HHS Public Access

Author manuscript

Neurobiol Lang (Camb). Author manuscript; available in PMC 2022 January 01.

Published in final edited form as:

Neurobiol Lang (Camb). 2021 ; 2(2): 280–307. doi:10.1162/nol_a_00035.

Emerging native-similar neural representations underlie non-native speech category learning success

Gangyi Feng[1,2,*], **Yu Li**[1,2], **Shen-Mou Hsu**[3], **Patrick C.M. Wong**[1,2], **Tai-Li Chou**[3,4], **Bharath Chandrasekaran**[5,*]

[1]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China.

[2]Brain and Mind Institute, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China.

[3]Imaging Center for Integrated Body, Mind and Culture Research, National Taiwan University, Taipei 10617, Taiwan.

[4]Department of Psychology, National Taiwan University, Taipei 10617, Taiwan.

[5]Department of Communication Sciences and Disorders, School of Health and Rehabilitation Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA.

Abstract

Learning non-native phonetic categories in adulthood is an exceptionally challenging task, characterized by large inter-individual differences in learning speed and outcomes. The neurobiological mechanisms underlying the inter-individual differences in the learning efficacy are not fully understood. Here we examined the extent to which training-induced neural representations of non-native Mandarin tone categories in English listeners (n = 53) are increasingly similar to those of the native listeners (n = 33) who acquired these categories early in infancy. We particularly assessed whether the neural similarities in representational structure between non-native learners and native listeners are robust neuromarkers of inter-individual differences in learning success. Using inter-subject neural representational similarity (IS-NRS) analysis and predictive modeling on two functional magnetic resonance imaging (fMRI) datasets, we examined the neural representational mechanisms underlying speech category learning success. Learners' neural representations that were significantly similar to the native listeners emerged in brain regions mediating speech perception following training; the extent of the emerging neural similarities with native listeners significantly predicted the learning speed and outcome in learners. The predictive power of IS-NRS outperformed models with other neural representational measures. Furthermore, neural representations underlying successful learning are multidimensional but cost-efficient in nature. The degree of the emergent native-similar neural

[*]Corresponding authors: Gangyi Feng, Ph.D., Brain and Mind Institute, Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China, +852-3943 3190, g.feng@cuhk.edu.hk, Bharath Chandrasekaran, Ph.D., Department of Communication Science and Disorders, University of Pittsburgh 6074 Forbes Tower, Pittsburgh, PA 15260, (412) 383-6565, b.chandra@pitt.edu.

Conflict of interest statement

Patrick C. M. Wong is a founder of a company in Hong Kong supported by a Hong Kong SAR government startup scheme for universities.

representations was closely related to the robust neural sensitivity to feedback in the frontostriatal network. These findings provide important insights on experience-dependent representational neuroplasticity underlying successful speech learning in adulthood and could be leveraged in designing individualized feedback-based training paradigms that maximize learning efficiency.

Keywords

individual differences; non-native speech learning; Mandarin tone category; predictive modeling; multivariate representation; neural feedback sensitivity

Introduction

During infancy, dramatic changes occur in the brain networks that support speech processing (Kuhl, 2004, 2010). Language-general perception narrows to become more selective to the statistical regularities of the native environment (Cheour et al., 1998; Garcia-Lazaro et al., 2011; Kuhl, 2004) promoting greater sensitivity to native speech sound categories (Kuhl, 2010; Nakahara et al., 2004; Vallabha et al., 2007). However, experience-dependent perceptual narrowing can also alter low-level perception and interfere in the acquisition of non-native speech categories in adulthood (Kuhl et al., 2008; Myers, 2014). Non-native speech categories can be acquired to native-like proficiency in adulthood when learners are provided some amount of feedback and with sufficient intensity of training (Lively et al., 1993; Reetzke et al., 2018). However, even in adults with similar language backgrounds, cognitive, socio-economic, motivation, and hearing status undergoing identical training paradigms, large inter-individual differences define speech learning performance (Ellis, 2004). Indeed, individual differences are ubiquitous in the acquisition of most sub-components of language (Kidd & Donnelly, 2020; Kidd et al., 2018). This is especially the case when adults with no tonal language experience learn to categorize acoustically-similar but lexical-relevant tone patterns (Chandrasekaran et al., 2010; Wong & Perrachione, 2007). Our goal here is to elucidate the neural mechanisms that underlie the extensive inter-individual variability in the non-native tone-category learning success. We examine the following questions: first, are the emerging neural representations of linguistic-tone categories in the successful adult learners fundamentally similar or dissimilar to the neural representations that are acquired in early infancy? That is, is the similarity in the neural representations between adult learners and native listeners a robust neural neuromarkers of learning success? Second, is the feedback sensitivity in the corticostriatal systems a key indicator of individual differences in learning success and the degree of the putative ‘nativeness’ of neural representations?

