Mice can learn phonetic categories

Jonny L. Saunders[1] and Michael Wehr[1, a)]

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

(Dated: 4 February 2019)

We perceive speech as a series of relatively invariant phonemes despite extreme variability in the acoustic signal. To be perceived as nearly-identical phonemes, speech sounds that vary continuously over a range of acoustic parameters must be perceptually discretized by the auditory system. Such many-to-one mappings of undifferentiated sensory information to a finite number of discrete categories are ubiquitous in perception. Although many mechanistic models of phonetic perception have been proposed, they remain largely unconstrained by neurobiological data. Current human neurophysiological methods lack the necessary spatiotemporal resolution to provide it: speech is too fast and the neural circuitry involved is too small. Here we demonstrate that mice are capable of learning generalizable phonetic categories, and can thus serve as a model for phonetic perception. Mice learned to discriminate consonants, and generalized consonant identity across novel vowel contexts and speakers, consistent with true category learning. A mouse model, given the powerful genetic and electrophysiological tools for probing neural circuits available for them, has the potential to powerfully augment our mechanistic understanding of phonetic perception.

[a] wehr@uoregon.edu

I. INTRODUCTION

A. Lack of acoustic invariance in phonemes

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

B. Generality of phonetic perception

The lack of a simple mapping between acoustic attributes and phoneme identity has had a deep influence on phonetics, in part motivating the hypothesis that speech is mechanistically unique to humans[?], and the development of non-acoustic theories of speech perception (most notably motor theories[?,?,?]). However, it has been clear for more than 30 years that at least some auditory components of speech perception are not unique to humans, suggesting that human speech perception exploits evolutionarily-preserved functions of the auditory system[?,?,?,?]. For example, nonhuman animals like quail[?,?], chinchillas[?], rats[?], macaques[?], and songbirds[?] are capable of learning phonetic categories that share some perceptual qualities with humans[?,?]. This is consistent with the idea that categorizing phonemes is just one instance of a more general problem faced by all auditory systems, which typically extract useable information from complex acoustic environments by reducing them to a small number of 'auditory objects' (for review, see[?]).

C. Neurolinguistic theories of phonetic perception

Many neurolinguistic theories of phonetic perception have been proposed[?,?,?,?,?], but neurophysiological evidence to support them is limited. One broad class of models follows the paradigm of hierarchical processing first described by Hubel and Weisel in the visual system[?,?,?]. In these models, successive processing stages in the auditory system extract acoustic features with progressively increasing complexity by combining the simpler representations present in preceding stages. Such hierarchical processing is relatively well-supported by experimental data. For example, the responses of neurons in primary auditory cortex (A1) to speech sounds are more diverse than those in inferior colliculus[?] (but see[?]). While phoneme identity can be classified post-hoc from population-level activity in A1[?,?,?], neurons in secondary auditory cortical regions explicitly encode higher-order properties of speech sounds[?,?,?,?,?].

Another class of models proposes that phonemes have no positive acoustic "prototype", and that we instead learn only the acoustic features useful for telling them apart[?]. Theoretically, these discriminative models provide better generalization and robustness to high variance[?]. Theories based on discrimination rather than prototype-matching have a long history in linguistics[?], but have rarely been implemented as neurolinguistic models. A possible neural implementation of discriminative perception is that informative contrast cues could evoke inhibition to suppress competing phonetic percepts, similar to predictive coding[?,?,?]. Neurophysiological evidence supports the existence of discriminative predictive coding, but its specific implementation is unclear[?,?].

These two very different classes of models illustrate a major barrier faced by phonetic research: both classes can successfully predict human categorization performance, making it difficult to empirically validate or refute either of them using psychophysical experiments alone. Mechanistic differences have deep theoretical consequences — for example, the characterizations made by the above two classes of models regarding what phonemes *are* precisely oppose one another: are they positive acoustic prototypes, or sets of negative acoustic contrasts? Perceptually, do listeners identify phonemes, or discriminate between them? Neurobiological evidence regarding how the brain actually solves these categorization problems could help overcome this barrier.

D. The utility of a mouse model for speech research

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

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

Second, it is difficult to isolate the purely auditory component of speech perception in humans. Humans can use contextual information from syntax, semantics or task structure to infer phoneme identity[?,?]. It is also difficult to rule out the contribution of multimodal information[?], or of motor simulation predicted by motor theories. Certainly, these and other non-auditory strategies are used during normal human speech perception. Nevertheless, speech perception is possible without these cues, so any neurocomputational theory of phonetic perception must be able to explain the purely auditory case. Animal models allow straightforward isolation of purely auditory phonetic categorization without interference from motor, semantic, syntactic, or other non-auditory cues.

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

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

E. The utility of phonetics for auditory neuroscience

Conversely, auditory neuroscience stands to benefit from the framework provided by phonetics for studying how sound is transformed to meaning. Understanding how complex sounds are encoded and processed by the auditory system, ultimately leading to perception and behavior, remains a challenge for auditory neuroscience. For example, it has been difficult to extrapolate from simple frequency/amplitude receptive fields to understand the hierarchical organization of complex feature selectivity across brain areas. A great strength of neuroethological model systems such as the songbird is that both the stimulus (e.g., the bird's own song) and the behavior (song perception and production) are well understood. This has led to significant advances in understanding the hierarchical organization and function of the song system[?,?]. The long history of of speech research in humans has produced a deep understanding of the relationships between acoustic features and phonetic perception[?]. These insights have enabled specific predictions about what kinds of neuronal selectivity for features (and combinations of features) might underlie phonetic perception[?]. Although recognizing human speech sounds is not a natural ethological behavior for mice, phonetics nevertheless provides a valuable framework for studying how the brain encodes and transforms complex sounds into perception and behavior.

