

**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*<sup>a</sup>

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

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

(Dated: 25 July 2024)

1 Human speech perception tends to achieve robust speech recognition, despite sub-  
2 stantial cross-talker variability. Believed to be critical to this ability are auditory  
3 normalization mechanisms whereby listeners adapt to individual differences in vocal  
4 tract physiology in vowel perception. This study asks what types of computations  
5 are involved in such normalization. Two 8-way alternative forced-choice experiments  
6 assessed L1 listeners' categorizations across the entire US English vowel space—both  
7 for unaltered and for synthesized stimuli. Listeners' responses in these experiments  
8 were compared against the predictions of twenty influential normalization accounts  
9 that differ starkly in the inference and memory capacities they imply for speech  
10 perception. Listeners' responses were best explained by *extrinsic* normalization ac-  
11 counts, suggesting that listeners learn and store distributional properties of talkers'  
12 speech. Of the extrinsic accounts, it was the *computationally least complex* variants  
13 that best fit listeners' responses, using a single parameter. These findings have conse-  
14 quences for any research that aims to investigate the perceptual, social, and linguistic  
15 information of vowel productions. This includes research in phonetics and phonol-  
16 ogy, sociolinguistics, and language acquisition. In these fields, it remains common  
17 to employ normalization accounts that the present study confirms to be inadequate  
18 models of human perception (e.g., Lobanov normalization).

---

<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). These studies  
 35 have decoded effects of talker identity from subcortical brain areas like the brain stem, and  
 36 thus prior to the cortical regions believed to encode linguistic categories (e.g., Sjerps *et al.*,  
 37 2019; Tang *et al.*, 2017). This includes brain responses that lag the acoustic signal by as  
 38 little as 20-50 msec (Lee, 2009), suggesting very fast and highly automatic processes. By  
 39 removing talker-specific variability from the phonetic signal early, auditory normalization

40 offers elegant and effective solutions to cross-talker variability, that might reduce the need  
 41 for more complex adaptation of individual phonetic categories further upstream (Apfelbaum  
 42 and McMurray, 2015; Xie *et al.*, 2023).<sup>1</sup>

43 While it is relatively uncontroversial *that* normalization contributes to robust speech  
 44 perception, it is still unclear what types of computations this implicates. We address this  
 45 question for the perception of vowels, which cross-linguistically relies on peaks in the distri-  
 46 bution of spectral energy over acoustic frequencies (formants). Vowel perception has long  
 47 been a focus in research on normalization (e.g., Bladon *et al.*, 1984; Fant, 1975; Gerstman,  
 48 1968; Johnson, 2020; Joos, 1948; Lobanov, 1971; Miller, 1989; Nearey, 1978; Nordström  
 49 and Lindblom, 1975; Syrdal and Gopal, 1986; Traunmüller, 1981; Watt and Fabricius, 2002;  
 50 Zahorian and Jagharghi, 1991; for review, see Barreda, 2020), with some reviews citing over  
 51 100 competing proposals (Carpenter and Govindarajan, 1993). Importantly, these accounts  
 52 differ in the types and complexity of computations they assume to take place during nor-  
 53 malization. On the lower end of computational complexity, comparatively simple static  
 54 transformations of the acoustic signal might suffice to achieve invariance in the mapping  
 55 from cues to phonetic categories. For example, there is evidence that a transformation of  
 56 acoustic frequencies (measured in Hz) into the psycho-acoustic Mel-space better describes  
 57 how listeners perceive differences in the frequency of sine tones (e.g., Fastl and Zwicker,  
 58 2007; Stevens and Volkmann, 1940; for a critique, see Greenwood, 1997; Siegel, 1965). It  
 59 is thus possible that cross-talker variability in vowel pronunciations is effectively reduced  
 60 when formants are represented in Mel, rather than Hz. Similar arguments have been made  
 61 about other psycho-acoustic transformations (e.g., Bark, Traunmüller, 1990; ERB, Glasberg

62 and Moore, 1990; or semitones, Fant *et al.*, 2002). Most of these accounts share that they  
63 log-transform acoustic frequencies—in line with neurophysiological evidence that the audi-  
64 tory representations in the brain seem to follow a roughly logarithmic organization, so that  
65 auditory perception is (up to a point) more sensitive to differences between lower frequen-  
66 cies than to the same difference between higher frequencies (e.g., Merzenich *et al.*, 1975;  
67 for review, see Saenz and Langers, 2014). If such static psycho-acoustic transformations are  
68 sufficient for formant normalization, this would offer a particularly parsimonious account of  
69 vowel perception as listeners would not have to infer talker-specific properties.

70 The parsimony of psycho-acoustic transformations contrasts with the majority of accounts  
71 for vowel normalization, which introduce additional computations. This includes accounts  
72 that normalize formants relative to other information that is available at the same point in  
73 the acoustic signal (intrinsic normalization, e.g., Miller, 1989; Peterson, 1961; Syrdal and  
74 Gopal, 1986). For example, according to one proposal, listeners normalize vowel formants  
75 by the vowel’s fundamental frequency or other formants estimated at the same point in  
76 time (Syrdal and Gopal, 1986). To the extent that the fundamental frequency is correlated  
77 with the talkers’ vocal tract size (for review, see Vorperian and Kent, 2007), this allows  
78 the removal of physiologically-conditioned cross-talker variability in formant realizations.  
79 While such intrinsic accounts arguably entail more computational complexity than static  
80 transformations, they do not require that listeners *maintain* talker-specific estimates over  
81 time. This distinguishes intrinsic from extrinsic accounts, which introduce additional com-  
82 putational complexity.

83 According to extrinsic accounts, normalization mechanisms infer and store estimates of  
84 talker-specific properties that then are used to normalize subsequent speech from that talker  
85 (Gerstman, 1968; Lobanov, 1971; Nearey, 1978; Nordström and Lindblom, 1975; Watt and  
86 Fabricius, 2002; for review, see also Weatherholtz and Jaeger, 2016). At the upper end of  
87 computational complexity, some accounts hold that listeners continuously infer and maintain  
88 both talker-specific means for each formant and talker-specific estimates of each formant's  
89 variability (Gerstman, 1968; Lobanov, 1971). These estimates are then used to normalize  
90 formants, e.g., by centering and standardizing them (essentially z-scoring formants, Lobanov,  
91 1971), removing cross-talker variability in the distribution of formant values. There are,  
92 however, more parsimonious extrinsic accounts that require inference and maintenance of  
93 fewer talker-specific properties. The most parsimonious of these is Nearey's *uniform scaling*  
94 account, which assumes that listeners infer and maintain a single talker-specific parameter.  
95 This parameter ( $\Psi$ ) can be thought of as capturing the effects of the talkers' vocal tract  
96 length on the spectral scaling applied to the formant pattern produced by a talker (Nearey,  
97 1978).<sup>2</sup> Uniform scaling deserves particular mention here as it is arguably one of the most  
98 developed normalization accounts, and rooted in principled considerations about the physics  
99 of sound and the evolution of auditory systems (for review, see Barreda, 2020).

100 In summary, hypotheses about the computations implied by formant normalization differ  
101 in the flexibility they afford as well as the inference and memory complexity they entail.  
102 Considerations about the complexity of inferences—essentially the number of parameters  
103 that listeners are assumed to estimate at any given moment in time—arguably gain in  
104 importance in light of the speed at which normalization seems to unfold. In the present

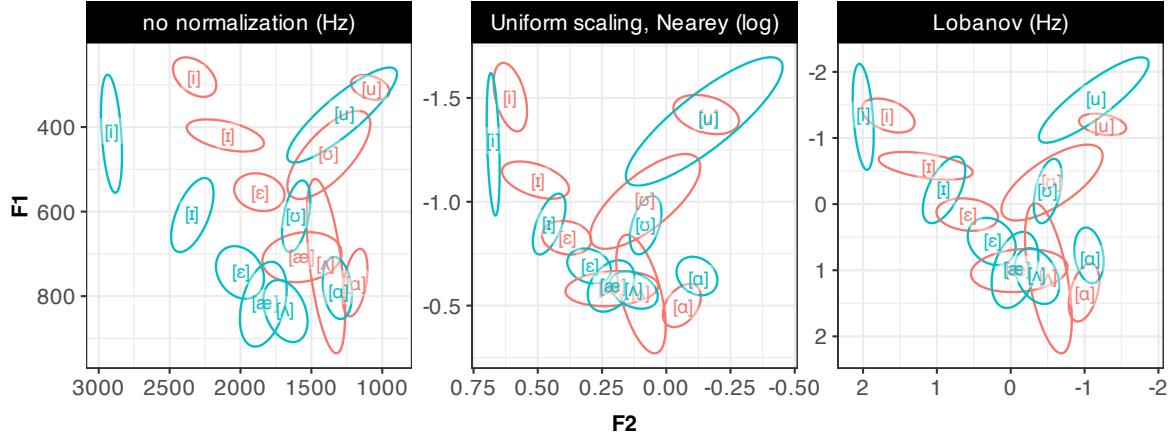


FIG. 1. Illustration of how height, which is positively correlated with vocal tract size, affects vowels' F1 and F2, and how normalization can partially remove this effect. Shown here are realizations of all 8 monophthong vowels of US English by a short (cyan) and a tall native talker (red). **Panel 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 three of these accounts, along with 17 other normalization accounts.

105 study, we thus ask whether computationally simple accounts are sufficient to explain human  
 106 vowel perception.

107 While previous research has compared normalization accounts across languages, most of  
 108 this work has evaluated proposals in terms of how well the normalized phonetic space sup-  
 109 ports the separability of vowel categories (Adank *et al.*, 2004; Carpenter and Govindarajan,  
 110 1993; Cole *et al.*, 2010; Escudero and Bion, 2007; Johnson and Sjerps, 2021; Syrdal, 1985).

111 This approach is illustrated in Figure 1. These studies have found that computationally

more complex accounts—which also afford more flexibility—tend to achieve higher category separability and higher categorization accuracy (for review, see Persson and Jaeger, 2023). This includes Lobanov normalization, which continues to be highly influential in, for example, variationist and sociolinguistic research because of its effectiveness in removing cross-talker variability (for a critique, see Barreda, 2021). It is, however, by no means clear that human speech perception employs the same computations that achieve the best category separability or accuracy (see also discussion in Barreda, 2021; Nearey and Assmann, 2007).

A substantially smaller body of research has addressed this question by comparing normalization accounts against *listeners' perception* (Barreda and Nearey, 2012; Barreda, 2021; Nearey, 1989; Richter *et al.*, 2017; for a review, see Whalen, 2016). Interestingly, these works seem to suggest that computationally simpler accounts might provide a better fit against human speech perception than the influential Lobanov model (Barreda, 2021; Richter *et al.*, 2017). For example, Barreda (2021) compared the predictions of uniform scaling and Lobanov normalization against listeners' categorization responses in a forced-choice categorization task over parts of the US English vowel space. In his experiment, listeners' categorization responses were better predicted by uniform scaling than by Lobanov normalization. Findings like these suggest that comparatively simple corrections for vocal tract size—such as uniform scaling—might provide a better explanation of human perception than more computationally complex accounts (see also Johnson, 2020; Richter *et al.*, 2017).

This motivates the present work. We take a broad-coverage approach by comparing the 20 normalization accounts in Table I against the perception of all 8 monophthongs

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

<sub>149</sub> Next, we motivate and describe the two experiments we conducted. Then we compare  
<sub>150</sub> the normalization accounts in Table I against listeners responses from these experiments.

<sub>151</sub> **A. Open Science Statement**

<sub>152</sub> All stimulus recordings, results, and the code for the experiment, data analysis, and  
<sub>153</sub> computational modeling for this article can be downloaded from OSF at <https://osf.io/zemwn/>. The OSF repo also include extensive supplementary information (SI). Both the ar-  
<sub>155</sub> ticle and SI are written in R markdown, allowing readers to replicate our analyses with the  
<sub>156</sub> click of a button, using freely available software (R Core Team, 2024; RStudio Team, 2020).

<sub>157</sub> Readers can revisit the assumptions we committed to for the present project—for example,  
<sub>158</sub> by substituting alternative normalization accounts or categorization models. Researchers  
<sub>159</sub> can also substitute their own experiments on vowel normalization for our Experiments 1a  
<sub>160</sub> and 1b, to see whether our findings generalize to novel data. We see this as an important  
<sub>161</sub> contribution of the present work, as it should make it substantially easier to consider ad-  
<sub>162</sub> ditional normalization accounts—including variants to the accounts we considered—and to  
<sub>163</sub> assess the generalizability of the conclusions we reach based on the present data.

<sub>164</sub> **II. EXPERIMENTS 1A AND 1B**

<sub>165</sub> To compare the performance of different normalization accounts against listeners' percep-  
<sub>166</sub> tion, we conducted two small web-based experiments on US English listeners' perception of  
<sub>167</sub> US English vowels. The two experiments employ the same 8-alternative forced-choice vowel  
<sub>168</sub> categorization task (Figure 2), and differ only in the whether they employed 'natural' (Ex-

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
	n/a	Hz	n/a	n/a
transformation	—	log	—	$F_n^{log} = \ln(F_n)$
	Bark	—	Traunmüller (1990)	$F_n^{Bark} = \frac{26.81 \times F_n}{1960 + F_n} - 0.53$
	ERB	—	Glasberg & Moore (1990)	$F_n^{ERB} = 21.4 \times \log_{10}(1 + F_n) \times 0.00437$
	Mel	—	Stevens & Volkmann (1940)	$F_n^{Mel} = 2595 \times \log_{10}(1 + \frac{F_n}{700})$
	Semitones conversion	—	Fant et al. (2002)	$F_n^{ST} = 12 \times \ln(\frac{F_n}{100})$
	Bark	Syrdal & Gopal (1986)	—	$F1_{SyrdalGopal1} = F1_{Bark} - F0_{Bark}$
	Syrdal & Gopal 1 (Bark-distance model)	—	—	$F2_{SyrdalGopal1} = F2_{Bark} - F1_{Bark}$
	Syrdal & Gopal 2 (Bark-distance model)	—	—	$F1_{SyrdalGopal2} = F1_{Bark} - F0_{Bark}$
	Miller (formant-ratio)	log	Miller (1989)	$F2_{SyrdalGopal2} = F3_{Bark} - F2_{Bark}$
	Miller (formant-ratio)	—	—	$SR = k(\frac{G_{Miller}}{k})^{1/3}$
	—	—	—	$F1_{Miller} = \log(\frac{F1}{SR})$
	—	—	—	$F2_{Miller} = \log(\frac{F2}{F1})$
	—	—	—	$F3_{Miller} = \log(\frac{F3}{F2})$
	log	Nearey (1978)	Nearey (1978)	$F_n^{Nearey} = \ln(F_n) - \text{mean}(\ln(F))$
intrinsic	Uniform scaling, Nearey	Hz	Nordström & Lindblom (1975)	$F_n^{NordströmLindblom} = \frac{F_n}{\text{mean}(\frac{F_{n,1} \dots F_{n,600}}{2.5})}$
	Uniform scaling, Nordström & Lindblom	—	—	$F_n^{Johnson} = \frac{F_n}{\text{mean}(\frac{F1}{0.5}, \frac{F2}{1.5}, \frac{F3}{2.5})}$
	Uniform scaling, Johnson	Hz	Johnson (2020)	$F_n^{Nearey} = \ln(F_n) - \text{mean}(\ln(F_n))$
	Nearey's formantwise log-mean	log	Nearey (1978)	$F_n^{C-CuRE} = F_n - \text{mean}(F_n)$
	C-CuRE	Hz	McMurray & Jongman (2011)	$F_n^{Gerstman} = 999 \times \frac{F_n - F_n^{min}}{F_n^{max} - F_n^{min}}$
	—	Bark	—	$F_n^{Lobanov} = \frac{F_n - \text{mean}(F_n)}{sd(F_n)}$
	—	ERB	—	
	—	Mel	—	
	Semitones conversion	—	Gerstman (1968)	
extrinsic	standardizing	Hz	—	
	Gerstman (range normalization)	—	—	
	Lobanov (z-score)	Hz	Lobanov (1971)	

heed who'd hood

hid  hud

head had hod

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

<sup>169</sup> periment 1a) or synthesized stimuli (Experiment 1b). To the best of our knowledge, these  
<sup>170</sup> two experiments are the first designed to compare normalization accounts against listeners'  
<sup>171</sup> perception over the entire monophthong inventory of a language.

