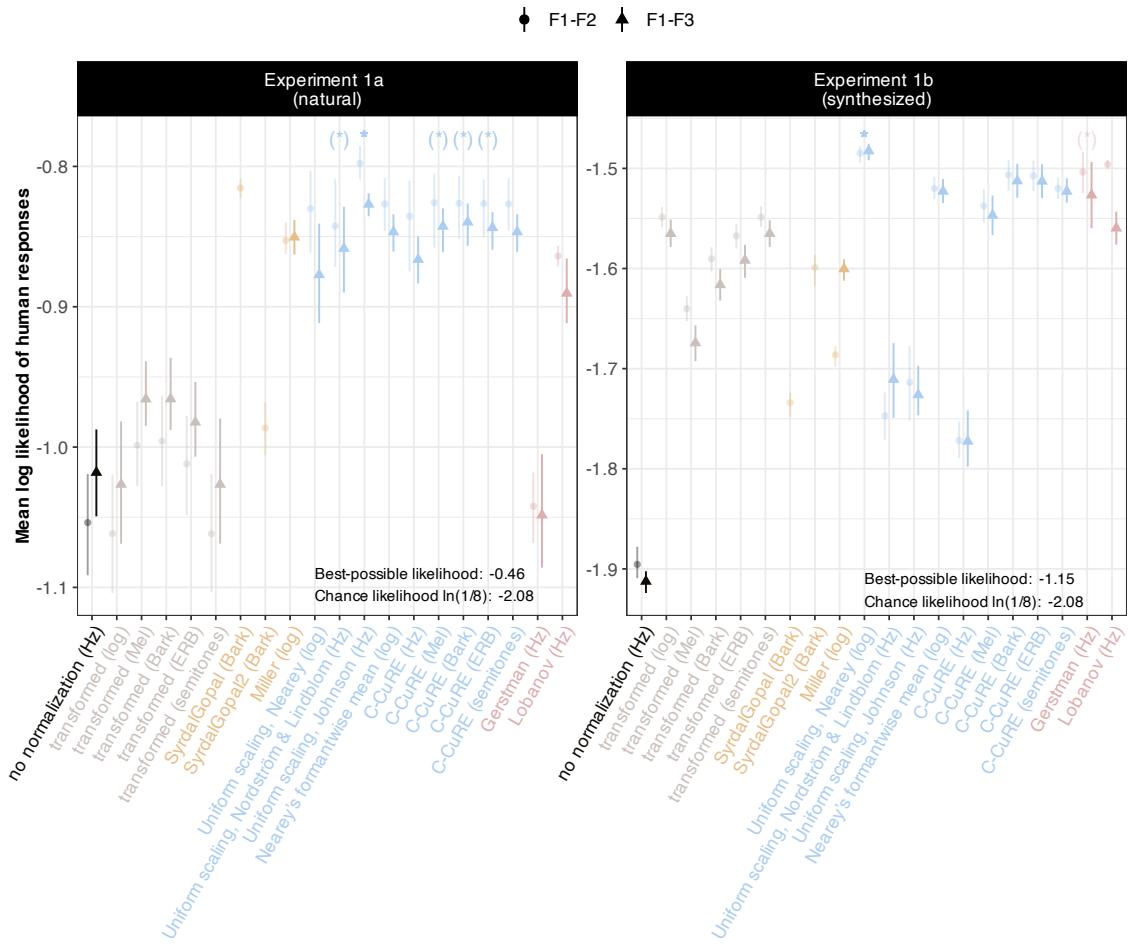


FIG. S16. Results of ideal observer models trained on F1, F2 and F3 as cues to vowel identity. As in Figure 9 in the main text, pointranges indicate mean and 95% bootstrapped CIs of the log-likelihood summarized over the five training sets, and normalized by the number of responses in each experiment (higher is better). For comparison, results from the F1-F2 models are included (more transparent circles). Note that the Syrdal & Gopal accounts are not included in the F1-F3 evaluation as they do not provide normalization for F3.

1552 The grid searches confirmed the pattern described in the main text, as did additional grid  
 1553 searches beyond the values shown here. For all normalization accounts, all combinations of  
 1554 cues, and both experiments, the goodness of fit of the ideal observers initially improved  
 1555 with increasing  $\lambda$  and increasing  $\tau^{-1}$ s, and then decreased once  $\lambda$ s or  $\tau^{-1}$ s reached the  
 1556 best-fitting values (which depended on the combination of normalization account, cues, and  
 1557 experiment). The grid search further indicated that Nearey's uniform scaling, together with  
 1558 the other uniform scaling accounts and some of the C-CuRE accounts (Experiment 1a) and

## Comparing normalization against perception

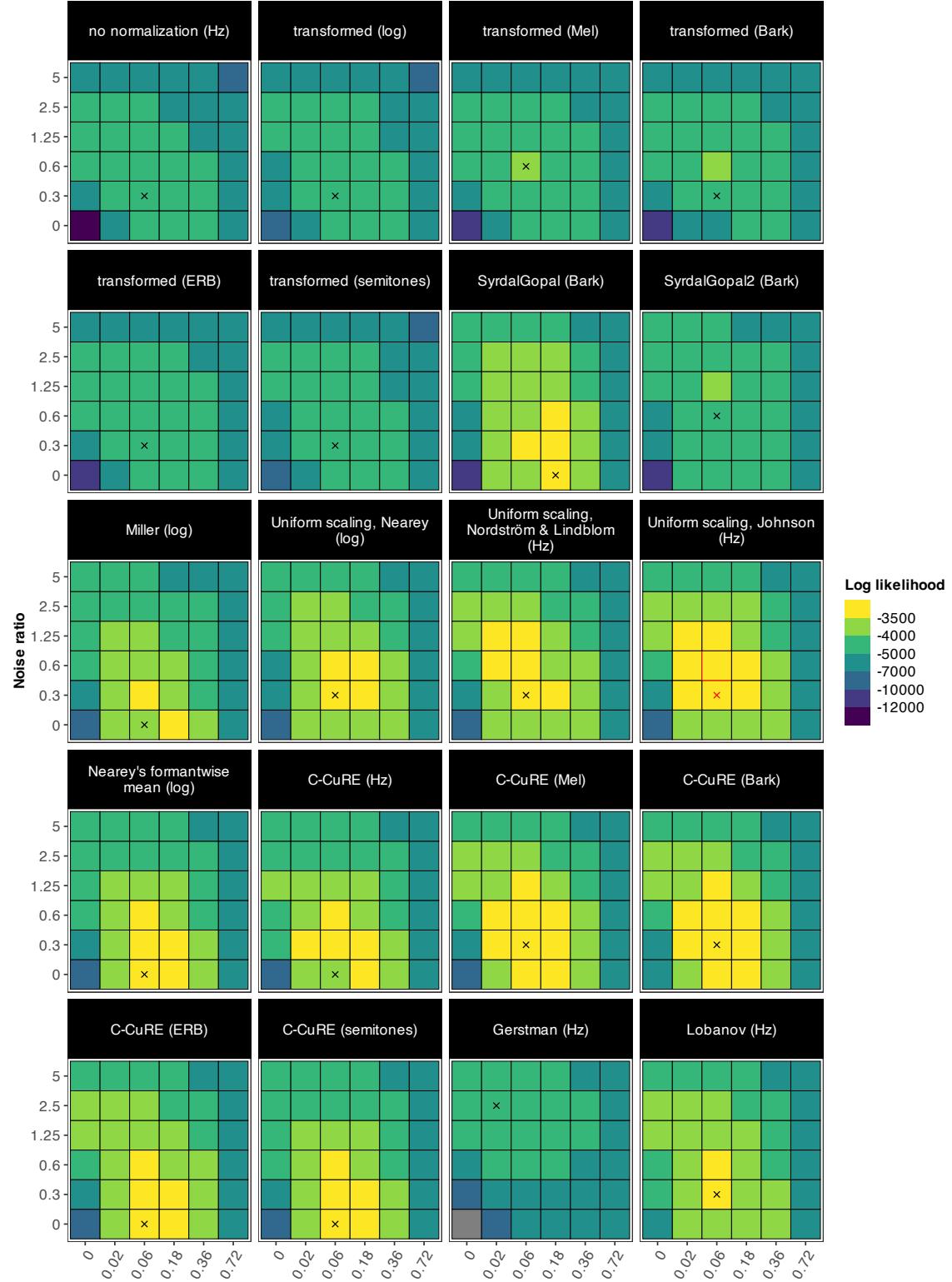


FIG. S17. Predicted likelihoods of ideal observers trained on F1-F2 for human vowel responses in Experiment 1a, under different normalization accounts, different  $\lambda$ s and different  $\tau^{-1}$ s. Likelihood is aggregated across vowels. Crosses are placed at the combination of parameters for which the maximum likelihood for an account was found. The red cross indicates the maximum likelihood achieved for a single training set and account across the entire grid search.