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

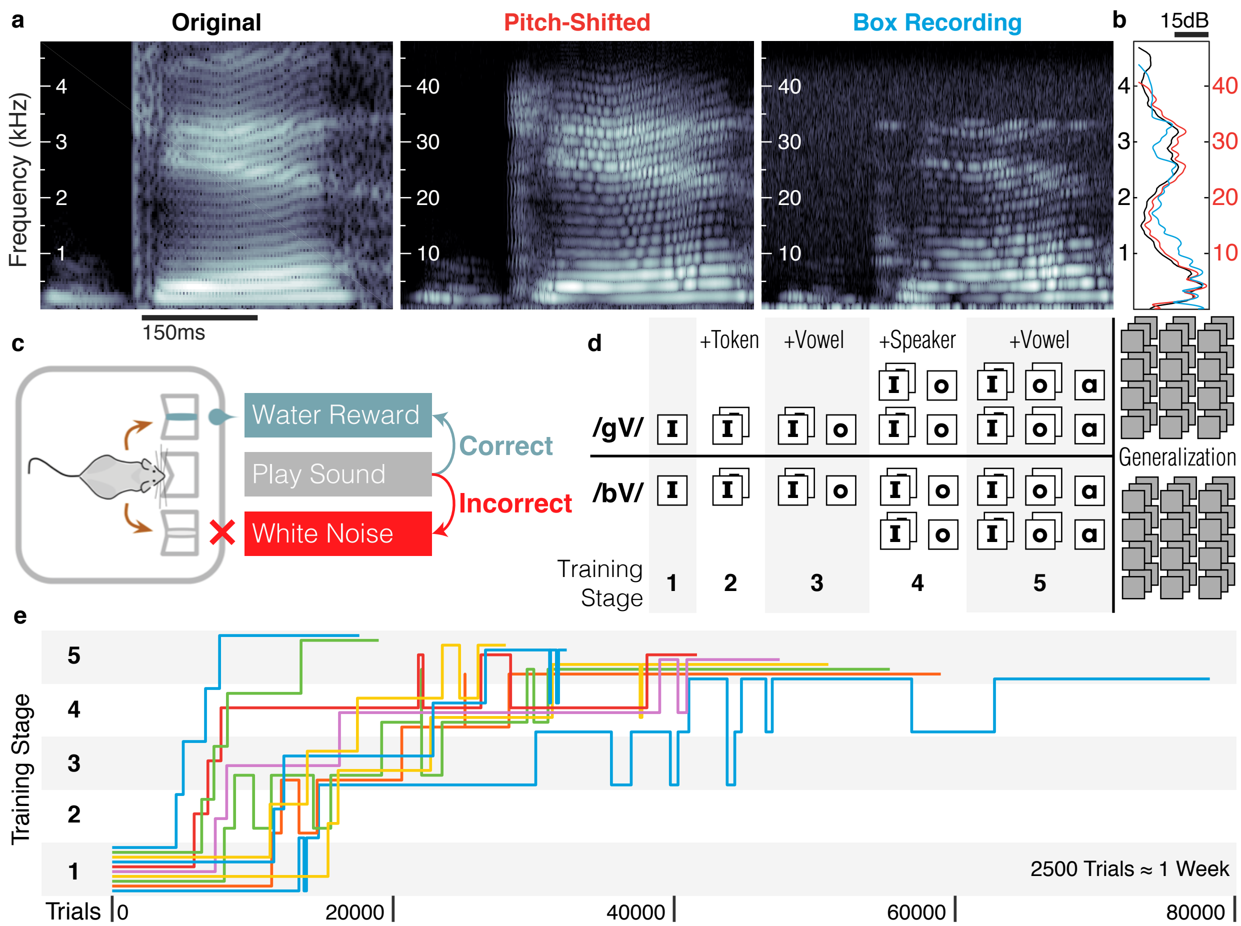

FIG. 1. **Stimuli and Task Design. a)** Spectrograms of stimuli. Left: Example of an original recording of an isolated consonant-vowel token (/gI/). Center: the same token pitch-shifted upwards by 10x (3.3 octaves) into the mouse hearing range. Right: Recording of the pitch-shifted token presented in 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.

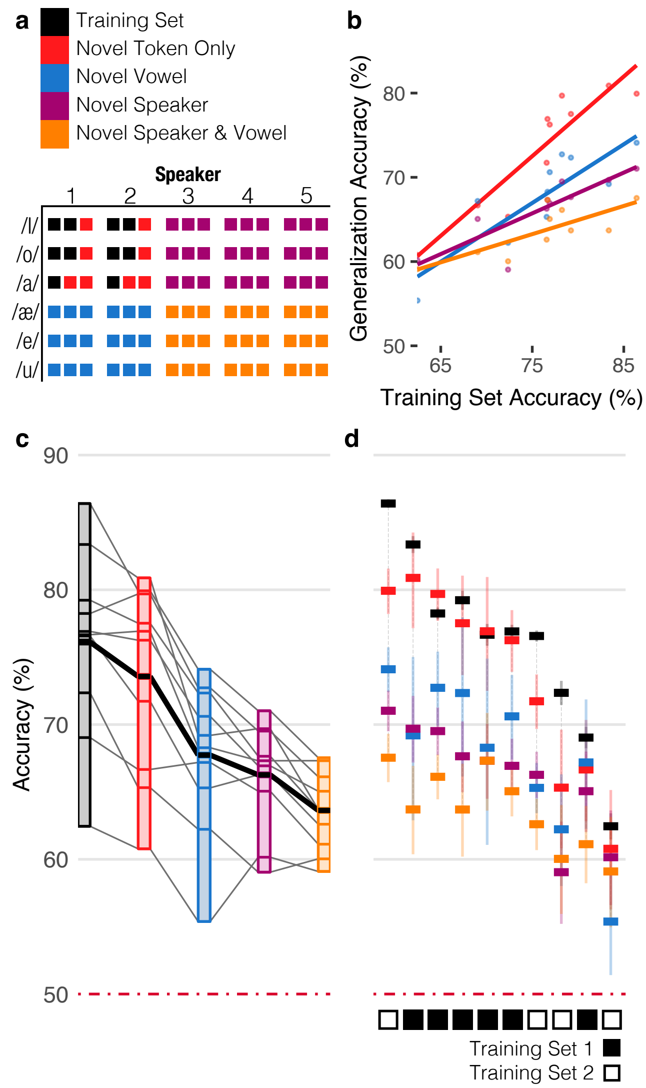

FIG. 2. **Generalization accuracy by novelty class.** Mice generalized stop consonant discrimination to novel CV recordings. **a)** Four types of novelty are possible with our stimuli: novel tokens from the speakers and vowels used in the training set (red), novel vowels (blue), novel speakers (purple), 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 on 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**.