<sup>172</sup> Experiment 1a employs recordings of *hVd* word productions from a female talker of US  
<sup>173</sup> English, these recordings are ‘natural’ in the sense that they were not synthesized or other-  
<sup>174</sup> wise phonetically manipulated. One consequence of this is that the formant values of these  
<sup>175</sup> recordings are clustered around the category means, and thus span only a comparatively  
<sup>176</sup> small part of the phonetic space. This can limit the statistical power to distinguish between  
<sup>177</sup> competing accounts. Natural recordings furthermore vary not only along the primary cues  
<sup>178</sup> to vowel quality in US English (F1, F2) but also along potential secondary cues (e.g., F0,  
<sup>179</sup> F3, and vowel duration) as well as other unknown properties, which can make it difficult to  
<sup>180</sup> discern whether the performance of a normalization model is due to the normalization itself  
<sup>181</sup> or other reasons, e.g., because a normalized cue happens to correlate with another cue that  
<sup>182</sup> listeners are sensitive to but that is not included in the model.

183 Experiment 1b thus adopts an alternative approach and uses synthesized vowels. Unlike  
184 most previous work, which has used isolated vowels as stimuli (Barreda, 2021; Barreda and  
185 Nearey, 2012; Nearey, 1989; Richter *et al.*, 2017), Experiment 1b uses synthesized *hVd* words  
186 to facilitate comparison to Experiment 1a. This allowed us to sample larger parts of the F1-  
187 F2 space, which has two advantages. First, it allowed us to collect responses over parts of the  
188 formant space for which we expect listeners to have more uncertainty, and thus exhibit more  
189 variable responses. This can increase the statistical power to distinguish between competing  
190 accounts. Second, differences in the predictions of competing normalization accounts will  
191 tend to become more pronounced with increasing distance from the category centers. By  
192 collecting responses at those locations, we can thus increase the contrast between competing  
193 accounts.

194 The use of synthesized stimuli does, however, also come with potential disadvantages.  
195 Synthesized stimuli can suffer in ecological validity, lacking correlations between cues, and  
196 across the speech signal (e.g., due to co-articulation) that are characteristic of human speech.  
197 This raises questions about the extent to which processing of such stimuli engages the same  
198 mechanisms as everyday speech perception. Additionally, it is possible that the use of robotic  
199 sounding synthesized speech affects listener engagement. This can lead to an increased rate  
200 of attentional lapses, and thus a decrease in the proportion of trials on which listeners'  
201 responses are based on the acoustics of the speech stimulus rather than random guessing  
202 (compare, e.g., Kleinschmidt, 2020; Tan and Jaeger, 2024). By comparing normalization  
203 accounts against both natural and synthesized stimuli, we investigate the extent to which

204 the accounts that best describe human perception depend on the type of stimuli used in the  
205 experiment.

206 **A. Methods**

207 **1. Participants**

208 We recruited 24 (Experiment 1a) and 24 (Experiment 1b) participants from Amazon's  
209 Mechanical Turk. Participants were paid \$6/hour prorated by the duration of the exper-  
210 iments (15 minutes). Participants only saw the experiment advertised, and could only  
211 participate in it, if (i) they were located within the US, (ii) had an approval rating of 99%  
212 or higher, (iii) met the software requirements (a recent version of the Chrome browser en-  
213 gine), and (iv) had not previously completed any other experiments on vowel perception in  
214 our lab. Before the experiment could be accepted, participants had to confirm that they  
215 were (i) native speakers of US English (defined as having spent their childhood until the  
216 age of 10 speaking English and living in the United States), (ii) in a quiet room without  
217 distractions, (ii) wearing over-the-ear headphones. Participants' responses were collected via  
218 Javascript developed by the Human Language Processing Lab at the University of Rochester  
219 ([Kleinschmidt \*et al.\*, 2021](#)).

220 An optional post-experiment survey recorded participant demographics using NIH pre-  
221 scribed categories, including participant sex (Male: 27, Female: 20), age (mean = 35.5 years;  
222 SD = 11.4; 95% quantiles = 24-63.25 years), race (White: 36, Asian: 3, Black: 6, multiple:

<sup>223</sup> 1, declined to report: 1), and ethnicity (Non-Hispanic: 42, Hispanic: 4, declined to report:  
<sup>224</sup> 1).

<sup>225</sup> **2. Materials**

<sup>226</sup> Experiment 1a employed *hVd* word recordings by one adult female talker from a photo-  
<sup>227</sup> netically annotated database of L1-US English vowel productions (Xie and Jaeger, 2020).  
<sup>228</sup> Specifically, we used all 9 recordings of each of the eight *hVd*-words—*heed*, *hid*, *head*, *had*,  
<sup>229</sup> *hut*, *odd*, *who'd*, *hood* (the use of “hut” and “odd” rather than “hud” and “hod” follows  
<sup>230</sup> Assmann *et al.*, 2008; but see Hillenbrand *et al.*, 1995).

<sup>231</sup> The stimuli for Experiment 1b were synthesized from a single *had* recording used in  
<sup>232</sup> Experiment 1a. Specifically, we used a script (based on descriptions in Wade *et al.*, 2007)  
<sup>233</sup> in Praat (Boersma and Weenink, 2022) to concatenate the original /h/ with a synthesized  
<sup>234</sup> vowel and the original /d/ recording. Unlike in Experiment 1a, all eight words thus had an  
<sup>235</sup> *hVd* context (including “hud” and “hod”, rather than “hut” and “odd”). The Praat script  
<sup>236</sup> first segmented the original *had* token into /h/, /ae/ and /d/ portions. It then filtered  
<sup>237</sup> the /h/ sound inversely with its LPC, and concatenated this neutral fricative sound with  
<sup>238</sup> a complex waveform generated from the pitch and intensity patterns of the original vowel,  
<sup>239</sup> to create a neutral hV-section that did not reflect any vocal tract resonances. The script  
<sup>240</sup> then created a formant grid that filtered the hV-section to create the intended vowel, and  
<sup>241</sup> finally concatenated this segment to the final /d/ to create an *hVd* word. For each *hVd*  
<sup>242</sup> word, the formant grid was populated with the F1, F2 and F3 values that we handed to the  
<sup>243</sup> script at five time-points transitioning from the /h/ to the vowel, to the final /d/ through

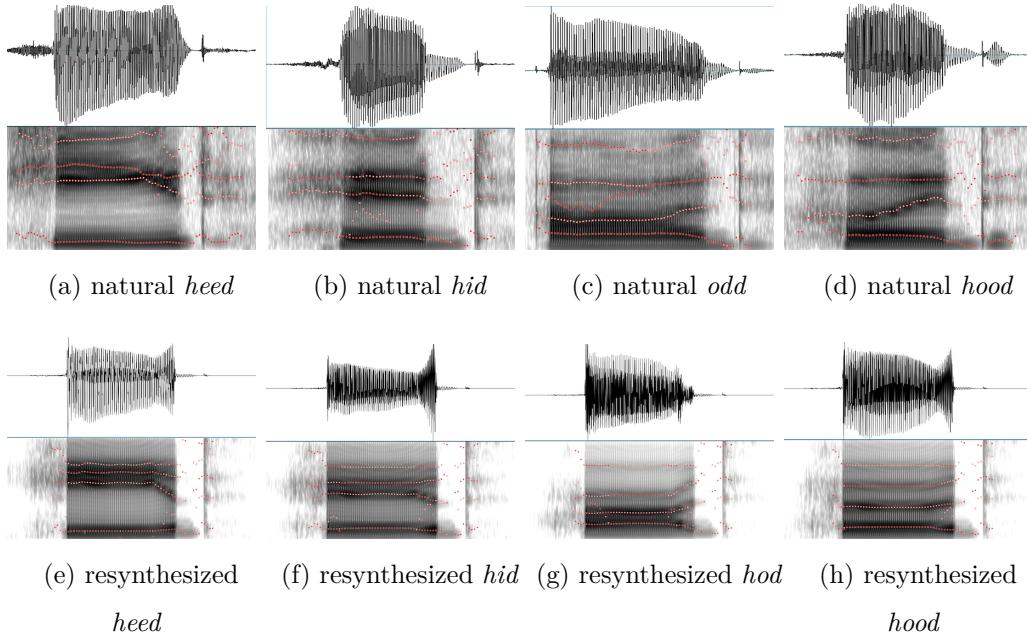


FIG. 3. **Top:** Spectrograms of four natural recordings from Experiment 1a. **Bottom:** Same for four synthesized tokens with similar formant values from Experiment 1b.

244 linear interpolation. Formant bandwidths were 500 Hz at the initial two time-points (the  
 245 /h/ and beginning of transition to vowel), and then decreased linearly during vowel onset  
 246 and throughout the final three time-points to 50 Hz (F1), 100 Hz (F2), 200 Hz (F3), 300  
 247 Hz (F4), and 400 Hz (F5-F8, following Wade *et al.*, 2007). The bandwidth manipulation  
 248 implied that formants became stronger as the vowel unfolded (see Figure 3). We used this  
 249 approach to create synthesized vowels for arbitrary F1-F2 combinations. F3 was set based  
 250 on those F1-F2 values. Specifically, we ran a linear regression over the natural productions  
 251 of the talker from Experiment 1a, predicting F3 from F1, F2 and their interaction. We then  
 252 used that regression to predict F3 values for any F1-F2 combination in Experiment 1b. F4  
 253 to F8, as well as vowel duration, were held identical across all tokens (using the same values  
 254 as Wade *et al.*, 2007).

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

264 **3. Procedure**

265 The procedure for both experiments was identical. Live instances of each experiment  
266 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  
267 met all requirements and provided consent, as per the Research Subjects Review Board of  
268 the University of Rochester. Before starting the experiment, participants performed a sound  
269 check. Participants were then instructed to listen to a female talker saying words, and click  
270 on the word on screen to report what word they heard. On each trial, all eight *hVd*-words  
271 were displayed on screen. Half of the participants in each experiment saw the response  
272 options organized as in Figure 2 (resembling the IPA representation of a vowel space), half  
273 saw the response options in the opposite order (flipping top and bottom and left and right  
274 in Figure 2). Each trial started with the response grid on screen, together with a light green  
275 in Figure 2).

276 dot centered on screen. After 1000 ms, an *hVd* recording played, and participants indicated  
277 their response by a mouse-click. After a 1000 ms intertrial interval, the screen reset, and  
278 the next trial started.

279 In both experiments, participants heard two blocks of the materials described in the  
280 previous sections, for a total of 144 trials in Experiment 1a and 292 trials in Experiment 1b.  
281 Presentation within each block was randomized for each participant. Participants were not  
282 informed about the block structure of the experiment.

283 After completing the experiment, participants filled out a language background question-  
284 naire and the optional demographic survey. On average, participants took 10.3 minutes to  
285 complete Experiment 1a ( $SD = 6.6$ ) and 18.4 minutes for Experiment 1b ( $SD = 7.3$ ).

286 **4. Exclusions**

287 We excluded participants who failed to follow instructions and did not wear over-the-ear  
288 headphones (as indicated in the post-experiment survey). We also excluded participants  
289 with mean (log-transformed) reaction times that were unusually slow or fast (absolute z-  
290 score over by-participant means  $> 3$ ), or if they clearly did not do the task (e.g., by answering  
291 randomly). This excluded 6 participants from Experiment 1a and 2 from Experiment 1b  
292 (for details, see [§2 A](#)).

293 We further excluded all trials that were unusually fast or slow. Specifically, we first z-  
294 scored the log-transformed response times *within each participant* and then z-scored these  
295 z-scores *within each trial* across participants. Trials with absolute z-scores  $> 3$  were removed  
296 from analysis. This double-scaling approach was necessary as participants' response times

decreased substantially over the first few trials and then continued to decrease less rapidly throughout the remainder of the experiment. The approach removes response times that are unusually fast or slow *for that participant at that trial*, while avoiding specific assumptions about the shape of the speed up in response times across trials. This excluded 1.2% of the trials in Experiment 1a and 1.1% in Experiment 1b. This left for analysis 2565 observations from 18 participants in Experiment 1a, and 6354 observations from 22 participants in Experiment 1b.

## 304 B. Results

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

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

Unsurprisingly, participants in both experiments divided the F1-F2 space into the eight vowel categories in ways that qualitatively resembled each other (after taking into account that Experiment 1b covers a larger range of F1-F2 values). Also unsurprisingly, there were some differences between participants' responses across the two experiments, at least when plotted in Hz. For example, [u] rarely was the most frequent response in Experiment 1b, even for stimuli that were predominantly categorized as [u] in Experiment 1a. There are at

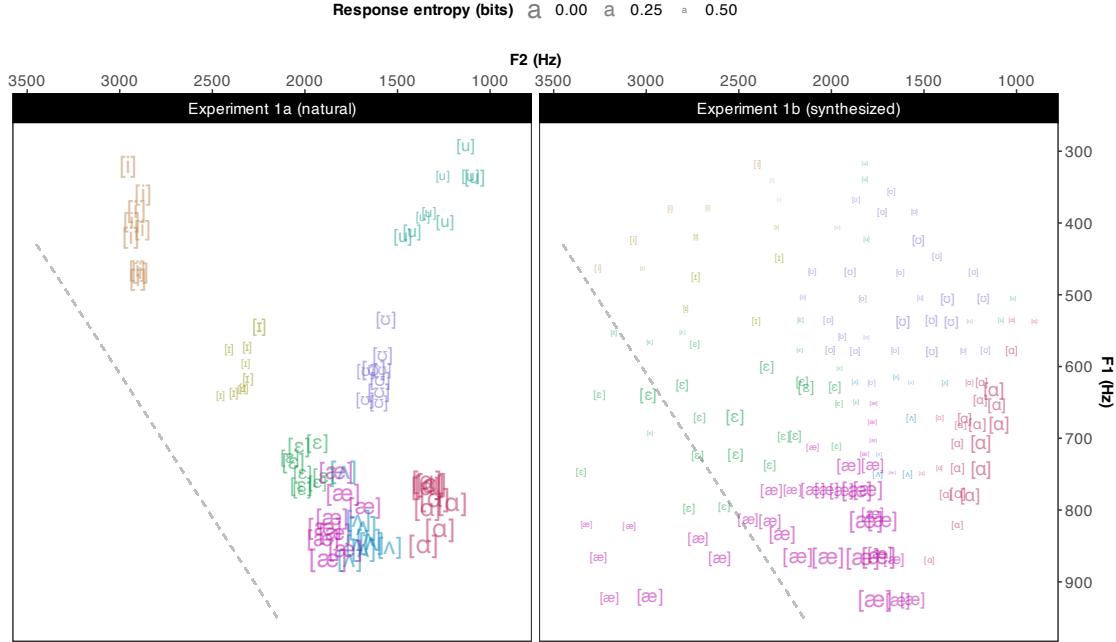


FIG. 4. Summary of listeners' categorization responses in Experiments 1a and 1b in F1-F2 space. The vowel label indicates the most frequent response provided across participants on each test location. Size indicates how consistent responses were across participants, which larger symbols indicating more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

317 least two reasons to expect such differences. First, stimuli with similar F1-F2 values across  
318 the two experiments still differed in other acoustic properties (e.g. vowel duration or F3).  
319 These acoustic differences might have affected participants' responses. Second, it is possible  
320 that *formant normalization* affected participants' responses—i.e., the very mechanism we  
321 seek to investigate in the remainder of the paper. The two experiments differ in the means,  
322 variances, and other statistical properties that some normalization accounts predict to affect

323 perception. As a consequence, Hz might not be the space in which we should expect identical  
 324 responses across experiments.

