

Mice can learn phonetic categories

Jonny L. Saunders¹ and Michael Wehr^{1, a)}

*University of Oregon, Institute of Neuroscience and Department of Psychology,
Eugene, OR 97403, United States*

(Dated: 13 February 2019)

1 We perceive speech as a series of relatively invariant phonemes despite extreme vari-
2 ability in the acoustic signal. To be perceived as nearly-identical phonemes, speech
3 sounds that vary continuously over a range of acoustic parameters must be percep-
4 tually discretized by the auditory system. Such many-to-one mappings of undiffer-
5 entiated sensory information to a finite number of discrete categories are ubiquitous
6 in perception. Although many mechanistic models of phonetic perception have been
7 proposed, they remain largely unconstrained by neurobiological data. Current human
8 neurophysiological methods lack the necessary spatiotemporal resolution to provide
9 it: speech is too fast and the neural circuitry involved is too small. Here we demon-
10 strate that mice are capable of learning generalizable phonetic categories, and can
11 thus serve as a model for phonetic perception. Mice learned to discriminate conso-
12 nants, and generalized consonant identity across novel vowel contexts and speakers,
13 consistent with true category learning. A mouse model, given the powerful genetic
14 and electrophysiological tools for probing neural circuits available for them, has the
15 potential to powerfully augment our mechanistic understanding of phonetic percep-
16 tion.

a) wehr@uoregon.edu

¹⁷ I. INTRODUCTION

¹⁸ A. Lack of acoustic invariance in phonemes

¹⁹ We perceive speech as a series of relatively invariant phonemes despite extreme variability
²⁰ in the acoustic signal. This lack of order within phonemic categories remains one of the
²¹ fundamental problems of speech perception⁵². Plosive stop consonants (such as /b/ or
²² /g/) are the paradigmatic example of phonemes with near-categorical perception^{1–3} without
²³ invariant acoustic structure^{4,5}. The problem is not just that phonemes are acoustically
²⁴ variable, but rather that there is a fundamental lack of invariance in the relation between
²⁵ phonemes and the acoustic signal⁵. Despite our inability to find a source of invariance in
²⁶ the speech signal, the auditory system learns some acoustic-perceptual mapping such that
²⁷ a plosive stop like /b/ is perceived as nearly identical across phonetic contexts. A key
²⁸ source of variability is coarticulation, which causes the sound of a spoken consonant to be
²⁹ strongly affected by neighboring segments, such as vowels. Coarticulation occurs during stop
³⁰ production because the articulators (such as the tongue or lips) have not completely left the
³¹ positions from the preceding phoneme, and are already moving to anticipate the following
³² phoneme^{6,7}. Along with many other sources of acoustic variation like speaker identity, sex,
³³ accent, or environmental noise; coarticulation guarantees that a given stop consonant does
³⁴ not have a uniquely invariant acoustic structure across phonetic contexts. In other words,
³⁵ there is no canonical acoustic /b/^{1,6}. Phonetic perception therefore cannot be a simple,
³⁶ linear mapping of some continuous feature space to a discrete phoneme space. Instead it
³⁷ requires a mapping that flexibly uses evidence from multiple, imperfect cues depending on

³⁸ context^{1,8}. This invariant perception of phonemes, despite extreme variability in the physical
³⁹ speech signal, is referred to as the non-invariance problem⁹.

⁴⁰ **B. Generality of phonetic perception**

⁴¹ The lack of a simple mapping between acoustic attributes and phoneme identity has had a
⁴² deep influence on phonetics, in part motivating the hypothesis that speech is mechanistically
⁴³ unique to humans¹⁰, and the development of non-acoustic theories of speech perception
⁴⁴ (most notably motor theories^{6,8,11}). However, it has been clear for more than 30 years
⁴⁵ that at least some auditory components of speech perception are not unique to humans,
⁴⁶ suggesting that human speech perception exploits evolutionarily-preserved functions of the
⁴⁷ auditory system^{5,12–14}. For example, nonhuman animals like quail^{5,16}, chinchillas¹⁸, rats¹⁹,
⁴⁸ macaques²⁰, and songbirds²¹ are capable of learning phonetic categories that share some
⁴⁹ perceptual qualities with humans^{15,17}. This is consistent with the idea that categorizing
⁵⁰ phonemes is just one instance of a more general problem faced by all auditory systems, which
⁵¹ typically extract useable information from complex acoustic environments by reducing them
⁵² to a small number of 'auditory objects' (for review, see²²).

⁵³ **C. Neurolinguistic theories of phonetic perception**

⁵⁴ Many neurolinguistic theories of phonetic perception have been proposed^{11,23–26}, but neu-
⁵⁵ rophysiological evidence to support them is limited. One broad class of models follows
⁵⁶ the paradigm of hierarchical processing first described by Hubel and Weisel in the visual
⁵⁷ system^{23,24,27}. In these models, successive processing stages in the auditory system ex-

58 tract acoustic features with progressively increasing complexity by combining the simpler
59 representations present in preceding stages. Such hierarchical processing is relatively well-
60 supported by experimental data. For example, the responses of neurons in primary auditory
61 cortex (A1) to speech sounds are more diverse than those in inferior colliculus²⁸ (but see²⁹).
62 While phoneme identity can be classified post-hoc from population-level activity in A1^{30–32},
63 neurons in secondary auditory cortical regions explicitly encode higher-order properties of
64 speech sounds^{33–37}.

65 Another class of models proposes that phonemes have no positive acoustic “prototype”,
66 and that we instead learn only the acoustic features useful for telling them apart²⁵. Theo-
67 retically, these discriminative models provide better generalization and robustness to high
68 variance³⁸. Theories based on discrimination rather than prototype-matching have a long
69 history in linguistics³⁹, but have rarely been implemented as neurolinguistic models. A
70 possible neural implementation of discriminative perception is that informative contrast
71 cues could evoke inhibition to suppress competing phonetic percepts, similar to predictive
72 coding^{25,40,41}. Neurophysiological evidence supports the existence of discriminative predic-
73 tive coding, but its specific implementation is unclear^{42,43}.

74 These two very different classes of models illustrate a major barrier faced by phonetic
75 research: both classes can successfully predict human categorization performance, mak-
76 ing it difficult to empirically validate or refute either of them using psychophysical exper-
77 iments alone. Mechanistic differences have deep theoretical consequences — for example,
78 the characterizations made by the above two classes of models regarding what phonemes
79 *are* precisely oppose one another: are they positive acoustic prototypes, or sets of negative

80 acoustic contrasts? Perceptually, do listeners identify phonemes, or discriminate between
81 them? Neurobiological evidence regarding how the brain actually solves these categorization
82 problems could help overcome this barrier.

83 **D. The utility of a mouse model for speech research**

84 Neurolinguistic research in humans faces several limitations that could be overcome using
85 animal models.

86 First, most current human neurophysiological methods lack the spatiotemporal resolution
87 to probe the fine spatial scale of neuronal circuitry and the millisecond timescale of speech
88 sounds. A causal, mechanistic understanding of computation in neural circuits is also greatly
89 aided by the ability to manipulate individual neurons or circuit components, which is difficult
90 in humans. Optogenetic methods available in mice provide the ability to activate, inactivate,
91 or record activity from specific types of neurons at the millisecond timescales of speech
92 sounds.

93 Second, it is difficult to isolate the purely auditory component of speech perception in
94 humans. Humans can use contextual information from syntax, semantics or task structure
95 to infer phoneme identity^{44,45}. It is also difficult to rule out the contribution of multimodal
96 information⁴⁶, or of motor simulation predicted by motor theories. Certainly, these and
97 other non-auditory strategies are used during normal human speech perception. Neverthe-
98 less, speech perception is possible without these cues, so any neurocomputational theory of
99 phonetic perception must be able to explain the purely auditory case. Animal models al-

¹⁰⁰ low straightforward isolation of purely auditory phonetic categorization without interference
¹⁰¹ from motor, semantic, syntactic, or other non-auditory cues.

¹⁰² Third, it is difficult to control for prior language experience in humans. Experience-
¹⁰³ dependent effects on phonetic perception are present from infancy⁴⁷. It can therefore be
¹⁰⁴ challenging to separate experience-driven effects from innate neurocomputational constraints
¹⁰⁵ imposed by the auditory system. Completely language-naive subjects (such as animals)
¹⁰⁶ allow the precise control of language exposure, permitting phonetics and phonology to be
¹⁰⁷ disentangled in neurolinguistics.

¹⁰⁸ Animal models of phonetic perception are a useful way to avoid these confounds, and
¹⁰⁹ provide an important alternative to human studies for empirically grounding the develop-
¹¹⁰ ment of neurolinguistic theories. The mouse is particularly well-suited to serve as such a
¹¹¹ model. A growing toolbox of powerful electrophysiological and optogenetic methods in mice
¹¹² has allowed unprecedented precision in characterizing neural circuits and the computations
¹¹³ they perform.

¹¹⁴ **E. The utility of phonetics for auditory neuroscience**

¹¹⁵ Conversely, auditory neuroscience stands to benefit from the framework provided by
¹¹⁶ phonetics for studying how sound is transformed to meaning. Understanding how complex
¹¹⁷ sounds are encoded and processed by the auditory system, ultimately leading to perception
¹¹⁸ and behavior, remains a challenge for auditory neuroscience. For example, it has been
¹¹⁹ difficult to extrapolate from simple frequency/amplitude receptive fields to understand the
¹²⁰ hierarchical organization of complex feature selectivity across brain areas. A great strength

121 of neuroethological model systems such as the songbird is that both the stimulus (e.g., the
122 bird's own song) and the behavior (song perception and production) are well understood.
123 This has led to significant advances in understanding the hierarchical organization and
124 function of the song system^{49,50}. The long history of speech research in humans has
125 produced a deep understanding of the relationships between acoustic features and phonetic
126 perception⁵¹. These insights have enabled specific predictions about what kinds of neuronal
127 selectivity for features (and combinations of features) might underlie phonetic perception⁵².
128 Although recognizing human speech sounds is not a natural ethological behavior for mice,
129 phonetics nevertheless provides a valuable framework for studying how the brain encodes
130 and transforms complex sounds into perception and behavior.

131 Here we trained mice to discriminate between pitch-shifted recordings of naturally pro-
132 duced consonant-vowel (CV) pairs beginning with either /g/ or /b/. Mice demonstrated the
133 ability to generalize consonant identity across novel vowel contexts and speakers, consistent
134 with true category learning. To our knowledge this is the first demonstration that any ani-
135 mal can generalize consonant identity across both novel vowel contexts and novel speakers.
136 These results indicate that mice can solve the non-invariance problem, and suggest that
137 mice are a suitable model for studying phonetic perception.