II. RESULTS

A. Generalization performance

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

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

Generalization is an essential feature of categorical perception. By testing whether mice can generalize their phonetic categorization to novel stimuli, we can distinguish whether mice actually learn phonetic categories or instead just memorize the reward contingency for each training token. Four types of novelty are possible with our stimuli: new tokens from the speakers and vowel contexts used in the training set, new vowels, new speakers, and new vowels from new speakers (colored groups in Fig. **??**a). In the final generalization stage, we randomly interleaved tokens from each of these novelty classes on 20% of trials, with the remaining 80% consisting of tokens from the training set. We interleaved novel tokens with training tokens for two reasons: (1) to avoid a sudden increase in task difficulty, which can degrade performance, and (2) to minimize the possibility that mice could learn each new token by widely separating them in time (on average, generalization tokens were repeated only once every five days).

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

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

Performance on all generalization types was strongly and positively correlated with performance on the training set (Fig. **??**b, adj. R^2=0.74, F(4, 5) = 7.4, $p < 0.05$). If mice were "overfitting," that is, memorizing the training tokens rather than learning categories, then we would expect the opposite (i.e., above some threshold, mice that performed better on the training set would perform correspondingly worse on the generalization set). It appears instead that better prototypes or decision boundaries learned in the training stages allowed better generalization to novel tokens.

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

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

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

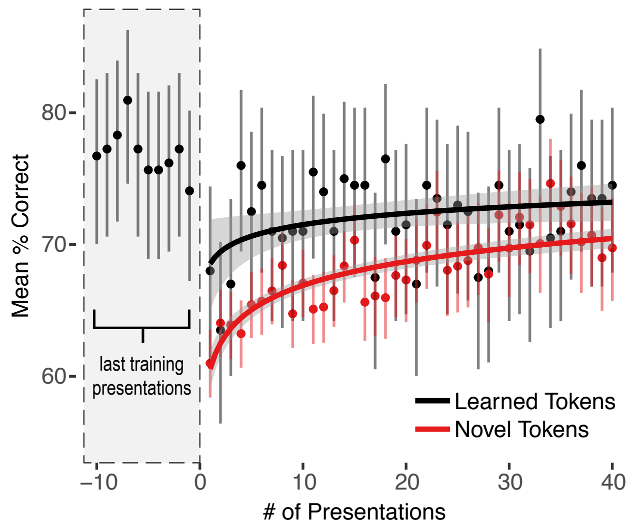

FIG. 3. **Learning curve for novel tokens.** Performance for both novel and training set tokens dropped transiently and recovered similarly after the transition to the generalization stage. Presentation 0 corresponds to the transition to the generalization stage. The final ten trials before the transition 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).

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

Performance on the first presentation of novel tokens was significantly greater than chance (Fig. 3, all mice, all tokens from all novelty classes: one-sided binomial test, n=1410, $P_{correct}$ = 0.61, lower 95% CI = 0.588, p $\ll$ 0.001; all mice, worst novelty class: n=458, $P_{correct}$ = 0.581, lower 95% CI = 0.541, p < 0.001). This demonstrates that mice were able to generalize immediately without additional reinforcement. Although performance on novel tokens did increase with repetition, so did performance on training tokens (Fig. **??**). We noted that performance on all tokens (both novel and previously learned tokens) transiently dropped after each transition between task stages, suggesting a non-specific effect of an increase in task difficulty. To distinguish an increase in performance due to learning from an increase due to acclimating to a change in the task, we compared performance on generalization and training tokens over the first 40 presentations of each token. If the mice were learning the generalization tokens, the increase in performance with repeated presentations should be significantly greater than that of the already trained tokens.

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

In summary, the behavior of the mice is consistent with an ability to generalize some learned acoustic criteria to novel stimuli. It is unlikely that the mice rapidly learned the novel tokens because (1) performance on the first presentation of novel tokens was significantly above chance, (2) performance on subsequent presentations of novel tokens did not improve compared to trained tokens, and (3) learning each token would have to take place over unrealistically long timescales: there were an average of 2355 trials (5 days) between the first and second presentation of each novel token.

B. Training Set Differences

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

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

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

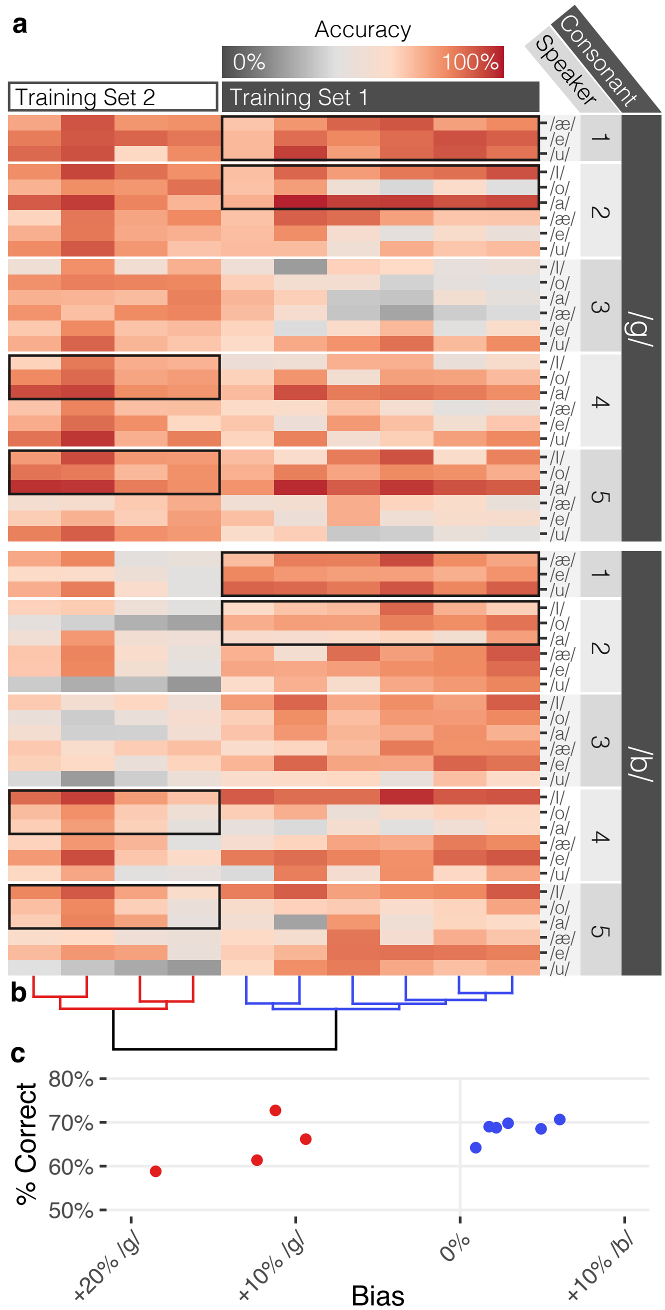

FIG. 4. **Patterns of individual and group variation.** **a)**Mean accuracy (color, scale at top) for each mouse (columns) on tokens grouped by consonant, speaker, and vowel (rows). The different training sets (cells outlined with black boxes) led to different patterns of accuracy on the generalization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

