Learning nonnative speech sounds changes local encoding in the adult human cortex

Han G. Yi[a,b], Bharath Chandrasekaran[c], Kirill V. Nourski[d], Ariane E. Rhone[d], William L. Schuerman[a,b], Matthew A. Howard III[d], Edward F. Chang[a,b,1], and Matthew K. Leonard[a,b,1,2]

[a]Department of Neurological Surgery, University of California, San Francisco, CA 94143; [b]Weill Institute for Neurosciences, University of California, San Francisco, CA 94143; [c]Department of Communication Sciences and Disorders, University of Pittsburgh, Pittsburgh, PA 15260; and [d]Department of Neurosurgery, Roy J. and Lucille A. Carver College of Medicine, The University of Iowa, Iowa City, IA 52242-1089

Edited by Patricia K. Kuhl, University of Washington, Seattle, WA, and approved July 12, 2021 (received for review January 30, 2021)

Adults can learn to identify nonnative speech sounds with training, albeit with substantial variability in learning behavior. Increases in behavioral accuracy are associated with increased separability for sound representations in cortical speech areas. However, it remains unclear whether individual auditory neural populations all show the same types of changes with learning, or whether there are heterogeneous encoding patterns. Here, we used high-resolution direct neural recordings to examine local population response patterns, while native English listeners learned to recognize unfamiliar vocal pitch patterns in Mandarin Chinese tones. We found a distributed set of neural populations in bilateral superior temporal gyrus and ventrolateral frontal cortex, where the encoding of Mandarin tones changed throughout training as a function of trial-by-trial accuracy ("learning effect"), including both increases and decreases in the separability of tones. These populations were distinct from populations that showed changes as a function of exposure to the stimuli regardless of trial-by-trial accuracy. These learning effects were driven in part by more variable neural responses to repeated presentations of acoustically identical stimuli. Finally, learning effects could be predicted from speech-evoked activity even before training, suggesting that intrinsic properties of these populations make them amenable to behavior-related changes. Together, these results demonstrate that nonnative speech sound learning involves a wide array of changes in neural representations across a distributed set of brain regions.

learning | speech | neurophysiology | perception

Humans are finely attuned to the sounds in their native language (1, 2), driven by extensive experience hearing these sounds in many different contexts from different speakers (3–5). However, for nonnative sounds in unfamiliar languages, adult listeners often struggle to learn to recognize relatively simple contrasts (6–9). For example, although native English listeners understand how changes in vocal pitch indicate intonational prosody (e.g., statements versus questions; refs. 10 and 11), this does not translate to the ability to easily identify the syllable-level pitch patterns that define lexical tones in Mandarin Chinese (12, 13). Fundamentally, this difficulty may reflect a trade-off between maintaining stable representations of deeply engrained speech sounds and retaining enough plasticity to be able to continue to learn behaviorally relevant information throughout the lifespan (14–19). Learning to identify nonnative speech sounds often requires long and intense periods of active training (12, 20–22), consistent with the observation that speech circuits in the human brain are resistant to change following developmental critical periods (15, 17).

However, even brief training periods can lead to an increased ability to identify novel speech sounds, albeit with highly variable performance across individuals (23–25). Behavioral evidence has further shown that the way listeners perceive relevant auditory cues changes after speech training (14, 25–27), which has led to the hypothesis that learning is rooted in more distinct neural representations of those sounds (17, 27). Consistent with this hypothesis, previous neuroimaging studies have shown that activation in frontotemporal areas increases following identification or discrimination tasks (13, 19, 28–30). These increases in the magnitude of activation are further associated with greater neural separability among sound categories for both speech (31–33) and nonspeech sounds (34–36). However, recent evidence suggests a highly diverse set of speech representations even within areas like the superior temporal gyrus (STG; ref. 37). Currently, the extent to which learning-related changes vary at the level of local populations remains unknown due to the broad spatial scale of noninvasive methods, which may obscure more complex dynamics. In addition, it is unclear how learning-related changes evolve on a trial-by-trial basis, as listeners initially learn to use the stimulus dimensions that allow them to achieve increased accuracy on the task, since most previous work examines neural activity only at early and late stages of the task.

Here, we examined the relationship between behavioral performance during the initial stage of nonnative speech sound learning and the trial-by-trial encoding of speech content in local neural populations in the human brain. English-speaking participants listened to unfamiliar Mandarin syllables and learned to identify tone categories (22, 38), while neural activity was recorded from electrocorticography (ECoG) arrays placed over lateral cortical areas. We hypothesized that, as listeners heard

Significance

Speech sound learning in adulthood is a highly dynamic process. Here, we used direct neurophysiology of the human brain to examine learning-associated changes in neural activity with unprecedented spatiotemporal detail. While native English listeners were trained to identify unfamiliar pitch patterns in Mandarin, local neural populations throughout the cortex showed a diverse set of encoding properties for Mandarin sounds that tracked behavioral performance. While previous neuroimaging studies have focused on highlighting general differences across broad cortical regions, we demonstrate that these functionally heterogeneous populations are spatially interspersed with one another. These findings provide insight into how the human brain strikes a balance between stability and plasticity during learning in adulthood.

Author contributions: H.G.Y., B.C., E.F.C., and M.K.L. designed research; H.G.Y., K.V.N., A.E.R., W.L.S., M.A.H., E.F.C., and M.K.L. performed research; H.G.Y. and M.K.L. analyzed data; and H.G.Y., B.C., E.F.C., and M.K.L. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]E.F.C. and M.K.L. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: matthew.leonard@ucsf.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2101777118/-/DCSupplemental.

Published September 2, 2021.

the same stimuli across multiple exposures, some neural populations would show responses to Mandarin speech sounds that track the trial-by-trial fluctuations in participants' behavioral performance during learning, and we also asked whether these changes would be uniformly reflected in increased separability among tones. We further hypothesized that these learning-related neural populations would be distinct from other potential patterns of change across trials that do not directly correlate with learning (e.g., the number of exposures to a given token, independent from accuracy) and neural populations that show stable activity patterns across trials. To address the relationship to stimulus feature encoding (e.g., pitch representations for English intonational prosody; ref. 39), we also measured the extent to which neural responses to unfamiliar Mandarin speech sounds prior to training can be used to predict the emergence of learning-related changes during training.

We found a subset of local populations across the cortical surface that track trial-by-trial accuracy, even when learning performance is relatively low and variable. These learning-related effects manifest as both increases and decreases in the amplitude of neural responses to specific speech sounds and are spatially interspersed and dissociable from those that arise simply as a function of repeated exposure. Furthermore, learning-related changes are associated with higher variability of the response amplitude across repeated exposures to the same acoustic stimulus, suggesting less robust emergent neural representations. Finally, we show that intrinsic properties of these neural populations are associated with whether they show learning-related effects during training, allowing us to predict whether these effects will occur based on responses to the novel speech sounds prior to training. Together, these results demonstrate that learning to identify novel speech sounds scaffolds on existing sensitivities to relevant features and that the initial stages of learning a new language involve a specific set of processes to fine-tune local speech representations in the brain. We propose that the learning-induced increased neural separability in frontotemporal regions arises from heterogeneous changes among local populations, which comprise those regions.

Results

A total of 10 native English speakers performed a Mandarin tone identification training task while we recorded ECoG from the cortical surface (*SI Appendix*, Fig. S1). On each trial, participants listened to a Mandarin syllable, which varied according to speaker (male/female), phonemes (/bu/, /di/, /lu/, /ma/, and /mi/) and lexical tone (pitch contour; T1: high-flat; T2: high-falling; T3: low-dipping; and T4: low-rising; Fig. 1*A* and *SI Appendix*, Fig. S2). After hearing each syllable, participants pressed a button to indicate which tone they heard and received visual feedback to indicate accuracy (22, 31, 38). Six participants performed a four-alternative forced choice task, three of whom also performed two-alternative forced choice tasks with two pairs of tones (T1 and T3, or T2 and T4). The remaining four participants only performed two-alternative forced choice tasks (*SI Appendix*, Fig. S3).

Participants showed variable learning curves over the course of the task and across tones. A representative participant showed increases from chance to above chance accuracy across 130 exposures to T3 (mixed-effects model; $B = 0.047$, $SE = 0.012$, $Z = 3.73$, and $P < 0.001$) and T4 ($B = 0.064$, $SE = 0.018$, $Z = 3.49$, and $P < 0.001$; Fig. 1*B*). In this example, other tones showed nonmonotonic changes in accuracy: While T1 showed above-chance performance in the middle of the task, accuracy did not change monotonically with exposure ($B = 0.002$, $SE = 0.007$, $Z = 0.031$, and $P = 0.76$), and accuracy for T2 decreased with exposure ($B = -0.030$, $SE = 0.011$, $Z = -2.66$, and $P = 0.008$), showing above-chance performance during early trials only.