325 Similarly, the two experiments differed in the extent to which participants agreed with  
 326 each other. Participants in Experiment 1b exhibited overall less agreement in their responses  
 327 (mean by-item response entropy = 0.45 bits, SE = 0.01) than participants in Experiment  
 328 1a (mean by-item response entropy = 0.23 bits, SE = 0.02). This was expected given that  
 329 Experiment 1b explored the entire F1-F2 space, including—by design—formant combina-  
 330 tions located *between* the centers of the natural vowel categories. Experiment 1b therefore  
 331 achieved its goal of eliciting less categorical response distributions, which is expected to  
 332 facilitate comparison of competing normalization accounts.<sup>5</sup>

333 Auxiliary analyses presented in the SI (§2C) suggest that *some but not all* of the dif-  
 334 ferences in response entropy between the two experiments were caused by the placement  
 335 of the stimuli in formant space: when comparing categorization responses for tokens from  
 336 the two experiments with similar acoustic properties (differences of  $\leq 30$  Hz along F1 and  
 337 F2), response entropies still differed substantially (for  $N = 40$  acoustically similar tokens,  
 338 mean by-item response entropy for Experiment 1a = 0.18 bits, SE = 0.03; Experiment 1b  
 339 = 0.39 bits, SE = 0.03). We see two mutually compatible explanations. First, similar to  
 340 the differences between experiments in the dominant response pattern discussed above, dif-  
 341 ferences in the degree of agreement between participants might originate in *normalization*.  
 342 Second, it is possible that the relation between formants in the synthesized stimuli or some  
 343 other unknown acoustic-phonetic differences between the experiments explain the difference  
 344 in response. For example, the absence of vowel inherent spectral change (VISC) or differ-

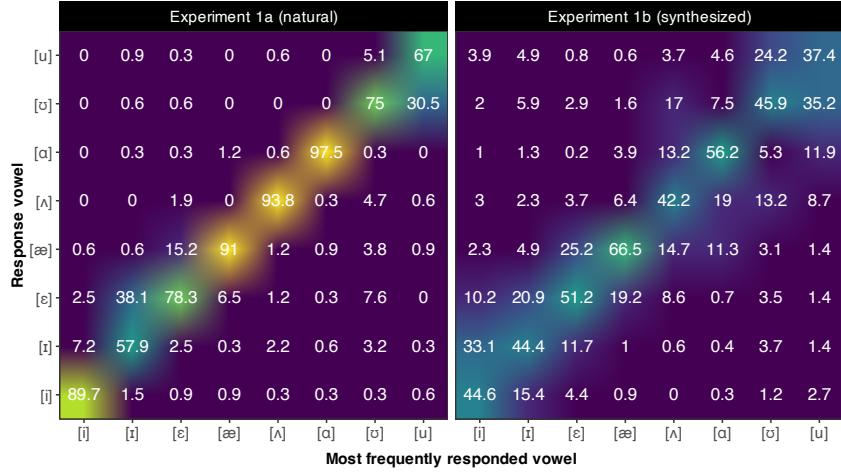
345 ences in tilt in the synthesized stimuli might have deprived listeners of information that is  
 346 actually crucial for establishing phonemic identity (Hillenbrand and Nearey, 1999). This  
 347 would result in increased uncertainty on each trial, leading to increased entropy of listeners'  
 348 responses. The computational study we present below shed some light on these two mutually  
 349 compatible possibilities.

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

351 Since the intended category was known for Experiment 1a, it was possible to calculate  
 352 participants' recognition accuracy. As also evident in the left panel of Figure 4, participants'  
 353 most frequent response *always* matched the intended vowel in Experiment 1a. Overall,  
 354 participants' responses matched the intended vowel on 81.2% (SE = 4.8%) of all trials  
 355 (Experiment 1b had no such ground truth). This is much higher than chance (12.5%). It is,  
 356 however, also quite a bit lower than 100%. To better understand the reasons for this, Figure  
 357 5A plots the confusion matrix. This suggests that participants' performance was largely  
 358 affected by confusions between [i]-to-[ɛ] (*hid-to-head*), [ɛ]-to-[æ] (*head-to-had*), and [u]-to-[ʊ]  
 359 (*who'd-to-hood*).

360 One plausible explanation for this pattern of vowel confusions lies in the substantial  
 361 variation that exists across US English dialects (Labov *et al.*, 2006). Differences in the  
 362 realization of vowel categories, and associated representations, across dialects will directly  
 363 affect the expected classification for any given token. In addition, listeners might differ in  
 364 terms of experience with different dialects, or in the dialect they attribute to the talker who  
 365 produced the stimuli. To test this hypothesis, we calculated the [i]-to-[ɛ], [ɛ]-to-[æ], and

A



B

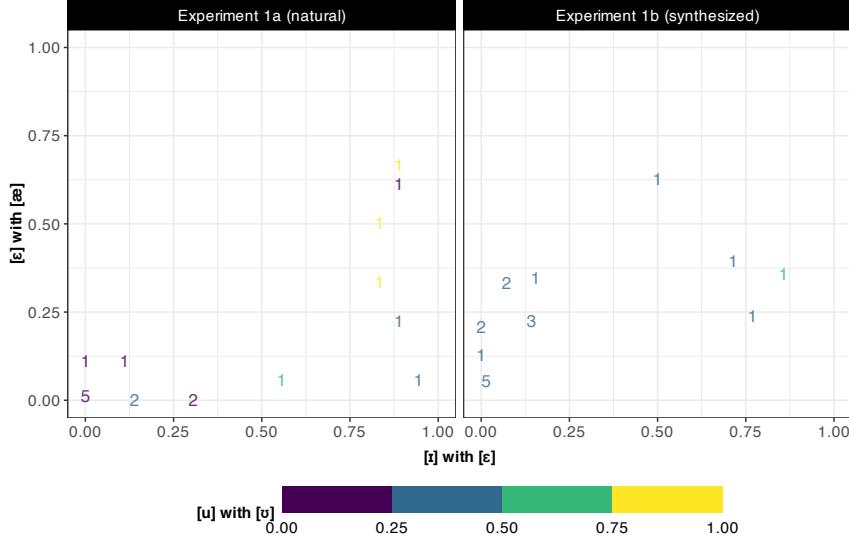


FIG. 5. Category confusability in Experiments 1a and 1b. **Panel A** summarizes the category confusability. Since correct responses were not defined for Experiment 1b, we grouped items along the x-axis based on most frequent response that listeners provided (for Experiment 1a, this was always identical 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 [o] (color fill).

366 [u]-to-[u] confusion rates for each participant in Experiment 1a. These data are summarized

367 in the left panel of Figure 5B. The data in the left panel suggest that most participants in

368 Experiment 1a either heard [i] tokens consistently as the intended [i] (clustering on the left  
 369 side of the panel) or as [ɛ] (clustering on the right side of the panel). Only a few participants  
 370 exhibited mixed responses for items intended to be [i]. Tellingly, many of the participants  
 371 who exhibited increased [i]-to-[ɛ] confusion *also* exhibited increased [ɛ]-to-[æ] confusion. This  
 372 is precisely what would be expected by listeners who assume a dialect in which these vowels  
 373 are articulated lower (with higher F1) than in the dialect of the talker in Experiment 1a.

374 A similar, but less pronounced, pattern was also found with regard to [u]-to-[ʊ] confusions.<sup>6</sup>  
 375 Finally, a qualitatively similar relation between [i]-to-[ɛ], [ɛ]-to-[æ], and [u]-to-[ʊ] confusions  
 376 was also observed in Experiment 1b (right panel of Figure 5B), though the pattern was  
 377 unsurprisingly less pronounced given that the stimuli in Experiment 1b by design often  
 378 fell into the ambiguous region *between* vowels. Taken together, Experiments 1a and 1b  
 379 thus suggest that systematic dialectal differences between participants may be a substantial  
 380 contributor of the relatively low correct classification rate observed for Experiment 1a.

381 This highlights two important points. First, the data from Experiment 1a demonstrate  
 382 the perceptual challenges associated with an unfamiliar talker: in the absence of lexical or  
 383 other context to distinguish between the eight available response options, listeners can only  
 384 rely on the acoustic information in the input. In such a scenario, even listeners who are  
 385 in principle familiar with the dialect spoken by the talker have comparatively little infor-  
 386 mation to determine the talker's dialect, making apparent what Matt Winn (2018) aptly  
 387 summarizes as “speech [perception] is not as acoustic as [we] think”. Second, when dialect  
 388 variability is taken into account, listeners' recognition accuracy improved substantially. Af-  
 389 ter removing 7 listeners who heard more than 50% of the [i] items as [ɛ], *all* vowels were

390 correctly recognized at least 88.3% of the time (overall accuracy = 95.9%). This suggests  
391 that dialect differences affected the recognition of all vowels. This aspect of our results serves  
392 as an important reminder that formant normalization is only expected to erase inter-talker  
393 variability associated with *physiological* differences: variation in dialect, sociolect, or other  
394 non-physiologically-conditioned variation pose separate challenges to human perception, and  
395 require additional mechanisms (see discussion in [Barreda, 2021](#); [Weatherholtz and Jaeger,](#)  
396 [2016](#)). This introduces noise—variability in listeners' responses that cannot be accounted  
397 for by normalization—to any comparison of normalization accounts, potentially reducing  
398 the power to detect differences between accounts.

399 **III. COMPARISON OF NORMALIZATION ACCOUNTS**

400 In order to evaluate normalization accounts against speech perception, it is necessary to  
401 map the phonetic properties of stimuli—under different hypotheses about normalization—  
402 onto listeners' responses in Experiments 1a and 1b. Previous work has done so by directly  
403 predicting listeners' responses from the raw or normalized phonetic properties of stimuli  
404 ([Apfelbaum and McMurray, 2015](#); [Barreda, 2021](#); [Crinnion \*et al.\*, 2020](#); [McMurray and](#)  
405 [Jongman, 2011](#); [Nearey, 1989](#)). For example, McMurray and Jongman used multinomial  
406 logistic regression to predict 8-way fricative categorization responses in US English (see also  
407 [Barreda, 2021](#)).

408 Here we pursued an alternative approach by committing to a core assumption common to  
409 contemporary theories of speech perception: that listeners acquire implicit knowledge about  
410 the probabilistic mapping from acoustic inputs to linguistic categories, and draw on this

411 knowledge during speech recognition (e.g., TRACE, McClelland and Elman, 1986; exem-  
 412 plar theory, Johnson, 1997; Bayesian accounts, Luce and Pisoni, 1998; Nearey, 1990; Norris  
 413 and McQueen, 2008; ASR-inspired models like DIANA or EARSHOT, ten Bosch *et al.*,  
 414 2015; Magnuson *et al.*, 2020). Using a general computational framework for adaptive speech  
 415 perception (ASP, Xie *et al.*, 2023) we trained Bayesian ideal observers to capture the expec-  
 416 tations that a ‘typical’ L1 adult listener might have about the formant-to-vowel mappings of  
 417 US English. We approximated these expectations using a database of L1-US English vowel  
 418 productions (Xie and Jaeger, 2020)—transformed to reflect the different normalization ac-  
 419 counts. We then ask which of the different ideal observer models—corresponding to different  
 420 hypotheses about formant normalization—best predicts listeners’ responses in Experiments  
 421 1a and 1b.

422 A welcome side effect of this is that far fewer degrees of freedom (DFs) are required  
 423 to predict listeners’ responses. For example, using ordinary multinomial logistic regression  
 424 trained on our perceptual data to predict 8-way vowel categorization as a function of F1,  
 425 F2 and their interaction would require up to 28 DFs. This problem increases with the  
 426 number of cues considered. Because the model is trained on data that is independent of  
 427 our perceptual data, the ASP-based approach we employ instead uses only 2 DFs (i.e.,  
 428 parameters estimated based on our perceptual data) to mediate the mapping from stimuli  
 429 properties to listeners’ responses, regardless of the number of cues considered. Over the  
 430 next few sections, we describe how this parsimony is made possible through a commitment  
 431 to strong linking hypotheses motivated by theories of speech perception.

432

## A. Methods

433

### 1. A general-purpose categorization model for J-AFC categorization tasks

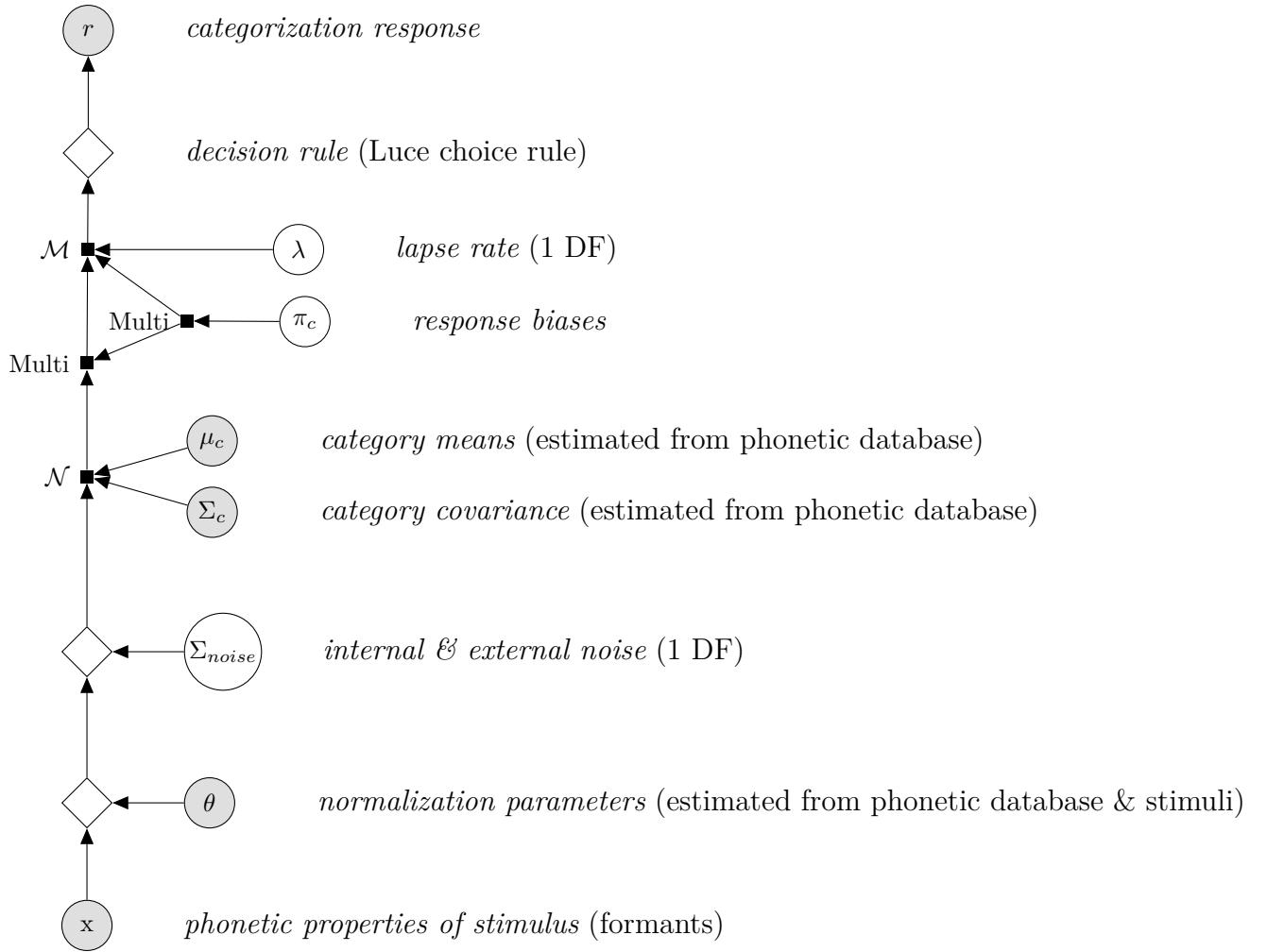


FIG. 6. Graphical model of ASP's general categorization framework (adapted for the current purpose from Xie et al., 2023, Figure 4). Here  $J = 8$  (the eight vowel response options in Experiments 1a and 1b). We use this framework to compare normalization accounts against listeners' categorization responses 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), Multi(nomial), and  $\mathcal{M}$ (ixture) distributions.

434 Figure 6 summarises ASP's categorization model for a  $J$ -alternative forced-choice task  
 435 (for an in-depth description, we refer to [Xie et al., 2023](#)). The model combines Bayesian ideal  
 436 observers (as used in e.g., [Clayards et al., 2008](#); [Feldman et al., 2009](#); [Norris and McQueen, 2008](#);  
 437 [Xie et al., 2021](#); for a closely related approach, see also [Nearey and Hogan, 1986](#)) with  
 438 psychometric lapsing models ([Wichmann and Hill, 2001](#)). To reduce researchers' degrees of  
 439 freedom, we adopt all assumptions made in [Xie et al. \(2023\)](#), and do not introduce additional  
 440 assumptions.