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

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

C. Acoustic-behavioral correlates

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

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

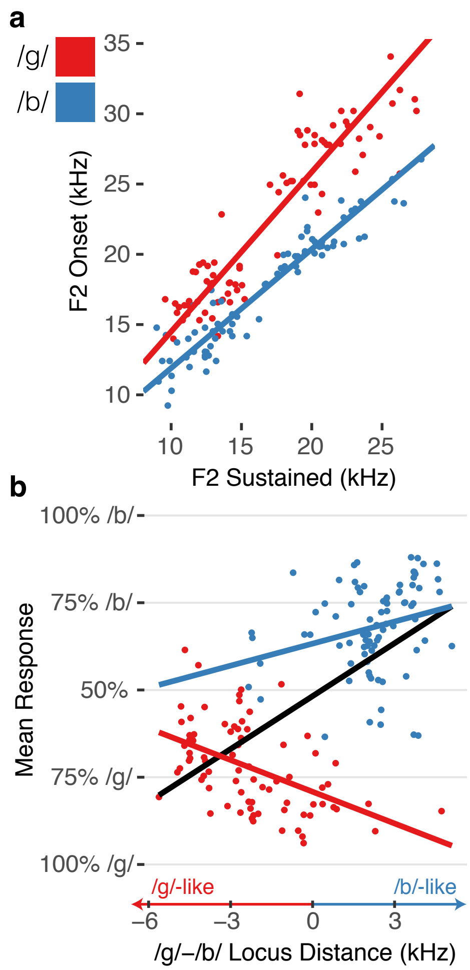

FIG. 5. **Acoustic-Behavior Correlates** F2 Onset-Vowel transitions do not explain observed response patterns. **a)** Locus equations relating F2 at burst onset and vowel steady state (sustained) for each token (points), split by consonant (colors, same as **b)**). **b)** As the difference of a token's 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. **??**b). However, it is important to note that this correlation does not necessarily demonstrate that mice relied on this acoustic cue. Because multiple acoustic features are correlated with consonant identity, performance that is correlated with one such cue would also be correlated with all the others. The mice learned some acoustic property of the consonant classes, and since the acoustic features are all highly correlated with one another, they are all likely to correlate with mean responses. To distinguish whether mice specifically relied on F2 locus distance, we therefore measured the marginal effect of this acoustic cue within a consonant class. This is shown by the slopes of the red and blue lines in Fig. **??**b. For example, is a /g/ token that is further away from the blue /b/ line more likely to be identified as a /g/ than one very near the /b/ line? Mean responses to /g/ tokens were negatively correlated with locus distance (Mean response /g/ to /b/ between 0 and 1, b=-0.028kHz, 95% CI = [-0.035, -0.022], $p \ll 0.001$). In other words, tokens that should have been more frequently confused with /b/ were actually more likely to be classified as /g/. Note the red points at locus distance of zero in Fig. **??**b: these tokens have an equal distance from both the /b/ and /g/ locus equation prototypes but are some of the most accurately categorized /g/ tokens. /b/ tokens obeyed the predicted direction of locus distance (b=0.049, 95% CI = [0.039, 0.06], $p \ll 0.001$), but the effect was very small: moving one standard deviation ($\sigma_{/b/}$=1.618kHz) towards the /g/ line only changed responses by 7.9%. These results suggest that mice did not rely on F2 transitions to categorize these consonants.

We repeated this analysis separately for each training cohort to test whether the two cohorts could have developed different acoustic templates that better explained their response patterns. We derived cohort-specific locus-equation lines and distances using only the tokens from each of their respective training sets. These models were qualitatively similar to the model that included all tokens and mice and did not improve the model fit (Cohort 1: /g/: $b = -0.051, 95\%CI = [-0.064, -0.038]$, /b/: $b = 0.041, 95\%CI = [0.022, 0.059]$; Cohort 2: /g/: $b = -0.022, 95\%CI = [-0.031, -0.014]$, /b/: $b = 0.055, 95\%CI = [0.042, 0.069]$).

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

III. DISCUSSION

These results demonstrate that mice are capable of learning and generalizing phonetic categories. Indeed, this is the first time to our knowledge that mice have been trained to discriminate between any classes of natural, non-species-specific sounds. Thus mice join a number of model organisms that have demonstrated categorical learning with speech sounds[?,?,?,?,?,?,?], making a new suite of genetic and electrophysiological tools available for phonetic research.

Two subgroups of our mice that were trained using different sets of speech tokens demonstrated distinct patterns of consonant identification, presumably reflecting differences in underlying acoustic prototypes. The ability to precisely control exposure to speech sounds provides an opportunity to probe the neurocomputational constraints that govern the possible solutions to consonant identification.

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

Future experiments using parametrically synthesized speech sounds are a critical next step, and will support a qualitatively different set of inferences. Being able to carefully manipulate reduced speech sounds is useful to probe the acoustic cue structure of learned phonetic categories, but the reduction in complexity that makes them useful also makes it correspondingly more difficult to probe the learning and recognition mechanisms for a perceptual category that is defined by multiple imperfect, redundant cues. It is possible that the complexity of natural speech may have caused our attrition rate to be higher, and task performance lower, than other sensory-driven tasks. Neither of those concerns, however, detracts from the possibility for the mouse to shed mechanistic insight on phonetic perception. Indeed, error trials may provide useful neurophysiological data about how and why the auditory system fails to learn or perceive phonetic categories.

We hope in future experiments to directly test predictions made by neurolinguistic models regarding phonetic acquisition and discrimination. For example, one notable model proposes that consonant perception relies on combination-sensitive neurons that selectively respond to specific combinations of acoustic features[?] . This model predicts that mice trained to discriminate stop consonants would have neurons selective for the feature combinations that drive phoneme discrimination, perhaps in primary or higher auditory cortical areas. Combination-selective neurons have been observed in A1[? ?] , and speech training can alter the response properties of A1 neurons in rats[?] , but it is unclear whether speech training induces combination-selectivity that would facilitate phonetic discrimination.