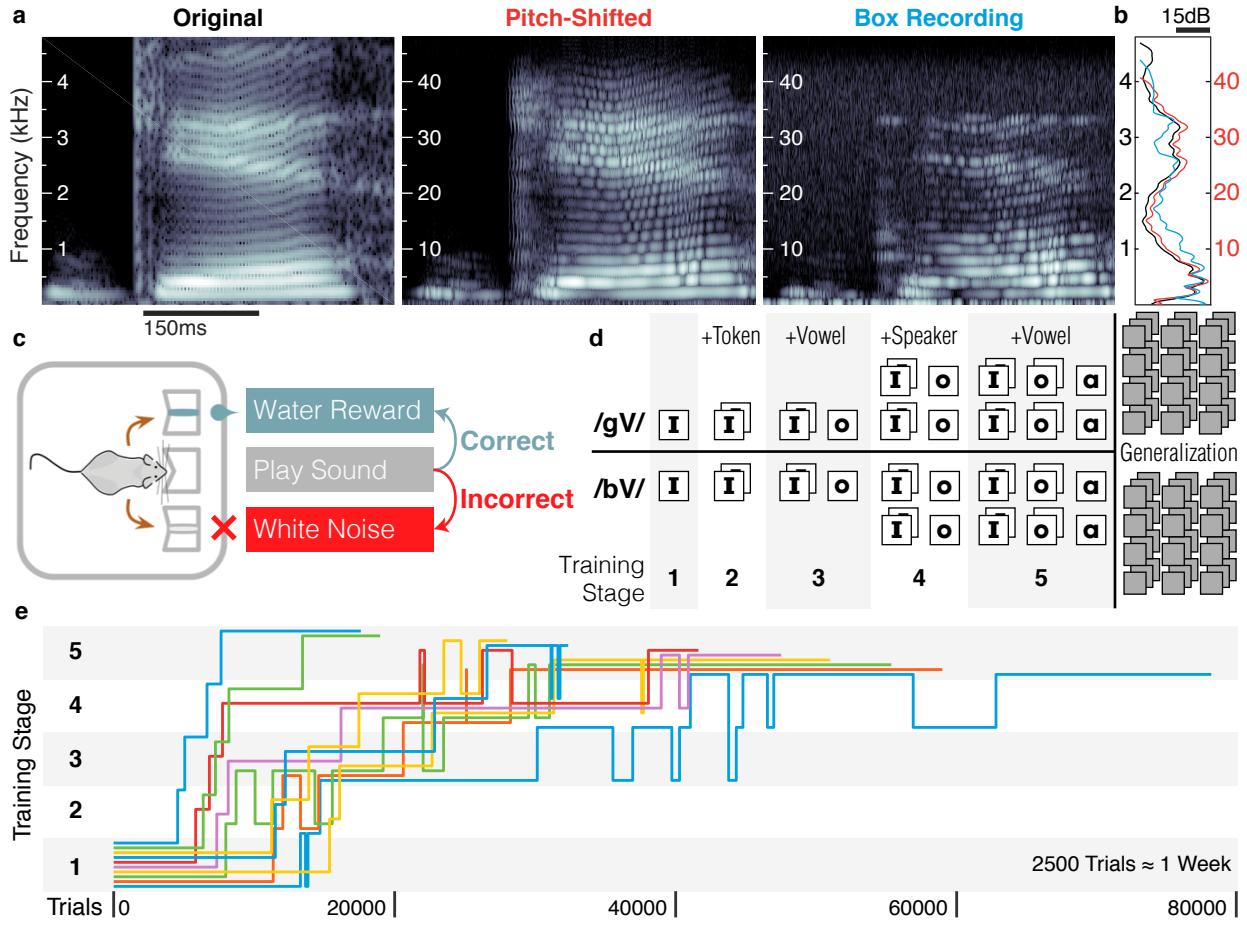


FIG. 1. Stimuli and Task Design. **a)** Spectrograms of stimuli. Left: Example of an original recording of an isolated consonant-vowel token (/gV/). Center: the same token pitch-shifted upwards by 10x (3.3 octaves) into the mouse hearing range. Right: Recording of the pitch-shifted token presented in the behavior box. Stimuli retained their overall acoustic structure below 34kHz (the upper limit of the speaker frequency response). **b)** Power spectra (dB, Welch's method) of tokens in **a**. Black: Original (left frequency axis), red: Pitch-shifted (right frequency axis), blue: Box Recording (right frequency axis). **c)** Mice initiated a trial by licking in a center port and responded by licking on one of two side ports. Correct responses were rewarded with water and incorrect responses were punished with a mildly-aversive white noise burst. **d)** The difficulty of the task was gradually expanded by adding more tokens (squares), vowels (labels), and speakers (rows) before the mice were tested on novel tokens in a generalization task. **e)** Mice (colored lines) varied widely in the duration of training required to reach the generalization phase. Mice were returned to previous levels if they remained at chance performance after reaching a new stage.

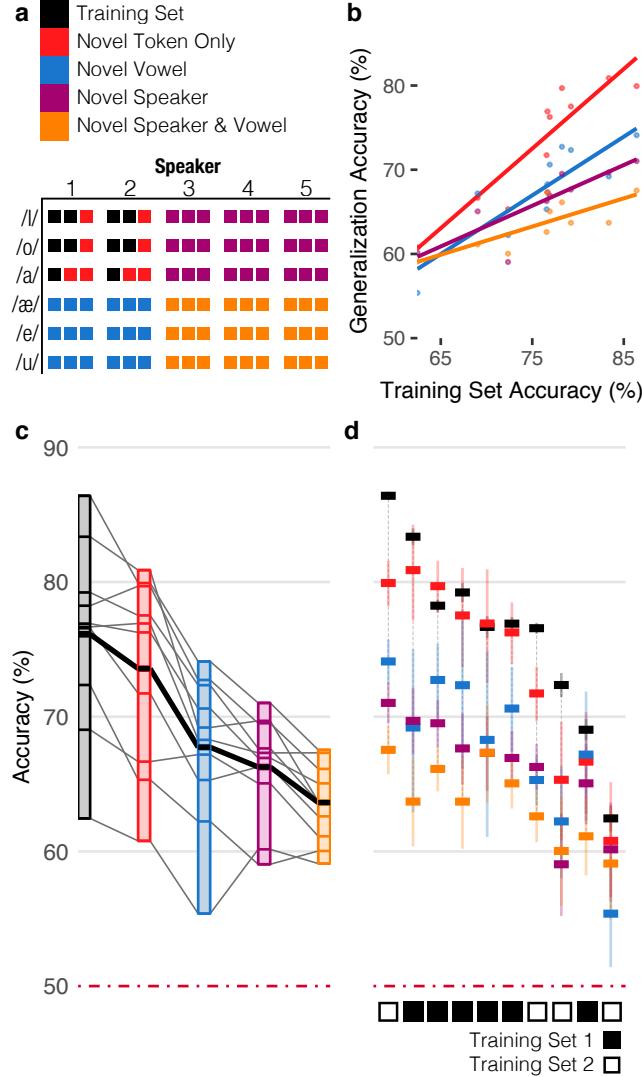


FIG. 2. Generalization accuracy by novelty class. Mice generalized stop consonant discrimination to novel CV recordings. **a)** Four types of novelty are possible with our stimuli: novel tokens from the speakers and vowels used in the training set (red), novel vowels (blue), novel speakers (purple), and novel speakers with novel vowels (orange). Tokens in the training set are indicated in black. Colors same throughout. **b)** Mice that performed better on the training set were better at generalization. Each point shows the performance for a single mouse on a given novelty class, plotted against that mouse's performance on training tokens presented during the generalization phase (both averaged across the entire generalization phase). Lines show linear regression for each novelty class. **c)** Mean accuracy for each novelty class (gray lines indicate individual mice). **d)** Mean accuracy for individual mice (colored bars indicate each novelty class). Error bars in **d** are 95% binomial confidence intervals. Mice were assigned one of two sets of training tokens, black and white boxes in **d**.

¹³⁸ II. RESULTS

¹³⁹ A. Generalization performance

¹⁴⁰ We began training 23 mice to discriminate between consonant-vowel (CV) pairs beginning
¹⁴¹ with either /b/ or /g/ in a two-alternative forced choice task. CV tokens were pitched-shifted
¹⁴² up into the mouse hearing range (Fig. 1a-b). Each mouse began training with a pair of
¹⁴³ tokens (individual recordings) in a single vowel context (ie. /bI/ and /gI/) from a single
¹⁴⁴ speaker, and then advanced through stages that progressively introduced new tokens, vowels,
¹⁴⁵ and speakers (Fig. 1c-d, see Methods). Training was discontinued in 13 (56.5%) of these
¹⁴⁶ mice because their performance on the first stage was not significantly better than chance
¹⁴⁷ after two months. The remaining 10 (43.5%) mice progressed through all the training stages
¹⁴⁸ to reach a final generalization task, on average in 14.9 ($\sigma \pm 7.8$) weeks (Fig. 1e). This success
¹⁴⁹ rate and training duration suggest that the task is difficult but achievable.

¹⁵⁰ We note that this training time is similar to that reported previously for rats (14 ± 0.3
¹⁵¹ weeks¹⁹). Previous studies have not generally reported success rates. Human infants also
¹⁵² vary in the rate and accuracy of their acquisition of phonetic categories⁹⁴, so we did not
¹⁵³ expect perfect accuracy from every mouse. The cause of such differences in ability is itself
¹⁵⁴ an opportunity for future study.

¹⁵⁵ Generalization is an essential feature of categorical perception. By testing whether mice
¹⁵⁶ can generalize their phonetic categorization to novel stimuli, we can distinguish whether
¹⁵⁷ mice actually learn phonetic categories or instead just memorize the reward contingency for
¹⁵⁸ each training token. Four types of novelty are possible with our stimuli: new tokens from

159 the speakers and vowel contexts used in the training set, new vowels, new speakers, and new
160 vowels from new speakers (colored groups in Fig. 2a). In the final generalization stage, we
161 randomly interleaved tokens from each of these novelty classes on 20% of trials, with the
162 remaining 80% consisting of tokens from the training set. We interleaved novel tokens with
163 training tokens for two reasons: (1) to avoid a sudden increase in task difficulty, which can
164 degrade performance, and (2) to minimize the possibility that mice could learn each new
165 token by widely separating them in time (on average, generalization tokens were repeated
166 only once every five days).

167 We looked for 4 hallmarks of generalization: (1) Mice should be able to accurately categorize
168 novel tokens, (2) performance should reflect the quality of the acoustic-phonetic criteria
169 learned in training, (3) performance on novel tokens should be correspondingly worse for
170 tokens that differ more from those in the training set, and (4) accurate categorization of
171 novel tokens should not require additional reinforcement.

172 All 10 mice were able to categorize tokens of all generalization types with an accuracy
173 significantly greater than chance. We estimated the impact of each generalization class on
174 performance as a fixed factor nested within each mouse as a random factor in a mixed-
175 effects logistic regression (see Methods). The predicted accuracy for each generalization
176 class is shown in Table I, each providing an estimate of the difficulty of that class after
177 accounting for the random effects of individual mice.

178 Performance on all generalization types was strongly and positively correlated with per-
179 formance on the training set (Fig. 2b, adj. $R^2=0.74$, $F(4, 5) = 7.4$, $p < 0.05$). If mice were
180 “overfitting,” that is, memorizing the training tokens rather than learning categories, then

181 we would expect the opposite (i.e., above some threshold, mice that performed better on
182 the training set would perform correspondingly worse on the generalization set). It appears
183 instead that better prototypes or decision boundaries learned in the training stages allowed
184 better generalization to novel tokens.

185 Mice were better at some types of generalization than others (Fig. 2c). The estimates of
186 their relative difficulty (Fig. 2c) provide a ranking of the perceptual novelty of the stimulus
187 classes based on their similarity to the training tokens. From easiest to hardest, these were:
188 novel token, novel vowel, novel speaker (which was not significantly more difficult than novel
189 vowel), novel speaker & vowel. The effects of generalizing to novel vowels and novel speakers
190 were not significantly different from each other, but pairwise comparisons between each of
191 the other types of generalization were (Tukey's method, all $p < 0.001$, also see confidence
192 intervals in Table I).