## Comparing normalization against perception

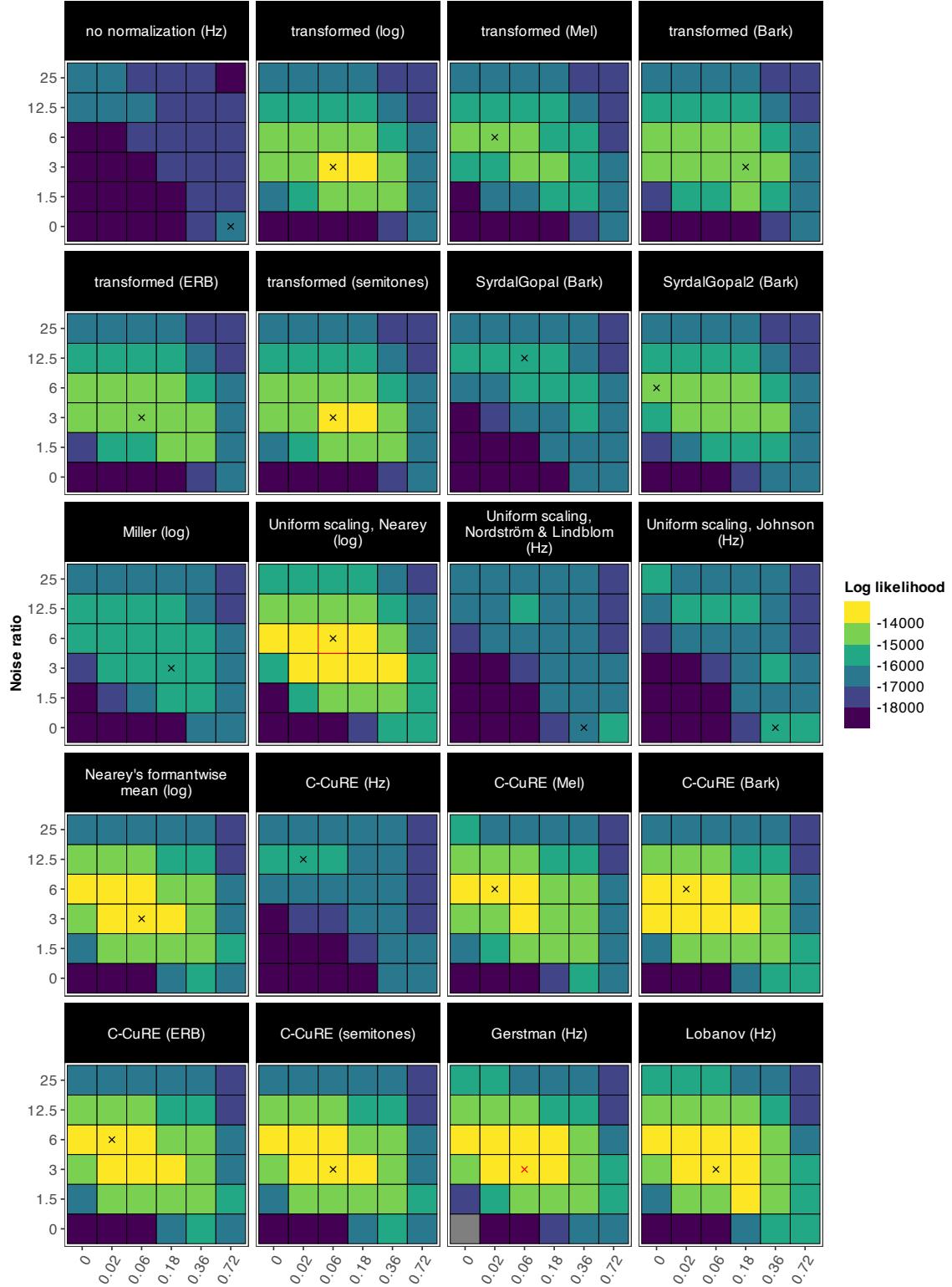


FIG. S18. Predicted likelihoods of ideal observers trained on F1-F2 for human vowel responses in Experiment 1b, under different normalization accounts, different  $\lambda$ s and different  $\tau^{-1}$ s. Likelihood is aggregated across vowels. Crosses are placed at the combination of parameters for which the maximum likelihood for an account was found. The red cross indicates the maximum likelihood achieved for a single training set and account across the entire grid search. <sup>37</sup>

## Comparing normalization against perception

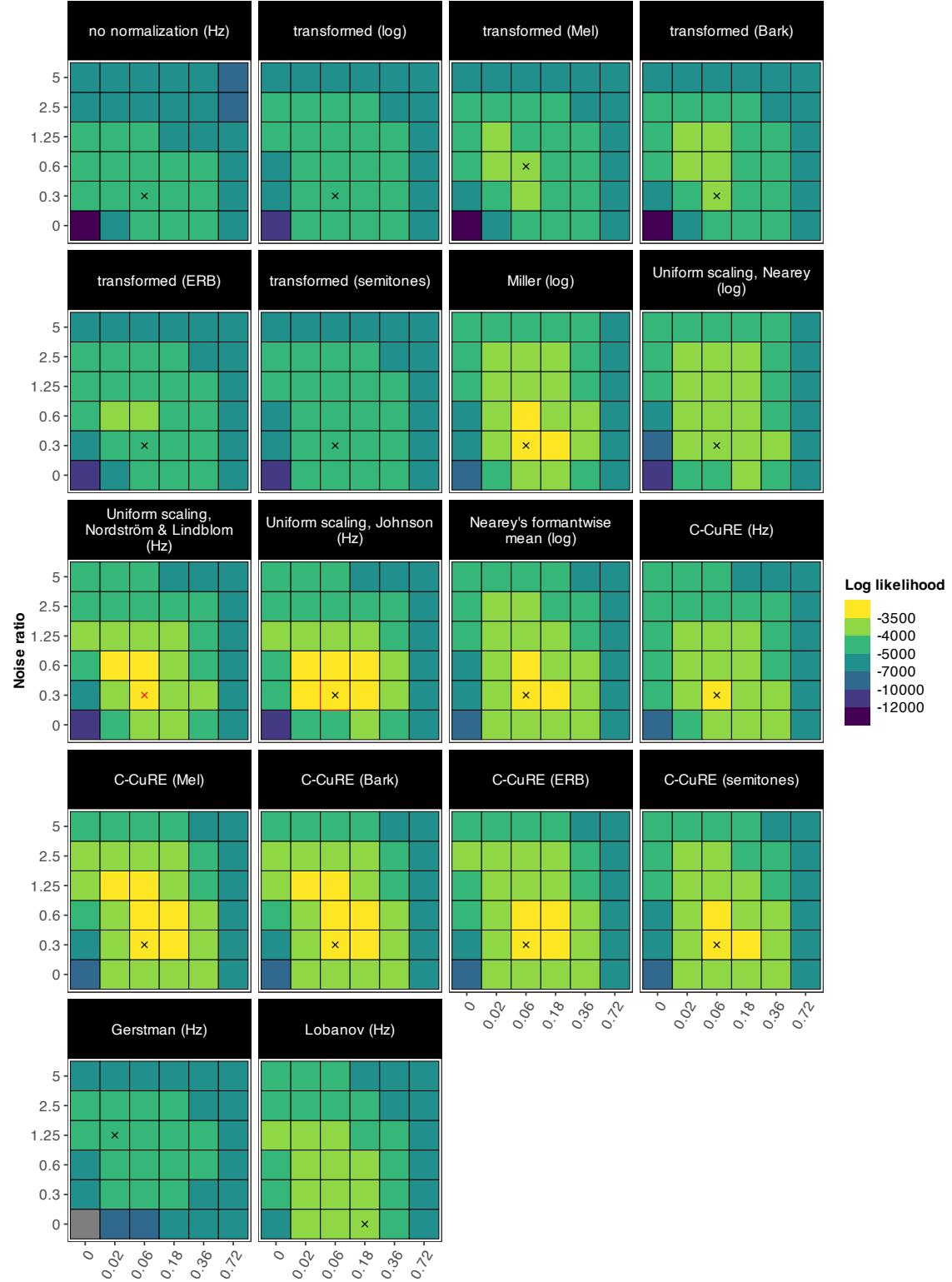


FIG. S19. Predicted likelihoods of ideal observers trained on F1-F3 for human vowel responses in Experiment 1a, under different normalization accounts, different  $\lambda$ s and different  $\tau^{-1}$ s. Likelihood is aggregated across vowels. Crosses are placed at the combination of parameters for which the maximum likelihood for an account was found. The red cross indicates the maximum likelihood achieved for a single training set and account across the entire grid search.