441 Starting at the bottom of the figure, the phonetic input  $x$  is normalized. Here,  $x =$   
 442 the F1 and F2 of our stimuli (the SI, §3 E reports additional analyses that instead employ  
 443 F1-F3; these analyses support the same conclusion presented here, and we mention them  
 444 below where relevant). The specific computations applied to the input  $x$  depend on the  
 445 normalization accounts (see Table I). We use  $\theta$  to refer to the parameters required by the  
 446 normalization account. For example, for the uniform scaling account ([Nearey, 1978](#)),  $\theta$  is the  
 447 overall mean of all log-transformed formants. For Lobanov normalization ([Lobanov, 1971](#)),  
 448  $\theta$  is a vector of means and standard deviations for each formant (in Hz).

449 The normalized input is then perturbed by perceptual and environmental noise. Following  
 450 [Feldman et al. \(2009\)](#), this noise is assumed to be Gaussian distributed centered around the  
 451 transformed stimulus with noise variances that are independent and identical for all formants  
 452 (i.e.,  $\Sigma_{noise}$  is a diagonal matrix, and all diagonal entries have the same value). Next, the  
 453 likelihood of the normalized percept under each of the eight vowel categories is calculated,  
 454  $p(F1, F2|vowel)$ . This requires specifying listeners' expectations about the cue-to-category  
 455 mapping (listeners' likelihood function). We followed [Xie et al. \(2023\)](#) and previous work and

456 assume that each vowel maps onto a multivariate Gaussian distribution over the phonetic  
 457 cues, here bivariate Gaussians over F1 and F2 (cf. [Clayards \*et al.\*, 2008](#); [Feldman \*et al.\*, 2009](#);  
 458 [Kleinschmidt and Jaeger, 2015](#); [Norris and McQueen, 2008](#); [Xie \*et al.\*, 2021](#)). The posterior  
 459 probability of each vowel is obtained by combining its likelihood with its prior probability  
 460 or response bias  $\pi_c$ , according to Bayes theorem:<sup>7</sup>

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

461 Up to this point, the model is identical to a standard Bayesian ideal observer over noisy  
 462 input ([Feldman \*et al.\*, 2009](#); [Kronrod \*et al.\*, 2016](#)) for which the input has been transformed  
 463 based on the normalization account. ASP's categorization model adds to this the potential  
 464 that participants experience attentional lapses—or for other reasons do not respond based  
 465 on the input—on some proportion of all trials ( $\lambda$ , as in standard psychometric lapsing  
 466 models, [Wichmann and Hill, 2001](#)). On those trials, the posterior probability of a category  
 467 is determined solely by participants' response bias, which we assume to be identical to the  
 468 response bias on non-lapsing trials (following [Xie \*et al.\*, 2023](#)). This results in a posterior  
 469 that is described by weighted mixture of two components, describing participants' posterior  
 470 on non-lapsing and lapsing trials, respectively:

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

471 Finally, a decision rule is applied to the posterior to determine the response of the model,  
 472 conditional on the input (one of the eight vowels in Experiments 1a and 1b). We followed

473 the gross of research on speech perception and assume Luce's choice rule (Luce, 1959; for  
 474 discussion, see Massaro and Friedman, 1990). Under this choice rule, the model can be  
 475 seen as sampling from the posterior, responding with each category proportional to that  
 476 category's posterior probability.

477 Next, we describe how we estimated the  $\theta$ s,  $\mu_c$ s and  $\Sigma_c$ s for each normalization account  
 478 from a phonetic database. We use this database as a—very coarse-grained—approximation  
 479 of a the speech input a ‘typical’ listener might have experienced previously. By fixing  $\theta$ ,  $\mu_c$   
 480 and  $\Sigma_c$  based on the distribution of phonetic cues in the database, we substantially reduce  
 481 the DFs that are allowed to mediate the mapping from stimulus properties to listeners’  
 482 responses (following Xie *et al.*, 2023). In addition, this approach naturally penalizes overly  
 483 complex models by validating these against out-of-sample data. Finally, we describe how  
 484 we fit the remaining parameters as DFs to participants’ responses from Experiments 1a and  
 485 1b.

486 **2. Modeling listeners’ prior experience (and guarding against overfitting):  $\theta$ ,  $\mu_c$ ,**  
 487 **and  $\Sigma_c$**

488 By fixing  $\theta$ ,  $\mu_c$ , and  $\Sigma_c$  based on a database of vowel *productions*, we impose strong  
 489 constraints on the functional flexibility of the model in predicting listeners’ responses. This  
 490 benefit is made possible by committing to a strong linking hypothesis—that listeners’ cate-  
 491 gories are learned from, and reflect, the distributional mapping from formants to vowels in  
 492 previously experienced speech input (e.g., Abramson and Lisker, 1973; Massaro and Fried-  
 493 man, 1990; Nearey and Hogan, 1986). The database we use to approximate listeners’ prior

494 experience was originally developed to compare the production of L1 and L2 speakers (Xie  
 495 and Jaeger, 2020). It contains 9-10 recordings of the 8 *hVd* words from each of 17 (5 fe-  
 496 male) L1 talkers of a Northeastern dialect of US English (ages 18 to 35 years old). Since  
 497 Experiments 1a and 1b used recordings of one of these talkers, we excluded that talker prior  
 498 to fitting training ideal observers on the data. In total, this yields 5842 recordings that are  
 499 annotated for F0, F1-F3, and vowel duration. The SI (§3 A 1) summarizes the distribution  
 500 of these cues, and how the different normalization accounts affect those distributions.

501 To avoid over-fitting the ASP model to the database, we used 5-fold cross-validation:  
 502 we randomly split the Xie and Jaeger (2020) database into five approximately evenly-sized  
 503 folds (following Persson and Jaeger, 2023). This split was performed within each vowel to  
 504 guarantee that all five folds had the same relative amount of data for each vowel category.  
 505 These splits were combined into five training sets, each containing one of the folds (20% of  
 506 the data). This way, each training set was different from the others, increasing the variability  
 507 between sets.<sup>8</sup>

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

513 This yielded 100 ideal observer models, 5 for each of the 20 normalization accounts in  
 514 Table I. Of note, the 20 ideal observers fit on each fold differ *only* in the assumptions  
 515 they make about the normalization that is applied to cues before they are mapped onto

516 the eight vowel categories. Figure 7 visualizes the resulting bivariate Gaussian categories  
 517 for four of the 20 normalization accounts. This illustrates one advantage of the cross-  
 518 validation approach: it takes a modest step towards simulating differences across listeners'  
 519 prior experience (represented by the five different folds).

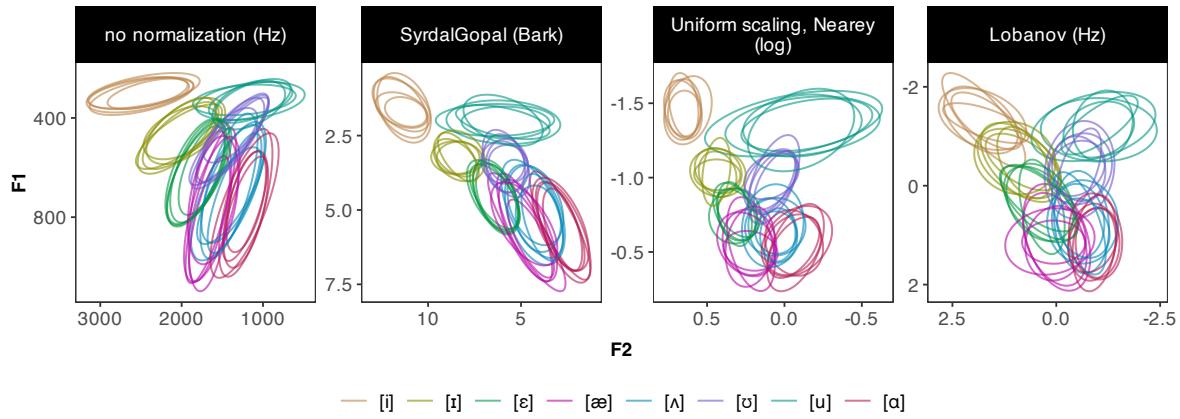


FIG. 7. Visualizing the bivariate Gaussian categories (prior to adding  $\Sigma_{noise}$ ) of four example normalization accounts in F1-F2 space. Separate ellipses are shown for each of the five training sets (each set corresponds to one set of eight ellipses). The relative stability of the category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

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

522 Next, we transformed the stimuli of Experiments 1a and 1b into the formant space defined  
 523 by the 20 normalization accounts in Table I. This requires estimating the required normal-

524 ization parameters  $\theta$  for each experiment and normalization account. We calculated these  $\theta$ s  
 525 over all stimuli (of each experiment and normalization account). For example, for the uni-  
 526 form scaling account (Nearey, 1978), we calculated the overall mean of all log-transformed  
 527 formants over all stimuli. For Lobanov normalization (Lobanov, 1971), we calculated the  
 528 mean and standard deviation of each formant (in Hz) over all stimuli. For each combina-  
 529 tion of experiment and normalization account, we then normalized the stimuli using those  
 530 parameter estimates.

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

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

534 Finally, we describe the two parameters of the ASP model that we fit against listeners'  
 535 responses in Experiments 1a and 1b. These two parameters constitute the only DFs that  
 536 mediate the link from ideal observers' predictions to listeners' responses, and which are  
 537 specifically tuned to these. The first DF ( $\Sigma_{noise}$ ) models the effects of internal (perceptual)  
 538 and external (environmental) noise on listeners' perception. While previous work provides  
 539 estimates of the internal noise in formant perception, these estimates were obtained under  
 540 *assumptions* about the relevant formant space. For example, Feldman *et al.* (2009) estimated  
 541 the internal noise variance to be about 15% of the average category variance along F1 and F2.  
 542 This estimate was based on the assumption that human speech perception transforms vowel  
 543 formants into Mel, without further normalization. Since we aim to *test* which normalization  
 544 account best explains speech perception, we cannot rely on this or other internal noise

545 estimates obtained under a single specific assumption. Additionally, internal noise can vary  
 546 across individuals and external noise can vary across environments (a point particularly  
 547 noteworthy, given that we conducted Experiments 1a and 1b over the web). We thus allowed  
 548 the noise variance  $\Sigma_{noise}$  to vary in fitting participants' responses. Following [Feldman \*et al.\* \(2009\)](#), we assumed that perceptual noise had identical effects on all formants in the phonetic  
 549 space defined by the normalization account (see also [Kronrod \*et al.\*, 2016](#)). This reduces  
 550  $\Sigma_{noise}$  to a single DF, regardless of the normalization account (for details, see SI [§3 A 3](#)).

552 The magnitude of  $\Sigma_{noise}$  affects the slope of the categorization functions that predict  
 553 listeners' responses from stimulus properties (here, F1 and F2): higher  $\Sigma_{noise}$  imply more  
 554 shallow categorization slopes. To facilitate comparison of  $\Sigma_{noise}$  values across normaliza-  
 555 tion accounts, we report results in terms of the best-fitting *noise ratios* ( $\tau^{-1}$ ), rather than  
 556  $\Sigma_{noise}$ s. Specifically,  $\Sigma_{noise}$  is best understood *relative* to the inherent variability of the  
 557 vowel categories ( $\Sigma_c$ ). This variability in turn depends on the phonetic space defined by  
 558 the normalization account. We thus divide  $\Sigma_{noise}$  by the mean of the diagonals of all  $\Sigma_c$ s to  
 559 obtain the *noise ratio*  $\tau^{-1}$ . For example, noise ratio of 0 corresponds to the absence of any  
 560 noise, and a noise ratio of 1 corresponds to noise variance of the same magnitude as the av-  
 561 erage category variance along F1 and F2 in the phonetic space defined by the normalization  
 562 account.<sup>10</sup> Figure 8B illustrates the effects of this noise ratio for Nearey's uniform scaling  
 563 account.

564 Second, participants can attentionally lapse or for other reasons reply without considering  
 565 the speech input. We thus allowed lapse rates ( $\lambda$ ) to vary while fitting human responses. This  
 566 introduces a second DF, which we fit against listeners' responses. Together, the inclusion

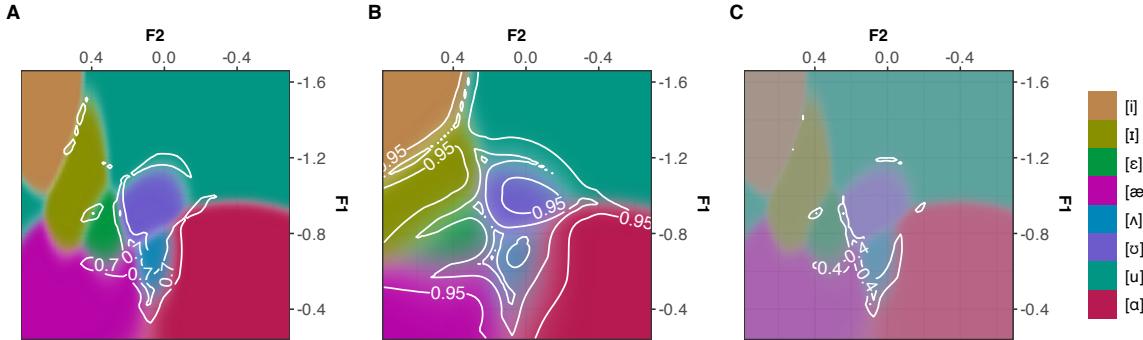


FIG. 8. Illustrating the consequences of perceptual and external noise ( $\Sigma_{noise}$ ) and attentional lapse rates ( $\lambda$ ) on the predicted posterior distribution of vowel categorizations. Shown are the average predicted posteriors across all five folds for Nearey's uniform scaling account. **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).

567 of freely varying lapse rates and a uniform response bias allows the ASP models to capture  
 568 that some unknown proportion of listeners' responses might be more or less random, rather  
 569 than reflecting properties of the vowel stimuli. This is illustrated in Figure 8C.

570 Finally, participants can have response biases that reflect their beliefs about the prior  
 571 probability of each category. However, to reduce the DFs fit to participants' responses, we  
 572 did *not* fit this response bias against listeners' responses (thus avoiding  $J - 1 = 7$  additional  
 573 DFs). Instead, we assumed uniform response biases—i.e., that listeners believed all eight

574 response options in the experiments to be equally likely ( $\forall c \pi_c = .125$ ). This decision implies  
 575 that our models would not be able to capture any potential non-uniformity in listeners'  
 576 response biases—including potential effects of additional acoustic differences (the absence  
 577 of [h] in *odd* or the coda [t], rather than [d] in *hut*) and orthographically particular response  
 578 options in Experiment 1a (“who’d”, “odd”, and “hut”). We do, however, see no reasons to  
 579 expect this decision to bias the comparison of normalization accounts.

580 **5. *Fitting normalization accounts to listeners’ responses***

581 For each of the 200 combinations of experiment, normalization account, training set,  
 582 we used constrained quasi-Newton optimization (Byrd *et al.*, 1995, as implemented in R’s  
 583 `optim()` function) to find the  $\lambda$  and  $\tau^{-1}$  values that best described listener’s responses.  
 584 Specifically, we used the 100 ideal observers described in the previous sections, applied them  
 585 to the normalized stimuli of the experiment, and determined which  $\lambda$  and  $\tau^{-1}$  maximized  
 586 the likelihood of listener’s responses (for details, see SI §3 A 3). This procedure yielded five  
 587 maximum likelihood estimates for both  $\lambda$  and  $\tau^{-1}$  for each combination of experiment and  
 588 normalization account—one for each training set. All results presented below were validated  
 589 and confirmed by grid searches over the parameter spaces (SI, §3 F).

590 We compare normalization accounts in terms of the likelihood of listeners’ responses  
 591 under these maximum likelihood estimates of  $\lambda$  and  $\tau^{-1}$ . Comparing accounts in terms  
 592 of their data likelihood, rather than the accuracy of predicting intended productions (e.g.,  
 593 Johnson, 2020; Persson and Jaeger, 2023), or correlations with human response proportions  
 594 (e.g., Hillenbrand and Nearey, 1999; Nearey and Assmann, 1986), follows more recent work

595 (e.g., [Barreda, 2021](#); [McMurray and Jongman, 2011](#); [Richter \*et al.\*, 2017](#); [Xie \*et al.\*, 2023](#))

596 and parallels standard approaches to model comparison in contemporary data analysis. We

597 note that this approach puts normalization accounts to a stronger test. For example, a

598 model can exhibit high correlations with listeners' responses even when its predictions are

599 systematically 'off'. Similarly, a model can achieve high accuracy in predicting listeners'

600 responses simply because it always predicts the most frequent response, and that response

601 accounts for sufficiently much of the data. In contrast, the likelihood of listeners' responses

602 under a model is a direct measure of how well the model captures the distribution of listeners'

603 responses conditional on the stimulus properties. In particular, data likelihood will be

604 maximized if, and only if, the model-predicted posterior probabilities of each vowel for each

605 stimulus are identical to the proportion with which those vowels occur in listeners' responses.