193 Although the effect of each generalization type on performance was significantly different
194 between mice (Likelihood Ratio Test, $\chi^2(14) = 407.22, p \ll 0.001$), they were highly corre-
195 lated (see Table I). The relative consistency of novelty type difficulty across mice (ie. the
196 correlation of fixed effects, Fig. 2c) is striking, but our results cannot distinguish whether it
197 is due to the mice or the stimuli: it is unclear whether the acoustic/phonetic criteria learned
198 by all mice are similarly general, or whether the “cost” of each type of generalization is
199 similar across an array of possible acoustic/phonetic criteria.

200 True generalization requires that one set of discrimination criteria can be successfully
201 applied to novel cases without reinforcement. It is possible that the mice were instead able
202 to rapidly learn the reward contingency of novel tokens during the generalization stage. If

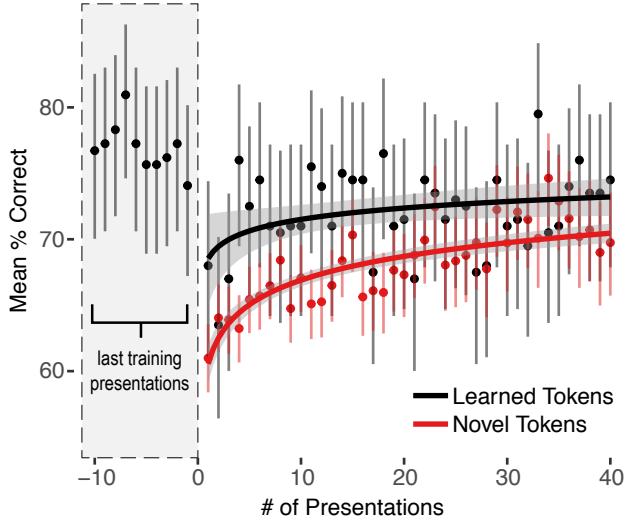


FIG. 3. Learning curve for novel tokens. Performance for both novel and training set tokens dropped transiently and recovered similarly after the transition to the generalization stage. Presentation 0 corresponds to the transition to the generalization stage. The final ten trials before the transition are shown in the gray dashed box. Mean accuracy and 95% binomial confidence intervals are collapsed across mice for novel (red, all novelty classes combined) or learned (black) tokens, by number of presentations in the generalization task. Logistic regression of binomial correct/incorrect responses fit to log-transformed presentation number (lines, shading is smoothed standard error).

203 mice were learning rapidly rather than generalizing, this would predict that novel token per-
 204 formance (1) would be indistinguishable from chance on the first presentation, and (2) would
 205 increase relative to performance on already-learned tokens with repeated presentations.

206 Performance on the first presentation of novel tokens was significantly greater than chance
 207 (Fig. 3, all mice, all tokens from all novelty classes: one-sided binomial test, $n=1410$, $P_{correct}$
 208 = 0.61, lower 95% CI = 0.588, $p \ll 0.001$; all mice, worst novelty class: $n=458$, $P_{correct} =$
 209 0.581, lower 95% CI = 0.541, $p < 0.001$). This demonstrates that mice were able to generalize
 210 immediately without additional reinforcement. Although performance on novel tokens did

211 increase with repetition, so did performance on training tokens (Fig. 3). We noted that
212 performance on all tokens (both novel and previously learned tokens) transiently dropped
213 after each transition between task stages, suggesting a non-specific effect of an increase in
214 task difficulty. To distinguish an increase in performance due to learning from an increase
215 due to acclimating to a change in the task, we compared performance on generalization and
216 training tokens over the first 40 presentations of each token. If the mice were learning the
217 generalization tokens, the increase in performance with repeated presentations should be
218 significantly greater than that of the already trained tokens.

219 Performance was well fit by a logistic regression of correct/incorrect responses from each
220 mouse against the novelty of a token (trained vs. novel tokens), and the number of times
221 it had been presented (Fig. 3). The effect of the number of presentations on accuracy was
222 not significantly different for novel tokens compared to trained tokens (interaction between
223 novelty and the number of presentations: Wald test, $z = 1.239$, 95% CI = [-0.022, 0.1],
224 $p=0.215$). This was also true when the model was fit with the generalization types
225 themselves rather than trained vs. novel tokens (most significant interaction, generalization
226 to novel speakers x number of presentations: Wald test, $z = 1.425$, 95% CI = [-0.018, 0.117],
227 $p=0.154$) and with different numbers of repetitions (10: $z = -0.219$, 95% CI = [-0.161,
228 0.13], $p=0.827$; 20: $z = -0.521$, 95% CI = [-0.116, 0.068], $p=0.602$). This indicates that the
229 asymptotic increase in performance on novel tokens was a general effect of adapting to a
230 change in the task rather than a learning period for the novel stimuli.

231 In summary, the behavior of the mice is consistent with an ability to generalize some
232 learned acoustic criteria to novel stimuli. It is unlikely that the mice rapidly learned the novel

233 tokens because (1) performance on the first presentation of novel tokens was significantly
234 above chance, (2) performance on subsequent presentations of novel tokens did not improve
235 compared to trained tokens, and (3) learning each token would have to take place over
236 unrealistically long timescales: there were an average of 2355 trials (5 days) between the
237 first and second presentation of each novel token.

238 **B. Training Set Differences**

239 One strength of studying phonetic perception in animal models is the ability to precisely
240 control exposure to speech sounds. To test whether and how the training history impacted
241 the pattern of generalization, we divided mice into two cohorts trained with different sets of
242 speech tokens. In the first cohort ($n = 6$ mice), mice were trained with tokens from speakers
243 1 and 2 (speaker number in Fig. 4a), whereas the second cohort ($n = 4$ mice) were trained
244 on speakers 4 and 5.

245 The two training cohorts had significantly different patterns of which tokens were accu-
246 rately categorized (Fig. 4a, Likelihood-Ratio test, regression of mean accuracy on tokens
247 with and without token x cohort interaction: χ^2_{161} , $p \ll 0.001$). Put another way, accuracy
248 patterns were markedly similar within training cohorts: cohort differences accounted for
249 fully 40.6% of all accuracy variance (sum of squared-error) between tokens.

250 Mice from the second training cohort were far more likely to report novel tokens as a /g/
251 than the first cohort (Fig. 4b), an effect that was not significantly related to their overall
252 accuracy ($b=0.351$, $t(8) = 2.169$, $p=0.062$). Since the only difference between these mice
253 were the tokens they were exposed to during training (they were trained contemporaneously

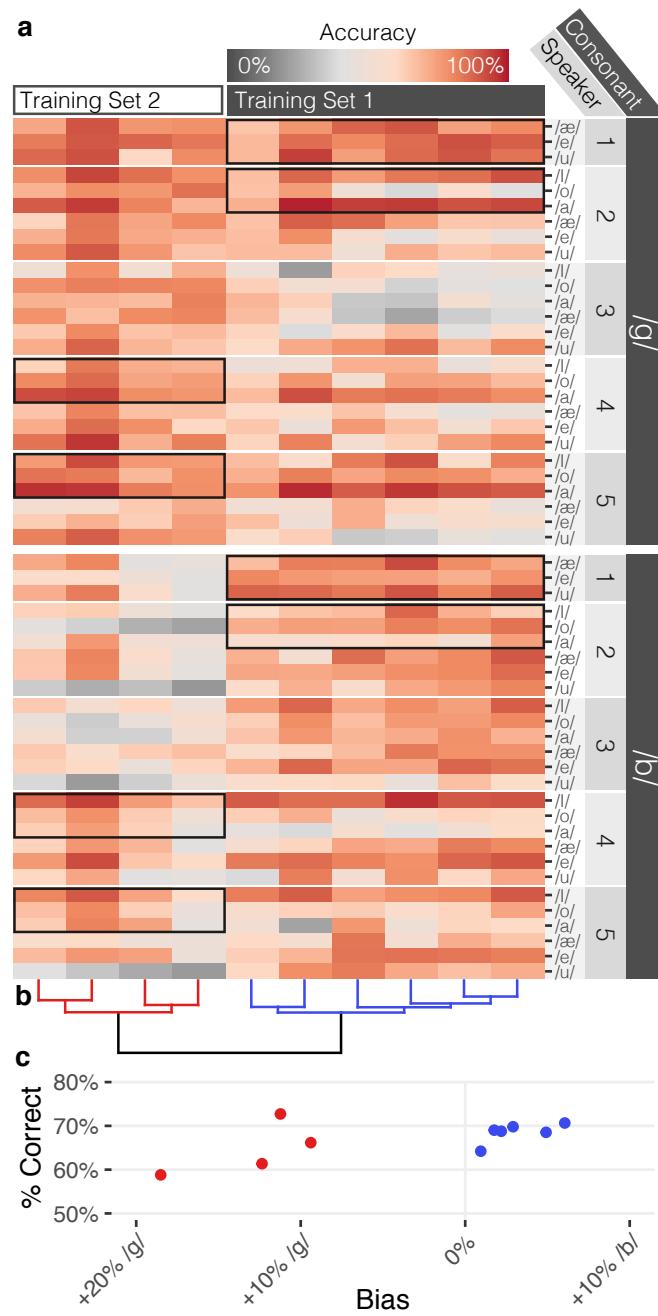


FIG. 4. Patterns of individual and group variation. **a)** Mean accuracy (color, scale at top) for each mouse (columns) on tokens grouped by consonant, speaker, and vowel (rows). The different training sets (cells outlined with black boxes) led to different patterns of accuracy on the generalization set. **b)** Ward clustering dendrogram, colored by cluster. **c)** Training set cohorts differed in bias but not mean accuracy.

254 in the same boxes), we interpret this response bias as the influence of the training tokens
255 on whatever acoustic cues the mice had learned in order to perform the generalization task.
256 This suggests that the acoustic properties of training set 2 caused the /g/ “prototype” to
257 be overbroad.

258 We searched for additional sub-cohort structure with hierarchical clustering (Ward’s
259 Method, dendrogram in fig 4b). Within each training cohort there appeared to be two
260 additional clusters of accuracy patterns. Though our sample size was too small to mean-
261 ingfully interpret these clusters, they raise the possibility that even when trained using the
262 same set of sounds mice might learn multiple sets of rules to distinguish between consonant
263 classes.

264 **C. Acoustic-behavioral correlates**

265 Humans can flexibly use several acoustic features such as burst spectra and formant
266 transitions to discriminate plosive consonants, and we wondered to what extent mice were
267 sensitive to these same features.

268 One dominant acoustic cue for place of articulation in stop consonants is the transition of
269 the second formant following the plosive burst^{52,55,98}. Formant transitions are complex and
270 dependent on vowel context, but tokens for a given place of articulation cluster around a line
271 – or “locus equation” – relating F2 frequency at release to its mid-vowel steady-state^{52,55}
272 (Fig. 5a). If mice were sensitive to this cue, the distance from both locus equation lines
273 should influence response. For example, a /g/ token between the locus equation lines should
274 have a greater rate of misclassification than a token at an equal distance above the red /g/

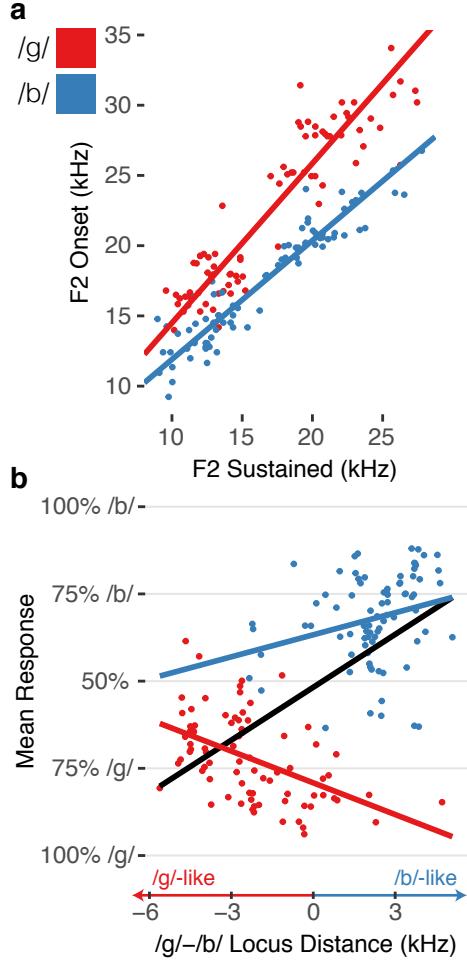


FIG. 5. Acoustic-Behavior Correlates F2 Onset-Vowel transitions do not explain observed response patterns. **a)** Locus equations relating F2 at burst onset and vowel steady state (sustained) for each token (points), split by consonant (colors, same as **b**). **b)** As the difference of a token's distance from the ideal /g/ and /b/ locus equation lines increased (x axis, greater distance from /g/, smaller distance from /b/), /b/ tokens obeyed the predicted categorization while /g/ tokens did not (slopes of colored lines).

line. Therefore we tested how classification depended on the difference of distances from each line ($/g/$ distance - $/b/$ distance, which we refer to as “locus difference”).