## Comparing normalization against perception

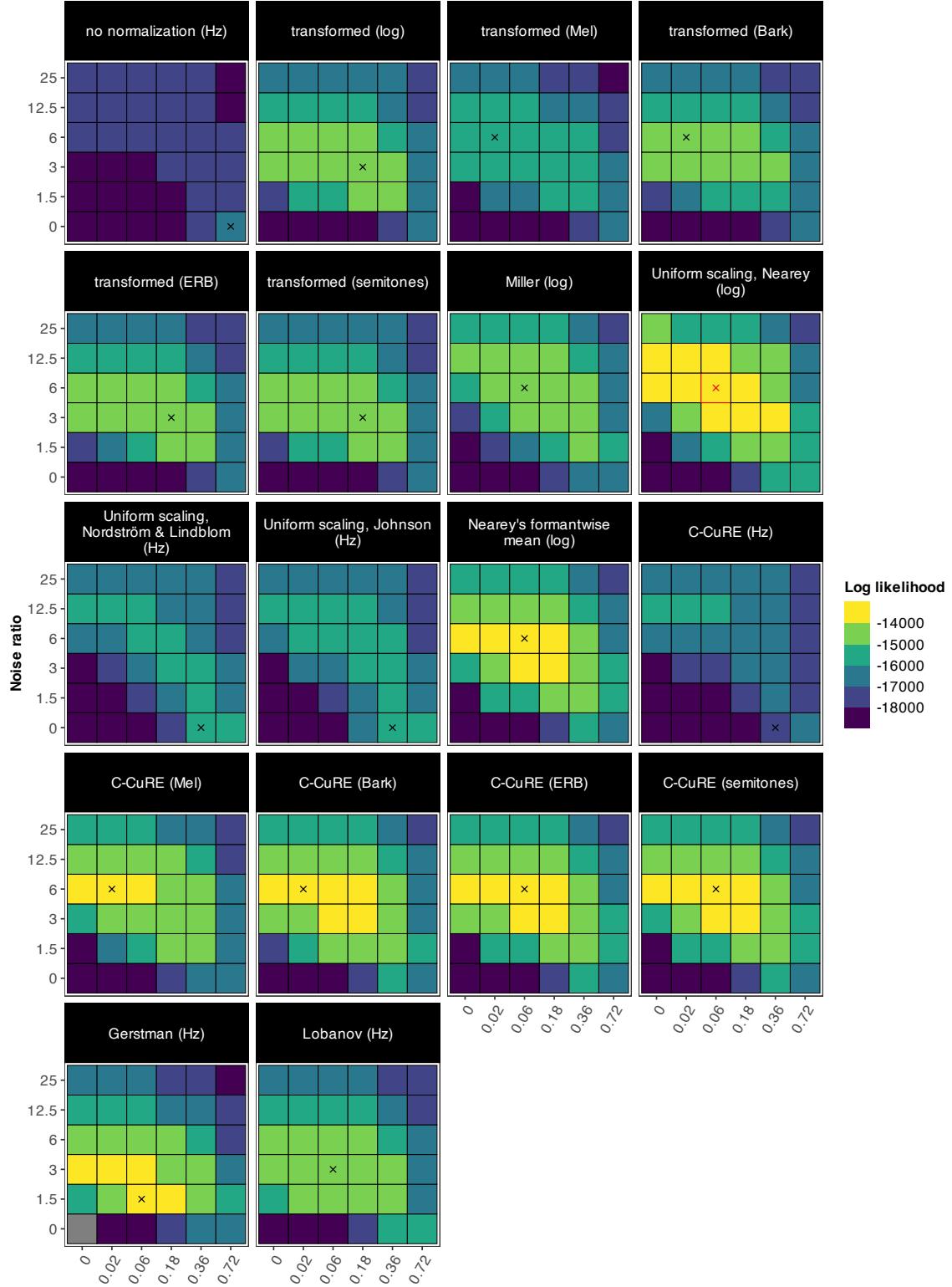


FIG. S20. Predicted likelihoods of ideal observers trained on F1-F3 for human vowel responses in Experiment 1b, under different normalization accounts, different  $\lambda$ s and different  $\tau^{-1}$ s. Likelihood is aggregated across vowels. Crosses are placed at the combination of parameters for which the maximum likelihood for an account was found. The red cross indicates the maximum likelihood achieved for a single training set and account across the entire grid search.

1559 the standardizing accounts (Experiment 1b), improved faster and performed consistently  
1560 well for a good range of parameters, even for high  $\tau^{-1}$ . Many of the other models were less  
1561 consistent and only performed well for a smaller range of estimates.

1562 The grid searches for models trained on F1-F3 largely replicate the results for models  
1563 trained on F1-F2 (Figures S19 and S20). However, for Experiment 1a, several of the C-  
1564 CuRE accounts achieve their maximum likelihood values with higher  $\tau^{-1}$ s. For Experiment  
1565 1b, several accounts reach their maximum likelihoods for a smaller range of values for both  
1566  $\tau^{-1}$ s and  $\lambda$ .

1567 **§4. REFERENCES**

1568 <sup>1</sup>Some hypotheses hold that robust speech perception does not require normalization, and that research  
 1569 on normalization has over-estimated its effectiveness because studies tend to consider only a fraction of  
 1570 the phonetic information available to listeners (for review, see [Strange and Jenkins, 2012](#)). For vowel  
 1571 recognition, for example, listeners might use cues other than just formants ([Hillenbrand \*et al.\*, 2006](#); [Nearey  
 1572 and Assmann, 1986](#)), and/or might use information about the dynamic development of formant trajectories  
 1573 over the entire vowel rather than just point estimates of formants at the vowel center (e.g., [Shankweiler  
 1574 \*et al.\*, 1978](#)). We return to this in the general discussion but note that even studies who use much richer  
 1575 inputs have found that normalization provides a better fit to listeners' perception ([Richter \*et al.\*, 2017](#)).

1576 <sup>2</sup>Under uniform scaling accounts, listeners essentially 'slide' the center of their category representations  
 1577 (e.g, the 'template' of vowel categories for a given dialect) along a single line in formant space, with  $\Psi$   
 1578 determining the target of this sliding. Later extensions of this account maintain its memory parsimony but  
 1579 increased its inference complexity by allowing both intrinsic (the current F0) and extrinsic information (the  
 1580 talker's single mean of log-transformed formants) to influence the inference of  $\Psi$  ([Nearey and Assmann,  
 1581 2007](#)).

1582 <sup>3</sup>We use Johnson's (2020) implementation of [Nordström and Lindblom \(1975\)](#). We group both [Nordström  
 1583 and Lindblom \(1975\)](#) and [Johnson \(2020\)](#) with the centering accounts, as they are essentially variants of  
 1584 uniform scaling, differing in their estimation of  $\Psi$ . We also include both versions of Syrdal & Gopal's  
 1585 Bark-distance model. The two versions differ only in their normalization of F2, and have not previously  
 1586 been compared against human perception.

1587 <sup>4</sup>[Shannon \(1948\)](#) response entropy is defined as  $H(x) = -\sum_{i=1}^n P(x_i) \log P(x_i)$ . The maximum possible  
 1588 response entropy for an eight-way response choice is 3 bits, which means that all eight vowels are responded  
 1589 equally often. The minimum response entropy = 0 bits, which means that the same vowel is responded all  
 1590 the time.

1591 <sup>5</sup>Note that participants in Experiment 1a exhibited high agreement on [ʌ], [æ], and [ɑ], despite the close  
 1592 proximity between, and partial overlap of, these vowels in F1-F2 space. To understand this pattern, it is  
 1593 important to keep in mind that the recordings for [ʌ] and [ɑ] differed from the recordings for other stimuli  
 1594 in their word onset ("odd" for [ɑ]) or offset ("hut" for [ʌ]).

1595 <sup>6</sup>[u] has been undergoing changes in many varieties of US English. Whereas the talker in Experiment 1a  
 1596 produces [u] with low F1 and F2 (high and back), other L1 talkers of US English produce this vowel  
 1597 considerably more forward (higher F2).

1598 <sup>7</sup>For Gaussian noise and Gaussian category likelihoods, the resulting noise-convolved likelihood is a Gaussian  
 1599 with variance equal to the sum of the noise and category variances ([Kronrod \*et al.\*, 2016](#)).

1600 <sup>8</sup>We intentionally did *not* split the data within talkers since normalization accounts are meant to make  
 1601 speech perception robust to cross-talker variability. Further, splitting the data by speaker rather than  
 1602 by vowel category avoids the potential for biases in the normalization parameter estimates for different  
 1603 speakers in the case of missing or unbalanced tokens across vowel categories, see (Barreda and Nearey,  
 1604 2018). Additional analyses not reported here confirmed that the same results are obtained when splits are  
 1605 performed within talkers and within vowels (except that this lead to smaller CIs, and thus *more* significant  
 1606 differences, in Figure 9). These analyses can be replicated by downloading the R markdown document this  
 1607 article is based on from our OSF (see comments in our code).

1608 <sup>9</sup>Alternatively, it would be possible to treat these parameters as DFs in the link to listeners' responses,  
 1609 and infer them from the responses in Experiments 1a and 1b (cf., Kleinschmidt and Jaeger, 2016). This  
 1610 approach would afford the model with a high degree of functional flexibility, regardless of which normal-  
 1611 ization approach is applied (similar to previous approaches that have employed, e.g., multinomial logistic  
 1612 regression).

1613 <sup>10</sup>This ratio is a generalization of the inverse of the “meaningful-to-noise variance ratio ( $\tau$ )” used in Kronrod  
 1614 *et al.* (2016). However, whereas Kronrod and colleagues committed to the simplifying assumption that  
 1615 all categories have identical variance (along all formants), we allowed category variances to differ between  
 1616 vowels, and between F1 and F2 (matching the empirically facts). We merely assume that the *noise* variance  
 1617 is identical across all formants (in the phonetic space defined by the normalization account, e.g., log-Hz for  
 1618 uniform scaling and Hz for Lobanov).

1619 <sup>11</sup>Additional analyses reported in the SI (§3 C) overall replicated this result for subsets of Experiments 1a  
 1620 and 1b, with Nearey's uniform scaling achieving the best fit to listeners' responses in both experiments.  
 1621 For Experiment 1a, we excluded responses to the two *hVd* stimuli that differed from the other stimuli in  
 1622 the preceding (*odd*) or following phonological context (*hut*). For Experiment 1b, we excluded responses  
 1623 to any stimuli that were physiologically implausible for the talker (stimuli below the diagonal dashed line  
 1624 in Figure 4). As requested by a reviewer, the SI §3 B 4 also reports the accuracy of predicting listeners'  
 1625 responses for all normalization accounts. The best performing accounts achieved 61.8% for Experiment 1a  
 1626 (Johnson normalization), and 29.2% for Experiment 1b (Nearey's uniform scaling), compared to 52.3% and  
 1627 16.9%, respectively, without normalization.