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

IV. METHODS

A. Animals

All procedures were performed in accordance with National Institutes of Health guidelines, as approved by the University of Oregon Institutional Animal Care and Use Comittee.

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

B. Speech stimuli

Speech stimuli were recorded in a sound-attenuating booth with a head-mounted microphone attached to a Tascam DR-100mkII handheld recorder sampling at 96kHz/24bit. Each speaker produced a set of 3 recordings (tokens) of each of 12 CV pairs beginning with either /b/ or /g/, and ending with /I/, /o/, /a/, /æ/, /ɛ/, /u/. To reduce a slight hiss that was present in the recordings, they were denoised using a Daubechies wavelet with two vanishing moments in MATLAB. The typical human hearing range is 20 Hz - 20 kHz, whereas the mouse hearing range is 1 kHz - 80 kHz[?] . The F_0 of our recorded speech sounds ranged from 100 - 200 Hz, which is well below the lower frequency limit of the mouse hearing range. We therefore pitch shifted all stimuli upwards by 10x (3.3 octaves) in MATLAB[?] . This shifted all spectral information equally upwards into an analogous part of mouse hearing range while preserving temporal information unaltered.

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

Formant frequency values were measured manually using Praat[?] . F2 at onset was measured at its center as soon as it was discernible, typically within 20 ms of burst onset, and at vowel steady-state, typically 150-200ms after burst onset.

C. Training

We trained mice to discriminate between CV pairs beginning with /b/ or /g/ in a two-alternative forced choice task. Training sessions lasted approximately 1 hour, 5 days a week. Each custom-built sound-attenuating training chamber contained two free-field JBL Duet speakers for stimulus presentation with a high-frequency rolloff of 34 kHz, and a smaller 15 x 30 cm plastic box with three "lick ports." Each lick port consisted of a water delivery tube and an IR beam-break sensor mounted above the tube. Beam breaks triggered water delivery by actuating a solenoid valve. Water-restricted mice were trained to initiate each trial with an unrewarded lick at the center port, which started playback of a randomly selected stimulus, and then to indicate their stimulus classification by licking at one of the ports on either side. Tokens beginning with /g/ were always on the left, with /b/ on the right. Two cohorts were trained on two separate sets of tokens. Training set 1 started with speaker 1 (Fig. 4a) and had speaker 2 introduced on the fourth stage, where Training set 2 started training with speaker 5 and had speaker 4 introduced on the fourth stage. Correct classifications received $\sim$10 µL water rewards, and incorrect classifications received a 5s time-out that included a mildly aversive 60 dB SPL white noise burst.

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

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

D. Data analysis

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

All analyses were performed in R (R version 3.5.1 (2018-07-02))[?] using RStudio (1.1.456)[?] . Generalization performance was modeled using a logistic generalized linear mixed model (GLMM) using the R package “lme4”[?] . Binary correct/incorrect responses were fit hierarchically to models of increasing complexity (see Table ??), with a final model consisting of the generalization class (as in Fig. 2a: training tokens, novel tokens from the speakers and vowels in the training set, novel speaker, novel vowel, and novel speaker and vowel) as a fixed effect with random slopes and intercepts nested within each mouse as a random effect. There was no evidence of overdispersion (i.e., deviance $\approx$ degrees of freedom, or less than $\sim$ 2 times degrees of freedom), and the profile of the model showed that the deviances by each fixed effect were approximately normal. Accordingly, we report Wald confidence intervals. We also computed bootstrapped confidence intervals, which had only minor disagreement with the Wald confidence intervals and agreed with our interpretation in the text.

Clustering was performed with the "cluster"[?] package. Ward clustering split the mice into two notable clusters, which are plotted in Fig. ??.

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

Spectrograms in Figure ??a were computed with the "spectrogram" function in MATLAB 2017b, and power spectra in Figure ??b were computed with the "pwelch" function in MATLAB 2018b with the same window and overlap as ??a spectrograms.

The remaining analyses are described in the text and used the "binom"[?], "reshape"[?], and "plyr"[?] packages. Data visualization and tabulation was performed with the "ggplot2,"[?] and "xtable"[?] packages.

REFERENCES

[1] L. L. Holt and A. J. Lotto, "Speech perception as categorization," Attention, Perception & Psychophysics **72**(5), 1218–1227 (2010) http://www.springerlink.com/index/10.3758/APP.72.5.1218 doi: 10.3758/APP.72.5.1218.

[2] Y. Kronrod, E. Coppess, and N. H. Feldman, "A unified account of categorical effects in 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.

[3] A. M. LIBERMAN, K. S. HARRIS, H. S. HOFFMAN, and B. C. GRIFFITH, "The discrimination of speech sounds within and across phoneme boundaries.," Journal of experimental psychology **54**(5), 358–68 (1957) http://www.ncbi.nlm.nih.gov/pubmed/13481283.

[4] J. L. Elman and D. Zipser, "Learning the hidden structure of speech.," The Journal of the Acoustical Society of America **83**(4), 1615–26 (1988) http://www.ncbi.nlm.nih.gov/pubmed/3372872.

[5] K. R. Kluender, R. L. Diehl, and P. R. Killeen, "Japanese quail can learn phonetic 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.

[6] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code.," Psychological review **74**(6), 431–61 (1967) http://www.ncbi.nlm.nih.gov/pubmed/4170865.

[7] E. Farnetani, “V-C-V Lingual Coarticulation and Its Spatiotemporal Domain,” in *Speech Production and Speech Modelling* (Springer Netherlands, Dordrecht, 1990), pp. 93–130, http://www.springerlink.com/index/10.1007/978-94-009-2037-8{_}5, doi: 10.1007/978-94-009-2037-8_5.

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

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

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

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

[12] K. M. Carbonell and A. J. Lotto, “Speech is not special... again.,” Frontiers in psychology **5**(June), 427 (2014) http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00427/abstracthttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4042079{&}tool=pmcentrez{&}rendertype=abstract doi: 10.3389/fpsyg.2014.00427.

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

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

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

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

[17] K. R. Kluender, "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.

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

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

doi: 10.1016/j.bbr.2015.03.044.

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

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

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

[23]J. P. Rauschecker and S. K. Scott, "Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing," Nature Neuroscience **12**(6), 718–724 (2009) http://www.nature.com/doifinder/10.1038/nn.2331 doi: 10.1038/nn.2331.