These questions relate to theoretical positions adopted in the domain of second language (L2) acquisition to explain individual variability in attainments. Much of the focus in this literature is on the learning of grammatical structures; however, this literature provides a theoretical scaffolding for learning non-native phonology. The shallow structure hypothesis posits that the representations underlying L2 acquisition have less detail relative to those underlying native language (L1) acquisition (Clahsen et al. 2006a, 2006b, 2018; also see Ullman, 2006). The fundamental difference hypothesis posits a lack of convergence between

non-native and native language representations and proposes that L2 learning necessitates relying on domain-general learning mechanisms, such as executive control functions and feedback processes (Bley-Vroman, 1990, 2009). These theoretical perspectives not only explain the differences between L1 and L2 acquisition and representation but also implicate the comparison between the representations of non-native learners' L2 and native speakers' L1 could potentially reflect the nativeness in L2 processing and attainment for the learners (Birdsong, 2018; Hartshorne et al., 2018). For example, using fMRI with traditional univariate activation-based analysis approaches, quantitative differences and similarities in brain activations have been found between L1 and L2 processes where the degree of similarity is dependent on the level of L2 proficiency and age of acquisition (Abutalebi, 2008; Chee et al., 1999; Feng et al., 2015; Perani & Abutalebi, 2005; Perani et al., 1996). Moving beyond the activation-based group-level comparisons between L1 and L2, here we focus on examining the multivariate representation-based neural mechanisms underlying the inter-individual variability in the acquisition of a new phonological structure not present in English—lexical tones.

In tone languages, pitch information plays a similar role as voice onset time and duration in altering word meaning (Yip, 2012). For native speakers of tone languages, extracting pitch patterns from the incoming auditory stream and mapping key pitch features to tone categories are critical for speech communication. In contrast, non-native listeners who do not have tonal language experience have great difficulty in discriminating tonal contrasts with similar pitch patterns (Reid et al., 2015; So & Best, 2010). This discrimination difficulty may be mainly due to the fact that tonal information is not linguistically relevant to non-native listeners' phonological systems (Best, 1995; Reid et al., 2015). The challenges in discrimination and learning have also been viewed from the perspective of feature-weighting (Chandrasekaran, Gandour, et al., 2007; Chandrasekaran et al., 2010). For contour-tone languages like Mandarin, at least two pitch-related dimensions define tone categories: pitch height and slope; both native and non-native listeners weight pitch height significantly during categorization. In contrast, the weighting of pitch slope or the combination of pitch height and direction (i.e., contour: time-varying height) is strongly modulated by language experience, with native listeners weighting this dimension more heavily than non-native listeners (Chandrasekaran, Krishnan, et al., 2007). The increasing weighting of pitch contour allows for a more stable mapping of tone categories across contexts and talkers with varying fundamental frequencies like native listeners, which could increase the probability of successful categorization and learning.

Successful learners must establish new representations of novel tone categories by mapping highly time-varying pitch patterns to stable tone categories (Feng et al., 2019), a key step for learning the lexicon. To achieve this, learners would likely need to update their internally emerging representations with an increased weighting of key dimensions that underlie the native tone perceptual space. Here, we assess the extent to which the emerging neural representations of tone categories and the underlying dimensions, acquired in adulthood are similar to the representations in native Mandarin listeners who use tones linguistically. We specifically compare the detailed representational structure (including all tonal contrasts under different syllable contexts) in the brain with a hypothesis that more successful learners

demonstrate emerging representational patterns that are more similar to those in native listeners.