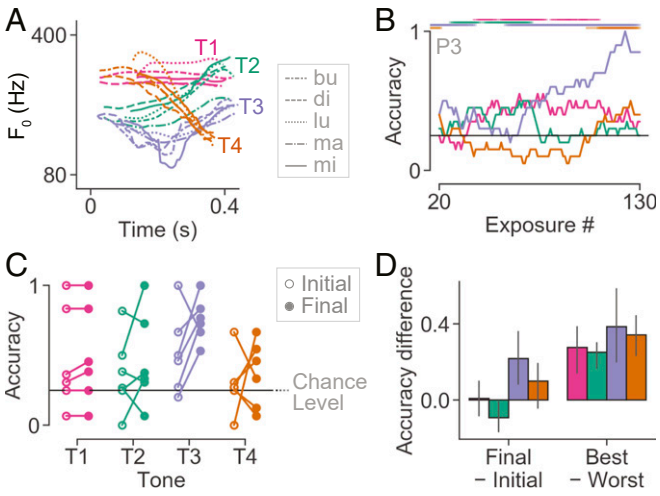

Fig. 1. Mandarin tone identification task and behavioral performance. (*A*) Mandarin tones are syllable-level fundamental frequency (F_0) patterns. Example stimuli from a female native speaker are shown for all four tones produced with five syllables. (*B*) Behavioral performance in an example participant (P3), showing sliding-window (20 trials) average identification accuracy as a function of tone-specific exposure. Horizontal lines indicate trial windows with above-chance accuracy (FDR–$P < 0.05$). (*C*) Tone-specific accuracy for the initial and final 10% of all trials across participants. (*D*) Tone-specific learning effects quantified as the difference between final versus initial trials (*Left*) and best and worst trial windows (*Right*) across all participants.

Across all participants, performance in later trial windows was better than in earlier trial windows ($B = 0.013$, $SE = 0.002$, $Z = 6.41$, and $P < 0.001$; Fig. 1*C* and *SI Appendix*, Fig. S3), demonstrating that, overall, performance increased with more exposures to each tone category. However, the rate, magnitude, and monotonicity of this learning effect was variable. Performance for T1 significantly improved ($B = 0.006$, $SE = 0.002$, $Z = 2.51$, and $P = 0.012$). All the other tones showed faster improvement of accuracy compared to T1: T2 ($B = 0.015$, $SE = 0.003$, $Z = 546$, and $P < 0.001$), T3 ($B = 0.019$, $SE = 0.004$, $Z = 4.44$, and $P < 0.001$), and T4 ($B = 0.025$, $SE = 0.004$, $Z = 5.67$, and $P < 0.001$).

Overall, T3 showed the largest improvement between initial and final trials (Fig. 1 *D*, *Left*), consistent with previous studies (40). However, as was evident from the example participant (Fig. 1*B*), some of the behavioral variability was due to nonmonotonic learning curves. Therefore, we also compared the amount of improvement between the best and worst trial windows, defined by 20-trial moving averages of the highest and lowest numbers of correct trials per tone in each participant. We found that all four tones showed significant improvement [T1: $T(5) = 3.09$, $P = 0.03$; T2: $T(5) = 4.33$, $P = 0.008$; T3: $T(5) = 3.07$, $P = 0.028$; and T4: $T(5) = 2.63$, $P = 0.046$; Fig. 1 *D*, *Right*]. This demonstrates that although many participants did not simply learn and retain the novel tone identities during the course of this short task, they exhibited periods of high accuracy that could be used to characterize the neural computations underlying speech sound learning on a trial-by-trial basis.

Next, we examined local neural population responses to each stimulus as participants performed the Mandarin tone identification training task. We focused on activity in the high-gamma range (70 to 150 Hz), which is correlated with multiunit activity (41, 42) and has been shown to encode auditory- and speech-relevant features (43, 44). Specifically, we assessed the neural encoding of two critical behavioral features that define learning: 1) trial-by-trial accuracy and 2) the number of exposures to a given tone category. We first identified electrodes that had a significant high-gamma amplitude (HGA) response averaged

across all trials following stimulus onset (n = 1,242 out of 2,816 total electrode contacts; paired t test; false discovery rate (FDR)–$P < 0.05$). These sound-sensitive electrodes were used in subsequent analyses that model trial-by-trial behavioral–neural relationships.

For all sound-sensitive electrodes, we modeled HGA using a time-dependent linear ridge (L2-regularized) regression analysis (see *Materials and Methods* for details). We found electrodes that had a significant interaction between accuracy, exposure, and tone identity, which we define as a "learning effect." In order to illustrate these types of changes, we first present single electrodes as representative examples, followed by quantitative analyses across the population of electrodes. This learning effect manifested as three distinct types of changes in HGA depending on trial-by-trial behavior. First, we found electrodes with increased HGA for specific tones only for correct later trials (Fig. 2*A*; evoked response plots collapsed into first versus last 50% of trials for visualization). Second, we found electrodes with increased HGA for specific tones only in incorrect later trials (Fig. 2*B*). Third, we found electrodes in which HGA for specific tones decreased for incorrect later trials (Fig. 2*C*).

Crucially, this learning effect was highly specific and distinct from other types of tone-specific changes in neural activity that occurred over the course of the task but did not depend on both accuracy and exposure. For example, we found electrodes in which HGA responses increased for specific tones on later trials independent of accuracy (exposure effect; Fig. 2*D*). We also observed a small number of electrodes that were sensitive to trial-by-trial accuracy but not number of exposures (accuracy effect). Likewise, there was a small number of electrodes that encoded tone identity in a similar manner across the task regardless of both trial-by-trial accuracy and exposure number (tone-stable electrodes).

To quantify these effects for all sites, we used the same linear ridge regression analysis across all electrodes. Across all participants, we found 141 electrodes with significant learning effects (FDR–$P < 0.05$), 108 of which had the most unique explained variance for the learning effect compared to all other effects tested (Fig. 2*E*). These electrodes were primarily located throughout the STG (n = 50; 16.1% of sound-sensitive electrodes in the region), with a smaller number on ventral sensorimotor cortex (n = 22; 8.8%) and inferior frontal cortex (n = 9; 7.0%). We also found 201 electrodes with significant exposure effects (FDR–$P < 0.05$), 150 of which had the most unique explained variance for the exposure effect (Fig. 2*E*). Learning, exposure, accuracy, and tone-stable electrodes were all intermixed throughout frontal, temporal, and parietal regions, with no clear spatial organization for specific effect types (Fig. 2*E*). These results show that neural populations that track various aspects of behavioral performance during nonnative speech sound learning are distributed throughout cortical regions that encode critical acoustic features for native speech sounds (45).

Next, we evaluated the relative magnitude of each type of effect. We directly compared the unique variance explained by each effect for all electrodes for which the fully specified encoding model (see *Materials and Methods*) explained a significant amount of variance [n = 433; FDR–$P < 0.05$; permutation testing (46)]. The learning effect explained an average of 35.0%