Mean responses to tokens (ranging from 100% $/g/-$ 100% $/b/-$) were correlated with locus differences (black line, Fig. 5b). However, it is important to note that this correlation does

279 not necessarily demonstrate that mice relied on this acoustic cue. Because multiple acoustic
280 features are correlated with consonant identity, performance that is correlated with one such
281 cue would also be correlated with all the others. The mice learned some acoustic property
282 of the consonant classes, and since the acoustic features are all highly correlated with one
283 another, they are all likely to correlate with mean responses. To distinguish whether mice
284 specifically relied on F2 locus distance, we therefore measured the marginal effect of this
285 acoustic cue within a consonant class. This is shown by the slopes of the red and blue lines
286 in Fig. 5b. For example, is a /g/ token that is further away from the blue /b/ line more
287 likely to be identified as a /g/ than one very near the /b/ line? Mean responses to /g/
288 tokens were negatively correlated with locus distance (Mean response /g/ to /b/ between
289 0 and 1, $b=-0.028\text{kHz}$, 95% CI = [-0.035, -0.022], $p \ll 0.001$). In other words, tokens
290 that should have been more frequently confused with /b/ were actually more likely to be
291 classified as /g/. Note the red points at locus distance of zero in Fig. 5b: these tokens
292 have an equal distance from both the /b/ and /g/ locus equation prototypes but are some
293 of the most accurately categorized /g/ tokens. /b/ tokens obeyed the predicted direction of
294 locus distance ($b=0.049$, 95% CI = [0.039, 0.06], $p \ll 0.001$), but the effect was very small:
295 moving one standard deviation ($\sigma_{/b}/=1.618\text{kHz}$) towards the /g/ line only changed responses
296 by 7.9%. These results suggest that mice did not rely on F2 transitions to categorize these
297 consonants.

298 We repeated this analysis separately for each training cohort to test whether the two co-
299 horts could have developed different acoustic templates that better explained their response
300 patterns. We derived cohort-specific locus-equation lines and distances using only the tokens

301 from each of their respective training sets. These models were qualitatively similar to the
302 model that included all tokens and mice and did not improve the model fit (Cohort 1: /g/:
303 $b = -0.051, 95\%CI = [-0.064, -0.038]$, /b/: $b = 0.041, 95\%CI = [0.022, 0.059]$; Cohort 2:
304 /g/: $b = -0.022, 95\%CI = [-0.031, -0.014]$, /b/: $b = 0.055, 95\%CI = [0.042, 0.069]$).

305 We conclude that while our stimulus set had the expected F2 formant transition structure,
306 this was unable to explain the behavioral responses we observed both globally and within
307 training cohorts. There are, of course, many more possible acoustic parameterizations to
308 test, but the failure of F2 transitions to explain our behavioral data is notable because of its
309 perceptual dominance in humans and its common use in parametrically synthesized speech
310 sounds. This demonstrates one advantage of using natural speech sounds: mice trained
311 on synthesized speech that varied parametrically only on F2 transitions would likely show
312 sensitivity to this cue, but this does not mean that mice show the same feature sensitivity
313 when trained with natural speech. Preserving the complexity of natural speech stimuli is
314 important for developing a general understanding of auditory category learning.

315 III. DISCUSSION

316 These results demonstrate that mice are capable of learning and generalizing phonetic
317 categories. Indeed, this is the first time to our knowledge that mice have been trained to
318 discriminate between any classes of natural, non-species-specific sounds. Thus mice join
319 a number of model organisms that have demonstrated categorical learning with speech
320 sounds^{5,15,17–21}, making a new suite of genetic and electrophysiological tools available for
321 phonetic research.

322 Two subgroups of our mice that were trained using different sets of speech tokens demon-
323 strated distinct patterns of consonant identification, presumably reflecting differences in
324 underlying acoustic prototypes. The ability to precisely control exposure to speech sounds
325 provides an opportunity to probe the neurocomputational constraints that govern the pos-
326 sible solutions to consonant identification.

327 Here we opted to use naturally recorded speech tokens in order to demonstrate that mice
328 could perform a “hard version” of phonetic categorization that preserves the full complexity
329 of the speech sounds and avoids *a priori* assumptions about the parameterization of phonetic
330 contrasts. Although our speech stimuli had the expected F2 formant transition structure,
331 that did not explain the response patterns of our mice. This suggests that the acoustic rules
332 that mice learned are different from those that would be learned from synthesized speech
333 varying only along specifically chosen parameters.

334 Future experiments using parametrically synthesized speech sounds are a critical next
335 step, and will support a qualitatively different set of inferences. Being able to carefully
336 manipulate reduced speech sounds is useful to probe the acoustic cue structure of learned
337 phonetic categories, but the reduction in complexity that makes them useful also makes
338 it correspondingly more difficult to probe the learning and recognition mechanisms for a
339 perceptual category that is defined by multiple imperfect, redundant cues. It is possible
340 that the complexity of natural speech may have caused our attrition rate to be higher,
341 and task performance lower, than other sensory-driven tasks. Neither of those concerns,
342 however, detracts from the possibility for the mouse to shed mechanistic insight on phonetic

343 perception. Indeed, error trials may provide useful neurophysiological data about how and
344 why the auditory system fails to learn or perceive phonetic categories.

345 We hope in future experiments to directly test predictions made by neurolinguistic models
346 regarding phonetic acquisition and discrimination. For example, one notable model proposes
347 that consonant perception relies on combination-sensitive neurons that selectively respond
348 to specific combinations of acoustic features⁵². This model predicts that mice trained to
349 discriminate stop consonants would have neurons selective for the feature combinations
350 that drive phoneme discrimination, perhaps in primary or higher auditory cortical areas.
351 Combination-selective neurons have been observed in A1^{63,64}, and speech training can alter
352 the response properties of A1 neurons in rats¹⁹, but it is unclear whether speech training
353 induces combination-selectivity that would facilitate phonetic discrimination.

354 The ability to record from hundreds of neurons in awake behaving animals using tetrode
355 electrophysiology or 2-photon calcium imaging presents exciting opportunities to test predic-
356 tions like these. Should some candidate population of cells be found with phonetic selectivity,
357 the ability to optogenetically activate or inactivate specific classes of neurons (such as ex-
358 citatory or inhibitory cell types, or specific projections from one region to another) could
359 shed light on the circuit computations and transformations that confer that selectivity.

³⁶⁰ **IV. METHODS**

³⁶¹ **A. Animals**

³⁶² All procedures were performed in accordance with National Institutes of Health guide-
³⁶³ lines, as approved by the University of Oregon Institutional Animal Care and Use Committee.

³⁶⁴ We began training 23 C57BL/6J mice to discriminate and generalize stop consonants
³⁶⁵ in CV (consonant-vowel) pairs. 13 mice failed to learn the task (see Training, below). 10
³⁶⁶ mice (43.5%) progressed through all training stages and reached the generalization task
³⁶⁷ in an average 14.9 ($\sigma = 7.8$) weeks. Mean age at training onset was 8.1 ($\sigma = 2$) weeks,
³⁶⁸ and at discontinuation of training was 50.6 ($\sigma = 11.2$) weeks. Sex did not significantly
³⁶⁹ affect the probability of passing or failing training (Fisher's Exact Test: $p = 0.102$), nor
³⁷⁰ did the particular behavioral chamber used for training ($p = 0.685$), nor age at the start of
³⁷¹ training (Logistic regression: $z = 1.071$, $p = 0.284$). Although this task was difficult, our
³⁷² training time (14±0.3 weeks as in¹⁹), and accuracy (generalization: 76%⁵, training tokens
³⁷³ only: 84.1%¹⁹) are similar to comparable experiments in other animals.

³⁷⁴ **B. Speech stimuli**

³⁷⁵ Speech stimuli were recorded in a sound-attenuating booth with a head-mounted micro-
³⁷⁶ phone attached to a Tascam DR-100mkII handheld recorder sampling at 96kHz/24bit. Each
³⁷⁷ speaker produced a set of 3 recordings (tokens) of each of 12 CV pairs beginning with either
³⁷⁸ /b/ or /g/, and ending with /ɪ/, /o/, /a/, /æ/, /ɛ/, /u/. To reduce a slight hiss that was
³⁷⁹ present in the recordings, they were denoised using a Daubechies wavelet with two vanishing

380 moments in MATLAB. The typical human hearing range is 20 Hz - 20 kHz, whereas the
381 mouse hearing range is 1 kHz - 80 kHz⁸³. The F_0 of our recorded speech sounds ranged
382 from 100 - 200 Hz, which is well below the lower frequency limit of the mouse hearing range.
383 We therefore pitch shifted all stimuli upwards by 10x (3.3 octaves) in MATLAB¹⁰⁰. This
384 shifted all spectral information equally upwards into an analogous part of mouse hearing
385 range while preserving temporal information unaltered.

386 Tokens from five speakers (one male - speaker 1 throughout, four female - speakers 2-
387 5 throughout) were used. Three vowel contexts (/æ/, /ɛ/, and /u/) were not recorded
388 from one speaker. It is unlikely that this had any effect on our results, as our primary
389 claims are based on the ability to generalize at all, rather than generalization to tokens
390 from a particular speaker. Tokens were normalized to a common mean amplitude, but were
391 otherwise unaltered to preserve natural variation between speakers — indeed, preserving
392 such variation was the reason for using naturally recorded rather than synthesized speech.

393 Formant frequency values were measured manually using Praat⁸⁴. F2 at onset was mea-
394 sured at its center as soon as it was discernible, typically within 20 ms of burst onset, and
395 at vowel steady-state, typically 150-200ms after burst onset.

396 **C. Training**

397 We trained mice to discriminate between CV pairs beginning with /b/ or /g/ in a two-
398 alternative forced choice task. Training sessions lasted approximately 1 hour, 5 days a week.
399 Each custom-built sound-attenuating training chamber contained two free-field JBL Duet
400 speakers for stimulus presentation with a high-frequency rolloff of 34 kHz, and a smaller 15

401 x 30 cm plastic box with three “lick ports.” Each lick port consisted of a water delivery
402 tube and an IR beam-break sensor mounted above the tube. Beam breaks triggered water
403 delivery by actuating a solenoid valve. Water-restricted mice were trained to initiate each
404 trial with an unrewarded lick at the center port, which started playback of a randomly
405 selected stimulus, and then to indicate their stimulus classification by licking at one of the
406 ports on either side. Tokens beginning with /g/ were always on the left, with /b/ on the
407 right. Two cohorts were trained on two separate sets of tokens. Training set 1 started
408 with speaker 1 (Fig. 4a) and had speaker 2 introduced on the fourth stage, where Training
409 set 2 started training with speaker 5 and had speaker 4 introduced on the fourth stage.
410 Correct classifications received ~10 µL water rewards, and incorrect classifications received
411 a 5s time-out that included a mildly aversive 60 dB SPL white noise burst.