606 **B. Results**

607 We begin by comparing the fit of different accounts against listeners' responses in Ex-

608 periments 1a and 1b. Given the comparatively large number of accounts compared here, we

609 provide initial conclusions based on the best-fitting accounts along with the description of

610 the results (more in-depth discussion is provided in the general discussion). Following this

611 comparison, we visualize how different normalization accounts predict the formant space to

612 be divided into the eight vowel categories.

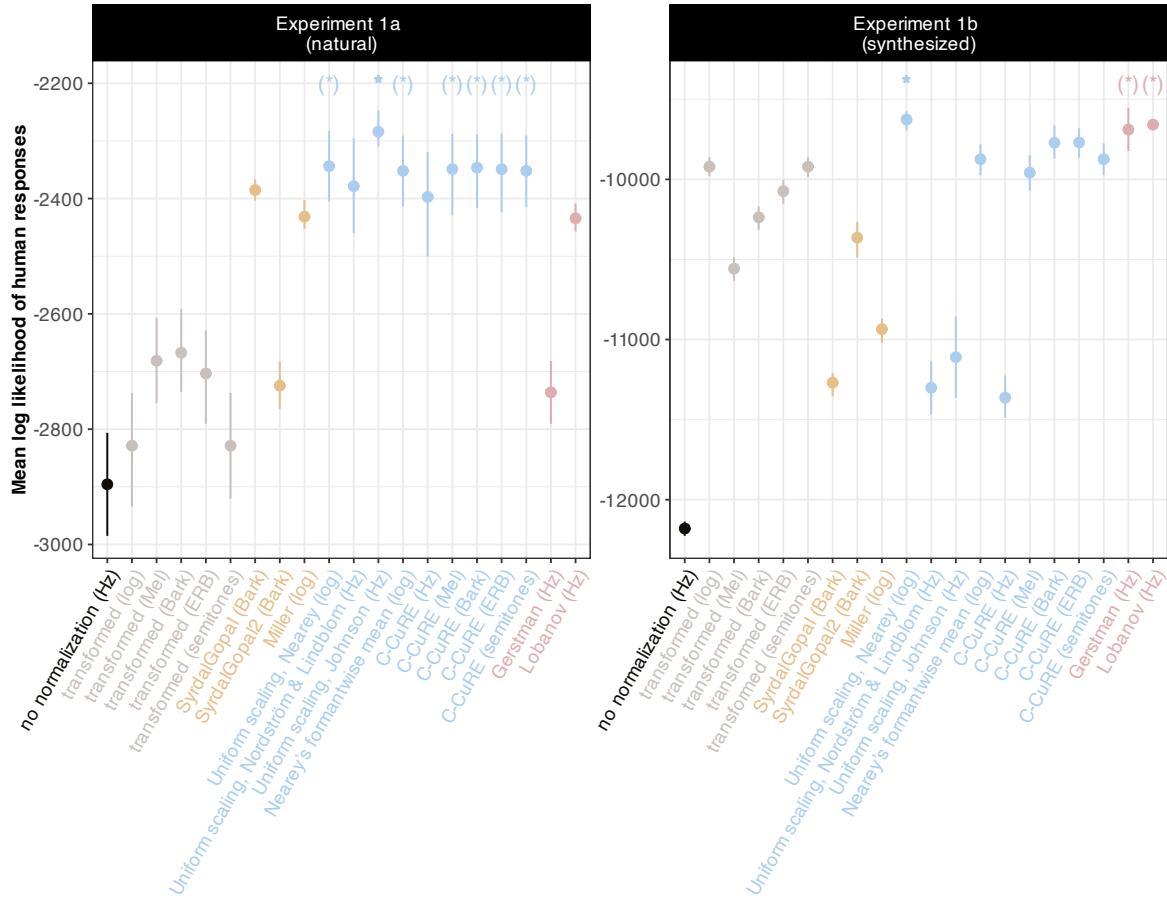


FIG. 9. Comparison of normalization accounts against listeners' responses. Pointranges indicate mean and 95% bootstrapped CIs of the log-likelihood summarized over the five training sets (higher is better). Accounts that fit listeners' responses to an extent that is statistically indistinguishable from the best-fitting account are marked by (\*). Note that y-axis range differs across panels, and that it is *not* meaningful to compare the absolute log-likelihood values across the two experiments (just as it is not meaningful to compare the data likelihood of regressions that are fit on two different data sets).

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

614      Figure 9 compares how well the different normalization accounts fit listeners' responses  
 615      in Experiments 1a and 1b. All accounts performed well above chance guessing (chance log

<sup>616</sup> likelihood in Experiment 1a: -5334; Experiment 1b: -13213) but also well below the highest  
<sup>617</sup> possible performance (in Experiment 1a, log-likelihood = -1348, in Experiment 1b: -7225).

<sup>618</sup> Normalization significantly improved the fit to listeners' responses relative to no normal-  
<sup>619</sup> ization. This was confirmed by paired one-sided *t*-tests comparing the maximum likelihood  
<sup>620</sup> values for each normalization account against those in the absence of normalization (all *ps*  
<sup>621</sup> < .05; see SI [§3 B 1](#)). Not all normalization accounts achieved equally good fits, however:  
<sup>622</sup> only some extrinsic accounts fit listeners' behavior well across both experiments. This sup-  
<sup>623</sup> ports two conclusions. First, it suggests that the normalization mechanisms operating during  
<sup>624</sup> human speech perception involve computations that go beyond static transformations into  
<sup>625</sup> psycho-acoustic spaces. Second, it suggests that the input to these computations is not  
<sup>626</sup> limited to intrinsic information—i.e., that the computations draw on information beyond  
<sup>627</sup> what is available in the acoustic signal *at that moment*. In particular, extrinsic normaliza-  
<sup>628</sup> tion requires the estimation and memory maintenance of talker-specific properties from the  
<sup>629</sup> speech signal.

<sup>630</sup> While the accounts that achieved the best fit against listeners' responses differed be-  
<sup>631</sup> tween experiments, both were variants of uniform scaling. For Experiment 1a, Johnson  
<sup>632</sup> normalization account provided the best fit (log likelihood = -2284, SD = 41 across the  
<sup>633</sup> five crossvalidation folds), while Nearey's uniform scaling account provided the best fit to  
<sup>634</sup> Experiment 1b (log likelihood = -9626, SD = 78). Both accounts essentially slide the repre-  
<sup>635</sup> sentational 'template' of a dialect—here the eight bivariate Gaussian categories of an ideal  
<sup>636</sup> observer—along a single line in the formant space. They differ only in *which* space this linear  
<sup>637</sup> 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, §3E).<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).

For both experiments, there were several accounts that fit listeners' responses similarly well as the best-fitting accounts ( $ps > .065$ ). All of these were extrinsic accounts, though the specific accounts differed between experiments. Notably, only Nearey's uniform scaling either provided the best fit (Experiment 1b) or achieved performance statistically indistinguishable from the best fit *for both experiments* (for Experiment 1a:  $p > 0.08$ , log likelihood = -2344,  $SD = 84$ ). Beyond the performance of Nearey's uniform scaling, there was little evidence of a correlation in relative ordering of accounts between experiments (Spearman rank  $r = 0.09$ ,  $p = 0.72$ ). Some accounts fit listeners' responses well for Experiment 1a, but not for Experiment 1b, and vice versa. Of note is the particularly variable performance of the centering accounts operating in Hertz space, i.e., C-CuRE Hz, Nordström & Lindblom and Johnson normalization. Similar variability across the two experiments is also observed for the two standardizing accounts, both of which operate in Hz space.

That an account operating over log-transformed formants—Nearey's uniform scaling—fits human behavior better should not be surprising. While questions remain about the exact organization of auditory formant representations, it is uncontroversial that the perceptual

660 sensitivity to acoustic frequency information is better approximated by a logarithmic scale  
 661 than by a linear scale (see [Moore, 2012](#)). As a result, a 30 Hz difference in an F1 of 300  
 662 Hz (a 10% change) is expected to be perceptually more salient than a 30 Hz change in an  
 663 F2 of 2500 Hz (a 1.2% change). In line with this reasoning, additional tests not reported  
 664 here found that Johnson normalization would provide the best fit to *both* experiments if  
 665 it was applied to log-transformed formants (instead of Hertz). In summary, variability in  
 666 how well different accounts predict human behavior across the two experiments highlights  
 667 the importance of psycho-acoustic transformations for human speech perception. This also  
 668 highlights the importance of comparing normalization accounts against multiple types of  
 669 data.

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

671 Before we turn to the general discussion, we briefly visualize how different normalization  
 672 mechanisms affect vowel categorization. This sheds light on *why* the accounts differ in  
 673 how well they fit listeners' responses. Figure 10 visualizes the categorization functions  
 674 predicted by four different normalization accounts, using the best-fitting  $\lambda$  and  $\tau^{-1}$  values  
 675 for each account (i.e., the values that lead to the fit shown in Figure 9). Figure 10 highlights  
 676 three points. First, a comparison across rows of Figure 10 shows how much the choice  
 677 of normalization can affect how the acoustic space gets carved up into vowel categories: a  
 678 comparison of the first (no normalization), third (Johnson), and fourth row (Lobanov) shows  
 679 that even normalization accounts operating over the same space can yield very different  
 680 categorization behavior.

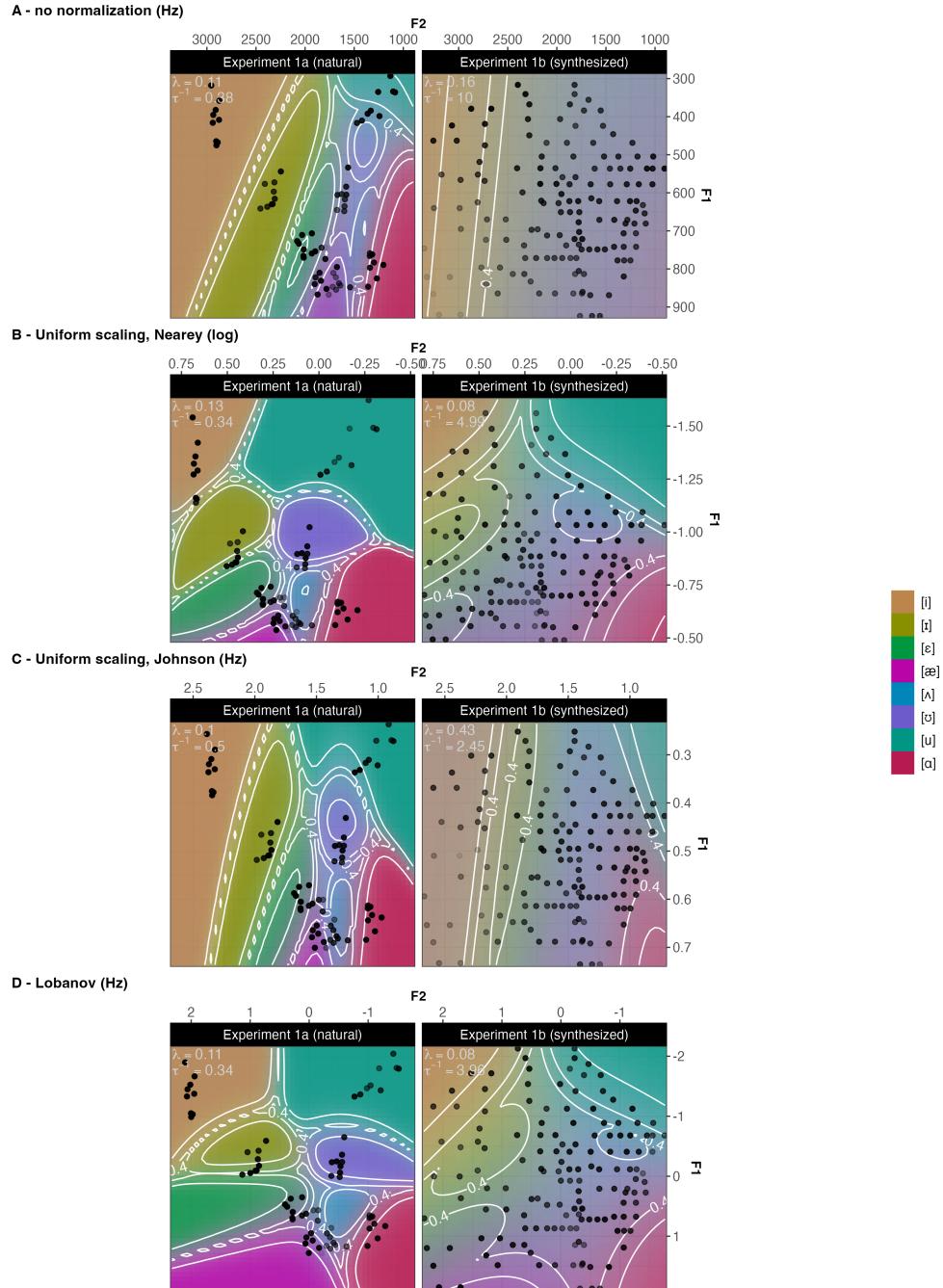


FIG. 10. Predicted categorization functions over the F1-F2 space under four different normalization accounts. For each account, we show the predicted posterior probabilities of all eight vowels obtained by averaging over the maximum likelihood parameterizations (of  $\lambda$  and  $\tau^{-1}$ ) for the five training sets (shown at top of each panel). **Top:** absence of normalization shown for reference. **2nd row:** the best-fitting account for Experiment 1a. **3rd row:** the best-fitting account for Experiment 1b (and second best for Experiment 1a). **Bottom:** the second best-fitting account in Experiment 1b. Contours indicate the highest posterior probability of any vowel. Points indicate location of test stimuli. The increasing opacity of points indicates a worse fit by the account (see text for detail).

681 Second, the best-fitting parameters (shown at the top of each panel) were relatively com-  
 682 parable across accounts but differed more substantially across experiments. Specifically,  
 683 the best-fitting estimates of lapse rates  $\lambda$  were generally comparable across the two exper-  
 684 iments (with the exception of Nordström & Lindblom and Johnson normalization, which  
 685 exhibited substantially higher lapse rates in Experiment 1b; SI §3B2). This suggests that  
 686 participants in both experiments were about equally likely to pay attention to the stimulus.  
 687 The best-fitting noise ratios  $\tau^{-1}$ , however, differed substantially across experiments, and  
 688 were 9 times larger for Experiment 1b (mean  $\tau^{-1} = 4.74$ , SD = 2.57 across normalization  
 689 accounts) than for Experiment 1a (mean  $\tau^{-1} = 0.52$ , SD = 0.49). This difference most likely  
 690 reflects the fact that the synthesized stimuli in Experiment 1b left listeners with substan-  
 691 tially more uncertainty about the intended category, as discussed during the description of  
 692 the experiments.

693 Since noise is assumed to be independent of category variability (see also Feldman *et al.*,  
 694 2009; Kronrod *et al.*, 2016), differences in noise ratios can substantially change the catego-  
 695 rization function. This is particularly evident for the accounts that had more variable per-  
 696 formance across the two experiments. For example, both Johnson (third row) and Lobanov  
 697 normalization (fourth row) resulted in very different best-fitting categorization functions for  
 698 Experiments 1a and 1b.

699 Third and finally, Figure 10 also shows how well accounts fit listeners' responses for each  
 700 test stimulus (opaqueness of the black points). This begins to explain *why* some accounts  
 701 fit listeners' responses in Experiment 1b less well. For example, the Johnson normalization  
 702 account (third row) predicts the responses to the test stimuli in Experiment 1a well, but

703 fails to predict the responses to the test stimuli in Experiment 1b. This drop in performance  
 704 seems to be primarily driven by stimuli that are unlikely to be articulated by the same talker  
 705 (lower left, cf. dashed line in Figure 4). This might suggest that this account was over-  
 706 engineered to explain naturally occurring productions—the type of data, it was originally  
 707 tested on (Johnson, 2020). A plausible account of normalization, however, should be able  
 708 to explain human perception to any type of stimulus, including synthesized stimuli. The SI  
 709 (§3 B 3) presents more detailed by-item comparisons of normalization accounts that might  
 710 be of interest to some readers.