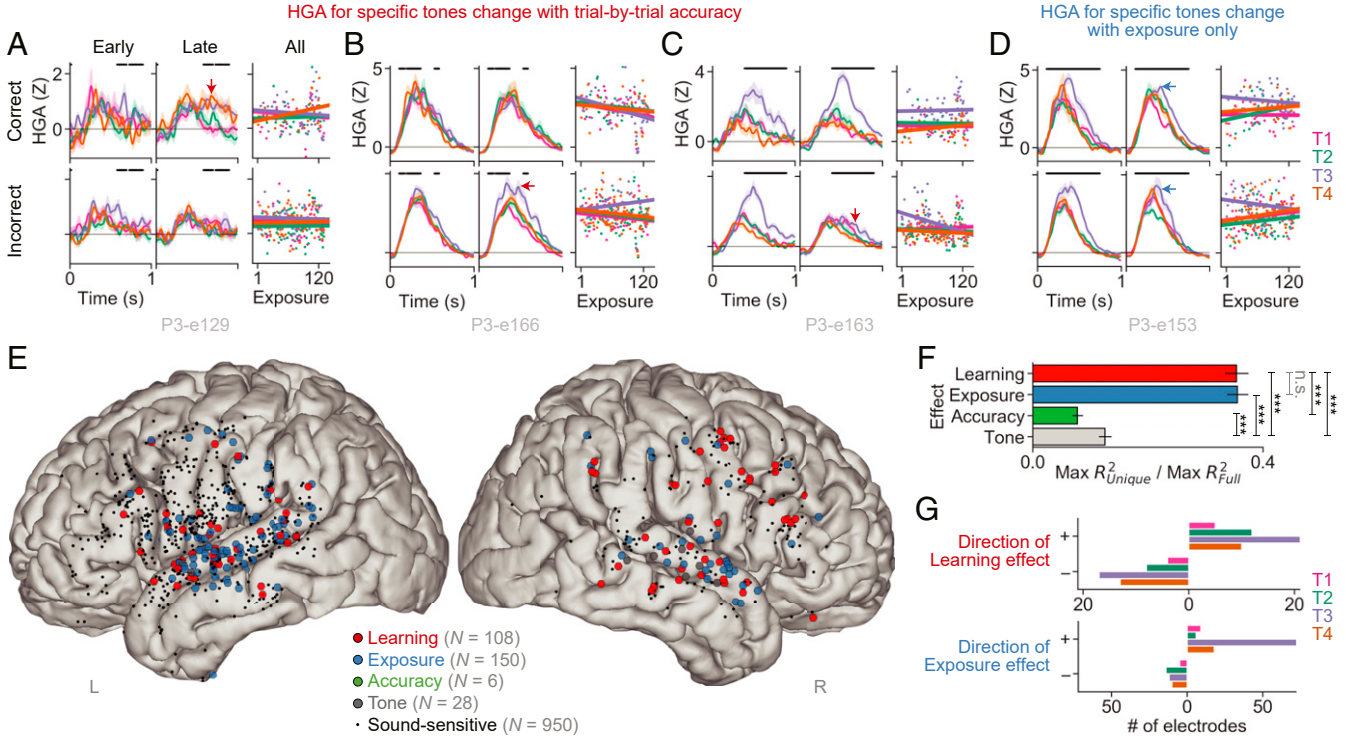

Fig. 2. Neural populations track trial-by-trial behavioral performance, dependent on number of prior exposures ("learning effect"). (*A*–*D*) Individual electrode HGA responses time locked to stimulus onset for first versus last 50% exposures per tone (columns) and correct versus incorrect trials (rows). Scatter plots show time-averaged HGA for correct versus incorrect trials (rows) as a function of exposure with regression fits for each tone. Learning effects can manifest as increased amplitude for late correct trials (*A*), increased amplitude for late incorrect trials (*B*), or decreased amplitude for late incorrect trials (*C*). (*D*) Learning effects contrast with simple exposure effects (e.g., changes in amplitude dependent on exposure but not accuracy). (*E*) Learning, exposure, accuracy (not visible), and stable tone-encoding electrodes are located primarily in the bilateral STG and ventrolateral frontal cortex, with no clear spatial organization for different effect types. (*F*) Unique explained variance for learning and exposure effects are larger than for accuracy and behavior-independent tone effects. (*G*) Learning effects can manifest as tone-specific increases and decreases in HGA, while exposure effects manifest more predominantly as increases, particularly for T3. *$P \leq 0.05$; **$P \leq 0.01$; ***$P \leq 0.001$.

(SD = 20.5) of the total explained variance, which was not significantly different from the contribution of the exposure effect (35.6%; SD = 19.6; B = 0.002, SE = 0.009, Z = 0.17, and P = 0.9). In contrast, the accuracy effect explained 7.8% (SD = 8.5) of the total explained variance, which was significantly lower than for the learning effect (B = −0.229, SE = 0.009, Z = −25.68, and P < 0.001). The proportion of variance explained by the tone-stable effect (12.5%; SD = 10.8) was also significantly lower than for the learning effect (B = −0.229, SE = 0.009, Z = −25.68, and P < 0.001) but higher than for the accuracy effect (B = −0.048, SE = 0.009, Z = −5.37, and P < 0.001; Fig. 2*F*). Given the small number of significant accuracy and tone-stable electrodes (Fig. 2*E*) and lower effect size (Fig. 2*F*), the remaining analyses are focused on understanding the characteristics of electrodes that change with trial-by-trial behavioral performance (learning electrodes) versus those that change with experience alone (exposure electrodes).

Several studies have observed that novel sound category learning is associated with increased neural activity in frontotemporal areas (29, 34, 35, 47, 48). Since we observed examples of both increases and decreases in activity over time with learning (Fig. 2 *A*–*D*), we next quantified the extent to which these patterns manifested differently between learning and exposure effects across the four Mandarin tones. We found that learning effects were equally distributed between positive (increase for correct later trials or decrease for incorrect later trials) and negative (increase for incorrect later trials or decrease for correct later trials) directions [$\chi^2(1, n = 91) = 1.41$ and $P = 0.70$; Fig. 2 *G*, *Top*). In contrast, exposure effects were relatively more likely to manifest in the positive direction [increase for later trials; $\chi^2(1, n = 150) = 28.12$ and $P < 0.001$; Fig. 2 *G*, *Bottom*). Finally, both learning and exposure effects differed across tones [learning: $\chi^2(3, n = 91) = 20.25$ and $P < 0.001$; exposure: $\chi^2(3, n = 150) = 93.31$ and $P < 0.001$), where T3 was most likely to show either effect (learning: 42.9%; exposure: 58.7% among all significant electrodes).

Together, these results demonstrate that there are distinct neural populations throughout the lateral human cortex that encode trial-by-trial behavioral performance during the response to novel auditory stimuli. These neural populations are separate from those that show changes across trials simply due to sensory experience. Furthermore, the trial-by-trial behavior encoding can manifest in several different ways, suggesting a complex set of a changes that occur as participants learn to identify new speech sounds, rather than simply an increase in evoked responses for better-learned stimuli. Indeed, individual electrodes demonstrate heterogeneous changes (e.g., increases in correct responses ["positive"; Fig. 2*A*], increases in incorrect responses ["negative"; Fig. 2*B*], and decreases in incorrect responses ["positive"; Fig. 2*C*]). We suggest that these learning electrodes encode subjectively perceived sounds as manifested by behavioral responses during training (*SI Appendix*, Fig. S4; ref. 49).

Thus far, we have demonstrated that changes in the mean amplitude of the high-gamma response encode trial-by-trial behavioral performance. It is also possible for learning effects to manifest as changes in neural variability separately from changes in the overall amplitude of evoked responses (50, 51). In assessing neural variability, it is important to examine the relationship between variance (σ^2) and mean (μ) of evoked activity across multiple trials of the same stimulus, as the two measures are positively correlated within a given recording unit (52, 53), including at the level of ECoG electrodes, where increases and decreases in mean activity can obscure potentially meaningful changes in variance (54).

To illustrate how variance and mean of HGA can differ across electrodes, we examined the mean and the SD of neural responses to multiple repetitions of a single unique stimulus (T4 produced by a male speaker in the syllable /di/) from an example learning electrode (Fig. 3 *A*, *Left*) and exposure electrode (Fig. 3 *A*, *Right*). In this example, the learning electrode had higher variance ($\sigma^2_{learning}$ = 0.73; averaged between 110 and 530 ms post-stimulus onset; peak time period determined from all sound-sensitive electrodes) and similar mean ($\mu_{learning}$ = 2.20) HGA compared to the exposure electrode ($\sigma^2_{exposure}$ = 0.14; $\mu_{exposure}$ = 2.07; Fig. 3*A*). These differences indicate that the evoked response of the example learning electrode was less reliable across trials relative to that of the example exposure electrode, despite the fact that the stimulus was acoustically identical each time. Across all unique stimuli, the variance was significantly higher for the learning electrode [t (39) = 6.53 and $P < 0.001$], while the mean was not significantly different [$t(39)$ = 1.64 and P = 0.11; Fig. 3*B*]. Together with the previous results, these findings suggest that behavior-related changes in neural activity are encoded in both the mean and variance of HGA differently for learning and exposure effects.

Next, we asked whether learning and exposure electrodes exhibit differences in the relationship between mean and variance

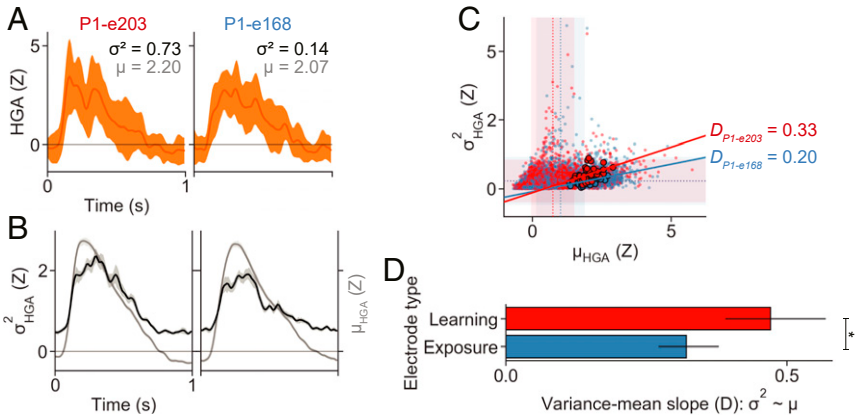

Fig. 3. Learning electrodes have higher variance-to-mean slopes compared to exposure electrodes. (*A*) Example responses (mean ± SD) for representative learning (*Left*) and exposure (*Right*) electrodes to repeated presentations of a single stimulus (T4; "di"; male speaker No. 1). Average variance (σ^2) and mean (μ) values within the peak time period (110 to 530 ms) are shown for each electrode. (*B*) Across all unique Mandarin tone stimuli, variance (σ^2; black) is larger for the learning electrode (*Left*), while mean (μ; gray) did not differ compared to the exposure electrode (*Right*). (*C*) Mean (*x*-axis) and variance (*y*-axis) for every unique stimulus for all learning (red; n = 108) and exposure (blue; n = 150) electrodes. Shaded regions correspond to mean ± SD for each electrode type. Regression lines and larger dots highlight the responses of two example electrodes shown in *A* and *B*. (*D*) Across all electrodes, the variance-mean slope of regression (*D*) is significantly higher for learning versus exposure electrodes.

of HGA, defined here as mean-normalized variance (D: variance-to-mean slope of regression). This metric is conceptually similar to the Fano factor, which is used where zero variance can be assumed for zero mean activity (50, 54). We calculated D for the example learning and exposure electrodes shown in Fig. 3 *A* and *B*. Variance and mean were positively correlated across all stimuli in the learning electrode [r (38) = 0.424, $P < 0.01$, and $D = 0.33$] as well as in the exposure electrode [r (38) = 0.483, $P < 0.005$, $D = 0.29$; Fig. 3*C*]. Across all learning and exposure electrodes, we found greater mean-normalized variance for learning electrodes (mean $D_{learning} = 0.47$ and $SD = 0.48$; mean $D_{exposure} = 0.32$, $SD = 0.31$; $B = -0.117$, $SE = 0.048$, $Z = -2.43$, and $P = 0.015$; Fig. 3*D*). Thus, in addition to differences in the amplitude of mean HGA, neural responses in learning electrodes were more variable across multiple presentations of an acoustically identical stimulus. Crucially, higher variability for learning electrodes is not simply due to larger changes across the learning task, since the effect size of the learning and exposure effects was not significantly different (Fig. 2*F*).