412 Training advanced in stages that progressively increased the number of tokens, vowel con-
413 texts, and speakers. Mice first learned a simple pure-tone frequency discrimination task to
414 familiarize them with the task and shape their behavior; the tones were gradually replaced
415 with the two CV tokens of the first training stage. CV discrimination training proceeded
416 in 5 stages outlined in Table 2. Mice automatically graduated from each stage when 75%
417 of the preceding 300 trials were answered correctly. In a few cases, a mouse was returned
418 to the previous stage if its performance fell to chance for more than a week after graduat-
419 ing. Training was discontinued after two to three months if performance in the first stage
420 never rose above chance. Mice that reached the final training stage were allowed to reach
421 asymptotic performance, and then advanced to a generalization task.

422 In the generalization task, stimuli from the set of all possible speakers, vowel contexts,
423 and tokens (140 total, not including the stage 5 stimulus set) were randomly presented
424 on 20% of trials and the stage 5 stimulus set was used on the remaining 80%. Training
425 tokens were drawn from a uniform random distribution so that each was equally likely to
426 occur during both the stage 5 training and generalization phases. Novel tokens were drawn
427 uniformly at random by their generalization class, but since there were unequal numbers
428 of tokens in each class (Novel token only: 16 tokens, Novel Vowel: 36, Novel Speaker: 54,
429 Novel Speaker + Vowel: 54), tokens in each class had an unequal number of presentations.
430 We note that the logistic regression analysis with restricted maximum likelihood that we
431 used is robust to unequal sample sizes⁸⁵.

432 D. Data analysis

433 Data were excluded from days on which a mouse had a > 10% drop in accuracy from
434 their mean performance on the previous day ($44/636 = 7\%$ of sessions). Anecdotally, mice
435 are sensitive to environmental conditions (e.g., thunderstorms), so even though all efforts
436 were made to minimize variation between days, even the best performing mice had “bad
437 days” where they temporarily fell to near-chance performance and exhibited strong response
438 bias. We thus assume these “bad days” were the result of temporary environmental or other
439 performance issues, and were unrelated to the difficulty of the task itself.

440 All analyses were performed in R (R version 3.5.1 (2018-07-02))⁸⁶ using RStudio (1.1.456)⁸⁷.
441 Generalization performance was modeled using a logistic generalized linear mixed model
442 (GLMM) using the R package “lme4”⁸⁸. Binary correct/incorrect responses were fit hierar-

443 chically to models of increasing complexity (see Table V), with a final model consisting of
444 the generalization class (as in Fig. 2a: training tokens, novel tokens from the speakers and
445 vowels in the training set, novel speaker, novel vowel, and novel speaker and vowel) as a
446 fixed effect with random slopes and intercepts nested within each mouse as a random effect.
447 There was no evidence of overdispersion (i.e., deviance \approx degrees of freedom, or less than \sim
448 2 times degrees of freedom), and the profile of the model showed that the deviances by each
449 fixed effect were approximately normal. Accordingly, we report Wald confidence intervals.
450 We also computed bootstrapped confidence intervals, which had only minor disagreement
451 with the Wald confidence intervals and agreed with our interpretation in the text.

452 Clustering was performed with the “cluster”⁸⁹ package. Ward clustering split the mice
453 into two notable clusters, which are plotted in Fig. 4.

454 We estimated locus equations relating F2 onset and F2 vowel using total least squares
455 linear regression. The locus equations of the /b/ and /g/ tokens accounted for 97.3% and
456 95.9% of the variance in the F2 measurements of our tokens, respectively.

457 Spectrograms in Figure 1a were computed with the “spectrogram” function in MAT-
458 LAB 2017b, and power spectra in Figure 1b were computed with the “pwelch” function in
459 MATLAB 2018b with the same window and overlap as 1a spectrograms.

460 The remaining analyses are described in the text and used the “binom”⁹³, “reshape”⁹⁶,
461 and “plyr”⁹⁵ packages. Data visualization and tabulation was performed with the “ggplot2,”⁹⁷
462 and “xtable”⁹⁹ packages.

463 REFERENCES

- 464 ¹L. L. Holt and A. J. Lotto, "Speech perception as categorization," Attention, Perception
465 & Psychophysics **72**(5), 1218–1227 (2010) <http://www.springerlink.com/index/10.3758/APP.72.5.1218> doi: [10.3758/APP.72.5.1218](https://doi.org/10.3758/APP.72.5.1218).
- 466 ²Y. Kronrod, E. Coppess, and N. H. Feldman, "A unified account of categorical effects in
467 phonetic perception," Psychonomic Bulletin & Review **23**(6), 1681–1712 (2016) <http://link.springer.com/10.3758/s13423-016-1049-y> doi: [10.3758/s13423-016-1049-y](https://doi.org/10.3758/s13423-016-1049-y).
- 468 ³A. M. LIBERMAN, K. S. HARRIS, H. S. HOFFMAN, and B. C. GRIFFITH, "The
469 discrimination of speech sounds within and across phoneme boundaries.,," Journal of ex-
470 perimental psychology **54**(5), 358–68 (1957) <http://www.ncbi.nlm.nih.gov/pubmed/13481283>.
- 471 ⁴J. L. Elman and D. Zipser, "Learning the hidden structure of speech.,," The Journal of the
472 Acoustical Society of America **83**(4), 1615–26 (1988) <http://www.ncbi.nlm.nih.gov/pubmed/3372872>.
- 473 ⁵K. R. Kluender, R. L. Diehl, and P. R. Killeen, "Japanese quail can learn phonetic
474 categories.,," Science (New York, N.Y.) **237**(4819), 1195–1197 (1987) <http://science/sciencemag.org/content/237/4819/1195.abstract> doi: [10.1126/science.3629235](https://doi.org/10.1126/science.3629235).
- 475 ⁶A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception
476 of the speech code.,," Psychological review **74**(6), 431–61 (1967) <http://www.ncbi.nlm.nih.gov/pubmed/4170865>.

483 ⁷E. Farnetani, “V-C-V Lingual Coarticulation and Its Spatiotemporal Domain,” in
484 *Speech Production and Speech Modelling* (Springer Netherlands, Dordrecht, 1990),
485 pp. 93–130, http://www.springerlink.com/index/10.1007/978-94-009-2037-8_5,
486 doi: [10.1007/978-94-009-2037-8_5](https://doi.org/10.1007/978-94-009-2037-8_5).

487 ⁸R. L. Diehl, A. J. Lotto, and L. L. Holt, “Speech Perception,” Annual Review of Psy-
488 chology **55**(1), 149–179 (2004) <http://www.ncbi.nlm.nih.gov/pubmed/14744213>
489 <http://www.annualreviews.org/doi/10.1146/annurev.psych.55.090902.142028> doi: [10.1146/annurev.psych.55.090902.142028](https://doi.org/10.1146/annurev.psych.55.090902.142028).

490 ⁹J. S. Perkell, D. H. Klatt, K. N. Stevens, and Symposium on Invariance and Variabil-
491 ity of Speech Processes (1983 : Massachusetts Institute of Technology), *Invariance and*
492 *variability in speech processes* (Lawrence Erlbaum Associates, 1986), p. 604.

493 ¹⁰P. Lieberman, *The biology and evolution of language* (Harvard University Press, 1984),
494 p. 138-193].

495 ¹¹A. M. Liberman and I. G. Mattingly, “The motor theory of speech perception revised,”
496 Cognition **21**(1), 1–36 (1985) <http://www.ncbi.nlm.nih.gov/pubmed/4075760> doi: [10.1016/0010-0277\(85\)90021-6](https://doi.org/10.1016/0010-0277(85)90021-6).

497 ¹²K. M. Carbonell and A. J. Lotto, “Speech is not special... again.,” Frontiers
498 in psychology **5**(June), 427 (2014) <http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00427/abstract>
499 <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4042079/>
500 tool=pmcentrez&rendertype=abstract
501 articlerender.fcgi?artid=4042079&tool=pmcentrez&rendertype=abstract
502 doi: [10.3389/fpsyg.2014.00427](https://doi.org/10.3389/fpsyg.2014.00427).

⁵⁰⁴ ¹³A. A. Ghazanfar and M. D. Hauser, "The neuroethology of primate vocal communication:
⁵⁰⁵ Substrates for the evolution of speech," (1999), doi: [10.1016/S1364-6613\(99\)01379-0](https://doi.org/10.1016/S1364-6613(99)01379-0).

⁵⁰⁶ ¹⁴I. Bornkessel-Schlesewsky, M. Schlesewsky, S. L. Small, and J. P. Rauschecker, “Neurobiological roots of language in primate audition: common computational properties,”
⁵⁰⁷ Trends in Cognitive Sciences **19**(3), 142–150 (2015) <http://linkinghub.elsevier.com/retrieve/pii/S1364661314002757> doi: 10.1016/j.tics.2014.12.008.
⁵⁰⁸

⁵¹⁰ ¹⁵A. Lotto, K. Kluender, and L. Holt, "Animal models of speech perception phenomena," Chicago Linguistic Society **33**, 357–367 (1997).

512 ¹⁶K. R. Kluender and a. J. Lotto, "Effects of first formant onset frequency on [-voice]
513 judgments result from auditory processes not specific to humans." The Journal of the
514 Acoustical Society of America **95**(2), 1044–52 (1994) [http://www.ncbi.nlm.nih.gov/
515 pubmed/8132898](http://www.ncbi.nlm.nih.gov/pubmed/8132898) doi: [10.1121/1.408466](https://doi.org/10.1121/1.408466).

⁵¹⁶ ¹⁷K. R. Klunder, “Contributions of nonhuman animal models to understanding human speech perception,” The Journal of the Acoustical Society of America **107**(5), 2835–2835 (2000). <http://asa.scitation.org/doi/10.1121/1.429153> doi: [10.1121/1.429153](https://doi.org/10.1121/1.429153).

⁵¹⁹ ¹⁸P. K. Kuhl and J. D. Miller, "Speech perception by the chinchilla: Identification functions
⁵²⁰ for synthetic VOT stimuli," The Journal of the Acoustical Society of America **63**(3), 905–
⁵²¹ 917 (1978) doi: [10.1121/1.381770](https://doi.org/10.1121/1.381770).

⁵²² ¹⁹C. T. Engineer, K. C. Rahebi, E. P. Buell, M. K. Fink, and M. P. Kilgard, "Speech training
⁵²³ alters consonant and vowel responses in multiple auditory cortex fields," *Behavioural
524 Brain Research* **287**, 256–264 (2015) <http://dx.doi.org/10.1016/j.bbr.2015.03.044>

525 doi: [10.1016/j.bbr.2015.03.044](https://doi.org/10.1016/j.bbr.2015.03.044).

526 ²⁰P. K. Kuhl and D. M. Padden, “Enhanced discriminability at the phonetic boundaries for
527 the place feature in macaques.,” The Journal of the Acoustical Society of America **73**(3),
528 1003–1010 (1983) doi: [10.3758/BF03204208](https://doi.org/10.3758/BF03204208).