Finally, we test the hypothesis that the corticostriatal learning systems that are sensitive to feedback valence are critical neural driving sources of individual differences in learning efficacy and the training-related neural representational plasticity in adulthood. An emerging perspective of the neural functionality of feedback is that adult learners require feedback, processed by the corticostriatal networks to build novel speech category (i.e., sound-to-category) representations (Chandrasekaran, Yi, et al., 2014; Feng et al., 2019; Lim et al., 2019; Yi et al., 2016). Per the dual-learning systems (DLS) model (Chandrasekaran, Koslov, et al., 2014; Chandrasekaran, Yi, et al., 2014), a reflective (sound-to-rule mapping) system and a reflexive (sound-to-reward mapping) system operate on a trial-by-trial basis to assist sound-to-category learning. The reflective system involves the frontoparietal attentional network and the hippocampus, which operates by generating and testing hypotheses based on corrective feedback; the reflexive system, on the other hand, involves the striatum in mapping stimuli to motor responses that result in rewards. This DLS model focuses on speech category learning in adulthood. The corticostriatal systems that subserve category learning in the DLS model may be ubiquitous to the acquisition of other language sub-components for adult learners. Indeed, previous studies have proposed comparable cortical-subcortical systems that drive reward-dependent acquisition of language components, e.g., word learning (Ripolles et al., 2016; Ripolles et al., 2014).

To test the two hypotheses, we analyzed data from a tone-category learning experiment that leveraged previously collected functional magnetic resonance imaging (fMRI) data to assess emerging representations in English-speaking learners ($n = 53$) as they learned to categorize non-native tone categories with feedback (Feng et al., 2019; Yi et al., 2016). To quantify the degree of the nativeness in neural representational structure for each learner, we conducted a new fMRI study in which a group of native Mandarin speakers ($n = 33$) performed the same tone categorization task with the same set of stimuli as the non-native learners but without feedback. The behavior response patterns of the learners were modeled with representational models that are informed by the acoustics and non-acoustic category-related features as well as the neural representational patterns from the native Mandarin speakers to examine the degree of emerging nativeness in representations. Importantly, using inter-subject neural representational similarity (IS-NRS) analysis (Chen et al., 2017; Diedrichsen & Kriegeskorte, 2017), we measured the extent of shared patterns in neural representational structure between learners and native listeners (i.e., IS-NRS). To test our hypotheses, we then used a predictive modeling approach (Gabrieli et al., 2015; Rosenberg et al., 2016) with learners' IS-NRSs as neural predictors to predict their behavioral learning efficacy (i.e., speed and outcome). To further evaluate the predictability of IS-NRS, we compared the predictive power of IS-NRS with those of models with other neural representational measures. To assess the detailed representational structure underlying successful learning, we combined the predictive analytics and a data-driven single vector decomposition procedure that estimated the relationship between dimensionality and learning efficacy. Finally, to evaluate the underlying driving factors of the inter-individual variability in learning success and emerging representations, we calculated a neural feedback sensitivity

index to predict learning speed and outcome as well as the degree of learners' nativeness in neural representations (i.e., IS-NRS).

Materials and Methods

Participants

Native speakers of Mandarin (n = 33; 18 females; right-handed; age range, 20–37 years; mean age = 25.5 years) were recruited from the communities around the National Taiwan University, Taipei. These native participants were highly proficient at listening and speaking in standard Mandarin. They were recruited to participate in a tone categorization fMRI experiment designed specifically for the current study. This experiment was approved by the Research Ethics Committee at National Taiwan University. Native speakers of English were recruited from the communities around The University of Texas at Austin (n = 53; 39 females; right-handed; age range, 18–35 years; mean age = 21.8 years). These English-speaking participants did not have tonal language experience and had minimal formal music training experience (< 3 years). All the participants reported normal hearing ability which was confirmed by audiological testing (pure tone thresholds < 25 dB HL at 1, 2, and 4 kHz). They had normal or corrected-to-normal visions and did not have any neurological impairments. Training protocols and materials were approved by the Institutional Review Board of The University of Texas at Austin. All participants provided written informed consent and were monetarily compensated for their time.