Thus far, we have demonstrated that a subset of cortical neural populations tracks trial-by-trial behavioral effects that relate to Mandarin tone category learning. Next, we asked whether these populations exhibit characteristic neurophysiological properties prior to training, which predict whether they will emerge as learning or exposure electrodes during the training task. We hypothesized that any pretraining activity that differentiates these unfamiliar sounds would be related to representations of pitch height and pitch change, which are the two critical acoustic features that distinguish Mandarin tones (Fig. 4*A*).

Seven participants listened to the same Mandarin tone stimuli passively prior to the training task, which evoked significant but smaller HGA responses relative to responses from the training task (*SI Appendix*, Fig. S5). We examined the encoding of unfamiliar speech sounds by comparing the separability of tone categories from single-trial neural responses. We first compared the pretraining neural representations of each stimulus for example learning and exposure electrodes by projecting the average HGA into a two-dimensional linear discriminant (LD) space (see *Materials and Methods*). In the learning electrode, the two-dimensional neural representation of the four tones was highly overlapping, with only T3 showing clear separability from the other tones (Fig. 4*B*; tone separability quantified with logistic regression area under the curve [AUC], see *Materials and Methods*). In contrast, prior to training, an example exposure electrode demonstrated clearer separability for all four tones (Fig. 4*C*).

Across all electrodes from participants who performed the pretraining passive listening task, learning electrodes ($n = 23$) had lower separability (mean AUC = 0.531; $SD = 0.051$) compared to exposure electrodes ($n = 33$; mean AUC = 0.624; $SD = 0.074$; and $B = 0.073$, $SE = 0.019$, $Z = 3.93$, and $P < 0.001$; Fig. 4*D*). These results demonstrate that electrodes that exhibit learning effects over the course of training are characterized by poorer tone separability prior to training, compared to electrodes that exhibit exposure effects (i.e., electrodes that show similar magnitudes of change during training but are not related to trial-by-trial behavior).

Next, we examined whether these pretraining tone representations were associated with the characteristic pitch features that define Mandarin tone categories. We observed that, in the two example electrodes (Fig. 4 *B* and *C*), the two-dimensional neural representation of the tone categories was qualitatively similar to the stimulus-based pitch height and pitch change representation (Fig. 4*A*). Indeed, the organization of these neural spaces was significantly correlated with the acoustic pitch space in both electrodes (learning: $R^2 = 0.401$ and $P < 0.001$; exposure: $R^2 = 0.538$ and $P < 0.001$), driven primarily by the pitch height dimension (learning: $R^2_{Height} = 0.502$, $P_{Height} < 0.001$, and $R^2_{Change} = 0.002$; $P_{Change} = 0.6$; and exposure: $R^2_{Height} = 0.595$, $P_{Height} < 0.001$, $R^2_{Change} = 0.022$, and $P_{Change} = 0.065$).

Across learning and exposure electrodes, while the pitch height feature was represented significantly better by exposure electrodes ($B = 107$, $SE = 0.041$, $Z = 2.58$, and $P = 0.01$), there was no significant difference for the representation of the pitch change feature between the two populations ($B = -0.026$, $SE = 0.015$, $Z = -1.75$, and $P = 0.08$). Both electrode types encoded pitch height more robustly than pitch change during pretraining passive listening (learning: $B = 0.126$, $SE = 0.027$, $Z = 4.62$, and $P < 0.001$; exposure: $B = 0.273$, $SE = 0.029$, $Z = 9.32$, and $P < 0.001$) (Fig. 4*E*). This demonstrates that prior to training, both learning and exposure populations are primarily sensitive to pitch height.

Finally, we directly tested the hypothesis that the representation of Mandarin tones during passive listening prior to training predicts the electrode-level encoding of behavioral changes during training. We used logistic regression to predict from passive listening data whether each electrode would exhibit significant learning or exposure effects in the training data. We used both Mandarin tone separability (AUC) and model fits for pitch height and pitch change (R^2) to perform classification.

Between learning and exposure electrodes ($n = 56$), the model could predict the presence of the learning effect (AUC = 0.730, $P = 0.005$, and permuted 95% CI [0.304, 0.680]), as well as the exposure effect (AUC = 0.778, $P = 0.004$, and permuted 95% CI [0.266, 0.713]), with no significant difference between the two predictions ($Z = 0.0857$, $P = 0.46$; permutation testing) (receiver operating characteristic [ROC] curves shown in Fig. 4*F*; AUC values shown in Fig. 4*G*). Furthermore, when the same analysis was expanded to include all sound-sensitive electrodes ($n = 475$), the presence of the learning effect was marginally above chance (AUC = 0.612, $P = 0.061$, and permuted chance level 95% CI [0.349, 0.630]). Likewise, the presence of the exposure effect was also significantly predicted above chance (AUC = 0.789, $P < 0.001$, and permuted 95% CI [0.369, 0.613]), with the exposure effect having a significantly better AUC effect than the learning effect ($Z = -3.47$ and $P < 0.001$; permutation testing). Together, these results demonstrate that during the initial exposure to unfamiliar speech sound categories, neural populations exhibit characteristic properties that predict their subsequent activity patterns as listeners learn to identify these categories.

Discussion

We demonstrate that speech sound encoding in the adult human cortex is modulated by trial-by-trial behavioral performance. Learning-related changes were highly specific, occurring in neural populations that showed either increasing or decreasing sensitivity to particular nonnative speech categories, and were distinct from neural populations that showed changes based simply on the number of exposures to a particular sound. These learning effects also manifested in differences in the mean-variance relationship of the neural response and were predictable based on neural responses to these sounds during a passive listening task that occurred prior to training. Together, these results demonstrate how local neural populations flexibly change their activity patterns in a manner that reflects trial-by-trial behavioral performance during the initial stages of learning to understand an unfamiliar language.

Previous work in human auditory learning has identified cortical networks that show changes in neural activity and stimulus representations during speech learning, typically at the level of comparing early versus late exposures, and with average across-subject behavioral performance that monotonically increases with training (29, 34, 35, 47, 48). While these studies suggest the existence of rapid modulation of cortical encoding during learning (31, 36, 55), how changes manifest on a single-subject, single-trial level has been previously unexplored. In line with results from large samples of healthy volunteers (23–25, 56), behavioral performance

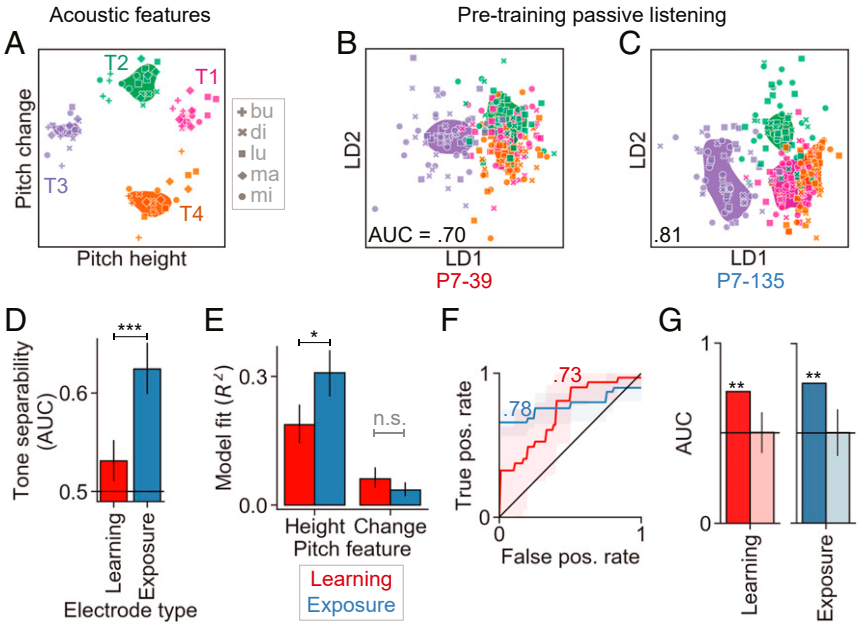

Fig. 4. Pretraining passive listening to Mandarin tones predicts different types of behavior-related neural changes. (*A*) Mandarin tone stimuli are highly separable based on acoustic features of pitch height and pitch change. (*B* and *C*) LD spaces for all trials during pretraining passive listening from two example STG electrodes, categorized as a learning (*B*) or exposure (*C*) electrode based on training trials. (*B*) Prior to training, the example learning electrode shows highly overlapping representations across tones (quantified with area under the curve from LD model; AUC). (*C*) In contrast, an example exposure electrode shows higher separability across all tones. (*D*) Across all electrodes, pretraining tone separability is lower for learning electrodes than for exposure electrodes. (*E*) Learning and exposure electrodes have similar representations of pitch features prior to training, driven primarily by pitch height. (*F*) Using tone separability and pitch representation during passive listening as input features, the presence of significant learning (red) or exposure (blue) effects in training can be predicted for each electrode significantly above chance level. (*G*) AUC values are shown for data in *F*. Lighter colors show permuted values ($n = 1{,}000$; error bars indicate SD). $^{*}P \leq 0.05$; $^{**}P \leq 0.01$; $^{***}P \leq 0.001$; $^{****}P \leq 0.0001$.