711 **IV. GENERAL DISCUSSION**

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

721 Our results support three theoretical insights. First, human speech perception draws on  
 722 more than psycho-acoustic transformations or intrinsic information, in line with previous  
 723 research on normalization (Adank *et al.*, 2004; Ladefoged and Broadbent, 1957; Nearey,

724 1989). Rather, formant normalization seems to involve the estimation and storing of talker-  
725 specific formant properties. Second, computationally simple uniform scaling accounts pro-  
726 vide the best fit to listeners' responses, suggesting comparatively parsimonious maintenance  
727 of talker-specific properties. This replicates and extends previous findings that uniform scal-  
728 ing or similarly simple corrections for vocal tract size provide a better explanation for human  
729 perception than more complex extrinsic accounts (Barreda, 2021; Richter *et al.*, 2017). It is  
730 impossible to rule out more complex approaches to perceptual normalization given the large  
731 number of possible alternatives. However, given that uniform scaling provides a parsimo-  
732 nious explanation for human formant normalization, and the current absence of empirical  
733 evidence for more complex computations, we submit that researchers ought to adapt uni-  
734 form scaling as our working hypothesis. Third, the psycho-acoustic representation assumed  
735 by different normalization accounts matter, as indicated by the comparison of otherwise  
736 computationally similar accounts (e.g. Nearey's vs. Johnson's uniform scaling).

737 These results contribute to a still comparatively small body of work that has evaluated  
738 competing normalization accounts against listeners' perception, whereas most previous work  
739 evaluates accounts against intended productions. Complementing previous work, we took  
740 a broad-coverage approach: the present study compared 20 of the most influential nor-  
741 malization accounts against listeners' perception of *hVd* words with all eight US English  
742 monophthongs in both natural and synthesized speech. This contrasts with previous work,  
743 which has typically focused on subsets of the vowel system, either using natural *or* synthe-  
744 sized speech, and considering a much smaller subset of accounts (typically 2-3 at a time). By  
745 considering a wider range of accounts, a wider range of formant values and vowel categories,

<sup>746</sup> and multiple types of speech, we aimed to contribute to a more comprehensive evaluation  
<sup>747</sup> of competing accounts.

<sup>748</sup> Next, we discuss the theoretical consequences of these findings for research beyond for-  
<sup>749</sup> mant normalization. Following that, we discuss limitations of the present work, and how  
<sup>750</sup> future research might overcome them.

<sup>751</sup> **A. Consequences for theories of speech perception and beyond**

<sup>752</sup> Understanding the perceptual space in which the human brain represents vowel categories—  
<sup>753</sup> i.e., the normalized formant space—has obvious consequences for research on speech percep-  
<sup>754</sup> tion. To illustrate how far reaching these consequences can be, we discuss a few examples.  
<sup>755</sup> For instance, research on *categorical perception* has found that vowels seem to be per-  
<sup>756</sup> ceived less categorically than some types of consonants. Recent work has offered an elegant  
<sup>757</sup> explanation for this finding: the perception of formants—relevant to the recognition of  
<sup>758</sup> vowels—might be more noisy than the perception of the acoustic cues that are critical to  
<sup>759</sup> the recognition of less categorically perceived consonants (Kronrod *et al.*, 2016). This is  
<sup>760</sup> a parsimonious explanation, potentially preempting the need for separate explanations for  
<sup>761</sup> the perception of different types of phonemic contrasts. Kronrod and colleagues based their  
<sup>762</sup> argument on estimates they obtained for the relative ratio of meaningful category variability  
<sup>763</sup> to perceptual noise ( $\tau$ , the inverse of our noise ratios,  $\tau^{-1}$ ). Critically, this ratio depends  
<sup>764</sup> both on (i) the perceptual space in which formants are assumed to be represented (Kronrod  
<sup>765</sup> et al used Mel-transformed formants), and on (ii) whether the meaningful category variabil-  
<sup>766</sup> ity is calculated prior to, or following, normalization (Kronrod et al assumed the former,

767 which increases estimates of category variability). Our point here is not to cast doubt on  
 768 the results of [Kronrod \*et al.\* \(2016\)](#) —the fact that the best-fitting noise ratios in our study  
 769 were relatively similar across accounts (while varying across experiments) suggests that the  
 770 result of Kronrod and colleagues are likely to hold even under different assumptions about  
 771 (i) and (ii)—but rather to highlight how research on the perception and recognition of  
 772 vowels depends on assumptions about formant normalization. For example, similar points  
 773 could be raised about experiments on statistical learning that manipulate formant or other  
 774 frequency statistics (e.g., [Chládková \*et al.\*, 2017](#); [Colby \*et al.\*, 2018](#); [Wade \*et al.\*, 2007](#); [Xie  
 775 \*et al.\*, 2021](#)). Such experiments, too, need to make assumptions about the space in which  
 776 formants are represented. If these assumptions are incorrect, this can affect whether the  
 777 experimental manipulations have the intended effects, increasing the chance of null effects  
 778 or misinterpretation of observed effects.

779 Understanding the perceptual space in which the human brain represents vowel cate-  
 780 gories also has consequences for research beyond speech perception, perhaps more so than is  
 781 sometimes recognized. For instance, in sociolinguistics and related fields, Lobanov remains  
 782 the norm for representing vowels due to its efficiency in removing cross-talker variability  
 783 (for review, see [Adank \*et al.\*, 2004](#); [Barreda, 2021](#)). However, as shown in the present study,  
 784 removing cross-talker variability is not the same as representing vowels in the perceptual  
 785 space that listeners actually employ. Here, we do *not* find Lobanov to describe human  
 786 perception particularly well. On the contrary, we find no support for the hypothesis that  
 787 human speech perception employs these more complex computations that have been found  
 788 to perform best at reducing category variability. This should worry sociolinguists. In order

789 to understand how listeners infer a talker's background or social identity, it is important  
790 to understand the perceptual space in which inferences are actually rooted. Critically, the  
791 representations resulting from formant normalization presumably form an important part of  
792 the information that listeners use to draw social and linguistic inferences. It should thus be  
793 obvious that the use of normalization accounts that do not actually correspond to human  
794 perception can both mask real markers of social identity, and hallucinate markers that are  
795 not actually present. For example, in order to determine how a talker's social identity influ-  
796 ences their vowel realizations, it is important to discount *all and only* effects that listeners'  
797 will attribute to physiology, rather than social identity (Disner, 1980; Hindle, 1978).

798 Similar concerns apply to dialectology, research on language change, second language  
799 acquisition research, etc. For example, the perceptual space in which vowels are represented  
800 is critical to well-formed tests of hypotheses about the factors shaping the organization of  
801 vowel inventories across languages of the world (Lindblom, 1986; Stevens, 1972, 1989). It is  
802 essential in testing hypotheses about the extent to which the cross-linguistic realization of  
803 those systems is affected by perceptual processes (Flemming, 2010; Steriade, 2001), or by  
804 preferences for communicatively efficient linguistic systems (e.g., Hall *et al.*, 2018; Lindblom,  
805 1990; Moulin-Frier *et al.*, 2015). Similarly, tests of the hypothesis that vowel *articulation*  
806 during natural interactions is shaped by communicative efficiency do in obvious ways depend  
807 on assumptions about the perceptual space in which talkers—by hypothesis—aim to reduce  
808 perceptual confusion (cf. Buz and Jaeger, 2016; Gahl *et al.*, 2012; Scarborough, 2010; Wedel  
809 *et al.*, 2018). The same applies to any other line of research that aims to understand the  
810 perceptual consequences of formant variation across talkers, including research on infant- or

811 child-directed speech (Eaves Jr *et al.*, 2016; Kuhl *et al.*, 1997), and research on whether non-  
 812 native talkers are inherently more variable than native talkers (Smith *et al.*, 2019; Vaughn  
 813 *et al.*, 2019; Xie and Jaeger, 2020). In short, the perceptual space in which vowels are  
 814 represented is a critical component of understanding the structure of vowel systems, the  
 815 factors that shape them, and the ways in which they are used in natural language.

816 **B. Limitations and future directions**

817 The present work shares a few limitations with previous work. Here we focus on limita-  
 818 tions that follow from the assumptions we made in our computational framework. While  
 819 theories and hypotheses often contain substantial vagueness, *quantitative tests* of those  
 820 theories—as we have done here—require assumptions about *every* aspect of the model.  
 821 Here, this included all the steps necessary to link properties of the stimuli to listeners’ re-  
 822 sponses. For this purpose, we adopted the ASP framework (Xie *et al.*, 2023), and visualized  
 823 the graphical model that links stimuli ( $x$ ) to responses ( $r$ ) in Figure 6.

824 Many of the assumptions we made should be quite uncontroversial—e.g., the decision to  
 825 include both external (environmental) and internal (perceptual) noise in our model. While  
 826 these noise sources are often ignored in modeling human behavior, it is uncontroversial that  
 827 they exist. Other assumptions we made were introduced as simplifying assumptions for  
 828 the sake of feasibility—e.g., we expressed the effect of both types of noise through a single  
 829 parameter that related the average within-category variability of formants to noise variability  
 830 in the transformed and normalized formant space. In reality, however, environment noise  
 831 can have effects that are independent of internal noise, and internal noise likely affects

832 information processing at multiple (or all) of the steps shown in Figure 6. Such simplifying  
833 assumptions are both inevitable, and not necessarily problematic: as long as they do not  
834 introduce systematic bias to the evaluation of normalization accounts, they should not limit  
835 the generalizability of our results.

836 Some of our assumptions, however, might be more controversial. For example, we as-  
837 sumed that category representations can be expressed as multivariate Gaussian distributions  
838 in the formant space. This assumption, too, is a simplifying assumption—it simplified the  
839 computation of likelihoods—rather than a critical feature of the ASP framework we em-  
840 ployed. While human category representations are unlikely to be Gaussians, the alternative,  
841 e.g., exemplar representations, would come with its own downsides, such as increased sen-  
842 sitivity to the limited size of phonetic databases and substantial increases in computation  
843 time (exemplar representations afford researchers with much larger degrees of freedom). For  
844 researchers curious how this and other assumptions we made affect our results, our data  
845 and code are shared on OSF. This includes the R markdown document that generates this  
846 PDF, making it comparatively easy to revisit any of our assumptions to then regenerate the  
847 entire study with a click of a button in RStudio.

848 Like previous work, we further assumed that all listeners in our experiments use the  
849 same underlying vowel representations—the same dialect template(s). However, as already  
850 discussed, it is rather likely that not all of our listeners employed the same dialect tem-  
851 plate(s). An additional analysis reported in the SI (§3D) thus compared normalization  
852 accounts against only the subset of listeners who employed the dialect template used by  
853 the majority of participants (see lower-left of Figure 5B). This left only 11 participants for

854 Experiment 1a (61.1%) and 14 for Experiment 1b (77.8%), substantially reducing statistical  
 855 power. Replicating the main analysis, uniform scaling accounts again fit listeners' behavior  
 856 well across both experiments. The best-performing accounts did, however, differ from the  
 857 ones obtained for the superset of data (see SI, §3D).

858 A related assumption was introduced by the use of a phonetic database to approximate  
 859 listeners' vowel representations. This deviates from most previous evaluations of normal-  
 860 ization accounts (McMurray and Jongman, 2011; Barreda, 2021; but see Richter *et al.*,  
 861 2017), and reflects our commitment to a strong assumption made by most theories of speech  
 862 perception: that listeners' representations reflect the formant statistics previously experi-  
 863 enced speech input. By using a phonetic database to estimate listeners' representations, we  
 864 *substantially* reduced the degrees of freedom in the evaluation of normalization accounts,  
 865 reducing the chance of over-fitting to the data from our experiments. Our approach does,  
 866 however, also introduce two new assumptions.

867 First, our approach assumes that the mixture of dialect template(s) used by talkers in the  
 868 database sufficiently closely approximates those of the listeners in our experiments. Some  
 869 validation for this assumption comes from the additional analysis reported in the preceding  
 870 paragraph: when we subset listeners to only those who used the majority dialect template,  
 871 this improved the fit of all normalization accounts—as expected, if the category representa-  
 872 tions we trained on the phonetic database primarily reflect those listeners' representations  
 873 (see SI, §3D). Future work could further address this assumption in a number of ways. On  
 874 the one hand, dialect analyses like the ones we presented for our listeners (in Figure 5B)  
 875 could compare listeners' templates against the templates used by talkers in the database.

876 Alternatively or additionally, researchers could see whether our results replicate if ideal  
877 observers are instead trained on other databases that have been hypothesized to reflect a  
878 ‘typical’ L1 listeners’ experience with US English.

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

884 Finally, the evaluation of normalization accounts in the present work shares with all previ-  
885 ous work (e.g., [Apfelbaum and McMurray, 2015](#); [Barreda, 2021](#); [Cole \*et al.\*, 2010](#); [McMurray](#)  
886 [and Jongman, 2011](#); [Nearey, 1989](#); [Richter \*et al.\*, 2017](#)) another simplifying assumption that  
887 is clearly wrong: the assumption that listeners *know* the talker-specific formant properties  
888 required for normalization. Specifically, we normalized the input for each ideal observer  
889 using the maximum likelihood estimates of the normalization parameters for the respective  
890 experiment. For example, for the evaluation of the ideal observer trained on Lobanov nor-  
891 malized formants against listeners’ responses in Experiment 1a, we used the formant means  
892 and standard deviations of the stimuli used in Experiment 1a to normalize F1 and F2.

893 While this follows previous work, it constitutes a problematic assumption for the evaluation  
894 of extrinsic normalization accounts. For these accounts, the approach adopted essentially  
895 assumes the ability to predict the future: even on the first trial of the experiment, the input  
896 to the ideal observers were formants that were normalized based on the maximum likelihood  
897 estimate of the normalization parameters given the acoustic properties of *all* stimuli. Lis-

898 teners instead need to *incrementally infer* talker-specific properties from the speech input  
 899 (Nearey and Assmann, 2007; Xie *et al.*, 2023). The development and testing of incremental  
 900 variants of formant normalization strikes us an important avenue for future research.

901 **C. Concluding remarks**

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

909 Comparing 20 of the most influential normalization accounts against L1 listeners' per-  
 910 ception of US English monophthongs, we found that the normalization accounts that best  
 911 describe listeners' perception share that they (1) learn and store talker-specific properties  
 912 and (2) that they seem to be computationally very simple—taking advantage of the physics  
 913 of sound generation to use as few as a single parameter to normalize inter-talker variability  
 914 in vocal tract size. While the number of studies that have compared normalization accounts  
 915 against *listeners'* behavior remains surprisingly small, these two results confirm the findings  
 916 from more targeted comparisons that were focused on 2-3 accounts at a time (Barreda, 2021;  
 917 Nearey, 1989; Richter *et al.*, 2017). Overall then, we submit that it is time for research in

918 speech perception and beyond to consider simple uniform scaling the most-likely candidate  
919 for human formant normalization.

920 **ACKNOWLEDGMENTS**

921 Earlier versions of this work were presented at 2023 ASA meeting, ExLing 2022, at the  
922 Department of Computational Linguistics at the University of Zürich and at the Depart-  
923 ment of Swedish language and multilingualism at Stockholm University. We are grateful to  
924 OMITTED FOR REVIEW.

925 **AUTHOR CONTRIBUTIONS**

926 AP designed the experiments and collected the data, with input from TFJ. TFJ pro-  
927 grammed the experiments with input from AP. AP analyzed the experiments, with input  
928 from TFJ. AP and TFJ wrote the code to implement and fit the normalization models, with  
929 input from SB. AP developed the visualizations within input from SB and TFJ. AP wrote  
930 the first draft of the manuscript with edits by SB and TFJ.

931 **AUTHOR DECLARATIONS**

932 **Conflict of Interest**

933 The authors declare that the research was conducted in the absence of any commercial  
934 or financial relationships that could be construed as a potential conflict of interest.