[24]T. J. Strauss, H. D. Harris, and J. S. Magnuson, "jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition," Behavior Research Methods **39**(1), 19–30 (2007) http://www.springerlink.com/index/10.3758/BF03192840 doi: 10.3758/BF03192840.

[25]K. R. Kluender, C. E. Stilp, M. Kiefte, K. R. Kluender, C. E. Stilp, and M. Kiefte, "Perception of Vowel Sounds Within a Biologically Realistic Model of Efficient Coding,"

in *Vowel Inherent Spectral Change*, pp. 117–151, doi: 10.1007/978-3-642-14209-3_6.

[26]M. G. Gaskell and W. D. Marslen-Wilson, "Integrating Form and Meaning: A Distributed Model of Speech Perception," Language and Cognitive Processes **12**(5-6), 613–656 (1997) http://www.tandfonline.com/doi/abs/10.1080/016909697386646 doi: 10.1080/016909697386646.

[27]D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," The Journal of Physiology **160**(1), 106–154.2 (1962) http://www.ncbi.nlm.nih.gov/pubmed/14449617http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1359523 doi: 10.1523/JNEUROSCI.1991-09.2009.

[28]K. G. Ranasinghe, W. A. Vrana, C. J. Matney, and M. P. Kilgard, "Increasing diversity of neural responses to speech sounds across the central auditory pathway," Neuroscience **252**, 80–97 (2013) http://dx.doi.org/10.1016/j.neuroscience.2013.08.005 doi: 10.1016/j.neuroscience.2013.08.005.

[29]E. L. Bartlett, "The organization and physiology of the auditory thalamus and its role in processing acoustic features important for speech perception.," Brain and language **126**(1), 29–48 (2013) http://dx.doi.org/10.1016/j.bandl.2013.03.003http://linkinghub.elsevier.com/retrieve/pii/S0093934X13000722 doi: 10.1016/j.bandl.2013.03.003.

[30]T. M. Centanni, A. M. Sloan, A. C. Reed, C. T. Engineer, R. L. Rennaker, and M. P. Kilgard, "Detection and identification of speech sounds using cortical activity patterns," Neu-

roscience **258**, 292–306 (2013) http://dx.doi.org/10.1016/j.neuroscience.2013.11.030 doi: 10.1016/j.neuroscience.2013.11.030.

[31] C. T. Engineer, C. A. Perez, Y. H. Chen, R. S. Carraway, A. C. Reed, J. A. Shetake, V. Jakkamsetti, K. Q. Chang, and M. P. Kilgard, "Cortical activity patterns predict speech discrimination ability.," Nature neuroscience **11**(5), 603–8 (2008) http://www.ncbi.nlm.nih.gov/pubmed/18425123 doi: 10.1038/nn.2109.

[32] M. Steinschneider, Y. I. Fishman, and J. C. Arezzo, "Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex of the awake monkey.," The Journal of the Acoustical Society of America **114**(1), 307–321 (2003) doi: 10.1121/1.1582449.

[33] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus.," Science (New York, N.Y.) **343**(6174), 1006–10 (2014) http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4350233{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1126/science.1245994.

[34] P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike, "Voice-selective areas in human auditory cortex.," Nature **403**(6767), 309–312 (2000) doi: 10.1038/35002078.

[35] E. F. Chang, J. W. Rieger, K. Johnson, M. S. Berger, N. M. Barbaro, and R. T. Knight, "Categorical speech representation in human superior temporal gyrus.," Nature neuroscience **13**(11), 1428–32 (2010) http://www.ncbi.nlm.nih.gov/pubmed/20890293http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2967728 doi: 10.1038/nn.2641.

[36]B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, “Reconstructing speech from human auditory cortex,” PLoS Biology **10**(1) (2012) doi: 10.1371/journal.pbio.1001251.

[37]G. M. Bidelman, S. Moreno, and C. Alain, “Tracing the emergence of categorical speech 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.

[38]A. Ng and M. I. Jordan, “On generative vs. discriminative classifiers: A comparison of logistic regression and naive bayes,” Proceedings of Advances in Neural Information Processing **28**(3), 169–187 (2002) http://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-a pdf.

[39]F. de Saussure, *Cours de linguistique gnrale.* (Payot, Lausanne, Paris, 1916).

[40]U. Rutishauser, J. J. Slotine, and R. Douglas, “Computation in Dynamically Bounded Asymmetric Systems,” PLoS Computational Biology **11**(1), e1004039 (2015) http://dx.plos.org/10.1371/journal.pcbi.1004039 doi: 10.1371/journal.pcbi.1004039.

[41]B. E. Dresher, “The contrastive hierarchy in phonology,” Contrast in phonology: theory, perception, acquisition **13**, 11 (2008) doi: 10.1017/CBO9780511642005.

[42]H. Blank and M. H. Davis, “Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception.,” PLoS Biology **14**(11), e1002577 (2016) http://www.ncbi.nlm.nih.gov/pubmed/27846209 doi: 10.1371/journal.pbio.

1002577.

[43]P. Gagnepain, R. Henson, and M. Davis, “Temporal Predictive Codes for Spoken Words in Auditory Cortex,” (2012), doi: 10.1016/j.cub.2012.02.015.

[44]N. P. Fox and S. E. Blumstein, “Top-down effects of syntactic sentential context on phonetic processing.,” Journal of Experimental Psychology: Human Perception and Performance **42**(5), 730–741 (2016) http://www.ncbi.nlm.nih.gov/pubmed/26689310 doi: 10.1037/a0039965.

[45]B. Schouten, E. Gerrits, and A. Van Hessen, “The end of categorical perception as we know it,” in *Speech Communication* (2003), Vol. 41, pp. 71–80, doi: 10.1016/S0167-6393(02)00094-8.

[46]L. D. Rosenblum, “Speech Perception as a Multimodal Phenomenon,” Current Directions 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.

[47]P. Kuhl, K. Williams, F. Lacerda, K. Stevens, and B. Lindblom, “Linguistic experience alters phonetic perception in infants by 6 months of age,” Science **255**(5044) (1992).

[48]L. L. Holt, A. J. Lotto, and K. R. Kluender, “Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement?,” 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.

[49]E. A. Brenowitz, D. Margoliash, and K. W. Nordeen, “An introduction to birdsong and the avian song system,” (1997).