was highly variable and nonmonotonic. Using direct neurophysiological recordings in awake human participants during behavior, we discovered that neural populations exhibit specific changes associated with learning on a trial-by-trial basis, which provides evidence that the process of learning to identify novel speech sounds involves fine-scale tuning of neural representations with each exposure.

We found a highly specific learning effect, in which neural activity tracked the trial-by-trial interaction between behavioral accuracy and number of exposures to a given stimulus. Prior work has shown that similar learning effects manifest as responses to preferred stimuli that increase on later correct trials (57–59). We observed this type of learning effect in a subset of neural populations, but we also observed decreases for incorrect trials later during training. These findings are consistent with the hypothesis that the stimulus being learned becomes overall more discriminable in neural activity (60–62).

However, some populations showed other unexpected patterns, such as increased responses for incorrect trials later during training, which may reflect other cognitive components that are associated with perceptual learning but do not necessarily contribute to increased discriminability at the population level, such as reward prediction error (63) or attention to task-irrelevant features (64). These varied learning effects were highly distributed across both hemispheres of the cortex, including both temporal auditory and ventrolateral prefrontal areas (45, 65–68), suggesting a complex manifold of computations that involve both enhancement and suppression of speech representations or dimensions during learning. Specifically, this manifold could reflect shaping of the population-level tuning to pitch features of Mandarin tones as the result of training.

Previous studies have demonstrated the flexibility of tuning of responses to acoustic features in the auditory cortex in varying behavioral contexts (58, 59, 61, 69). Indeed, a recent human ECoG study showed that tuning to pitch features in natural Mandarin speech is dependent on linguistic experience and impacts the representation of tone categories at the neural population level (70). It is possible that these heterogeneous patterns of learning-related changes are present in prior work that has shown more homogeneous effects using noninvasive methods and that the fine-grained spatiotemporal resolution of direct cortical recordings is necessary to uncover them. Further work is necessary to understand the relationship between the local neural population learning effects ofmbserved here and the effects that are observed with magneto/electroencephalography and functional magnetic resonance imaging.

In the present study, we examined neural activity time-aligned to stimulus presentation. We focused on the time period in which the responses reflect local encoding of spectrotemporally complex acoustic–phonetic features of the speech signal (37, 45, 71–73), and we hypothesized that learning effects could modulate these representations. Whereas learning effects observed with functional neuroimaging could reflect multiple stages of processing due to temporal smearing (74), the present results demonstrate that neural responses to auditory stimuli track behavior as listeners learn to identify novel speech sounds.

We found that learning-related neural populations were functionally distinguished from other types of changes that occurred in neural responses across the task. In addition to these changes manifesting in the mean amplitude of the high-gamma response, we also observed distinctions between learning- and exposure-related effects based on the variability of the neural response. The extent to which response amplitudes differ across repeated presentations of the same stimulus offers meaningful insights into how neural activity encodes the relevant features of that stimulus (50, 75), including the relative separability of

learned representations in the context of perceptual learning (51, 60, 61, 76). In the present study, a higher mean-variance relationship in learning populations may indicate that acoustic encoding is more flexible than in populations that do not track trial-by-trial behavior (75, 77, 78). One possibility is that following extended training during which listeners gain high levels of expertise in identifying the Mandarin tones, the neural variability in the learning populations decreases as tuning to pitch features become sharpened (51, 61). These results thus provide an example of distinct processes for encoding stimulus content in both variability and mean amplitude of neural activity (50, 75).

Finally, we assessed how speech sound learning affects or leverages preexisting acoustic representations in the brain. Specifically, we asked whether intrinsic properties of neural populations make them more or less amenable to learning-related changes that occur subsequently during the learning task. We found that many neural populations distinguished Mandarin tone stimuli prior to training, possibly relying on representations of pitch features used for intonational prosody in English (39, 79). These representations reflected critical pitch features for the perceptual distinction of Mandarin tones, namely pitch height and pitch change (20, 80–82). However, specifically in the populations that subsequently showed a learning effect, we observed poorer pretraining separability for Mandarin tones compared to other populations. Like the difference in response variability, this may reflect specific properties of learning populations that allow for flexible representation of stimulus dimensions according to changes in their perceptual relevance (49, 83–86), especially compared to the other populations that may change in their encoding but do not necessarily reflect behavior. Thus, these results demonstrate how the postcritical-period brain maintains the ability to selectively modify behaviorally salient representations in specific neural populations with short-term training while keeping the representation of native speech sounds intact. We further attempted to examine the hypothesis that Mandarin tone training specifically modulates neural representations for the learned pitch dimensions [e.g., height versus direction (31)], but we were unable to collect sufficient data to test pre- versus posttraining effects in this patient cohort.

We leveraged the high spatiotemporal resolution of direct cortical recordings to understand the fine-grained changes that underlie the initial stages of learning novel speech sounds in an unfamiliar language. There is a long-standing debate regarding the nature and extent of the ability of the adult brain to remain plastic enough to learn unfamiliar speech sounds (14–19). A particularly challenging question is how stability of native speech sound representations is preserved in the face of perceptual changes brought on by relatively short periods of intense training with novel speech sounds (14, 22, 25–27). In the present study, we found that neural encoding of these speech sounds in the adult human cortex is highly amenable to fine-tuning through behavioral training in a subset of neural populations. These populations were distinct from those in which encoding changed only as a function of repeated exposure. These results are an example of how the brain balances the trade-off between stability and plasticity in learning (87, 88). This trade-off reflects the need to learn and maintain robust representations of behaviorally relevant stimuli such as speech, both during development (3, 7, 14) and in adulthood (21, 25). Learning novel speech sound categories as an adult is a crucial first step toward acquiring a second language (89), and our results support the notion that the stability–plasticity trade-off is managed by a collection of diverse neural populations that exhibit different kinds and degrees of change during the initial stages of learning.

Materials and Methods

Participants. A total of 10 human patients with epilepsy (five female; mean age: 32.7 y; SD = 12.9; range: 19 to 59) participated in this study. Seven participants (P1 through 7) were recruited at the University of California, San Francisco (UCSF) Medical Center, and three participants (P8 through 10) were recruited at the University of Iowa. All participants had normal hearing, were native speakers of English, and did not report any experience with a tonal language. For the clinical purpose of localizing seizure foci, ECoG arrays were surgically implanted on the cortical surface of one hemisphere for each participant (seven left hemisphere; see *SI Appendix*, Table S1). Electrode positions were extracted from postimplantation computed tomography scans and coregistered with the patient's preimplantation MRI (90). The research protocol was approved by the UCSF Committee on Human Research and The University of Iowa Institutional Review Board. Prior to surgery, each patient gave written informed consent to participate in this research.

Neural Data Processing. Cortical local field potentials were recorded and amplified with a multichannel amplifier optically connected to a digital signal acquisition system (Tucker-Davis Technologies). The stimuli were presented from loudspeakers at a comfortable level. The ambient audio (recorded with a microphone aimed at the participant) along with a direct audio signal of stimulus presentation were simultaneously recorded with the ECoG signals to allow for precise stimulus–neural data alignment. Signals were referenced to a subgaleal electrode. ECoG data were amplified, filtered, and digitized with a sampling rate of 3,052 Hz for data collected at UCSF and at 2,034.5 Hz for data collected at the University of Iowa. Neural data were preprocessed offline by down-sampling to 400 Hz, notch-filtering line noise at 60, 120, and 180 Hz, and excluding bad channels and bad time intervals (determined by visual inspection). From the remaining electrodes and time points, we extracted the analytic amplitude in the high-gamma frequency range using eight band-pass filters (Gaussian filters, logarithmically increasing center frequencies [70 to 150 Hz] with semilogarithmically increasing bandwidths) with the Hilbert transform. HGA was calculated by averaging the analytic amplitude across these eight bands. Lastly, the HGA was down-sampled to 100 Hz and *Z*-scored relative to the mean and SD of the data within each experimental block. For each stimulus presentation trial, we analyzed the neural data from 500 ms before stimulus onset to 1,000 ms after stimulus onset.