529 ²¹R. J. Dooling, C. T. Best, and S. D. Brown, “Discrimination of synthetic full-formant
530 and sinewave /rala/ continua by budgerigars (*Melopsittacus undulatus*) and zebra finches
531 (*Taeniopygia guttata*),” (1995), <http://scitation.aip.org/content/asa/journal/jasa/97/3/10.1121/1.412058>, doi: [10.1121/1.412058](https://doi.org/10.1121/1.412058).

533 ²²J. K. Bizley and Y. E. Cohen, “The what, where and how of auditory-
534 object perception.,” Nature reviews. Neuroscience **14**(10), 693–707 (2013)
535 <http://www.ncbi.nlm.nih.gov/pubmed/24052177><http://www.ncbi.nlm.nih.gov/articlerender.fcgi?artid=PMC4082027> doi: [10.1038/nrn3565](https://doi.org/10.1038/nrn3565).

536 ²³J. P. Rauschecker and S. K. Scott, “Maps and streams in the auditory cortex: nonhu-
537 man primates illuminate human speech processing,” Nature Neuroscience **12**(6), 718–724
538 (2009) <http://www.nature.com/doifinder/10.1038/nn.2331> doi: [10.1038/nn.2331](https://doi.org/10.1038/nn.2331).

539 ²⁴T. J. Strauss, H. D. Harris, and J. S. Magnuson, “jTRACE: A reimplementation and
540 extension of the TRACE model of speech perception and spoken word recognition,” Be-
541 havior Research Methods **39**(1), 19–30 (2007) <http://www.springerlink.com/index/10.3758/BF03192840.html> doi: [10.3758/BF03192840](https://doi.org/10.3758/BF03192840).

542 ²⁵K. R. Kluender, C. E. Stilp, M. Kiefte, K. R. Kluender, C. E. Stilp, and M. Kiefte,
543 “Perception of Vowel Sounds Within a Biologically Realistic Model of Efficient Coding,”

- 546 in *Vowel Inherent Spectral Change*, pp. 117–151, doi: [10.1007/978-3-642-14209-3_6](https://doi.org/10.1007/978-3-642-14209-3_6).
- 547 ²⁶M. G. Gaskell and W. D. Marslen-Wilson, “Integrating Form and Meaning: A Distributed
548 Model of Speech Perception,” *Language and Cognitive Processes* **12**(5-6), 613–656
549 (1997) <http://www.tandfonline.com/doi/abs/10.1080/016909697386646> doi: [10.1080/016909697386646](https://doi.org/10.1080/016909697386646).
- 550
- 551 ²⁷D. H. Hubel and T. N. Wiesel, “Receptive fields, binocular interaction and
552 functional architecture in the cat’s visual cortex,” *The Journal of Physiology*
553 **160**(1), 106–154.2 (1962) <http://www.ncbi.nlm.nih.gov/pubmed/14449617>
554 <http://www.ncbi.nlm.nih.gov/articlerender.fcgi?artid=PMC1359523> doi: [10.1523/JNEUROSCI.1991-09.2009](https://doi.org/10.1523/JNEUROSCI.1991-09.2009).
- 555
- 556 ²⁸K. G. Ranasinghe, W. A. Vrana, C. J. Matney, and M. P. Kilgard, “Increasing diver-
557 sity of neural responses to speech sounds across the central auditory pathway,” *Neuro-
558 science* **252**, 80–97 (2013) <http://dx.doi.org/10.1016/j.neuroscience.2013.08.005>
559 doi: [10.1016/j.neuroscience.2013.08.005](https://doi.org/10.1016/j.neuroscience.2013.08.005).
- 560 ²⁹E. L. Bartlett, “The organization and physiology of the auditory thalamus and
561 its role in processing acoustic features important for speech perception.,” *Brain
562 and language* **126**(1), 29–48 (2013) <http://dx.doi.org/10.1016/j.bandl.2013.03.003>
563 doi: [10.1016/j.bandl.2013.03.003](https://doi.org/10.1016/j.bandl.2013.03.003).
- 564
- 565 ³⁰T. M. Centanni, A. M. Sloan, A. C. Reed, C. T. Engineer, R. L. Rennaker, and M. P. Kil-
566 gard, “Detection and identification of speech sounds using cortical activity patterns,” Neu-

567 roscience **258**, 292–306 (2013) <http://dx.doi.org/10.1016/j.neuroscience.2013.11.030>.
568
569 ³¹C. T. Engineer, C. A. Perez, Y. H. Chen, R. S. Carraway, A. C. Reed, J. A. Shetake,
570 V. Jakkamsetti, K. Q. Chang, and M. P. Kilgard, “Cortical activity patterns predict
571 speech discrimination ability.” *Nature neuroscience* **11**(5), 603–8 (2008) <http://www.ncbi.nlm.nih.gov/pubmed/18425123> doi: [10.1038/nn.2109](https://doi.org/10.1038/nn.2109).
572
573 ³²M. Steinschneider, Y. I. Fishman, and J. C. Arezzo, “Representation of the voice onset
574 time (VOT) speech parameter in population responses within primary auditory cortex of
575 the awake monkey.” *The Journal of the Acoustical Society of America* **114**(1), 307–321
576 (2003) doi: [10.1121/1.1582449](https://doi.org/10.1121/1.1582449).
577
578 ³³N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, “Phonetic fea-
579 ture encoding in human superior temporal gyrus.” *Science* (New York, N.Y.)
580 **343**(6174), 1006–10 (2014) <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4150233/> doi:
581 [10.1126/science.1245994](https://doi.org/10.1126/science.1245994).
582
583 ³⁴P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike, “Voice-selective areas in human
584 auditory cortex.” *Nature* **403**(6767), 309–312 (2000) doi: [10.1038/35002078](https://doi.org/10.1038/35002078).
585
586 ³⁵E. F. Chang, J. W. Rieger, K. Johnson, M. S. Berger, N. M. Barbaro, and R. T. Knight,
587 “Categorical speech representation in human superior temporal gyrus.” *Nature neuro-
science* **13**(11), 1428–32 (2010) <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2967728/> doi: [10.1038/nn.2967](https://doi.org/10.1038/nn.2967).

- 588 1038/nn.2641.
- 589 ³⁶B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T.
590 Knight, and E. F. Chang, “Reconstructing speech from human auditory cortex,” PLoS
591 Biology **10**(1) (2012) doi: [10.1371/journal.pbio.1001251](https://doi.org/10.1371/journal.pbio.1001251).
- 592 ³⁷G. M. Bidelman, S. Moreno, and C. Alain, “Tracing the emergence of categorical speech
593 perception in the human auditory system,” NeuroImage **79**, 201–212 (2013) <http://dx.doi.org/10.1016/j.neuroimage.2013.04.093> doi: [10.1016/j.neuroimage.2013.04.093](https://doi.org/10.1016/j.neuroimage.2013.04.093).
- 595 ³⁸A. Ng and M. I. Jordan, “On generative vs. discriminative classifiers: A compar-
596 ison of logistic regression and naive bayes,” Proceedings of Advances in Neural
597 Information Processing **28**(3), 169–187 (2002) <http://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-pdf.pdf>.
- 600 ³⁹F. de Saussure, *Cours de linguistique gnrale*. (Payot, Lausanne, Paris, 1916).
- 601 ⁴⁰U. Rutishauser, J. J. Slotine, and R. Douglas, “Computation in Dynamically Bounded
602 Asymmetric Systems,” PLoS Computational Biology **11**(1), e1004039 (2015) <http://dx.doi.org/10.1371/journal.pcbi.1004039> doi: [10.1371/journal.pcbi.1004039](https://doi.org/10.1371/journal.pcbi.1004039).
- 605 ⁴¹B. E. Dresher, “The contrastive hierarchy in phonology,” Contrast in phonology: theory,
606 perception, acquisition **13**, 11 (2008) doi: [10.1017/CBO9780511642005](https://doi.org/10.1017/CBO9780511642005).
- 607 ⁴²H. Blank and M. H. Davis, “Prediction Errors but Not Sharpened Signals Sim-
608 ulate Multivoxel fMRI Patterns during Speech Perception.,” PLoS Biology **14**(11),

- 609 e1002577 (2016) <http://www.ncbi.nlm.nih.gov/pubmed/27846209> doi: [10.1371/journal.pbio.1002577](https://doi.org/10.1371/journal.pbio.1002577).
- 610
611 43 P. Gagnepain, R. Henson, and M. Davis, "Temporal Predictive Codes for Spoken Words
612 in Auditory Cortex," (2012), doi: [10.1016/j.cub.2012.02.015](https://doi.org/10.1016/j.cub.2012.02.015).
- 613 44 N. P. Fox and S. E. Blumstein, "Top-down effects of syntactic sentential context on
614 phonetic processing.,," Journal of Experimental Psychology: Human Perception and Per-
615 formance **42**(5), 730–741 (2016) <http://www.ncbi.nlm.nih.gov/pubmed/26689310> doi:
616 [10.1037/a0039965](https://doi.org/10.1037/a0039965).
- 617 45 B. Schouten, E. Gerrits, and A. Van Hessen, "The end of categorical perception as
618 we know it," in *Speech Communication* (2003), Vol. 41, pp. 71–80, doi: [10.1016/S0167-6393\(02\)00094-8](https://doi.org/10.1016/S0167-6393(02)00094-8).
- 619
620 46 L. D. Rosenblum, "Speech Perception as a Multimodal Phenomenon," Current Directions
621 in Psychological Science **17**(6), 405–409 (2008) <http://journals.sagepub.com/doi/10.1111/j.1467-8721.2008.00615.x> doi: [10.1111/j.1467-8721.2008.00615.x](https://doi.org/10.1111/j.1467-8721.2008.00615.x).
- 622
623 47 P. Kuhl, K. Williams, F. Lacerda, K. Stevens, and B. Lindblom, "Linguistic experience
624 alters phonetic perception in infants by 6 months of age," Science **255**(5044) (1992).
- 625 48 L. L. Holt, A. J. Lotto, and K. R. Kluender, "Influence of fundamental frequency on stop-
626 consonant voicing perception: A case of learned covariation or auditory enhancement?,"
627 The Journal of the Acoustical Society of America **109**(2), 764–774 (2001) <http://asa.scitation.org/doi/10.1121/1.1339825> doi: [10.1121/1.1339825](https://doi.org/10.1121/1.1339825).

- 629 ⁴⁹E. A. Brenowitz, D. Margoliash, and K. W. Nordeen, “An introduction to birdsong and
630 the avian song system,” (1997).
- 631 ⁵⁰F. E. Theunissen and J. E. Elie, “Neural processing of natural sounds.,” Nature reviews.
632 Neuroscience **15**(6), 355–66 (2014) <http://www.ncbi.nlm.nih.gov/pubmed/24840800>
633 doi: [10.1038/nrn3731](https://doi.org/10.1038/nrn3731).
- 634 ⁵¹G. E. Peterson and H. L. Barney, “Control methods used in a study of the vowels,” The
635 Journal of the Acoustical Society of America **24**(2), 175–184 (1952) doi: [10.1121/1.1906875](https://doi.org/10.1121/1.1906875).
- 637 ⁵²H. M. Sussman, D. Fruchter, J. Hilbert, and J. Sirosh, “Linear correlates in the speech
638 signal: the orderly output constraint.,” The Behavioral and brain sciences **21**(2), 241–
639 59; discussion 260–99 (1998) <http://www.ncbi.nlm.nih.gov/pubmed/10097014> doi:
640 [10.1017/S0140525X98001174](https://doi.org/10.1017/S0140525X98001174).
- 641 ⁵³K. N. Stevens, *Acoustic phonetics* (MIT Press, 1998), p. 607.
- 642 ⁵⁴K. N. Stevens and S. E. Blumstein, “Invariant cues for place of articulation in stop
643 consonants,” The Journal of the Acoustical Society of America **64**(5), 1358–1368 (1978)
644 <http://asa.scitation.org/doi/10.1121/1.382102> doi: [10.1121/1.382102](https://doi.org/10.1121/1.382102).
- 645 ⁵⁵B. Lindblom and H. M. Sussman, “Dissecting coarticulation: How locus equations hap-
646 pen,” Journal of Phonetics **40**(1), 1–19 (2012) doi: [10.1016/j.wocn.2011.09.005](https://doi.org/10.1016/j.wocn.2011.09.005).
- 647 ⁵⁶M. F. Dorman, M. Studdert-Kennedy, and L. J. Raphael, “Stop-consonant recognition:
648 Release bursts and formant transitions as functionally equivalent, context-dependent
649 cues,” Perception & Psychophysics **22**(2), 109–122 (1977) <http://www.springerlink.com>.