1628 <sup>12</sup>In line with this reasoning, additional tests found that Johnson normalization would provide the best fit to  
 1629 Experiment 1b if it was applied to log-transformed formants (instead of Hertz).

1630

1631 Abramson, A. S., and Lisker, L. (1973). “Voice-timing perception in Spanish word-initial  
 1632 stops,” Journal of Phonetics 1(1), 01–08, doi: [10.1016/S0095-4470\(19\)31372-5](https://doi.org/10.1016/S0095-4470(19)31372-5).  
 1633 Adank, P., Smits, R., and van Hout, R. (2004). “A comparison of vowel normalization pro-  
 1634 cedures for language variation research,” The Journal of the Acoustical Society of America

- 1635      **116**(5), 3099–3107, doi: [10.1121/1.1795335](https://doi.org/10.1121/1.1795335).
- 1636      Allen, J. S., Miller, J. L., and DeSteno, D. (2003). “Individual talker differences in voice-  
1637      onset-time,” *Journal of the Acoustical Society of America* **113**(1), 544–552, doi: [10.1121/1.1528172](https://doi.org/10.1121/1.1528172).
- 1639      Apfelbaum, K., and McMurray, B. (2015). “Relative cue encoding in the context of sophisti-  
1640      cated models of categorization: Separating information from categorization,” *Psychonomic  
1641      Bulletin and Review* **22**(4), 916–943, doi: [10.3758/s13423-014-0783-2](https://doi.org/10.3758/s13423-014-0783-2).
- 1642      Assmann, P. F., and Katz, W. F. (2005). “Synthesis fidelity and time-varying spectral  
1643      change in vowels,” *The Journal of the Acoustical Society of America* **117**(2), 886–895, doi:  
1644      [10.1121/1.1852549](https://doi.org/10.1121/1.1852549).
- 1645      Assmann, P. F., Nearey, T. M., and Bharadwaj, S. (2008). “Analysis of a vowel database,”  
1646      *Canadian Acoustics* **36**(3), 148–149.
- 1647      Bache, S. M., and Wickham, H. (2022). *magrittr: A Forward-Pipe Operator for R*, <https://CRAN.R-project.org/package=magrittr>, r package version 2.0.3.
- 1649      Baese-Berk, M. M., Walker, K., and Bradlow, A. (2018). “Variability in speaking rate of  
1650      native and non-native speakers,” *The Journal of the Acoustical Society of America* **144**(3),  
1651      1717–1717, doi: [10.1121/1.5067612](https://doi.org/10.1121/1.5067612).
- 1652      Balzano, G. J. (1982). “The pitch set as a level of description for studying musical pitch per-  
1653      ception,” in *Music, mind, and brain: The neuropsychology of music*, edited by M. Clynes  
1654      (Springer), pp. 321–351.
- 1655      Barreda, S. (2020). “Vowel normalization as perceptual constancy,” *Language* **96**(2), 224–  
1656      254, doi: [10.1353/lan.2020.0018](https://doi.org/10.1353/lan.2020.0018).
- 1657      Barreda, S. (2021). “Perceptual validation of vowel normalization methods for vari-  
1658      ationist research,” *Language Variation and Change* **33**(1), 27–53, doi: [10.1017/S0954394521000016](https://doi.org/10.1017/S0954394521000016).
- 1660      Barreda, S. (2023). “phontools: Functions for phonetics in r” R package version 0.2-2.2.
- 1661      Barreda, S., and Jaeger, T. F. (submitted). “Re-introducing the probabilistic sliding tem-  
1662      plate model of vowel perception,” *Linguistic Vanguard* .
- 1663      Barreda, S., and Nearey, T. M. (2012). “The direct and indirect roles of fundamental  
1664      frequency in vowel perception,” *The Journal of the Acoustical Society of America* **131**(1),  
1665      466–477, doi: [10.1121/1.3662068](https://doi.org/10.1121/1.3662068).
- 1666      Barreda, S., and Nearey, T. M. (2018). “A regression approach to vowel normalization for  
1667      missing and unbalanced data,” *The Journal of the Acoustical Society of America* **144**(1),

- 1668 500–520, doi: [10.1121/1.5047742](https://doi.org/10.1121/1.5047742).
- 1669 Bengtsson, H. (2021). “A unifying framework for parallel and distributed processing in r us-  
1670 ing futures,” The R Journal **13**(2), 208–227, <https://doi.org/10.32614/RJ-2021-048>,  
1671 doi: [10.32614/RJ-2021-048](https://doi.org/10.32614/RJ-2021-048).
- 1672 Bladon, A., Henton, C., and Pickering, J. (1984). “Towards an auditory theory of speaker  
1673 normalization,” Language and Communication **4**, 59–69.
- 1674 Boersma, P., and Weenink, D. (2022). “Praat: Doing phonetics by computer [Computer  
1675 program]” .
- 1676 Buz, E., and Jaeger, T. F. (2016). “The (in) dependence of articulation and lexical planning  
1677 during isolated word production,” Language, Cognition and Neuroscience **31**(3), 404–424.
- 1678 Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). “A limited memory algorithm for  
1679 bound constrained optimization,” SIAM Journal on Scientific Computing **16**(5), 1190–  
1680 1208, doi: [10.1137/0916069](https://doi.org/10.1137/0916069).
- 1681 Bürkner, P.-C. (2017). “brms: An R package for Bayesian multilevel models using Stan,”  
1682 Journal of Statistical Software **80**(1), 1–28, doi: [10.18637/jss.v080.i01](https://doi.org/10.18637/jss.v080.i01).
- 1683 Bürkner, P.-C. (2018). “Advanced Bayesian multilevel modeling with the R package brms,”  
1684 The R Journal **10**(1), 395–411, doi: [10.32614/RJ-2018-017](https://doi.org/10.32614/RJ-2018-017).
- 1685 Bürkner, P.-C. (2021). “Bayesian item response modeling in R with brms and Stan,” Journal  
1686 of Statistical Software **100**(5), 1–54, doi: [10.18637/jss.v100.i05](https://doi.org/10.18637/jss.v100.i05).
- 1687 Campitelli, E. (2024). *ggnewscale: Multiple Fill and Colour Scales in 'ggplot2'*, <https://CRAN.R-project.org/package=ggnewscale>, r package version 0.5.0.
- 1688 Carpenter, G. A., and Govindarajan, K. K. (1993). “Neural Network and Nearest Neighbor  
1689 Comparison of Speaker Normalization Methods for Vowel Recognition,” in *ICANN '93.  
1690 Proceedings of the International Conference on Artificial Neural Networks, Amsterdam,  
1691 the Netherlands, 13-16 September*, edited by S. Gielen and B. Kappen (Springer London,  
1692 London), pp. 412–415, doi: [10.1007/978-1-4471-2063-6\\_98](https://doi.org/10.1007/978-1-4471-2063-6_98).
- 1693 Chládková, K., Podlipský, V. J., and Chionidou, A. (2017). “Perceptual adaptation of  
1694 vowels generalizes across the phonology and does not require local context.,” Journal of  
1695 Experimental Psychology: Human Perception and Performance **43**(2), 414.
- 1696 Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (2008). “Perception of  
1697 speech reflects optimal use of probabilistic speech cues,” Cognition **108**(3), 804–809, doi:  
1698 [10.1016/j.cognition.2008.04.004](https://doi.org/10.1016/j.cognition.2008.04.004).