Mandarin Tone Stimuli. Natural exemplars (n = 80) of the four Mandarin tones (T1: high-flat, T2: low-rising, T3: low-dipping, and T4: high-falling) were produced in citation form by four native Mandarin speakers (originally from Beijing; two female) in the context of five monosyllabic Mandarin Chinese words (/bu/, /di/, /lu/, /ma/, and /mi/). These syllables were chosen because they also exist in the American English phonetic inventory. The root-mean-square amplitude of the stimuli was normalized to 70 dB and the duration to 0.442 s (56, 91). Five independent native Mandarin speakers correctly identified the four tones (categorization accuracy >95%) and rated the stimuli as highly natural. These stimuli were identical to the stimuli that were used in previous experiments (22, 31, 38).

Behavioral Procedures. In each active training trial, participants were presented with a single Mandarin tone stimulus and were asked to provide an identification response by pressing the number keys (1, 2, 3, or 4) on a keyboard, corresponding to T1, T2, T3, and T4, respectively. Corrective feedback ("right"/"wrong") was provided on a screen 0.5 s following the participant response. Participants had unlimited time to respond, and the task moved on to the next trial once the participant input a response. Because of variable participant behavior in this patient population, we used two different versions of the task. For P1 through 3, training procedures closely followed a previous study with healthy young adults (22). In each recording block for these participants, stimuli were presented once, first with a male speaker (four tones and five syllables in a randomized order), then with a female speaker. For P4 through 10, the procedures were modified to improve behavioral performance while introducing stimuli in a gradual, performance-based manner. For these participants, training was divided into three stages based on the set of tones to be learned: 1) T2 and T4, 2) T1 and T3, and 3) all four tones. Each recording block used one speaker with three syllables (/bu/, /di/, and /ma/) repeated five times in a randomized order. Participants were instructed to categorize each stimulus by clicking on visually displayed arrow symbols corresponding to pitch changes. In this version of the task, participants moved onto subsequent training stages if they had identified more than 80% of the trials after at least two blocks or after having completed four training blocks for that stage. Six participants (P4, 5, and 7 through 10) were also presented with the same set of stimuli (all tones; five syllables; male and female speakers; randomized order; 1-s intertrial interval) without any task prior to training.

Data Analysis. All behavioral (except for mixed-effects modeling for behavioral data) and neural analyses were performed using custom software written in Python. Open-source scientific Python packages used included numpy, scipy, pandas, scikit-learn, statsmodels, h5py, and tensorflow. Figures were created using matplotlib and seaborn, except for brain reconstruction images, which were created using MATLAB. Mixed-effects modeling analyses for the behavioral data were performed using R, using packages lme4 and lmerTest.

Behavioral Data Analysis. In order to determine whether accuracy improved as a function of exposure, the proportion of accurately identified trials in a sliding window (20 exposures) was compared to the corresponding window for permuted participant responses (n = 1,000). For participants who performed pretraining passive listening (n = 6), calculation of tone-specific exposure for each trial included passive listening blocks, but assessment of learning only included trials from active training blocks. A P value was calculated in each window, then FDR-corrected at alpha = 0.05. The significance testing outcome was treated as a binomial dependent variable. A mixed-effects modeling analysis was performed with exposure as the fixed effect. Additional analyses were performed by also including the tone identity and its interaction effect with exposure as fixed effects. The model was corrected for random intercepts for syllable and speaker. All behavioral findings reported in the main text are based on recording blocks in which the participants identified stimuli among all four Mandarin tones (P1 through 4, 6, and 8).

Individual Electrode Encoding Analysis.

Electrode selection. We identified sound-sensitive electrodes for which there was a significant increase in mean HGA (70 to 150 Hz) across all tones following stimulus onset (paired t test; FDR–P < 0.05). HGA was averaged for 500-ms periods immediately preceding and following the stimulus onset. For each electrode, a paired t test was conducted across trials between the pre- and poststimulus onset period. For an electrode to qualify as a sound-sensitive electrode, T > 0 and FDR–P < 0.05. All subsequent analyses were subset to a [0, 1] s post–stimulus-onset time period from these electrodes.

Model-fitting approach. We modeled HGA using a time-dependent linear regression analysis for all sound-sensitive electrodes. The fully specified encoding model was constructed as follows:

$$\begin{aligned}\mathrm{HGA}_{\mathrm{Time,Electrode}} \sim\ & \mathrm{Tone{:}Exposure{:}Accuracy} + \mathrm{Tone{:}Exposure} \\ & + \mathrm{Tone{:}Accuracy} + \mathrm{Tone} + \mathrm{Exposure} + \mathrm{Accuracy} + \mathrm{Syllable} + \mathrm{Speaker}\ , \\ & + \mathrm{Training\ Stage} + (\mathrm{Intercept})\end{aligned}$$

where the Tone:Exposure:Accuracy interaction models the learning effect, the Tone:Exposure interaction models the exposure effect, the Tone:Accuracy interaction models the accuracy effect, and the Tone simple effect models the stable-tone encoding effect. The interaction terms included in the encoding model target stimulus-specific task-related effects; therefore, we did not include the nonspecific Exposure:Accuracy term. The regression analysis was performed only for the data collected during active training blocks. Categorical variables were dummy coded as binary regressors such that T1; T2; T3; T4 was coded as 1 0 0 0; 0 1 0 0; 0 0 1 0; 0 0 0 1. Exposure was defined as the number of preceding trials with the same tone (across speakers and syllables). L2 regularization (ridge) was performed to minimize the effects of correlated variables. All models were fourfold cross-validated using randomized selection, for which 80% of the data were used to train the model, the ridge parameter (alpha) was optimized based on the first remaining 10%, and the performance of the model was tested on the remaining 10%.

Variance partitioning and significance testing. For each electrode, we first calculated the total variance explained by the fully specified encoding model (R^2_{Full}). Then, unique variance explained by each effect of interest (ΔR^2_{Unique}) was calculated as the following:

$$\Delta R^2_{Unique} = R^2_{Full} - R^2_{Control},$$

where $R^2_{Control}$ corresponds to the variance explained by a model that excludes the effect of interest. The significance of each model was evaluated using permutation testing (n = 1,000). The resulting P values were FDR corrected across all time points from all electrodes for the fully specified encoding model. Only contiguous significant time points were included. Significance testing for the variance-partitioned models was FDR corrected within these significant time points and electrodes. The predominant effect type for each electrode was determined by calculating the maximum among significant values of $\Delta R^2_{Learning}$, $\Delta R^2_{Exposure}$, $\Delta R^2_{Accuracy}$, and $\Delta R^2_{Stable\text{-}tone\ encoding}$. Effect size was determined as follows:

$$\max(\Delta R^2_{Unique}) \,/\, \max(\Delta R^2_{Full}).$$

These electrode-specific effect sizes were compared using linear mixed-effects modeling, corrected for the random effects of participants and electrodes. For the purposes of characterizing the direction of each effect at the level of individual electrodes, tone-specific beta weights for each effect were averaged within significant time periods for each effect. Only the data from the recording blocks in which all four tones were presented were used (P1 through 4, 6, and 8). For P4, 6, and 8, separate regression analyses were performed within these recording blocks to yield valid estimates of regression weights. Maximum absolute value was used to choose the best-characterizing tone for each combination of electrode and effect. The direction of the effect was determined per the sign of the averaged beta value for the tone. χ^2 tests were performed to reject the null hypotheses that, for each effect, positive and negative directions and all four tones were equally likely to occur across electrodes.

Neural Variability Analysis.

Mean and variance calculation. A peak time period was determined using the full-width half-maximum of the average response across all sound-sensitive electrodes from all participants (0.11 to 0.53 s post-stimulus onset). The stimulus-specific mean and variance for HGA was calculated for each electrode within this peak time period.

Variance-to-mean slope of regression. Once the mean and variance were calculated for all unique stimuli, the variance-to-mean slope (D) was calculated using the following formula:

$$\sigma^2 = D\mu + C,$$

where σ^2 refers to variance, μ to mean, and C the intercept, which was ignored (54). Conceptually, D is equivalent to the mean-normalized variance (Fano factor), which is used to characterize neural variability when zero variance is assumed for zero mean activity, such as in spike count data. Comparison of D across learning and exposure electrodes was performed using linear mixed-effects modeling, corrected for the random effect of participants.

Pretraining Passive Listening.

Stimulus space for pitch height and pitch change features. Time-varying pitch patterns for the Mandarin tone stimuli were estimated as F_0 measurements (40 to 400 Hz) using Praat. To avoid edge effects, each sound file was padded with 500 ms of silence before the onset and after the offset. Resulting F_0 estimates were manually corrected and log transformed. Pitch height was calculated by Z-scoring the raw values within each speaker and averaging all valid values. Pitch change was calculated by first computing the difference in log absolute F_0 value in Hz in neighboring time points, dividing this value by the time differential, then finally smoothing these values across 100 ms, corresponding to 16 time points for F_0 estimation.