650 com/index/10.3758/BF03198744 doi: [10.3758/BF03198744](https://doi.org/10.3758/BF03198744).

651 ⁵⁷R. Shepard, “Toward a universal law of generalization for psychological science,” Science **237**(4820), 1317–1323 (1987) <http://www.sciencemag.org/cgi/doi/10.1126/science.3629243> doi: [10.1126/science.3629243](https://doi.org/10.1126/science.3629243).

654 ⁵⁸A. S. Buchman, D. C. Garron, J. E. Trost-Cardamone, M. D. Wichter, and M. Schwartz,
655 “Word deafness: one hundred years later..,” Journal of Neurology, Neurosurgery & Psychiatry
656 **49**(5), 489–499 (1986) <http://jnnnp.bmjjournals.org/cgi/doi/10.1136/jnnnp.49.5.489>
657 doi: [10.1136/jnnnp.49.5.489](https://doi.org/10.1136/jnnnp.49.5.489).

658 ⁵⁹G. Hickok and D. Poeppel, “The cortical organization of speech processing,” Nature Reviews Neuroscience **8**(5), 393–402 (2007) <http://www.nature.com/doifinder/10.1038/nrn2113> doi: [10.1038/nrn2113](https://doi.org/10.1038/nrn2113).

661 ⁶⁰G. Hickok, “The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model,” Journal of Communication Disorders
662 **45**(6), 393–402 (2012) doi: [10.1016/j.jcomdis.2012.06.004](https://doi.org/10.1016/j.jcomdis.2012.06.004).

664 ⁶¹F. W. Ohl, W. Wetzel, T. Wagner, A. Rech, and H. Scheich, “Bilateral ablation
665 of auditory cortex in Mongolian gerbil affects discrimination of frequency modulated tones but not of pure tones..,” Learning & memory (Cold Spring Harbor,
666 N.Y.) **6**(4), 347–62 (1999) <http://www.ncbi.nlm.nih.gov/pubmed/10509706> <http://www.ncbi.nlm.nih.gov/articlerender.fcgi?artid=PMC311295> doi: [10.1101/LM.6.4.347](https://doi.org/10.1101/LM.6.4.347).

670 ⁶²E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe,
671 M. S. Boguski, K. S. Brockway, E. J. Byrnes, L. Chen, T. M. Chen, M. C. Chin,
672 J. Chong, B. E. Crook, A. Czaplinska, C. N. Dang, S. Datta, N. R. Dee, A. L. De-
673 saki, T. Desta, E. Diep, T. A. Dolbeare, M. J. Donelan, H. W. Dong, J. G. Dougherty,
674 B. J. Duncan, A. J. Ebbert, G. Eichele, L. K. Estin, C. Faber, B. A. Facer, R. Fields,
675 S. R. Fischer, T. P. Fliss, C. Frensley, S. N. Gates, K. J. Glattfelder, K. R. Halver-
676 son, M. R. Hart, J. G. Hohmann, M. P. Howell, D. P. Jeung, R. A. Johnson, P. T.
677 Karr, R. Kawal, J. M. Kidney, R. H. Knapik, C. L. Kuan, J. H. Lake, A. R. Laramee,
678 K. D. Larsen, C. Lau, T. A. Lemon, A. J. Liang, Y. Liu, L. T. Luong, J. Michaels,
679 J. J. Morgan, R. J. Morgan, M. T. Mortrud, N. F. Mosqueda, L. L. Ng, R. Ng, G. J.
680 Orta, C. C. Overly, T. H. Pak, S. E. Parry, S. D. Pathak, O. C. Pearson, R. B. Puchal-
681 ski, Z. L. Riley, H. R. Rockett, S. A. Rowland, J. J. Royall, M. J. Ruiz, N. R. Sarno,
682 K. Schaffnit, N. V. Shapovalova, T. Sivisay, C. R. Slaughterbeck, S. C. Smith, K. A.
683 Smith, B. I. Smith, A. J. Sodt, N. N. Stewart, K. R. Stumpf, S. M. Sunkin, M. Sutram,
684 A. Tam, C. D. Teemer, C. Thaller, C. L. Thompson, L. R. Varnam, A. Visel, R. M. Whit-
685 lock, P. E. Wohnoutka, C. K. Wolkey, V. Y. Wong, M. Wood, M. B. Yaylaoglu, R. C.
686 Young, B. L. Youngstrom, X. F. Yuan, B. Zhang, T. A. Zwingman, and A. R. Jones,
687 “Genome-wide atlas of gene expression in the adult mouse brain,” *Nature* **445**(7124), 168–
688 176 (2007) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&doct=Citation&list_uids=17151600 doi: [10.1038/nature05453](https://doi.org/10.1038/nature05453).

690 ⁶³S. Sadagopan and X. Wang, “Nonlinear Spectrotemporal Interactions Underly-
691 ing Selectivity for Complex Sounds in Auditory Cortex,” *Journal of Neuro-*

- 692 science **29**(36), 11192–11202 (2009) <http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1286-09.2009>. doi: [10.1523/JNEUROSCI.1286-09.2009](https://doi.org/10.1523/JNEUROSCI.1286-09.2009).
- 693
694 64 X. Wang, T. Lu, R. K. Snider, and L. Liang, “Sustained firing in auditory cortex evoked
695 by preferred stimuli.,” Nature **435**(7040), 341–6 (2005) <http://www.ncbi.nlm.nih.gov/pubmed/15902257> doi: [10.1038/nature03565](https://doi.org/10.1038/nature03565).
- 696
697 65 P. D. Eimas and J. D. Corbit, “Selective adaptation of linguistic feature detectors,”
698 Cognitive Psychology **4**(1), 99–109 (1973) <http://www.sciencedirect.com/science/article/pii/0010028573900066> doi: [10.1016/0010-0285\(73\)90006-6](https://doi.org/10.1016/0010-0285(73)90006-6).
- 700
701 66 K. G. Estes and C. Lew-Williams, “Listening through voices: Infant statistical word
702 segmentation across multiple speakers.,” Developmental Psychology **51**(11), 1517–1528
703 (2015) <http://www.ncbi.nlm.nih.gov/pubmed/26389607> doi: [10.1037/a0039725](https://doi.org/10.1037/a0039725).
- 704
705 67 V. Vapnik, “An overview of statistical learning theory,” IEEE Transactions on Neural
706 Networks **10**(5), 988–999 (1999) <http://ieeexplore.ieee.org/document/788640/> doi:
707 [10.1109/72.788640](https://doi.org/10.1109/72.788640).
- 708
709 68 D. F. Kleinschmidt and T. F. Jaeger, “Robust speech perception: Recognize the
710 familiar, generalize to the similar, and adapt to the novel.,” Psychological Review
711 **122**(2), 148–203 (2015) <http://doi.apa.org/getdoi.cfm?doi=10.1037/a0038695> doi:
712 [10.1037/a0038695](https://doi.org/10.1037/a0038695).
- 713
714 69 M. A. Mines, B. F. Hanson, and J. E. Shoup, “Frequency of Occurrence
715 of Phonemes in Conversational English,” Language and Speech **21**(3), 221–241
716 (1978) <http://journals.sagepub.com/doi/abs/10.1177/002383097802100302> doi:
717 [10.1177/002383097802100302](https://doi.org/10.1177/002383097802100302)

- 713 [10.1177/002383097802100302](https://doi.org/10.1177/002383097802100302).
- 714 ⁷⁰S. Jayaraman, C. M. Fausey, and L. B. Smith, “The Faces in Infant-Perspective Scenes
715 Change over the First Year of Life,” PLOS ONE **10**(5), e0123780 (2015) <http://dx.plos.org/10.1371/journal.pone.0123780>. doi: [10.1371/journal.pone.0123780](https://doi.org/10.1371/journal.pone.0123780).
- 716 ⁷¹E. M. Clerkin, E. Hart, J. M. Rehg, C. Yu, and L. B. Smith, “Real-world visual statistics
717 and infants’ first-learned object names,” Philosophical Transactions of the Royal Society
718 B: Biological Sciences **372**(1711) (2016) [http://rstb.royalsocietypublishing.org/
719 content/372/1711/20160055.article-info](http://rstb.royalsocietypublishing.org/content/372/1711/20160055.article-info).
- 720 ⁷²A. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-law distributions in em-
721 pirical data,” (2007) <http://arxiv.org/abs/0706.1062><http://dx.doi.org/10.1137/070710111>. doi: [10.1137/070710111](https://doi.org/10.1137/070710111).
- 722 ⁷³R. Salakhutdinov, A. Torralba, and J. Tenenbaum, “Learning to share visual appearance
723 for multiclass object detection,” in *CVPR 2011*, IEEE (2011), pp. 1481–1488, <http://ieeexplore.ieee.org/document/5995720/>, doi: [10.1109/CVPR.2011.5995720](https://doi.org/10.1109/CVPR.2011.5995720).
- 724 ⁷⁴G. Weiss and F. Provost, “The effect of class distribution on classifier learning: an
725 empirical study,” Rutgers Univ (2001) [ftp://ftp.cs.rutgers.edu/http/cs/cs/pub/
726 technical-reports/work/ml-tr-44.pdf](ftp://ftp.cs.rutgers.edu/http/cs/cs/pub/technical-reports/work/ml-tr-44.pdf).
- 727 ⁷⁵D. F. Kleinschmidt and T. F. Jaeger, “Re-examining selective adaptation: Fatigu-
728 ing feature detectors, or distributional learning?,” Psychonomic Bulletin & Review
729 **23**(3), 678–691 (2016) <http://link.springer.com/10.3758/s13423-015-0943-z> doi:
730 [10.3758/s13423-015-0943-z](https://doi.org/10.3758/s13423-015-0943-z).