- 1700 Colby, S., Clayards, M., and Baum, S. (2018). “The role of lexical status and individual dif-  
 1701 ferences for perceptual learning in younger and older adults,” *Journal of Speech, Language,*  
 1702 and *Hearing Research* **61**(8), 1855–1874, doi: [10.1044/2018\\_JSLHR-S-17-0392](https://doi.org/10.1044/2018_JSLHR-S-17-0392).
- 1703 Cole, J., Linebaugh, G., Munson, C., and McMurray, B. (2010). “Unmasking the acous-  
 1704 tic effects of vowel-to-vowel coarticulation: A statistical modeling approach,” *Journal of*  
 1705 *Phonetics* **38**(2), 167–184, doi: [10.1016/j.wocn.2009.08.004](https://doi.org/10.1016/j.wocn.2009.08.004).
- 1706 Crinnion, A. M., Malmskog, B., and Toscano, J. C. (2020). “A graph-theoretic approach to  
 1707 identifying acoustic cues for speech sound categorization,” *Psychonomic Bulletin & Review*  
 1708 **27**(6), 1104–1125, doi: [10.3758/s13423-020-01748-1](https://doi.org/10.3758/s13423-020-01748-1).
- 1709 Csárdi, G., Hester, J., Wickham, H., Chang, W., Morgan, M., and Tenenbaum, D. (2024).  
 1710 *remotes: R Package Installation from Remote Repositories, Including 'GitHub'*, <https://CRAN.R-project.org/package=remotes>, r package version 2.5.0.
- 1711 Disner, S. F. (1980). “Evaluation of vowel normalization procedures,” *The Journal of the*  
 1712 *Acoustical Society of America* **67**(1), 253–261, doi: [10.1121/1.383734](https://doi.org/10.1121/1.383734).
- 1713 Eaves Jr, B. S., Feldman, N. H., Griffiths, T. L., and Shafto, P. (2016). “Infant-directed  
 1714 speech is consistent with teaching,” *Psychological Review* **123**(6), 758.
- 1715 Eddelbuettel, D., Francois, R., Allaire, J., Ushey, K., Kou, Q., Russell, N., Ucar, I., Bates,  
 1716 D., and Chambers, J. (2024). *Rcpp: Seamless R and C++ Integration*, <https://CRAN.R-project.org/package=Rcpp>, r package version 1.0.13-1.
- 1717 Escudero, P., and Bion, R. A. H. (2007). “Modeling vowel normalization and sound percep-  
 1718 tion as sequential processes,” *Proceedings of the 16th international congress of phonetic*  
 1719 *sciences*, Saarbrücken, Saarland University **XVI**, 1413–1416.
- 1720 Fant, G. (1975). “Non-uniform vowel normalization,” *STL-QPSR* **16**(2–3), 001–019.
- 1721 Fant, G., Kruckenberg, A., Gustafson, K., and Liljencrants, J. (2002). “A New Approach  
 1722 to Intonation Analysis and Synthesis of Swedish,” *Proceedings of Fonetik, TMH-QPSR*  
 1723 **44**(1), 161–164.
- 1724 Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). “The influence of categories  
 1725 on perception: Explaining the perceptual magnet effect as optimal statistical inference,”  
 1726 *Psychological Review* **116**(4), 752–782, doi: [10.1037/a0017196](https://doi.org/10.1037/a0017196).
- 1727 Flemming, E. (2010). “Modeling listeners: Comments on pluymakers et al. and scarbor-  
 1728 ough,” in *Laboratory Phonology*, edited by C. Fougeron, B. Kühnert, M. D’Imperio, and  
 1729 N. Vallée, **10** (De Gruyter Mouton), pp. 587–606.

- 1732 Fox, J., Venables, B., Damico, A., and Salverda, A. P. (2021). *english: Translate Integers*  
 1733 *into English*, <https://CRAN.R-project.org/package=english>, r package version 1.2-6.
- 1734 Gabry, J., Češnovar, R., and Johnson, A. (2024). *cmdstanr: R Interface to 'CmdStan'*,  
 1735 <https://mc-stan.org/cmdstanr/>, r package version 0.7.1, <https://discourse.mc-stan.org>.
- 1736 Gahl, S., Yao, Y., and Johnson, K. (2012). “Why reduce? Phonological neighborhood  
 1737 density and phonetic reduction in spontaneous speech,” *Journal of Memory and Language*  
 1738 **66**(4), 789–806, doi: [10.1016/j.jml.2011.11.006](https://doi.org/10.1016/j.jml.2011.11.006).
- 1739 Gerstman, L. (1968). “Classification of self-normalized vowels,” *IEEE Transactions on Au-*  
 1740 *dio and Electroacoustics* **16**(1), 78–80, doi: [10.1109/TAU.1968.1161953](https://doi.org/10.1109/TAU.1968.1161953).
- 1741 Glasberg, B. R., and Moore, B. C. J. (1990). “Derivation of auditory filter shapes from  
 1742 notched-noise data,” *Hearing Research* **47**(1), 103–138, doi: [10.1016/0378-5955\(90\)90170-T](https://doi.org/10.1016/0378-5955(90)90170-T).
- 1743 1744 Goldinger, S. D. (1996). “Words and voices: Episodic traces in spoken word identification  
 1745 and recognition memory.,” *Journal of Experimental Psychology: Learning Memory and*  
 1746 *Cognition* **22**(5), 1166–1183, doi: [10.1037/0278-7393.22.5.1166](https://doi.org/10.1037/0278-7393.22.5.1166).
- 1747 1748 Grolemund, G., and Wickham, H. (2011). “Dates and times made easy with lubridate,”  
 1749 *Journal of Statistical Software* **40**(3), 1–25, <https://www.jstatsoft.org/v40/i03/>.
- 1750 1751 Hall, K. C., Hume, E., Jaeger, T. F., and Wedel, A. (2018). “The role of predictability  
 1752 in shaping phonological patterns,” *Linguistics Vanguard* **4**(s2), 20170027, doi: [10.1515/lingvan-2017-0027](https://doi.org/10.1515/lingvan-2017-0027).
- 1753 1754 Hay, J., Podlubny, R., Drager, K., and McAuliffe, M. (2017). “Car-talk: Location-specific  
 1755 speech production and perception,” *Journal of Phonetics* **65**, 94–109, doi: [10.1016/j.wocn.2017.06.005](https://doi.org/10.1016/j.wocn.2017.06.005).
- 1756 1757 Hay, J., Walker, A., Sanchez, K., and Thompson, K. (2019). “Abstract social categories  
 1758 facilitate access to socially skewed words,” *PLoS ONE* **14**(2), 1–29, doi: [10.1371/journal.pone.0210793](https://doi.org/10.1371/journal.pone.0210793).
- 1759 1760 Henry, L., and Wickham, H. (2024). *rlang: Functions for Base Types and Core R and*  
 1761 *'Tidyverse' Features*, <https://CRAN.R-project.org/package=rlang>, r package version  
 1762 1.1.4.
- 1763 Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). “Acoustic charac-  
 1764 teristics of American English vowels,” *Journal of the Acoustical Society of America* **97**(5),  
 1765 3099–3111, doi: [10.1121/1.411872](https://doi.org/10.1121/1.411872).

- 1764 Hillenbrand, J. M., Houde, R. A., and Gayvert, R. T. (2006). “Speech perception based on  
 1765 spectral peaks versus spectral shape,” *The Journal of the Acoustical Society of America*  
 1766 **119**(6), 4041–4054, doi: [10.1121/1.2188369](https://doi.org/10.1121/1.2188369).
- 1767 Hillenbrand, J. M., and Nearey, T. M. (1999). “Identification of resynthesized /hvd/ utter-  
 1768 ances: Effects of formant contour,” *Journal of the Acoustical Society of America* **105**(6),  
 1769 3509–3523, doi: [10.1121/1.424676](https://doi.org/10.1121/1.424676).
- 1770 Hindle, D. (1978). “Approaches to Vowel Normalization in the Study of Natural Speech,”  
 1771 in *Linguistic Variation: Models and Methods*, edited by D. Sankoff (Academic Press, New  
 1772 York), pp. 161–171.
- 1773 Jaeger, T. F. (2024). *MVBeliefUpdatr: Fitting, Summarizing, and Visu-  
 1774 alizing of Multivariate Gaussian Ideal Observers and Adaptors*, [https://  
 1775 /github.com/hlplab/MVBeliefUpdatr](https://github.com/hlplab/MVBeliefUpdatr), r package version 0.0.1.0010, commit  
 1776 79ce50299b8c872ffb7f36c9a0547b57a47656d5.
- 1777 Johnson, K. (1997). “Speech perception without speaker normalization,” in *Talker Variabil-  
 1778 ity in Speech Processing*, edited by K. Johnson and W. Mullennix (CA: Academic Press,  
 1779 San Diego), pp. 146–165.
- 1780 Johnson, K. (2020). “The  $\Delta F$  method of vocal tract length normalization for vowels,”  
 1781 *Laboratory Phonology* **11**(1), doi: [10.5334/labphon.196](https://doi.org/10.5334/labphon.196).
- 1782 Johnson, K., and Sjerps, M. J. (2021). “Speaker normalization in speech perception,” in *The  
 1783 Handbook of Speech Perception*, edited by J. S. Pardo, L. C. Nygaard, R. E. Remez, and  
 1784 D. B. Pisoni (John Wiley & Sons, Inc), pp. 145–176, doi: [10.1002/9781119184096.ch6](https://doi.org/10.1002/9781119184096.ch6).
- 1785 Johnson, K., Strand, E. A., and D’Imperio, M. (1999). “Auditory–visual integration of  
 1786 talker gender in vowel perception,” *Journal of Phonetics* **27**(4), 359–384, doi: [10.1006/jpho.1999.0100](https://doi.org/10.1006/jpho.1999.0100).
- 1788 Joos, M. (1948). “Acoustic Phonetics,” *Language* **24**(2), 5–136, doi: [10.2307/522229](https://doi.org/10.2307/522229).
- 1789 Kay, M. (2024). *tidybayes: Tidy Data and Geoms for Bayesian Models*, [http://mjskay.github.io/tidybayes/](https://github.com/mjskay/tidybayes/), doi: [10.5281/zenodo.1308151](https://doi.org/10.5281/zenodo.1308151), r package version 3.0.7.
- 1791 Kleinschmidt, D. (2020). “What constrains distributional learning in adults?” doi: [10.31234/osf.io/6yhbe](https://doi.org/10.31234/osf.io/6yhbe), psyArXiv Preprint.
- 1793 Kleinschmidt, D., and Jaeger, T. F. (2015). “Robust speech perception: Recognize the  
 1794 familiar, generalize to the similar, and adapt to the novel,” *Psychological Review* **122**(2),  
 1795 148–203, doi: [10.1037/a0038695](https://doi.org/10.1037/a0038695).