Tone separability. For each electrode, a logistic regression model was trained to classify the four tones using HGA from the peak time period (110 to 530 ms poststimulus onset) as features, leave-one-out cross-validated across different syllables. To compare relative separability across different populations (e.g., learning versus exposure), the posterior probability from the classifier was used to calculate the AUC for each tone, which was then averaged across the four tones, resulting in a single value for a given electrode. Linear mixed-effects modeling was used to compare AUC across learning and exposure populations, corrected for random effects of participants.

Pitch representation. Linear discriminant analysis (LDA; number of components: 2) was performed within the peak time period per the procedures outlined above (e.g., see *Tone Separability*). Cross-validation was not performed because the goal for this analysis was to test the hypothesis that a low-dimensional representation of the neural data resembles the acoustic pitch features space of the stimuli. To obtain the average neural space, the resulting LDA-transformed values were averaged for each unique stimulus. Pairwise Euclidean distance was calculated within the two-dimensional LD space. Linear regression was performed between the neural distances and the acoustic distances as defined by pitch height and pitch change features to calculate the variance explained (R^2) for each acoustic feature. Significance testing was performed by permuting the average LD values across stimuli prior to the calculation of Euclidean distances (n = 1,000). Linear mixed-effects modeling was used to compare R^2 for either feature across learning and exposure populations and to compare R^2 for both features within each of the learning and exposure populations. All models were corrected for random effects of participants.

Classification analysis. To test the hypothesis that the type of behavioral effect encoded by each electrode during training can be predicted based on neural activity recorded prior to training, logistic regression was performed with the tone separability (AUC) measure and pitch representation (R^2_{Height}; R^2_{Change}) values. Logistic regression was fourfold cross-validated. In the first analysis, only the learning and exposure electrodes were used to test the hypothesis that the two sets of populations can be distinguished from one another. In the second analysis, all sound-sensitive electrodes were used to test the hypothesis that each set of populations can be detected without any a priori information. The posterior probability for each electrode in the testing set was used to calculate the AUC as well as to estimate the average receiver operating characteristic (ROC) curve. The mean ROC curve was visualized via linear interpolation of false positive rates across folds. Significance testing was performed by permuting the electrode labels prior to logistic regression (n = 1,000).

Data Availability. The data that support the findings of this study are available on request from the corresponding author.

ACKNOWLEDGMENTS. We thank Alia Shafi and Ben Speidel and the rest of the Chang Lab at UCSF. This work was supported by the Defence Advanced Research Projects Agency contract no. N66001-17-2-4008 (M.K.L.), NIH R01-DC012379 (E.F.C.), and NIH R01-DC0155004 (B.C.). E.F.C. is a New York Stem Cell Foundation Robertson Investigator. This research was also supported by the New York Stem Cell Foundation, the HHMI, the McKnight Foundation, the Shurl and Kay Curci Foundation, and the William K. Bowes Foundation.

1. R. L. Diehl, A. J. Lotto, L. L. Holt, Speech perception. *Annu. Rev. Psychol.* **55**, 149–179 (2004).
2. L. L. Holt, A. J. Lotto, Speech perception as categorization. *Atten. Percept. Psychophys.* **72**, 1218–1227 (2010).
3. M. Cheour *et al*., Development of language-specific phoneme representations in the infant brain. *Nat. Neurosci.* **1**, 351–353 (1998).
4. A. J. Doupe, P. K. Kuhl, Birdsong and human speech: Common themes and mechanisms. *Annu. Rev. Neurosci.* **22**, 567–631 (1999).
5. G. K. Vallabha, J. L. McClelland, F. Pons, J. F. Werker, S. Amano, Unsupervised learning of vowel categories from infant-directed speech. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 13273–13278 (2007).
6. P. Iverson *et al*., A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* **87**, B47–B57 (2003).
7. J. S. Johnson, E. L. Newport, Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognit. Psychol.* **21**, 60–99 (1989).
8. J. S. Logan, S. E. Lively, D. B. Pisoni, Training Japanese listeners to identify English /r/ and /l/: A first report. *J. Acoust. Soc. Am.* **89**, 874–886 (1991).
9. J. F. Werker, R. C. Tees, Phonemic and phonetic factors in adult cross-language speech perception. *J. Acoust. Soc. Am.* **75**, 1866–1878 (1984).
10. A. Cutler, D. Dahan, W. van Donselaar, Prosody in the comprehension of spoken language: A literature review. *Lang. Speech* **40**, 141–201 (1997).
11. S. Shattuck-Hufnagel, A. E. Turk, A prosody tutorial for investigators of auditory sentence processing. *J. Psycholinguist. Res.* **25**, 193–247 (1996).
12. Y. Wang, M. M. Spence, A. Jongman, J. A. Sereno, Training American listeners to perceive Mandarin tones. *J. Acoust. Soc. Am.* **106**, 3649–3658 (1999).
13. Y. Wang, A. Jongman, J. A. Sereno, Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *J. Acoust. Soc. Am.* **113**, 1033–1043 (2003).
14. C. T. Best, G. W. McRoberts, N. M. Sithole, Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *J. Exp. Psychol. Hum. Percept. Perform.* **14**, 345–360 (1988).
15. J. F. Werker, T. K. Hensch, Critical periods in speech perception: New directions. *Annu. Rev. Psychol.* **66**, 173–196 (2015).
16. P. K. Kuhl, Early language acquisition: Cracking the speech code. *Nat. Rev. Neurosci.* **5**, 831–843 (2004).
17. E. B. Myers, Emergence of category-level sensitivities in non-native speech sound learning. *Front. Neurosci.* **8**, 238 (2014).
18. P. C. Wong, T. K. Perrachione, T. B. Parrish, Neural characteristics of successful and less successful speech and word learning in adults. *Hum. Brain Mapp.* **28**, 995–1006 (2007).
19. Z. Qi *et al*., Speech processing and plasticity in the right hemisphere predict variation in adult foreign language learning. *Neuroimage* **192**, 76–87 (2019).
20. A. L. Francis, V. Ciocca, L. Ma, K. Fenn, Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *J. Phon.* **36**, 268–294 (2008).
21. C. Pallier, L. Bosch, N. Sebastián-Gallés, A limit on behavioral plasticity in speech perception. *Cognition* **64**, B9–B17 (1997).
22. R. Reetzke, Z. Xie, F. Llanos, B. Chandrasekaran, Tracing the trajectory of sensory plasticity across different stages of speech learning in adulthood. *Curr. Biol.* **28**, 1419–1427.e4 (2018).
23. P. Fuhrmeister, E. B. Myers, Desirable and undesirable difficulties: Influences of variability, training schedule, and aptitude on nonnative phonetic learning. *Atten. Percept. Psychophys.* **82**, 2049–2065 (2020).
24. N. Golestani, R. J. Zatorre, Individual differences in the acquisition of second language phonology. *Brain Lang.* **109**, 55–67 (2009).
25. K. Swan, E. Myers, Category labels induce boundary-dependent perceptual warping in learned speech categories. *Second Lang. Res.* **29**, 391–411 (2013).
26. A. R. Bradlow, D. B. Pisoni, R. Akahane-Yamada, Y. Tohkura, Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *J. Acoust. Soc. Am.* **101**, 2299–2310 (1997).
27. F. H. Guenther, F. T. Husain, M. A. Cohen, B. G. Shinn-Cunningham, Effects of categorization and discrimination training on auditory perceptual space. *J. Acoust. Soc. Am.* **106**, 2900–2912 (1999).
28. D. E. Callan *et al*., Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *Neuroimage* **19**, 113–124 (2003).
29. R. Desai, E. Liebenthal, E. Waldron, J. R. Binder, Left posterior temporal regions are sensitive to auditory categorization. *J. Cogn. Neurosci.* **20**, 1174–1188 (2008).
30. N. Golestani, R. J. Zatorre, Learning new sounds of speech: Reallocation of neural substrates. *Neuroimage* **21**, 494–506 (2004).
31. G. Feng, H. G. Yi, B. Chandrasekaran, The role of the human auditory corticostriatal network in speech learning. *Cereb. Cortex* **29**, 4077–4089 (2019).
32. S. Luthra *et al*., Brain-behavior relationships in incidental learning of non-native phonetic categories. *Brain Lang.* **198**, 104692 (2019).
33. E. B. Myers, K. Swan, Effects of category learning on neural sensitivity to non-native phonetic categories. *J. Cogn. Neurosci.* **24**, 1695–1708 (2012).
34. X. Jiang, M. A. Chevillet, J. P. Rauschecker, M. Riesenhuber, Training humans to categorize monkey calls: Auditory feature-and category-selective neural tuning changes. *Neuron* **98**, 405–416.e4 (2018).
35. A. Ley *et al*., Learning of new sound categories shapes neural response patterns in human auditory cortex. *J. Neurosci.* **32**, 13273–13280 (2012).
36. S.-J. Lim, J. A. Fiez, L. L. Holt, Role of the striatum in incidental learning of sound categories. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 4671–4680 (2019).
37. H. G. Yi, M. K. Leonard, E. F. Chang, The encoding of speech sounds in the superior temporal gyrus. *Neuron* **102**, 1096–1110 (2019).
38. B. Chandrasekaran, H.-G. Yi, N. J. Blanco, J. E. McGeary, W. T. Maddox, Enhanced procedural learning of speech sound categories in a genetic variant of FOXP2. *J. Neurosci.* **35**, 7808–7812 (2015).
39. C. Tang, L. S. Hamilton, E. F. Chang, Intonational speech prosody encoding in the human auditory cortex. *Science* **357**, 797–801 (2017).
40. F. Llanos *et al*., Non-invasive peripheral nerve stimulation selectively enhances speech category learning in adults. *NPJ Sci. Learn.* **5**, 12 (2020).
41. S. Ray, J. H. Maunsell, Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biol.* **9**, e1000610 (2011).
42. S. Ray, N. E. Crone, E. Niebur, P. J. Franaszczuk, S. S. Hsiao, Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *J. Neurosci.* **28**, 11526–11536 (2008).
43. V. L. Towle *et al*., ECoG gamma activity during a language task: Differentiating expressive and receptive speech areas. *Brain* **131**, 2013–2027 (2008).
44. K. J. Forseth, G. Hickok, P. S. Rollo, N. Tandon, Language prediction mechanisms in human auditory cortex. *Nat. Commun.* **11**, 5240 (2020).
45. N. Mesgarani, C. Cheung, K. Johnson, E. F. Chang, Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).
46. T. E. Nichols, A. P. Holmes, Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).
47. R. Leech, L. L. Holt, J. T. Devlin, F. Dick, Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *J. Neurosci.* **29**, 5234–5239 (2009).
48. E. Tricomi, M. R. Delgado, B. D. McCandliss, J. L. McClelland, J. A. Fiez, Performance feedback drives caudate activation in a phonological learning task. *J. Cogn. Neurosci.* **18**, 1029–1043 (2006).
49. S. Panzeri, C. D. Harvey, E. Piasini, P. E. Latham, T. Fellin, Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. *Neuron* **93**, 491–507 (2017).
50. M. M. Churchland *et al*., Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nat. Neurosci.* **13**, 369–378 (2010).
51. A. M. Ni, D. A. Ruff, J. J. Alberts, J. Symmonds, M. R. Cohen, Learning and attention reveal a general relationship between population activity and behavior. *Science* **359**, 463–465 (2018).
52. D. J. Tolhurst, J. A. Movshon, A. F. Dean, The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Res.* **23**, 775–785 (1983).
53. R. Vogels, W. Spileers, G. A. Orban, The response variability of striate cortical neurons in the behaving monkey. *Exp. Brain Res.* **77**, 432–436 (1989).
54. B. K. Dichter, K. E. Bouchard, E. F. Chang, Dynamic structure of neural variability in the cortical representation of speech sounds. *J. Neurosci.* **36**, 7453–7463 (2016).
55. S.-J. Lim, J. A. Fiez, L. L. Holt, How may the basal ganglia contribute to auditory categorization and speech perception? *Front. Neurosci.* **8**, 230 (2014).
56. T. K. Perrachione, J. Lee, L. Y. Ha, P. C. Wong, Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *J. Acoust. Soc. Am.* **130**, 461–472 (2011).
57. N. A. Francis, D. Elgueda, B. Englitz, J. B. Fritz, S. A. Shamma, Laminar profile of task-related plasticity in ferret primary auditory cortex. *Sci. Rep.* **8**, 16375 (2018).
58. J. Fritz, S. Shamma, M. Elhilali, D. Klein, Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci.* **6**, 1216–1223 (2003).