[50] F. E. Theunissen and J. E. Elie, "Neural processing of natural sounds.," Nature reviews. Neuroscience **15**(6), 355–66 (2014) http://www.ncbi.nlm.nih.gov/pubmed/24840800 doi: 10.1038/nrn3731.

[51] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," The Journal of the Acoustical Society of America **24**(2), 175–184 (1952) doi: 10.1121/1.1906875.

[52] H. M. Sussman, D. Fruchter, J. Hilbert, and J. Sirosh, "Linear correlates in the speech signal: the orderly output constraint.," The Behavioral and brain sciences **21**(2), 241–59; discussion 260–99 (1998) http://www.ncbi.nlm.nih.gov/pubmed/10097014 doi: 10.1017/S0140525X98001174.

[53] K. N. Stevens, *Acoustic phonetics* (MIT Press, 1998), p. 607.

[54] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," The Journal of the Acoustical Society of America **64**(5), 1358–1368 (1978) http://asa.scitation.org/doi/10.1121/1.382102 doi: 10.1121/1.382102.

[55] B. Lindblom and H. M. Sussman, "Dissecting coarticulation: How locus equations happen," Journal of Phonetics **40**(1), 1–19 (2012) doi: 10.1016/j.wocn.2011.09.005.

[56] M. F. Dorman, M. Studdert-Kennedy, and L. J. Raphael, "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," Perception & Psychophysics **22**(2), 109–122 (1977) http://www.springerlink.com/index/10.3758/BF03198744 doi: 10.3758/BF03198744.

[57]R. 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.

[58]A. S. Buchman, D. C. Garron, J. E. Trost-Cardamone, M. D. Wichter, and M. Schwartz, "Word deafness: one hundred years later.," Journal of Neurology, Neurosurgery & Psychiatry **49**(5), 489–499 (1986) http://jnnp.bmj.com/cgi/doi/10.1136/jnnp.49.5.489 doi: 10.1136/jnnp.49.5.489.

[59]G. 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.

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

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

[62]E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, L. Chen, T. M. Chen, M. C. Chin, J. Chong, B. E. Crook, A. Czaplinska, C. N. Dang, S. Datta, N. R. Dee, A. L. De-

saki, T. Desta, E. Diep, T. A. Dolbeare, M. J. Donelan, H. W. Dong, J. G. Dougherty, B. J. Duncan, A. J. Ebbert, G. Eichele, L. K. Estin, C. Faber, B. A. Facer, R. Fields, S. R. Fischer, T. P. Fliss, C. Frensley, S. N. Gates, K. J. Glattfelder, K. R. Halverson, M. R. Hart, J. G. Hohmann, M. P. Howell, D. P. Jeung, R. A. Johnson, P. T. Karr, R. Kawal, J. M. Kidney, R. H. Knapik, C. L. Kuan, J. H. Lake, A. R. Laramee, K. D. Larsen, C. Lau, T. A. Lemon, A. J. Liang, Y. Liu, L. T. Luong, J. Michaels, J. J. Morgan, R. J. Morgan, M. T. Mortrud, N. F. Mosqueda, L. L. Ng, R. Ng, G. J. Orta, C. C. Overly, T. H. Pak, S. E. Parry, S. D. Pathak, O. C. Pearson, R. B. Puchalski, Z. L. Riley, H. R. Rockett, S. A. Rowland, J. J. Royall, M. J. Ruiz, N. R. Sarno, K. Schaffnit, N. V. Shapovalova, T. Sivisay, C. R. Slaughterbeck, S. C. Smith, K. A. Smith, B. I. Smith, A. J. Sodt, N. N. Stewart, K. R. Stumpf, S. M. Sunkin, M. Sutram, A. Tam, C. D. Teemer, C. Thaller, C. L. Thompson, L. R. Varnam, A. Visel, R. M. Whitlock, P. E. Wohnoutka, C. K. Wolkey, V. Y. Wong, M. Wood, M. B. Yaylaoglu, R. C. Young, B. L. Youngstrom, X. F. Yuan, B. Zhang, T. A. Zwingman, and A. R. Jones, "Genome-wide atlas of gene expression in the adult mouse brain," Nature **445**(7124), 168–176 (2007) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{&}db=PubMed{&}dopt=Citation{&}list{_}uids=17151600 doi: 10.1038/nature05453.

[63]S. Sadagopan and X. Wang, "Nonlinear Spectrotemporal Interactions Underlying Selectivity for Complex Sounds in Auditory Cortex," Journal of Neuroscience **29**(36), 11192–11202 (2009) http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1286-09.2009 doi: 10.1523/JNEUROSCI.1286-09.2009.

[64]X. Wang, T. Lu, R. K. Snider, and L. Liang, "Sustained firing in auditory cortex evoked by preferred stimuli.," Nature **435**(7040), 341–6 (2005) http://www.ncbi.nlm.nih.gov/pubmed/15902257 doi: 10.1038/nature03565.

[65]P. D. Eimas and J. D. Corbit, "Selective adaptation of linguistic feature detectors," Cognitive Psychology **4**(1), 99–109 (1973) http://www.sciencedirect.com/science/article/pii/0010028573900066 doi: 10.1016/0010-0285(73)90006-6.

[66]K. G. Estes and C. Lew-Williams, "Listening through voices: Infant statistical word segmentation across multiple speakers.," Developmental Psychology **51**(11), 1517–1528 (2015) http://www.ncbi.nlm.nih.gov/pubmed/26389607 doi: 10.1037/a0039725.

[67]V. Vapnik, "An overview of statistical learning theory," IEEE Transactions on Neural Networks **10**(5), 988–999 (1999) http://ieeexplore.ieee.org/document/788640/ doi: 10.1109/72.788640.

[68]D. F. Kleinschmidt and T. F. Jaeger, "Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel.," Psychological Review **122**(2), 148–203 (2015) http://doi.apa.org/getdoi.cfm?doi=10.1037/a0038695 doi: 10.1037/a0038695.

[69]M. A. Mines, B. F. Hanson, and J. E. Shoup, "Frequency of Occurrence of Phonemes in Conversational English," Language and Speech **21**(3), 221=241 (1978) http://journals.sagepub.com/doi/abs/10.1177/002383097802100302 doi: 10.1177/002383097802100302.

[70]S. Jayaraman, C. M. Fausey, and L. B. Smith, "The Faces in Infant-Perspective Scenes 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.