- 1796 Kleinschmidt, D., and Jaeger, T. F. (2016). “What do you expect from an unfamiliar  
 1797 talker?,” Proceedings of the 38th Annual Meeting of the Cognitive Science Society, CogSci  
 1798 2016 2351–2356.
- 1799 Kleinschmidt, D., Liu, L., Bushong, W., Burchill, Z., Xie, X., Tan, M., Karboga, G., and  
 1800 Jaeger, F. (2021). “JSEXP” <https://github.com/hlplab/JSEXP>.
- 1801 Kronrod, Y., Coppess, E., and Feldman, N. H. (2016). “A unified model of categorical  
 1802 effects in consonant and vowel perception,” Psychological Bulletin and Review 1681–1712,  
 1803 doi: [10.3758/s13423-016-1049-y](https://doi.org/10.3758/s13423-016-1049-y).
- 1804 Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V.,  
 1805 Ryskina, V. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F. (1997). “Cross-language  
 1806 analysis of phonetic units in language addressed to infants,” Science 277(5326), 684–686,  
 1807 doi: [10.1126/science.277.5326.684](https://doi.org/10.1126/science.277.5326.684).
- 1808 Labov, W., Ash, S., and Boberg, C. (2006). *The atlas of North American English: Phonetics, phonology, and sound change* (De Gruyter Mouton, Berlin; New York).
- 1809
- 1810 Ladefoged, P., and Broadbent, D. E. (1957). “Information conveyed by vowels,” Journal of  
 1811 the Acoustical Society of America 29, 98–104, doi: [10.1121/1.1908694](https://doi.org/10.1121/1.1908694).
- 1812 Lee, C.-Y. (2009). “Identifying isolated, multispeaker mandarin tones from brief acoustic  
 1813 input: A perceptual and acoustic study,” The Journal of the Acoustical Society of America  
 1814 125(2), 0001–4966, doi: [10.1121/1.3050322](https://doi.org/10.1121/1.3050322).
- 1815 Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967).  
 1816 “Perception of the speech code,” Psychological review 74(6), 431–461, doi: [10.1037/h0020279](https://doi.org/10.1037/h0020279).
- 1817
- 1818 Lindblom, B. (1986). “Phonetic universals in vowel systems,” in *Experimental Phonology*,  
 1819 edited by J. J. Ohala and J. J. Jaeger (Academic Press, Orlando), pp. 13–44.
- 1820 Lindblom, B. (1990). “Explaining phonetic variation: A sketch of the H&H theory,” in  
 1821 *Speech Production and Speech Modeling*, edited by W. J. Hardcastle and A. Marchal (Dordrecht: Kluwer), pp. 403–439.
- 1822
- 1823 Lobanov, B. M. (1971). “Classification of Russian vowels spoken by different speakers,” The  
 1824 Journal of the Acoustical Society of America 49(2B), 606–608, doi: [10.1121/1.1912396](https://doi.org/10.1121/1.1912396).
- 1825 Luce, P. A., and Pisoni, D. B. (1998). “Recognizing spoken words: The neighborhood acti-  
 1826 vation model,” Ear and Hearing 19(1), 1–36, doi: [10.1097/00003446-199802000-00001](https://doi.org/10.1097/00003446-199802000-00001).
- 1827 Luce, R. D. (1959). *Individual Choice Behavior* (John Wiley, Oxford).

- 1828 Magnuson, J. S., and Nusbaum, H. C. (2007). “Acoustic differences, listener expectations,  
 1829 and the perceptual accommodation of talker variability,” *Journal of Experimental Psy-  
 1830 chology: Human Perception and Performance* **33**(2), 391–409, doi: [10.1037/0096-1523.  
 1831 33.2.391](https://doi.org/10.1037/0096-1523.33.2.391).
- 1832 Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna,  
 1833 P. D., Theodore, R., Monto, N., and Rueckl, J. G. (2020). “EARSHOT: A minimal neural  
 1834 network model of incremental human speech recognition,” *Cognitive Science* **44**(4), 1–17,  
 1835 doi: [10.1111/cogs.12823](https://doi.org/10.1111/cogs.12823).
- 1836 Massaro, D. W., and Friedman, D. (1990). “Models of integration given multiple sources of  
 1837 information.,” *Psychological Review* **97**(2), 225–252, doi: [10.1037/0033-295X.97.2.225](https://doi.org/10.1037/0033-295X.97.2.225).
- 1838 McClelland, J. L., and Elman, J. L. (1986). “The TRACE model of speech perception,”  
 1839 *Cognitive Psychology* **18**(1), 1–86, doi: [10.1016/0010-0285\(86\)90015-0](https://doi.org/10.1016/0010-0285(86)90015-0).
- 1840 McCloy, D. R. (2016). *phonR: tools for phoneticians and phonologists*, r package version  
 1841 1.0-7.
- 1842 McGowan, K. B. (2015). “Social expectation improves speech perception in noise,” *Lang-  
 1843 uage and Speech* **58**(4), 502–521, doi: [10.1177/0023830914565191](https://doi.org/10.1177/0023830914565191).
- 1844 McMurray, B., and Jongman, A. (2011). “What information is necessary for speech catego-  
 1845 rization?: Harnessing variability in the speech signal by integrating cues computed relative  
 1846 to expectations,” *Psychological Review* **118**(2), 219–246, doi: [10.1037/a0022325.What](https://doi.org/10.1037/a0022325.What).
- 1847 Merzenich, M. M., Knight, P. L., and Roth, G. L. (1975). “Representation of cochlea  
 1848 within primary auditory cortex in the cat,” *Journal of Neurophysiology* **38**(2), 231–249,  
 1849 doi: [10.1152/jn.1975.38.2.231](https://doi.org/10.1152/jn.1975.38.2.231).
- 1850 Miller, J. D. (1989). “Auditory-perceptual interpretation of the vowel,” *The Journal of*  
 1851 *Acoustical Society of America* **85**(5), 2114–2134, doi: [10.1121/1.397862](https://doi.org/10.1121/1.397862).
- 1852 Moore, B. C. (2012). *An Introduction to the Psychology of Hearing* (Brill, Bingley).
- 1853 Moulin-Frier, C., Diard, J., Schwartz, J.-L., and Bessière, P. (2015). “Cosmo (“communi-  
 1854 cating about objects using sensory–motor operations”): A bayesian modeling framework  
 1855 for studying speech communication and the emergence of phonological systems,” *Journal*  
 1856 *of Phonetics* **53**, 5–41, doi: [10.1016/j.wocn.2015.06.001](https://doi.org/10.1016/j.wocn.2015.06.001).
- 1857 Murdoch, D., and Chow, E. D. (2023). *ellipse: Functions for Drawing Ellipses and Ellipse-  
 1858 Like Confidence Regions*, <https://CRAN.R-project.org/package=ellipse>, r package  
 1859 version 0.5.0.

- 1860 Müller, K., and Wickham, H. (2023). *tibble: Simple Data Frames*, <https://CRAN.R-project.org/package=tibble>, r package version 3.2.1.
- 1862 Nearey, T. M. (1978). *Phonetic Feature Systems for Vowels* (Indiana University Linguistics Club, Indiana).
- 1864 Nearey, T. M. (1989). “Static, dynamic, and relational properties in vowel perception,” The Journal of the Acoustical Society of America **85**(5), 2088–2113, doi: [10.1121/1.397861](https://doi.org/10.1121/1.397861).
- 1866 Nearey, T. M. (1990). “The segment as a unit of speech perception,” Journal of Phonetics **18**(3), 347–373, doi: [10.1016/S0095-4470\(19\)30379-1](https://doi.org/10.1016/S0095-4470(19)30379-1).
- 1868 Nearey, T. M., and Assmann, P. F. (1986). “Modeling the role of inherent spectral change in vowel identification,” The Journal of the Acoustical Society of America **80**(5), 1297–1308, doi: [10.1121/1.394433](https://doi.org/10.1121/1.394433).
- 1871 Nearey, T. M., and Assmann, P. F. (2007). “Probabilistic ‘sliding template’ models for indirect vowel normalization,” in *Experimental approaches to phonology*, edited by J.-J. Solé, P. S. Beddor, and M. Ohala (Oxford University Press), pp. 246–270.
- 1874 Nearey, T. M., and Hogan, J. (1986). “Phonological contrast in experimental phonetics: Relating distributions of measurements production data to perceptual categorization curves,” in *Experimental Phonology*, edited by J. J. Ohala and J. Jaeger (Academic Press, New York), pp. 141–161.
- 1878 Newman, R. S., Clouse, S. A., and Burnham, J. L. (2001). “The perceptual consequences of within-talker variability in fricative production,” The Journal of the Acoustical Society of America **109**(3), 1181–1196, doi: [10.1121/1.1348009](https://doi.org/10.1121/1.1348009).
- 1881 Nordström, P., and Lindblom, B. (1975). “A normalization procedure for vowel formant data,” Proceedings of the 8th international congress of phonetic sciences, Leeds 212.
- 1883 Norris, D., and McQueen, J. M. (2008). “Shortlist B: A Bayesian model of continuous speech recognition.,” Psychological review **115**(2), 357–95, doi: [10.1037/0033-295X.115.2.357](https://doi.org/10.1037/0033-295X.115.2.357).
- 1885 Oganian, Y., Bhaya-Grossman, I., Johnson, K., and Chang, E. F. (2023). “Vowel and formant representation in the human auditory speech cortex,” Neuron **111**(13), 2105–2118.
- 1887 Patterson, R. D., and Irino, T. (2014). “Size matters in hearing: How the auditory system normalizes the sounds of speech and music for source size,” in *Perspectives on auditory research* (Springer), pp. 417–440.
- 1890 Pedersen, T. L. (2024a). *ggforce: Accelerating 'ggplot2'*, <https://CRAN.R-project.org/package=ggforce>, r package version 0.4.2.