59. D. B. Polley, E. E. Steinberg, M. M. Merzenich, Perceptual learning directs auditory cortical map reorganization through top-down influences. *J. Neurosci.* **26**, 4970–4982 (2006).
60. A. M. LeMessurier, D. E. Feldman, Plasticity of population coding in primary sensory cortex. *Curr. Opin. Neurobiol.* **53**, 50–56 (2018).
61. H. Makino, E. J. Hwang, N. G. Hedrick, T. Komiyama, Circuit mechanisms of sensorimotor learning. *Neuron* **92**, 705–721 (2016).
62. M. Sanayei *et al*., Perceptual learning of fine contrast discrimination changes neuronal tuning and population coding in macaque V4. *Nat. Commun.* **9**, 4238 (2018).
63. C.-T. Law, J. I. Gold, Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nat. Neurosci.* **12**, 655–663 (2009).
64. T. Watanabe, Y. Sasaki, Perceptual learning: Toward a comprehensive theory. *Annu. Rev. Psychol.* **66**, 197–221 (2015).
65. J. S. Arsenault, B. R. Buchsbaum, Distributed neural representations of phonological features during speech perception. *J. Neurosci.* **35**, 634–642 (2015).
66. W. A. de Heer, A. G. Huth, T. L. Griffiths, J. L. Gallant, F. E. Theunissen, The hierarchical cortical organization of human speech processing. *J. Neurosci.* **37**, 6539–6557 (2017).
67. S. Evans, M. H. Davis, Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis. *Cereb. Cortex* **25**, 4772–4788 (2015).
68. E. Formisano, F. De Martino, M. Bonte, R. Goebel, "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* **322**, 970–973 (2008).
69. N. M. Weinberger, R. Javid, B. Lepan, Long-term retention of learning-induced receptive-field plasticity in the auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 2394–2398 (1993).
70. Y. Li, C. Tang, J. Lu, J. Wu, E. F. Chang, Human cortical encoding of pitch in tonal and non-tonal languages. *Nat. Commun.* **12**, 1161 (2021).
71. A. M. Chan *et al*., Speech-specific tuning of neurons in human superior temporal gyrus. *Cereb. Cortex* **24**, 2679–2693 (2014).
72. B. Khalighinejad, G. Cruzatto da Silva, N. Mesgarani, Dynamic encoding of acoustic features in neural responses to continuous speech. *J. Neurosci.* **37**, 2176–2185 (2017).
73. M. Steinschneider *et al*., Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cereb. Cortex* **21**, 2332–2347 (2011).
74. T. J. Sejnowski, P. S. Churchland, J. A. Movshon, Putting big data to good use in neuroscience. *Nat. Neurosci.* **17**, 1440–1441 (2014).
75. A. K. Churchland *et al*., Variance as a signature of neural computations during decision making. *Neuron* **69**, 818–831 (2011).
76. Y. Gu *et al*., Perceptual learning reduces interneuronal correlations in macaque visual cortex. *Neuron* **71**, 750–761 (2011).
77. A. K. Dhawale, M. A. Smith, B. P. Ölveczky, The role of variability in motor learning. *Annu. Rev. Neurosci.* **40**, 479–498 (2017).
78. E. Y. Walker, R. J. Cotton, W. J. Ma, A. S. Tolias, A neural basis of probabilistic computation in visual cortex. *Nat. Neurosci.* **23**, 122–129 (2020).
79. E. S. Teoh, M. S. Cappelloni, E. C. Lalor, Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. *Eur. J. Neurosci.* **50**, 3831–3842 (2019).
80. Y. R. Chao, *Mandarin Primer: An Intensive Course in Spoken Chinese* (Harvard University Press, 1948).
81. J. Gandour, Tone dissimilarity judgments by Chinese listeners/声调异同辨别的测试. *J. Chin. Linguist.* **12**, 235–261 (1984).
82. D. W. Massaro, M. M. Cohen, C. Tseng, The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *J. Chin. Linguist.* **13**, 267–289 (1985).
83. M. Jazayeri, A. Afraz, Navigating the neural space in search of the neural code. *Neuron* **93**, 1003–1014 (2017).
84. W. J. Ma, M. Jazayeri, Neural coding of uncertainty and probability. *Annu. Rev. Neurosci.* **37**, 205–220 (2014).
85. D. A. Ruff, A. M. Ni, M. R. Cohen, Cognition as a window into neuronal population space. *Annu. Rev. Neurosci.* **41**, 77–97 (2018).
86. M. McCloskey, N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem" in *Psychology of Learning and Motivation*, G. H. Bower, Ed. (Elsevier, 1989), pp. 109–165.
87. M. Mermillod, A. Bugaiska, P. Bonin, The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Front. Psychol.* **4**, 504 (2013).
88. F. M. Richardson, M. S. Thomas, Critical periods and catastrophic interference effects in the development of self-organizing feature maps. *Dev. Sci.* **11**, 371–389 (2008).
89. L. L. Holt, A. J. Lotto, Cue weighting in auditory categorization: Implications for first and second language acquisition. *J. Acoust. Soc. Am.* **119**, 3059–3071 (2006).
90. L. S. Hamilton, D. L. Chang, M. B. Lee, E. F. Chang, Semi-automated anatomical labeling and inter-subject warping of high-density intracranial recording electrodes in electrocorticography. *Front. Neuroinform.* **11**, 62 (2017).
91. P. C. Wong, T. K. Perrachione, Learning pitch patterns in lexical identification by native English-speaking adults. *Appl. Psycholinguist.* **28**, 565–585 (2007).