[71]E. M. Clerkin, E. Hart, J. M. Rehg, C. Yu, and L. B. Smith, "Real-world visual statistics and infants' first-learned object names," Philosophical Transactions of the Royal Society B: Biological Sciences **372**(1711) (2016) http://rstb.royalsocietypublishing.org/content/372/1711/20160055.article-info.

[72]A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," (2007) http://arxiv.org/abs/0706.1062http://dx.doi.org/10.1137/070710111 doi: 10.1137/070710111.

[73]R. Salakhutdinov, A. Torralba, and J. Tenenbaum, "Learning to share visual appearance for multiclass object detection," in *CVPR 2011*, IEEE (2011), pp. 1481–1488, http://ieeexplore.ieee.org/document/5995720/, doi: 10.1109/CVPR.2011.5995720.

[74]G. Weiss and F. Provost, "The effect of class distribution on classifier learning: an empirical study," Rutgers Univ (2001) ftp://ftp.cs.rutgers.edu/http/cs/cs/pub/technical-reports/work/ml-tr-44.pdf.

[75]D. F. Kleinschmidt and T. F. Jaeger, "Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning?," Psychonomic Bulletin & Review **23**(3), 678–691 (2016) http://link.springer.com/10.3758/s13423-015-0943-z doi: 10.3758/s13423-015-0943-z.

[76]K. Idemaru and L. L. Holt, "Word recognition reflects dimension-based statistical learning.," Journal of Experimental Psychology: Human Perception and Performance **37**(6), 1939–1956 (2011) http://doi.apa.org/getdoi.cfm?doi=10.1037/a0025641 doi: 10.1037/a0025641.

[77]K. E. Moczulska, J. Tinter-Thiede, M. Peter, L. Ushakova, T. Wernle, B. Bathellier, and S. Rumpel, "Dynamics of dendritic spines in the mouse auditory cortex during memory formation and memory recall.," Proceedings of the National Academy of Sciences of the United States of America **110**(45), 18315–20 (2013) http://www.ncbi.nlm.nih.gov/pubmed/24151334 doi: 10.1073/pnas.1312508110.

[78]D. B. Polley, "Perceptual Learning Directs Auditory Cortical Map Reorganization through Top-Down Influences," Journal of Neuroscience **26**(18), 4970–4982 (2006) http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3771-05.2006 doi: 10.1523/JNEUROSCI.3771-05.2006.

[79]R. C. Froemke, M. M. Merzenich, and C. E. Schreiner, "A synaptic memory trace for cortical receptive field plasticity.," Nature **450**(7168), 425–429 (2007) http://dx.doi.org/10.1038/nature06289 doi: 10.1038/nature06289.

[80]J. Fritz, M. Elhilali, and S. Shamma, "Active listening: Task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex," Hearing Research **206**(1-2), 159–176 (2005) http://www.nature.com/doifinder/10.1038/nn1141 doi: 10.1016/j.heares.2005.01.015.

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

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

[83] K. E. Radziwon, K. M. June, D. J. Stolzberg, M. A. Xu-Friedman, R. J. Salvi, and M. L. Dent, "Behaviorally measured audiograms and gap detection thresholds in CBA/CaJ mice.," Journal of comparative physiology. A, Neuroethology, sensory, neural, and behavioral physiology **195**(10), 961–9 (2009) http://www.ncbi.nlm.nih.gov/pubmed/19756650http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2813807 doi: 10.1007/s00359-009-0472-1.

[84] P. Boersma, "Praat, a system for doing phonetics by computer," Glot International **5**(9/10), 341–347 (2001) doi: 10.1097/AUD.0b013e31821473f7.

[85] H. D. Patterson and R. Thompson, "Recovery of Inter-Block Information when Block Sizes are Unequal," Biometrika **58**(3), 545 (1971) http://www.jstor.org/stable/2334389?origin=crossref doi: 10.2307/2334389.

[86] R. C. Team, "R: A language and environment for statistical computing.," (2016).

[87] R. Team, "RStudio: Integrated Development for R.," (2015), http://www.rstudio.com/.

[88]D. Bates, M. Machler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," Journal of Statistical Software **67**(1), 1–48 (2015) doi: 10.18637/jss.v067.i01.

[89]M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, "cluster: Cluster Analysis Basics and Extensions," (2017).

[90]D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision **60**(2), 91–110 (2004).

[91]J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, "Fiji: an open-source platform for biological-image analysis," Nature Methods **9**(7), 676–682 (2012) http://www.nature.com/doifinder/10.1038/nmeth.2019 doi: 10.1038/nmeth.2019.

[92]J. Schindelin, C. T. Rueden, M. C. Hiner, and K. W. Eliceiri, "The ImageJ ecosystem: An open platform for biomedical image analysis," (2015), doi: 10.1002/mrd.22489.

[93]D.-R. Sundar, "binom: Binomial Confidence Intervals For Several Parameterizations," (2014).

[94]Werker, Janet F, Lalonde, Chris E, "Cross-language speech perception: Initial capabilities and developmental change.," Developmental psychology **24**(5) 672, (1988)

[95]H. Wickham, "The Split-Apply-Combine Strategy for Data Analysis," Journal of Statistical Software **40**(1), 1–29 (2011) http://www.jstatsoft.org/v40/i01/.

[96] H. Wickham, "Reshaping data with the reshape package," Journal of Statistical Software **21**(12), (2007) http://www.jstatsoft.org/v21/i12/paper

[97] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York, NY, 2009).

[98] R. Wright, *A review of perceptual cues and cue robustness* Phonetically based phonology 34–57 (2004).

[99] D. B. Dahl, "xtable: Export Tables to LaTeX or HTML," (2016).

[100] "Pitch Shifting and Time Dilation Using a Phase Vocoder in MATLAB," https://www.mathworks.com/help/audio/examples/pitch-shifting-and-time-dilation-using-a-phase-vocoder-in-matlab.html

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. **??**) 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}	$\Pr(>\chi^2)$
Mouse	2			
Mouse + Type	6	2534.46	4	$\ll 0.001$
Type | Mouse	20	407.22	14	$\ll 0.001$

VI. FIGURE CAPTIONS

Figure ??: **Stimuli and Task Design.** **a)** Spectrograms of stimuli. Left: Example of an original recording of an isolated consonant-vowel token (/gI/). 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.

Figure ??: 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 on 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**.

Figure ??: 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).

Figure ??: 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.

Figure ??: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/ in panel b), /b/ tokens obeyed the predicted categorization while /g/ tokens did not (slopes of colored lines).