935      **Ethics approval**

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

939 **V. REFERENCES**

940 <sup>1</sup>Normalization does not necessarily imply that *only* talker-normalized auditory percepts are available to  
 941 subsequent processing. There is ample evidence that subcategorical information can enter listeners' repre-  
 942 sentations of sound categories (e.g., [Hay \*et al.\*, 2017, 2019](#); [Johnson \*et al.\*, 1999](#); [McGowan, 2015](#); [Walker](#)  
 943 [and Hay, 2011](#)), in line with episodic ([Goldinger, 1996](#)) and exemplar theory of speech perception ([Johnson,](#)  
 944 [1997](#); [Sumner, 2011](#)).

945 <sup>2</sup>Under uniform scaling accounts, listeners essentially 'slide' the center of their category representations  
 946 (e.g., the 'template' of vowel categories for a given dialect) along a single line in formant space, with  $\Psi$   
 947 determining the target of this sliding. Later extensions of this account maintain its memory parsimony but  
 948 increased its inference complexity by allowing both intrinsic (the current F0) and extrinsic information (the  
 949 talker's single mean of log-transformed formants) to influence the inference of  $\Psi$  ([Nearey and Assmann,](#)  
 950 [2007](#)). We return to this extension in the general discussion.

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

956 <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  
 957 response entropy for an 8-way response choice is 3 bits, which means that all eight vowels are responded  
 958 equally often. The minimum response entropy = 0 bits, which means that the same vowel is responded all  
 959 the time.

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

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

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

969 <sup>8</sup>We intentionally did *not* split the data within talkers since normalization accounts are meant to make  
 970 speech perception robust to cross-talker variability. Further, splitting the data by speaker rather than  
 971 by vowel category avoids the potential for biases in the normalization parameter estimates for different  
 972 speakers in the case of missing or unbalanced tokens across vowel categories, see (Barreda and Nearey,  
 973 2018). Additional analyses not reported here confirmed that the same results are obtained when splits are  
 974 performed within talkers and within vowels (except that this lead to smaller CIs, and thus *more* significant  
 975 differences, in Figure 9). These analyses can be replicated by downloading the R markdown document this  
 976 article is based on from our OSF (see comments in our code).

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

982 <sup>10</sup>This ratio is a generalization of the inverse of the “meaningful-to-noise variance ratio ( $\tau$ )” used in Kronrod  
 983 *et al.* (2016). However, whereas Kronrod and colleagues committed to the simplifying assumption that

984 all categories have identical variance (along all formants), we allowed category variances to differ between  
 985 vowels, and between F1 and F2 (matching the empirically facts). We merely assume that the *noise* variance  
 986 is identical across all formants (in the phonetic space defined by the normalization account, e.g., log-Hz for  
 987 uniform scaling and Hz for Lobanov).

988 <sup>11</sup>Additional analyses reported in the SI (??) replicated this result for subsets of Experiments 1a and 1b.  
 989 For Experiment 1a, we excluded responses to the two *hVd* stimuli that differed from the other stimuli in  
 990 the preceding (*odd*) or following phonological context (*hut*). For Experiment 1b, we excluded responses to  
 991 any stimuli that were physiologically implausible for the talker (stimuli below the diagonal dashed line in  
 992 Figure 4).

993

994 Abramson, A. S., and Lisker, L. (1973). “Voice-timing perception in spanish word-initial  
 995 stops,” Journal of Phonetics 1(1), 1–8, doi: [10.1016/S0095-4470\(19\)31372-5](https://doi.org/10.1016/S0095-4470(19)31372-5).

996 Adank, P., Smits, R., and van Hout, R. (2004). “A comparison of vowel normalization pro-  
 997 cedures for language variation research,” The Journal of the Acoustical Society of America  
 998 116(5), 3099–3107, doi: [10.1121/1.1795335](https://doi.org/10.1121/1.1795335).

999 Allen, J. S., Miller, J. L., and DeSteno, D. (2003). “Individual talker differences in voice-  
 1000 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).

1002 Apfelbaum, K., and McMurray, B. (2015). “Relative cue encoding in the context of sophisti-  
 1003 cated models of categorization: Separating information from categorization,” Psychonomic  
 1004 Bulletin and Review 22(4), 916–943, doi: [10.3758/s13423-014-0783-2](https://doi.org/10.3758/s13423-014-0783-2).

1005 Assmann, P. F., Nearey, T. M., and Bharadwaj, S. (2008). “Analysis of a vowel database,”  
 1006 Canadian Acoustics 36(3), 148–149.

- 1007 Baese-Berk, M. M., Walker, K., and Bradlow, A. (2018). “Variability in speaking rate of  
 1008 native and non-native speakers,” *The Journal of the Acoustical Society of America* **144**(3),  
 1009 1717–1717, doi: [10.1121/1.5067612](https://doi.org/10.1121/1.5067612).
- 1010 Barreda, S. (2020). “Vowel normalization as perceptual constancy,” *Language* **96**(2), 224–  
 1011 254, doi: [10.1353/lan.2020.0018](https://doi.org/10.1353/lan.2020.0018).
- 1012 Barreda, S. (2021). “Perceptual validation of vowel normalization methods for vari-  
 1013 ationist research,” *Language Variation and Change* **33**(1), 27–53, doi: [10.1017/S0954394521000016](https://doi.org/10.1017/S0954394521000016).
- 1015 Barreda, S., and Nearey, T. M. (2012). “The direct and indirect roles of fundamental fre-  
 1016 quency in vowel perception,” *The Journal of the Acoustical Society of America* **131**(1),  
 1017 466–477, doi: [10.1121/1.3662068](https://doi.org/10.1121/1.3662068).
- 1018 Barreda, S., and Nearey, T. M. (2018). “A regression approach to vowel normalization for  
 1019 missing and unbalanced data,” *The Journal of the Acoustical Society of America* **144**(1),  
 1020 500–520, doi: [10.1121/1.5047742](https://doi.org/10.1121/1.5047742).
- 1021 Bladon, A., Henton, C., and Pickering, J. (1984). “Towards an auditory theory of speaker  
 1022 normalization,” *Language and Communication* **4**, 59–69.
- 1023 Boersma, P., and Weenink, D. (2022). “Praat: Doing phonetics by computer [Computer  
 1024 program]” .
- 1025 Buz, E., and Jaeger, T. F. (2016). “The (in) dependence of articulation and lexical planning  
 1026 during isolated word production,” *Language, Cognition and Neuroscience* **31**(3), 404–424.
- 1027 Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). “A limited memory algorithm for  
 1028 bound constrained optimization,” *SIAM Journal on Scientific Computing* **16**(5), 1190–

- 1029 1208, doi: [10.1137/0916069](https://doi.org/10.1137/0916069).
- 1030 Carpenter, G. A., and Govindarajan, K. K. (1993). “Neural Network and Nearest Neighbor
- 1031 Comparison of Speaker Normalization Methods for Vowel Recognition,” in *ICANN '93*,
- 1032 edited by S. Gielen and B. Kappen (Springer London, London), pp. 412–415, doi: [10.1007/978-1-4471-2063-6\\_98](https://doi.org/10.1007/978-1-4471-2063-6_98).
- 1033
- 1034 Chládková, K., Podlipský, V. J., and Chionidou, A. (2017). “Perceptual adaptation of
- 1035 vowels generalizes across the phonology and does not require local context.,” *Journal of*
- 1036 *Experimental Psychology: Human Perception and Performance* **43**(2), 414.
- 1037 Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (2008). “Perception of
- 1038 speech reflects optimal use of probabilistic speech cues,” *Cognition* **108**(3), 804–809, doi:
- 1039 [10.1016/j.cognition.2008.04.004](https://doi.org/10.1016/j.cognition.2008.04.004).
- 1040 Colby, S., Clayards, M., and Baum, S. (2018). “The role of lexical status and individual dif-
- 1041 ferences for perceptual learning in younger and older adults,” *Journal of Speech, Language,*
- 1042 *and Hearing Research* **61**(8), 1855–1874.
- 1043 Cole, J., Linebaugh, G., Munson, C., and McMurray, B. (2010). “Unmasking the acous-
- 1044 tic effects of vowel-to-vowel coarticulation: A statistical modeling approach,” *Journal of*
- 1045 *Phonetics* **38**(2), 167–184, doi: [10.1016/j.wocn.2009.08.004](https://doi.org/10.1016/j.wocn.2009.08.004).
- 1046 Crinnion, A. M., Malmskog, B., and Toscano, J. C. (2020). “A graph-theoretic approach to
- 1047 identifying acoustic cues for speech sound categorization,” *Psychonomic Bulletin & Review*
- 1048 **27**(6), 1104–1125, doi: [10.3758/s13423-020-01748-1](https://doi.org/10.3758/s13423-020-01748-1).
- 1049 Disner, S. F. (1980). “Evaluation of vowel normalization procedures,” *The Journal of the*
- 1050 *Acoustical Society of America* **67**(1), 253–261, doi: [10.1121/1.383734](https://doi.org/10.1121/1.383734).

- 1051 Eaves Jr, B. S., Feldman, N. H., Griffiths, T. L., and Shafto, P. (2016). “Infant-directed  
 1052 speech is consistent with teaching.,” *Psychological Review* **123**(6), 758.
- 1053 Escudero, P., and Bion, R. A. H. (2007). “Modeling vowel normalization and sound per-  
 1054 ception as sequential processes,” *ICPhS XVI*, 1413–1416.
- 1055 Fant, G. (1975). “Non-uniform vowel normalization,” *STL-QPSR* **16**(2-3), 001–019.
- 1056 Fant, G., Kruckenberg, A., Gustafson, K., and Liljencrants, J. (2002). “A New Approach  
 1057 to Intonation Analysis and Synthesis of Swedish,” *ISCA 2002* 283–286.
- 1058 Fastl, H., and Zwicker, E. (2007). *Psychoacoustics* (Springer, Berlin, Heidelberg).
- 1059 Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). “The influence of categories  
 1060 on perception: Explaining the perceptual magnet effect as optimal statistical inference,”  
 1061 *Psychological Review* **116**(4), 752–782.
- 1062 Flemming, E. (2010). “Modeling listeners: Comments on pluymakers et al. and scarbor-  
 1063 ough,” in *Laboratory Phonology*, edited by C. Fougeron, B. Kühnert, M. D’Imperio, and  
 1064 N. Vallée, **10**, pp. 587–606.
- 1065 Gahl, S., Yao, Y., and Johnson, K. (2012). “Why reduce? phonological neighborhood  
 1066 density and phonetic reduction in spontaneous speech,” *Journal of Memory and Language*  
 1067 **66**(4), 789–806.
- 1068 Gerstman, L. (1968). “Classification of self-normalized vowels,” *IEEE Transactions on Au-  
 1069 dio and Electroacoustics* **16**(1), 78–80, doi: [10.1109/TAU.1968.1161953](https://doi.org/10.1109/TAU.1968.1161953).
- 1070 Glasberg, B. R., and Moore, B. C. J. (1990). “Derivation of auditory filter shapes from  
 1071 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).

- 1073 Goldinger, S. D. (1996). "Words and voices: Episodic traces in spoken word identification  
1074 and recognition memory," *Journal of Experimental Psychology: Learning Memory and*  
1075 *Cognition* **22**(5), 1166–1183.
- 1076 Greenwood, D. D. (1997). "The mel scale's disqualifying bias and a consistency of pitch-  
1077 difference equisections in 1956 with equal cochlear distances and equal frequency ra-  
1078 tios," *Hearing Research* **103**(1), 199–224, <https://www.sciencedirect.com/science/article/pii/S037859559600175X>, doi: [https://doi.org/10.1016/S0378-5955\(96\)00175-X](https://doi.org/10.1016/S0378-5955(96)00175-X).
- 1081 Hall, K. C., Hume, E., Jaeger, T. F., and Wedel, A. (2018). "The role of predictability in  
1082 shaping phonological patterns," *Linguistics Vanguard* **4**(s2), 20170027.
- 1083 Hay, J., Podlubny, R., Drager, K., and McAuliffe, M. (2017). "Car-talk: Location-specific  
1084 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).
- 1086 Hay, J., Walker, A., Sanchez, K., and Thompson, K. (2019). "Abstract social categories  
1087 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).
- 1089 Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). "Acoustic charac-  
1090 teristics of American English vowels," *Journal of the Acoustical Society of America* **97**(5),  
1091 3099–3111.
- 1092 Hillenbrand, J. M., and Nearey, T. M. (1999). "Identification of resynthesized /hvd/ utter-  
1093 ances: Effects of formant contour," *Journal of the Acoustical Society of America* **105**(6),  
1094 3509–3523, doi: [10.1121/1.424676](https://doi.org/10.1121/1.424676).

- 1095 Hindle, D. (1978). "Approaches to Vowel Normalization in the Study of Natural Speech,"
- 1096 in *Linguistic Variation: Models and Methods*, edited by D. Sankoff (Academic Press, New
- 1097 York), pp. 161–171.
- 1098 Jaeger, T. F. (2024). *MVBeliefUpdatr: Fitting, Summarizing, and Visu-*
- 1099 *alizing of Multivariate Gaussian Ideal Observers and Adaptors*, <https://github.com/hlplab/MVBeliefUpdatr>, r package version 0.0.1.0010, commit
- 1100 c8dc91766d37bd7eee6561dd802088a6e5b93b9c.
- 1102 Johnson, K. (1997). "Speech perception without speaker normalization," in *Talker Variabil-*
- 1103 *ity in Speech Processing*, edited by K. Johnson and W. Mullennix (CA: Academic Press,
- 1104 San Diego), pp. 146–165.
- 1105 Johnson, K. (2020). "The  $\Delta f$  method of vocal tract length normalization for vowels," *Lab-*
- 1106 *oratory Phonology* 11(1), doi: [10.5334/labphon.196](https://doi.org/10.5334/labphon.196).
- 1107 Johnson, K., and Sjerps, M. J. (2021). "Speaker normalization in speech perception,"
- 1108 in *The Handbook of Speech Perception*, edited by J. S. Pardo, L. C. Nygaard, R. E.
- 1109 Remez, and D. B. Pisoni (John Wiley & Sons, Inc), Chap. 6, pp. 145–176, doi:
- 1110 [10.1002/9781119184096.ch6](https://doi.org/10.1002/9781119184096.ch6).
- 1111 Johnson, K., Strand, E. A., and D'Imperio, M. (1999). "Auditory–visual integration
- 1112 of talker gender in vowel perception," *Journal of Phonetics* 27(4), 359–384, <https://doi.org/10.1006/jpho.1999.0100>.
- 1113 [https://www.sciencedirect.com/science/article/pii/S0095447099901006](https://doi.org/10.1006/jpho.1999.0100), doi: <https://doi.org/10.1006/jpho.1999.0100>.
- 1114
- 1115 Joos, M. (1948). "Acoustic Phonetics," *Language* 24(2), 5–136, doi: [10.2307/522229](https://doi.org/10.2307/522229).
- 1116 Kleinschmidt, D. (2020). "What constrains distributional learning in adults?," .

- 1117 Kleinschmidt, D., and Jaeger, T. F. (2015). “Robust speech perception: Recognize the  
1118 familiar, generalize to the similar, and adapt to the novel,” Psychological Review **122**(2),  
1119 148–203, doi: [10.1037/a0038695](https://doi.org/10.1037/a0038695).
- 1120 Kleinschmidt, D., and Jaeger, T. F. (2016). “What do you expect from an unfamiliar  
1121 talker?,” Proceedings of the 38th Annual Meeting of the Cognitive Science Society, CogSci  
1122 2016 2351–2356.
- 1123 Kleinschmidt, D., Liu, L., Bushong, W., Burchill, Z., Xie, X., Tan, M., Karboga, G., and  
1124 Jaeger, F. (2021). “JSEXP” <https://github.com/hlplab/JSEXP>.
- 1125 Kronrod, Y., Coppess, E., and Feldman, N. H. (2016). “A unified model of categorical  
1126 effects in consonant and vowel perception,” Psychological Bulletin and Review 1681–1712,  
1127 doi: [10.3758/s13423-016-1049-y](https://doi.org/10.3758/s13423-016-1049-y).
- 1128 Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V.,  
1129 Ryskina, V. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F. (1997). “Cross-language  
1130 analysis of phonetic units in language addressed to infants,” Science **277**(5326), 684–686.
- 1131 Labov, W., Ash, S., and Boberg, C. (2006). *The atlas of North American English: Phonetics,*  
1132 *phonology, and sound change* (De Gruyter Mouton, Berlin • New York).
- 1133 Ladefoged, P., and Broadbent, D. E. (1957). “Information conveyed by vowels,” Journal of  
1134 the Acoustical Society of America **29**, 98–104.
- 1135 Lee, C.-Y. (2009). “Identifying isolated, multispeaker mandarin tones from brief acoustic  
1136 input: A perceptual and acoustic study,” The Journal of the Acoustical Society of America  
1137 **125**(2), 0001–4966, doi: [10.1121/1.3050322](https://doi.org/10.1121/1.3050322).