- 1892 Pedersen, T. L. (2024b). *patchwork: The Composer of Plots*, <https://CRAN.R-project.org/package=patchwork>, r package version 1.2.0.
- 1893
- 1894 Persson, A., and Jaeger, T. F. (2023). “Evaluating normalization accounts against the dense  
1895 vowel space of Central Swedish,” *Frontiers in Psychology* **14**, doi: [10.3389/fpsyg.2023.1165742](https://doi.org/10.3389/fpsyg.2023.1165742).
- 1896
- 1897 Peterson, G. E. (1961). “Parameters of vowel quality,” *Journal of Speech and Hearing  
1898 Research* **4**(1), 10–29, doi: [10.1044/jshr.0401.10](https://doi.org/10.1044/jshr.0401.10).
- 1899 Peterson, G. E., and Barney, H. L. (1952). “Control methods used in a study of the vowels,”  
1900 *Journal of the Acoustical Society of America* **24**(2), 175–184, doi: [10.1121/1.1906875](https://doi.org/10.1121/1.1906875).
- 1901 Pinheiro, J., Bates, D., and R Core Team (2023). *nlme: Linear and Nonlinear Mixed Effects  
1902 Models*, <https://CRAN.R-project.org/package=nlme>, r package version 3.1-164.
- 1903 R Core Team (2024). *R: A Language and Environment for Statistical Computing*, R Foun-  
1904 dation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>.
- 1905 Repp, B. H., and Crowder, R. G. (1990). “Stimulus order effects in vowel discrimination,”  
1906 *The Journal of the Acoustical Society of America* **88**(5), 2080–2090, doi: [10.1121/1.400105](https://doi.org/10.1121/1.400105).
- 1907
- 1908 Richter, C., Feldman, N. H., Salgado, H., and Jansen, A. (2017). “Evaluating low-level  
1909 speech features against human perceptual data,” *Transactions of the Association for Com-  
1910 putational Linguistics* **5**, 425–440, doi: [10.1162/tacl\\_a\\_00071](https://doi.org/10.1162/tacl_a_00071).
- 1911 Robinson, D. (2020). *fuzzyjoin: Join Tables Together on Inexact Matching*, <https://CRAN.R-project.org/package=fuzzyjoin>, r package version 0.1.6.
- 1912
- 1913 RStudio Team (2020). *RStudio: Integrated Development Environment for R*, RStudio,  
1914 PBC., Boston, MA.
- 1915 Saenz, M., and Langers, D. R. (2014). “Tonotopic mapping of human auditory cortex,”  
1916 *Hearing Research* **307**, 42–52, doi: [10.1016/j.heares.2013.07.016](https://doi.org/10.1016/j.heares.2013.07.016) human Auditory  
1917 NeuroImaging.
- 1918 Scarborough, R. (2010). “Lexical and contextual predictability: Confluent effects on the  
1919 production of vowels,” in *Laboratory Phonology*, edited by C. Fougeron, B. Kühnert,  
1920 M. D’Imperio, and N. Vallée, **10** (De Gruyter Mouton Berlin), pp. 557–586.
- 1921 Schertz, J., and Clare, E. J. (2020). “Phonetic cue weighting in perception and production,”  
1922 *Wiley Interdisciplinary Reviews: Cognitive Science* **11**(2), doi: [10.1002/wcs.1521](https://doi.org/10.1002/wcs.1521).
- 1923 Shankweiler, D., Verbrugge, R. R., and Studdert-Kennedy, M. (1978). “Insufficiency of the  
1924 target for vowel perception,” *The Journal of the Acoustical Society of America* **63**(S1),

- 1925 S4–S4, doi: [10.1121/1.2016686](https://doi.org/10.1121/1.2016686).
- 1926 Shannon, C. E. (1948). “A mathematical theory of communication,” The Bell System Technical Journal **27**(3), 379–423, doi: [10.1002/j.1538-7305.1948.tb01338.x](https://doi.org/10.1002/j.1538-7305.1948.tb01338.x).
- 1928 Siegel, R. J. (1965). “A replication of the mel scale of pitch,” The American Journal of 1929 Psychology **78**(4), 615–620, doi: [10.2307/1420924](https://doi.org/10.2307/1420924).
- 1930 Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny* 1931 (Chapman and Hall/CRC), <https://plotly-r.com>.
- 1932 Sjerps, M. J., Fox, N. P., Johnson, K., and Chang, E. F. (2019). “Speaker-normalized sound 1933 representations in the human auditory cortex,” Nature Communications **10**(1), 01–09, doi: 1934 [10.1038/s41467-019-10365-z](https://doi.org/10.1038/s41467-019-10365-z).
- 1935 Skoe, E., Krizman, J., Spitzer, E. R., and Kraus, N. (2021). “Auditory cortical changes 1936 precede brainstem changes during rapid implicit learning: Evidence from human EEG,” 1937 Frontiers in Neuroscience **15**, 01–09, doi: [10.3389/fnins.2021.718230](https://doi.org/10.3389/fnins.2021.718230).
- 1938 Smith, B. L., Johnson, E., and Hayes-Harb, R. (2019). “ESL learners’ intra-speaker 1939 variability in producing American English tense and lax vowels,” Journal of Second Language 1940 Pronunciation **5**(1), 139–164, doi: [10.1075/jslp.15050.smi](https://doi.org/10.1075/jslp.15050.smi).
- 1941 Smith, D. R., Patterson, R. D., Turner, R., Kawahara, H., and Irino, T. (2005). “The 1942 processing and perception of size information in speech sounds,” The Journal of the Acoustical 1943 Society of America **117**(1), 305–318, doi: [10.1121/1.1828637](https://doi.org/10.1121/1.1828637).
- 1944 Steriade, D. (2008). “The phonology of perceptibility effects: the P-map and its 1945 consequences for constraint organization,” in *The Nature of the Word: Studies in Honor of* 1946 *Paul Kiparsky*, edited by K. Hanson and S. Inkelas (MIT Press, UCLA), doi: [10.7551/mitpress/9780262083799.001.0001](https://doi.org/10.7551/mitpress/9780262083799.001.0001).
- 1948 Stevens, K. N. (1972). “The quantal nature of speech: Evidence from articulatory-acoustic 1949 data,” in *Human communication: a unified view* (McGraHill, New York), pp. 51–66.
- 1950 Stevens, K. N. (1989). “On the quantal nature of speech,” Journal of phonetics **17**(1-2), 1951 3–45.
- 1952 Stevens, S. S., and Volkmann, J. (1940). “The Relation of Pitch to Frequency: A Revised 1953 Scale,” The American Journal of Psychology **53**(3), 329–353, doi: [10.2307/1417526](https://doi.org/10.2307/1417526).
- 1954 Stilp, C. (2020). “Acoustic context effects in speech perception,” WIREs Cognitive Science 1955 **11**(1), 1–18, doi: [10.1002/wcs.1517](https://doi.org/10.1002/wcs.1517).
- 1956 Strange, W., and Jenkins, J. J. (2012). “Dynamic specification of coarticulated vowels: Re- 1957 search chronology, theory, and hypotheses,” in *Vowel Inherent Spectral Change* (Springer),

- 1958 pp. 87–115.
- 1959 Sumner, M. (2011). “The role of variation in the perception of accented speech,” *Cognition* 1960 **119**(1), 131–136, doi: [10.1016/j.cognition.2010.10.018](https://doi.org/10.1016/j.cognition.2010.10.018).
- 1961 Syrdal, A. K. (1985). “Aspects of a model of the auditory representation of American English 1962 vowels,” *Speech Communication* 4(1-3), 121–135, doi: [10.1016/0167-6393\(85\)90040-8](https://doi.org/10.1016/0167-6393(85)90040-8).
- 1963 Syrdal, A. K., and Gopal, H. S. (1986). “A perceptual model of vowel recognition based on 1964 the auditory representation of American English vowels,” *The Journal of the Acoustical Society of America* 1965 **79**(4), 1086–1100, doi: [10.1121/1.393381](https://doi.org/10.1121/1.393381).
- 1966 Tan, M., and Jaeger, T. F. (2024). “Incremental adaptation to an unfamiliar talker,” 1967 Manuscript, Stockholm University .
- 1968 Tang, C., Hamilton, L. S., and Chang, E. F. (2017). “Intonational speech prosody encoding 1969 in the human auditory cortex,” *Science* 357(6353), 797–801, doi: [10.1126/science.aam8577](https://doi.org/10.1126/science.aam8577).
- 1970 ten Bosch, L., Boves, L., Tucker, B., and Ernestus, M. (2015). “DIANA: towards computational 1971 modeling reaction times in lexical decision in north American English,” in *Proc. 1972 Interspeech 2015*, pp. 1576–1580, doi: [10.21437/Interspeech.2015-366](https://doi.org/10.21437/Interspeech.2015-366).
- 1973 Traunmüller, H. (1981). “Perceptual dimension of openness in vowels,” *The Journal of the 1974 Acoustical Society of America* 69(5), 1465–1475, doi: [10.1121/1.385780](https://doi.org/10.1121/1.385780).
- 1975 Traunmüller, H. (1990). “Analytical expressions for the tonotopic sensory scale,” *The Journal 1976 of the Acoustical Society of America* 88(1), 97–100, doi: [10.1121/1.399849](https://doi.org/10.1121/1.399849).
- 1977 Urbanek, S., and Horner, J. (2023). *Cairo: R Graphics Device using Cairo Graphics Library 1978 for Creating High-Quality Bitmap (PNG, JPEG, TIFF), Vector (PDF, SVG, PostScript) 1979 and Display (X11 and Win32) Output*, <https://CRAN.R-project.org/package=Cairo>, r 1980 package version 1.6-2.
- 1981 van den Brand, T. (2024). *ggh4x: Hacks for 'ggplot2'*, <https://CRAN.R-project.org/package=ggh4x>, r package version 0.2.8.
- 1982 van Rij, J. (2020). *plotfunctions: Various Functions to Facilitate Visualization of Data and 1983 Analysis*, <https://CRAN.R-project.org/package=plotfunctions>, r package version 1.4.
- 1984 van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2022). “itsadug: Interpreting 1985 time series and autocorrelated data using gamms” R package version 2.4.1.
- 1986 Vaughan, D., and Dancho, M. (2022). *furrr: Apply Mapping Functions in Parallel using 1987 Futures*, <https://CRAN.R-project.org/package=furrr>, r package version 0.3.1.