- 734 ⁷⁶K. Idemaru and L. L. Holt, "Word recognition reflects dimension-based statistical learn-
735 ing.," Journal of Experimental Psychology: Human Perception and Performance **37**(6),
736 1939–1956 (2011) <http://doi.apa.org/getdoi.cfm?doi=10.1037/a0025641> doi: [10.1037/a0025641](https://doi.org/10.1037/a0025641).
- 737
- 738 ⁷⁷K. E. Moczulska, J. Tinter-Thiede, M. Peter, L. Ushakova, T. Wernle, B. Bathellier, and
739 S. Rumpel, "Dynamics of dendritic spines in the mouse auditory cortex during memory
740 formation and memory recall.," Proceedings of the National Academy of Sciences of the
741 United States of America **110**(45), 18315–20 (2013) <http://www.ncbi.nlm.nih.gov/pubmed/24151334> doi: [10.1073/pnas.1312508110](https://doi.org/10.1073/pnas.1312508110).
- 742
- 743 ⁷⁸D. B. Polley, "Perceptual Learning Directs Auditory Cortical Map Reorganiza-
744 tion through Top-Down Influences," Journal of Neuroscience **26**(18), 4970–4982
745 (2006) <http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3771-05.2006> doi:
746 [10.1523/JNEUROSCI.3771-05.2006](https://doi.org/10.1523/JNEUROSCI.3771-05.2006).
- 747
- 748 ⁷⁹R. C. Froemke, M. M. Merzenich, and C. E. Schreiner, "A synaptic memory trace for
749 cortical receptive field plasticity.," Nature **450**(7168), 425–429 (2007) <http://dx.doi.org/10.1038/nature06289> doi: [10.1038/nature06289](https://doi.org/10.1038/nature06289).
- 750
- 751 ⁸⁰J. Fritz, M. Elhilali, and S. Shamma, "Active listening: Task-dependent plasticity of
752 spectrot temporal receptive fields in primary auditory cortex," Hearing Research **206**(1-2),
753 159–176 (2005) <http://www.nature.com/doifinder/10.1038/nn1141> doi: [10.1016/j.heares.2005.01.015](https://doi.org/10.1016/j.heares.2005.01.015).

754 ⁸¹J. R. Ison, P. D. Allen, and W. E. O'Neill, “Age-related hearing loss in C57BL/6J
755 mice has both frequency-specific and non-frequency-specific components that produce
756 a hyperacusis-like exaggeration of the acoustic startle reflex,” JARO - Journal of the As-
757 sociation for Research in Otolaryngology **8**(4), 539–550 (2007) <http://link.springer.com/10.1007/s10162-007-0098-3>.
758

759 ⁸²[Https://www.mathworks.com/help/audio/examples/pitch-shifting-and-time-dilation-](Https://www.mathworks.com/help/audio/examples/pitch-shifting-and-time-dilation-using-a-phase-vocoder-in-matlab.html)
760 using-a-phase-vocoder-in-matlab.html.

761 ⁸³K. E. Radziwon, K. M. June, D. J. Stolzberg, M. A. Xu-Friedman, R. J.
762 Salvi, and M. L. Dent, “Behaviorally measured audiograms and gap detection
763 thresholds in CBA/CaJ mice.,” Journal of comparative physiology. A, Neu-
764 roethology, sensory, neural, and behavioral physiology **195**(10), 961–9 (2009)
765 <Http://www.ncbi.nlm.nih.gov/pubmed/19756650><Http://www.ncbi.nlm.nih.gov/articlerender.fcgi?artid=PMC2813807> doi: <10.1007/s00359-009-0472-1>.

766

767 ⁸⁴P. Boersma, “Praat, a system for doing phonetics by computer,” Glot International
768 **5**(9/10), 341–347 (2001) doi: <10.1097/AUD.0b013e31821473f7>.

769

770 ⁸⁵H. D. Patterson and R. Thompson, “Recovery of Inter-Block Information when Block Sizes
771 are Unequal,” Biometrika **58**(3), 545 (1971) <Http://www.jstor.org/stable/2334389?origin=crossref> doi: <10.2307/2334389>.

772

773 ⁸⁶R. C. Team, “R: A language and environment for statistical computing.,” (2016).
774 ⁸⁷R. Team, “RStudio: Integrated Development for R.,” (2015), <Http://www.rstudio.com/>.

- 774 ⁸⁸D. Bates, M. Machler, B. Bolker, and S. Walker, “Fitting Linear Mixed-Effects Models
775 Using lme4,” Journal of Statistical Software **67**(1), 1–48 (2015) doi: [10.18637/jss.v067.i01](https://doi.org/10.18637/jss.v067.i01).
- 776 *i01.*
- 777 ⁸⁹M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, “cluster: Cluster
778 Analysis Basics and Extensions,” (2017).
- 779 ⁹⁰D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” International
780 Journal of Computer Vision **60**(2), 91–110 (2004).
- 781 ⁹¹J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch,
782 S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Harten-
783 stein, K. Eliceiri, P. Tomancak, and A. Cardona, “Fiji: an open-source platform for
784 biological-image analysis,” Nature Methods **9**(7), 676–682 (2012) <http://www.nature.com/doifinder/10.1038/nmeth.2019> doi: [10.1038/nmeth.2019](https://doi.org/10.1038/nmeth.2019).
- 785
- 786 ⁹²J. Schindelin, C. T. Rueden, M. C. Hiner, and K. W. Eliceiri, “The ImageJ ecosystem:
787 An open platform for biomedical image analysis,” (2015), doi: [10.1002/mrd.22489](https://doi.org/10.1002/mrd.22489).
- 788 ⁹³D.-R. Sundar, “binom: Binomial Confidence Intervals For Several Parameterizations,”
789 (2014).
- 790 ⁹⁴Werker, Janet F, Lalonde, Chris E, “Cross-language speech perception: Initial capabilities
791 and developmental change.,” Developmental psychology **24**(5) 672, (1988)
- 792 ⁹⁵H. Wickham, “The Split-Apply-Combine Strategy for Data Analysis,” Journal of Statis-
793 tical Software **40**(1), 1–29 (2011) <http://www.jstatsoft.org/v40/i01/>.

- 794 ⁹⁶H. Wickham, “Reshaping data with the reshape package,” Journal of Statistical Software
795 21(12), (2007) <http://www.jstatsoft.org/v21/i12/paper>
- 796 ⁹⁷H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York,
797 NY, 2009).
- 798 ⁹⁸R. Wright, *A review of perceptual cues and cue robustness* Phonetically based phonology
799 34–57 (2004).
- 800 ⁹⁹D. B. Dahl, “xtable: Export Tables to LaTeX or HTML,” (2016).
- 801 ¹⁰⁰“Pitch Shifting and Time Dilation Using a Phase Vocoder in
802 MATLAB,” <https://www.mathworks.com/help/audio/examples/pitch-shifting-and-time-dilation-using-a-phase-vocoder-in-matlab.html>
803

804 V. TABLES

TABLE I. **Impact of each generalization class on performance.** Accuracy values provide an estimate of the difficulty of that class after accounting for the random effects of individual mice. Accuracies are logistic GLMM coefficients transformed from logits, and model coefficients are logit differences from training set accuracy, which was used as an intercept. Correlation values are between fixed effects (novelty classes) across random effects (mice). * indicates significance ($p(>|z|) \ll .001$).

	Accuracy	95% Wald CI	Corr				
Learned	0.767*	[0.748, 0.785]					
Token	0.739*	[0.713, 0.763]	0.5				
Vowel	0.678*	[0.655, 0.701]	0.81	0.91			
Speaker	0.666*	[0.651, 0.68]	0.98	0.68	0.92		
Vow+Spk	0.637*	[0.624, 0.651]	0.98	0.64	0.9	1	

TABLE II. **Token structure of training stages**

Stage	Speakers	Vowels	Total Tokens
1	1	1	2
2	1	1	4
3	1	2	6
4	2	2	12
5	2	3	20
Generalization	5	6	160 (20 training, 140 novel)

TABLE III. **hierarchical GLMM:** To reach the appropriate complexity of model, we first modeled correct/incorrect answers as a function of each mouse as a fixed effect (row 1), then added the generalization type (as in Fig. 2) as a fixed effect (row 2), and finally modeled generalization type as a fixed effect nested within each mouse as a random effect (row 3). Since the final model had the best fit, it was used in all reported analyses related to the GLMM.

	DF	χ^2	DF_{χ^2}	$\text{Pr}(> \chi^2)$
Mouse	2			
Mouse + Type	6	2534.46	4	$\ll 0.001$
Type Mouse	20	407.22	14	$\ll 0.001$

805 **VI. FIGURE CAPTIONS**

806 **Figure 1: Stimuli and Task Design.** **a)** Spectrograms of stimuli. Left: Example of an
807 original recording of an isolated consonant-vowel token (/gI/). Center: the same token pitch-
808 shifted upwards by 10x (3.3 octaves) into the mouse hearing range. Right: Recording of
809 the pitch-shifted token presented in the behavior box. Stimuli retained their overall acoustic
810 structure below 34kHz (the upper limit of the speaker frequency response). **b)** Power spectra
811 (dB, Welch's method) of tokens in **a**. Black: Original (left frequency axis), red: Pitch-shifted
812 (right frequency axis), blue: Box Recording (right frequency axis). **c)** Mice initiated a trial
813 by licking in a center port and responded by licking on one of two side ports. Correct
814 responses were rewarded with water and incorrect responses were punished with a mildly-
815 aversive white noise burst. **d)** The difficulty of the task was gradually expanded by adding
816 more tokens (squares), vowels (labels), and speakers (rows) before the mice were tested on
817 novel tokens in a generalization task. **e)** Mice (colored lines) varied widely in the duration
818 of training required to reach the generalization phase. Mice were returned to previous levels
819 if they remained at chance performance after reaching a new stage.

820 **Figure 2: Generalization accuracy by novelty class.** Mice generalized stop conso-
821 nant discrimination to novel CV recordings. **a)** Four types of novelty are possible with our
822 stimuli: novel tokens from the speakers and vowels used in the training set (red), novel vow-
823 els (blue), novel speakers (purple), and novel speakers with novel vowels (orange). Tokens
824 in the training set are indicated in black. Colors same throughout. **b)** Mice that performed
825 better on the training set were better at generalization. Each point shows the performance

826 for a single mouse on a given novelty class, plotted against that mouse's performance on
827 training tokens presented on during the generalization phase (both averaged across the entire
828 generalization phase). Lines show linear regression for each novelty class. **c**) Mean accuracy
829 for each novelty class (gray lines indicate individual mice). **d**) Mean accuracy for individual
830 mice (colored bars indicate each novelty class). Error bars in **d** are 95% binomial confidence
831 intervals. Mice were assigned one of two sets of training tokens, black and white boxes in **d**.

832 **Figure 3: Learning curve for novel tokens.** Performance for both novel and training
833 set tokens dropped transiently and recovered similarly after the transition to the general-
834 ization stage. Presentation 0 corresponds to the transition to the generalization stage. The
835 final ten trials before the transition are shown in the gray dashed box. Mean accuracy
836 and 95% binomial confidence intervals are collapsed across mice for novel (red, all novelty
837 classes combined) or learned (black) tokens, by number of presentations in the generaliza-
838 tion task. Logistic regression of binomial correct/incorrect responses fit to log-transformed
839 presentation number (lines, shading is smoothed standard error).

840 **Figure 4: Patterns of individual and group variation.** **a**) Mean accuracy (color,
841 scale at top) for each mouse (columns) on tokens grouped by consonant, speaker, and vowel
842 (rows). The different training sets (cells outlined with black boxes) led to different patterns
843 of accuracy on the generalization set. **b**) Ward clustering dendrogram, colored by cluster.
844 **c**) Training set cohorts differed in bias but not mean accuracy.

845 **Figure 5: Acoustic-Behavior Correlates** F2 Onset-Vowel transitions do not explain
846 observed response patterns. **a**) Locus equations relating F2 at burst onset and vowel steady
847 state (sustained) for each token (points), split by consonant (colors, same as **b**)). **b**) As the

848 difference of a token's distance from the ideal /g/ and /b/ locus equation lines increased (x
849 axis, greater distance from /g/, smaller distance from /b/ in panel b), /b/ tokens obeyed
850 the predicted categorization while /g/ tokens did not (slopes of colored lines).