- <sub>1138</sub> Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967).  
<sub>1139</sub> “Perception of the speech code,” *Psychological review* **74**(6), 431–461.
- <sub>1140</sub> Lindblom, B. (1986). “Phonetic universals in vowel systems,” *Experimental phonology* 13–  
<sub>1141</sub> 44.
- <sub>1142</sub> Lindblom, B. (1990). “Explaining phonetic variation: A sketch of the H&H theory,” *Speech*  
<sub>1143</sub> *Production and Speech Modeling* 403–439.
- <sub>1144</sub> Lobanov, B. M. (1971). “Classification of Russian vowels spoken by different speakers,” *The*  
<sub>1145</sub> *Journal of the Acoustical Society of America* **49**(2B), 606–608, doi: [10.1121/1.1912396](https://doi.org/10.1121/1.1912396).
- <sub>1146</sub> Luce, P. A., and Pisoni, D. B. (1998). “Recognizing spoken words: The neighborhood acti-  
<sub>1147</sub> vation model,” *Ear and Hearing* **19**(1), 1–36, doi: [10.1097/00003446-199802000-00001](https://doi.org/10.1097/00003446-199802000-00001).
- <sub>1148</sub> Luce, R. D. (1959). *Individual Choice Behavior* (John Wiley, Oxford).
- <sub>1149</sub> Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna,  
<sub>1150</sub> P. D., Theodore, R., Monto, N., and Rueckl, J. G. (2020). “EARSHOT: A minimal neural  
<sub>1151</sub> network model of incremental human speech recognition,” *Cognitive Science* **44**(4), 1–17,  
<sub>1152</sub> doi: [10.1111/cogs.12823](https://doi.org/10.1111/cogs.12823).
- <sub>1153</sub> Massaro, D. W., and Friedman, D. (1990). “Models of integration given multiple sources of  
<sub>1154</sub> 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).
- <sub>1155</sub> McClelland, J. L., and Elman, J. L. (1986). “The TRACE model of speech perception,”  
<sub>1156</sub> *Cognitive Psychology* **18**(1), 1–86.
- <sub>1157</sub> McGowan, K. B. (2015). “Social expectation improves speech perception in noise,” *Lan-*  
<sub>1158</sub> *guage and Speech* **58**(4), 502–521, doi: [10.1177/0023830914565191](https://doi.org/10.1177/0023830914565191).

- 1159 McMurray, B., and Jongman, A. (2011). “What information is necessary for speech categorization?: Harnessing variability in the speech signal by integrating cues computed relative  
1160 to expectations,” *Psychological Review* **118**(2), 219–246, doi: [10.1037/a0022325.What](https://doi.org/10.1037/a0022325).
- 1162 Merzenich, M. M., Knight, P. L., and Roth, G. L. (1975). “Representation of cochlea  
1163 within primary auditory cortex in the cat,” *Journal of Neurophysiology* **38**(2), 231–249,  
1164 doi: [10.1152/jn.1975.38.2.231](https://doi.org/10.1152/jn.1975.38.2.231).
- 1165 Miller, J. D. (1989). “Auditory-perceptual interpretation of the vowel,” *The Journal of  
1166 Acoustical Society of America* **85**(5), 22.
- 1167 Moore, B. C. (2012). *An introduction to the psychology of hearing* (Brill).
- 1168 Moulin-Frier, C., Diard, J., Schwartz, J.-L., and Bessière, P. (2015). “Cosmo (“communi-  
1169 cating about objects using sensory–motor operations”): A bayesian modeling framework  
1170 for studying speech communication and the emergence of phonological systems,” *Journal  
1171 of Phonetics* **53**, 5–41.
- 1172 Nearey, T. M. (1978). Indiana University Linguistics Club *Phonetic Feature Systems for  
1173 Vowels*.
- 1174 Nearey, T. M. (1989). “Static, dynamic, and relational properties in vowel perception,” *The  
1175 Journal of the Acoustical Society of America* **85**(5), 2088–2113, doi: [10.1121/1.397861](https://doi.org/10.1121/1.397861).
- 1176 Nearey, T. M. (1990). “The segment as a unit of speech perception,” *Journal of Phonetics*  
1177 **18**(3), 347–373, doi: [10.1016/S0095-4470\(19\)30379-1](https://doi.org/10.1016/S0095-4470(19)30379-1).
- 1178 Nearey, T. M., and Assmann, P. F. (1986). “Modeling the role of inherent spectral change in  
1179 vowel identification,” *The Journal of the Acoustical Society of America* **80**(5), 1297–1308,  
1180 doi: [10.1121/1.394433](https://doi.org/10.1121/1.394433).

- 1181 Nearey, T. M., and Assmann, P. F. (2007). “Probabilistic ‘sliding template’ models for  
 1182 indirect vowel normalization,” in *Experimental approaches to phonology*, edited by J.-J.  
 1183 Solé, P. S. Beddor, and M. Ohala (Oxford University Press), pp. 246–270.
- 1184 Nearey, T. M., and Hogan, J. (1986). “Phonological contrast in experimental phonetics: Re-  
 1185 lating distributions of measurements production data to perceptual categorization curves,”  
 1186 in *Experimental Phonology*, edited by J. J. Ohala and J. Jaeger (Academic Press, New  
 1187 York), pp. 141–161.
- 1188 Newman, R. S., Clouse, S. A., and Burnham, J. L. (2001). “The perceptual consequences  
 1189 of within-talker variability in fricative production,” *The Journal of the Acoustical Society  
 1190 of America* **109**(3), 1181–1196.
- 1191 Nordström, P., and Lindblom, B. (1975). “A normalization procedure for vowel formant  
 1192 data,” *Proceedings of ICPHS VIII*, Leeds 212.
- 1193 Norris, D., and McQueen, J. M. (2008). “Shortlist B: A Bayesian model of continuous speech  
 1194 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).
- 1195 Oganian, Y., Bhaya-Grossman, I., Johnson, K., and Chang, E. F. (2023). “Vowel and  
 1196 formant representation in the human auditory speech cortex,” *Neuron* **111**(13), 2105–2118.
- 1197 Persson, A., and Jaeger, T. F. (2023). “Evaluating normalization accounts against the dense  
 1198 vowel space of Central Swedish,” *Frontiers in Psychology* **14**, <https://doi.org/10.3389/fpsyg.2023.1165742>, doi: [10.3389/fpsyg.2023.1165742](https://doi.org/10.3389/fpsyg.2023.1165742).
- 1200 Peterson, G. E. (1961). “Parameters of vowel quality,” *Journal of Speech and Hearing  
 1201 Research* **4**(1), 10–29, <https://doi.org/10.1044/jshr.0401.10>, doi:  
 1202 [10.1044/jshr.0401.10](https://doi.org/10.1044/jshr.0401.10).

- 1203 R Core Team (2024). *R: A Language and Environment for Statistical Computing*, R Foun-  
 1204 dation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>.
- 1205 Richter, C., Feldman, N. H., Salgado, H., and Jansen, A. (2017). “Evaluating low-level  
 1206 speech features against human perceptual data,” *Transactions of the Association for Com-  
 1207 putational Linguistics* 5, 425–440, doi: [10.1162/tacl\\_a\\_00071](https://doi.org/10.1162/tacl_a_00071).
- 1208 RStudio Team (2020). *RStudio: Integrated Development Environment for R*, RStudio,  
 1209 PBC., Boston, MA.
- 1210 Saenz, M., and Langers, D. R. (2014). “Tonotopic mapping of human auditory cortex,”  
 1211 *Hearing Research* 307, 42–52, <https://www.sciencedirect.com/science/article/pii/S0378595513001871>, doi: <https://doi.org/10.1016/j.heares.2013.07.016> hu-  
 1212 man Auditory NeuroImaging.
- 1213 Scarborough, R. (2010). “Lexical and contextual predictability: Confluent effects on the  
 1214 production of vowels,” *Laboratory Phonology* 10, 557–586.
- 1215 Schertz, J., and Clare, E. J. (2020). “Phonetic cue weighting in perception and production,”  
 1216 *Wiley Interdisciplinary Reviews: Cognitive Science* 11(2), doi: [10.1002/wcs.1521](https://doi.org/10.1002/wcs.1521).
- 1217 Shannon, C. E. (1948). “A mathematical theory of communication,” *The Bell System Tech-  
 1218 nical 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).
- 1219 Siegel, R. J. (1965). “A replication of the mel scale of pitch,” *The American Journal of  
 1220 Psychology* 78(4), 615–620, <http://www.jstor.org/stable/1420924>.
- 1221 Sjerps, M. J., Fox, N. P., Johnson, K., and Chang, E. F. (2019). “Speaker-normalized  
 1222 sound representations in the human auditory cortex,” *Nature Communications* 10(1), doi:  
 1223 [10.1038/s41467-019-10365-z](https://doi.org/10.1038/s41467-019-10365-z).

- 1225 Skoe, E., Krizman, J., Spitzer, E. R., and Kraus, N. (2021). “Auditory cortical changes  
 1226 precede brainstem changes during rapid implicit learning: Evidence from human EEG,”  
 1227 *Frontiers in Neuroscience* 1007.
- 1228 Smith, B. L., Johnson, E., and Hayes-Harb, R. (2019). “Esl learners’ intra-speaker vari-  
 1229 ability in producing American English tense and lax vowels,” *Journal of Second Language  
 1230 Pronunciation* 5(1), 139–164.
- 1231 Steriade, D. (2001). “The phonology of perceptibility effects: the P-map and its conse-  
 1232 quences for constraint organization” .
- 1233 Stevens, K. N. (1972). “The quantal nature of speech: Evidence from articulatory-acoustic  
 1234 data,” in *Human communication: a unified view* (McGraHill, New York), pp. 51–66.
- 1235 Stevens, K. N. (1989). “On the quantal nature of speech,” *Journal of phonetics* 17(1-2),  
 1236 3–45.
- 1237 Stevens, S. S., and Volkmann, J. (1940). “The Relation of Pitch to Frequency: A Revised  
 1238 Scale,” *The American Journal of Psychology* 53(3), 329–353, doi: [10.2307/1417526](https://doi.org/10.2307/1417526).
- 1239 Stilp, C. (2020). “Acoustic context effects in speech perception,” *WIREs Cognitive Science*  
 1240 11(1), 1–18, doi: [10.1002/wcs.1517](https://doi.org/10.1002/wcs.1517).
- 1241 Sumner, M. (2011). “The role of variation in the perception of accented speech,” *Cognition*  
 1242 119(1), 131–136, doi: [10.1016/j.cognition.2010.10.018](https://doi.org/10.1016/j.cognition.2010.10.018).
- 1243 Syrdal, A. K. (1985). “Aspects of a model of the auditory representation of American English  
 1244 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).
- 1245 Syrdal, A. K., and Gopal, H. S. (1986). “A perceptual model of vowel recognition based on  
 1246 the auditory representation of American English vowels,” *The Journal of the Acoustical*

- 1247 Society of America **79**(4), 1086–1100, doi: [10.1121/1.393381](https://doi.org/10.1121/1.393381).
- 1248 Tan, M., and Jaeger, T. F. (2024). “Incremental adaptation to an unfamiliar talker,”
- 1249 Manuscript, Stockholm University .
- 1250 Tang, C., Hamilton, L. S., and Chang, E. F. (2017). “Intonational speech prosody encod-
- 1251 ing in the human auditory cortex,” *Science* **357**(6353), 797–801, doi: [10.1126/science.aam8577](https://doi.org/10.1126/science.aam8577).
- 1252 ten Bosch, L., Boves, L., Tucker, B., and Ernestus, M. (2015). “DIANA: towards compu-
- 1253 tational modeling reaction times in lexical decision in north American English,” in *Proc.*
- 1254 *Interspeech 2015*, pp. 1576–1580, doi: [10.21437/Interspeech.2015-366](https://doi.org/10.21437/Interspeech.2015-366).
- 1255 Traunmüller, H. (1981). “Perceptual dimension of openness in vowels,” *The Journal of the*
- 1256 *Acoustical Society of America* **69**(5), 1465–1475, doi: [10.1121/1.385780](https://doi.org/10.1121/1.385780).
- 1257 Traunmüller, H. (1990). “Analytical expressions for the tonotopic sensory scale,” *The Jour-*
- 1258 *nal of the Acoustical Society of America* **88**(1), 97–100, doi: [10.1121/1.399849](https://doi.org/10.1121/1.399849).
- 1260 Vaughn, C., Baese-Berk, M., and Idemaru, K. (2019). “Re-examining phonetic variability
- 1261 in native and non-native speech,” *Phonetica* **76**(5), 327–358.
- 1262 Vorperian, H. K., and Kent, R. D. (2007). “Vowel acoustic space development in children: A
- 1263 synthesis of acoustic and anatomic data,” *Journal of Speech, Language & Hearing Research*
- 1264 **50**(6), 1510–1545, doi: [10.1044/1092-4388\(2007/104\)](https://doi.org/10.1044/1092-4388(2007/104)).
- 1265 Wade, T., Jongman, A., and Sereno, J. (2007). “Effects of acoustic variability in the per-
- 1266 ceptual learning of non-native-accented speech sounds,” *Phonetica* **64**(2-3), 122–144, doi:
- 1267 [10.1159/000107913](https://doi.org/10.1159/000107913).

- 1268 Walker, A., and Hay, J. (2011). “Congruence between ‘word age’ and ‘voice age’ facilitates  
1269 lexical access,” *Laboratory Phonology* **2**(1), 219–237, doi: [10.1515/labphon.2011.007](https://doi.org/10.1515/labphon.2011.007).
- 1270 Watt, D., and Fabricius, A. (2002). “Evaluation of a technique for improving the mapping  
1271 of multiple speakers’ vowel spaces in the F1 ~ F2 plane,” in *Leeds Working Papers in  
1272 Linguistics and Phonetics*, edited by D. Nelson, 9, pp. 159–173.
- 1273 Weatherholtz, K., and Jaeger, T. F. (2016). “Speech perception and generalization  
1274 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).
- 1275
- 1276 Wedel, A., Nelson, N., and Sharp, R. (2018). “The phonetic specificity of contrastive hy-  
1277 perarticulation in natural speech,” *Journal of Memory and Language* **100**, 61–88.
- 1278 Whalen, D. H. (2016). “A double-Nearey theory of vowel normalization: Approaching con-  
1279 sensus,” *The Journal of the Acoustical Society of America* **140**(4\_Supplement), 3163–3164,  
1280 doi: [10.1121/1.4969932](https://doi.org/10.1121/1.4969932).
- 1281 Wichmann, F. A., and Hill, N. J. (2001). “The psychometric function: I. Fitting, sampling,  
1282 and goodness of fit,” *Perception & psychophysics* **63**(8), 1293–1313.
- 1283 Winn, M. (2018). “Speech: It’s not as acoustic as you think,” *Acoustics Today* **12**(2), 43–49.
- 1284 Xie, X., Buxó-Lugo, A., and Kurumada, C. (2021). “Encoding and decoding of meaning  
1285 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).
- 1286
- 1287 Xie, X., and Jaeger, T. F. (2020). “Comparing non-native and native speech: Are L2  
1288 productions more variable?,” *The Journal of the Acoustical Society of America* **147**(5),  
1289 3322–3347, doi: [10.1121/10.0001141](https://doi.org/10.1121/10.0001141).

- 1290 Xie, X., Jaeger, T. F., and Kurumada, C. (2023). “What we do (not) know about the  
1291 mechanisms underlying adaptive speech perception: A computational review,” Cortex .  
1292 Zahorian, S. A., and Jagharghi, A. J. (1991). “Speaker normalization of static and dynamic  
1293 vowel spectral features,” The Journal of the Acoustical Society of America **90**(1), 67–75,  
1294 doi: [10.1121/1.402350](https://doi.org/10.1121/1.402350).