- 1990 Vaughn, C., Baese-Berk, M., and Idemaru, K. (2019). “Re-examining phonetic variability  
 1991 in native and non-native speech,” *Phonetica* **76**(5), 327–358, doi: [10.1159/000487269](https://doi.org/10.1159/000487269).
- 1992 Vorperian, H. K., and Kent, R. D. (2007). “Vowel acoustic space development in children: A  
 1993 synthesis of acoustic and anatomic data,” *Journal of Speech, Language & Hearing Research*  
 1994 **50**(6), 1510–1545, doi: [10.1044/1092-4388\(2007/104\)](https://doi.org/10.1044/1092-4388(2007/104)).
- 1995 Wade, T., Jongman, A., and Sereno, J. (2007). “Effects of acoustic variability in the per-  
 1996 ceptual learning of non-native-accented speech sounds,” *Phonetica* **64**(2-3), 122–144, doi:  
 1997 [10.1159/000107913](https://doi.org/10.1159/000107913).
- 1998 Walker, A., and Hay, J. (2011). “Congruence between ‘word age’ and ‘voice age’ facilitates  
 1999 lexical access,” *Laboratory Phonology* **2**(1), 219–237, doi: [10.1515/labphon.2011.007](https://doi.org/10.1515/labphon.2011.007).
- 2000 Watt, D., and Fabricius, A. (2002). “Evaluation of a technique for improving the mapping  
 2001 of multiple speakers’ vowel spaces in the F1 ~ F2 plane,” in *Leeds Working Papers in*  
 2002 *Linguistics and Phonetics*, edited by D. Nelson, 9 (University of Leeds), pp. 159–173.
- 2003 Weatherholtz, K., and Jaeger, T. F. (2016). “Speech perception and generalization  
 2004 across talkers and accents,” Oxford Research Encyclopedia of Linguistics doi: [10.1093/acrefore/9780199384655.013.95](https://doi.org/10.1093/acrefore/9780199384655.013.95).
- 2006 Wedel, A., Nelson, N., and Sharp, R. (2018). “The phonetic specificity of contrastive hy-  
 2007 perarticulation in natural speech,” *Journal of Memory and Language* **100**, 61–88, doi:  
 2008 [10.1016/j.jml.2018.01.001](https://doi.org/10.1016/j.jml.2018.01.001).
- 2009 Whalen, D. H. (2016). “A double-Nearey theory of vowel normalization: Approaching con-  
 2010 sensus,” *The Journal of the Acoustical Society of America* **140**(4\_Supplement), 3163–3164,  
 2011 doi: [10.1121/1.4969932](https://doi.org/10.1121/1.4969932).
- 2012 Wichmann, F. A., and Hill, N. J. (2001). “The psychometric function: I. Fitting, sam-  
 2013 pling, and goodness of fit,” *Perception & psychophysics* **63**(8), 1293–1313, doi: [10.3758/BF03194544](https://doi.org/10.3758/BF03194544).
- 2015 Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New  
 2016 York), <https://ggplot2.tidyverse.org>.
- 2017 Wickham, H. (2019). *assertthat: Easy Pre and Post Assertions*, <https://CRAN.R-project.org/package=assertthat>, r package version 0.2.1.
- 2019 Wickham, H. (2023a). *forcats: Tools for Working with Categorical Variables (Factors)*,  
 2020 <https://CRAN.R-project.org/package=forcats>, r package version 1.0.0.
- 2021 Wickham, H. (2023b). *modelr: Modelling Functions that Work with the Pipe*, <https://CRAN.R-project.org/package=modelr>, r package version 0.1.11.

- 2023 Wickham, H. (2023c). *stringr: Simple, Consistent Wrappers for Common String Operations*, <https://CRAN.R-project.org/package=stringr>, r package version 1.5.1.
- 2024 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemond, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). “Welcome to the tidyverse,” *Journal of Open Source Software* 4(43), 1686, doi: [10.21105/joss.01686](https://doi.org/10.21105/joss.01686).
- 2025 Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*, <https://CRAN.R-project.org/package=dplyr>, r package version 1.1.4.
- 2026 Wickham, H., and Henry, L. (2023). *purrr: Functional Programming Tools*, <https://CRAN.R-project.org/package=purrr>, r package version 1.0.2.
- 2027 Wickham, H., Hester, J., and Bryan, J. (2024a). *readr: Read Rectangular Text Data*, <https://CRAN.R-project.org/package=readr>, r package version 2.1.5.
- 2028 Wickham, H., Vaughan, D., and Girlich, M. (2024b). *tidyr: Tidy Messy Data*, <https://CRAN.R-project.org/package=tidyr>, r package version 1.3.1.
- 2029 Wilke, C. O., and Wiernik, B. M. (2022). *ggtext: Improved Text Rendering Support for 'ggplot2'*, <https://CRAN.R-project.org/package=ggtext>, r package version 0.1.2.
- 2030 Winn, M. (2018). “Speech: It’s not as acoustic as you think,” *Acoustics Today* 12(2), 43–49.
- 2031 Wood, S., Pya, and S”afken, B. (2016). “Smoothing parameter and model selection for general smooth models (with discussion),” *Journal of the American Statistical Association* 111, 1548–1575.
- 2032 Wood, S. N. (2003). “Thin-plate regression splines,” *Journal of the Royal Statistical Society (B)* 65(1), 95–114.
- 2033 Wood, S. N. (2004). “Stable and efficient multiple smoothing parameter estimation for generalized additive models,” *Journal of the American Statistical Association* 99(467), 673–686.
- 2034 Wood, S. N. (2011). “Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models,” *Journal of the Royal Statistical Society (B)* 73(1), 3–36.
- 2035 Xie, X., Buxó-Lugo, A., and Kurumada, C. (2021). “Encoding and decoding of meaning through structured variability in speech prosody,” *Cognition* 211, 1–27, doi: [10.1016/j.cognition.2021.104619](https://doi.org/10.1016/j.cognition.2021.104619).

- 2056 Xie, X., and Jaeger, T. F. (2020). “Comparing non-native and native speech: Are L2  
2057 productions more variable?,” The Journal of the Acoustical Society of America **147**(5),  
2058 3322–3347, doi: [10.1121/10.0001141](https://doi.org/10.1121/10.0001141).
- 2059 Xie, X., Jaeger, T. F., and Kurumada, C. (2023). “What we do (not) know about the  
2060 mechanisms underlying adaptive speech perception: A computational review,” Cortex  
2061 **166**, 377–424, doi: [10.1016/j.cortex.2023.05.003](https://doi.org/10.1016/j.cortex.2023.05.003).
- 2062 Xie, Y. (2024). *knitr: A General-Purpose Package for Dynamic Report Generation in R*,  
2063 <https://yihui.org/knitr/>, r package version 1.49.
- 2064 Yuan, J., and Liberman, M. (2008). “Speaker identification on the SCOTUS corpus,” The  
2065 Journal of the Acoustical Society of America **123**(5), 3878–3878, doi: [10.1121/1.2935783](https://doi.org/10.1121/1.2935783).
- 2066 Zahorian, S. A., and Jagharghi, A. J. (1991). “Speaker normalization of static and dynamic  
2067 vowel spectral features,” The Journal of the Acoustical Society of America **90**(1), 67–75,  
2068 doi: [10.1121/1.402350](https://doi.org/10.1121/1.402350).
- 2069 Zwicker, E. (1961). “Subdivision of the audible frequency range into critical bands (fre-  
2070 quenzgruppen),” The Journal of the Acoustical Society of America **33**(2), 248–248, doi:  
2071 [10.1121/1.1908630](https://doi.org/10.1121/1.1908630).
- 2072 Zwicker, E., Flottorp, G., and Stevens, S. S. (1957). “Critical band width in loudness  
2073 summation,” The Journal of the Acoustical Society of America **29**(5), 548–557, doi: [10.1121/1.1908963](https://doi.org/10.1121/1.1908963).
- 2075 Zwicker, E., and Terhardt, E. (1980). “Analytical expressions for critical-band rate and  
2076 critical bandwidth as a function of frequency,” The Journal of the Acoustical Society of  
2077 America **68**(5), 1523–1525, doi: [10.1121/1.385079](https://doi.org/10.1121/1.385079).