Stimulus

Natural exemplars (n = 40) of the four Mandarin tones (T1: high-flat; T2: low-rising; T3: low-dipping; T4: high-falling) were generated by two native Mandarin speakers (originally from Beijing; 1 female) in the context of five monosyllabic Mandarin Chinese words (/bu/, /di/, /lu/, /ma/, and /mi/) (see spectrograms of sample stimuli in Figure S1A, Supplementary Materials). These syllables were chosen because they also exist in the American English syllabic inventory. Therefore, the neural representations of native and non-native speech categories can be examined for the learners and compared with the native Mandarin speakers. The stimuli were normalized for an RMS amplitude of 70 dB and a duration of 442 ms (Perrachione et al., 2011). Both learners and native speakers heard the same set of stimuli during the experiments.

Experimental procedure

In the native tone-categorization fMRI experiment, Mandarin-speaking participants were required to categorize sounds into one of the four categories during scanning by pressing the "1", "2", "3", or "4" buttons using an in-scanner response box, with category-response mapping counterbalanced across participants. Native participants were not provided feedback following categorization responses. They briefly practiced categorization before scanning to establish category-response mapping. To reduce the interference of scanner noise to speech perception, we employed a customized sparse-sampling imaging sequence with an 800-ms silence gap between every two consecutive imaging acquisitions (Figure S1B, Supplementary Materials). Each sound was presented (duration = 442 ms) within the silence gap 100-ms after each imaging acquisition. Each sound was presented once in each

block and the order of stimuli randomly varied across blocks. To better estimate the hemodynamic responses for each trial, we added 20 null trials (i.e., silence, duration = 5 s) randomly between sound trials as jittered inter-trial intervals in each block. To accurately estimate the activation patterns of the sounds, native participants completed at least five blocks of 40 trials each of tone categorization (six participants completed five blocks [200 trials], and 27 completed six blocks [240 trials]). Each sound (e.g., /bu4/; collapsed across talkers) was repeated 10 to 12 times. The significant number of repetitions for the same item ensures a sufficient signal-to-noise ratio and accurate activation estimation.

The non-native sound-to-category training procedure has been extensively described in previous studies (Feng et al., 2019; Yi et al., 2016). Briefly, English-speaking participants performed a tone categorization task during scanning, in which they were required to learn to map the sounds onto four categories. The fMRI experiment consisted of six contiguous training blocks of 40 trials each. In each block, each trial started with a fixation cross and the auditory stimulus was presented for 442 ms. Participants were required to make a categorization response within two seconds. Following the stimulus presentation and categorization response, corrective feedback (i.e., “RIGHT” or “WRONG”) was displayed for 750 ms (see Figure 1A). If the participant did not respond within the two seconds, the response did not record and warning feedback was presented (i.e., “TIME”). To effectively model brain signals for stimulus and feedback presentation separately, we employed a jittered stimulus-feedback interval design (2–4 sec; feedback-stimulus interval = 1–3 sec; pooled from a uniform distribution) (Birn et al., 2002; Dale, 1999; Liu et al., 2001). Each sound stimulus was presented once within each block, with a total of 240 trials in the training experiment.

Imaging acquisition

For the native tone-categorization experiment, all MRI data were acquired using a Siemens 3T Magnetom Prisma MRI system with a 20-channel head coil at Imaging Center for Integrated Body, Mind, and Culture Research, National Taiwan University. Functional images were acquired using a T2*-weighted gradient echo-planar imaging (EPI) pulse sequence [repetition time (TR) = 2,500 ms with 800-ms silence gap, echo time (TE) = 30 ms, flip angle = 90°, 31 slices, field of view (FOV) = 224 × 224 mm^2, in-plane resolution = 3.5 × 3.5 mm^2, slice thickness = 3.5 mm with 1.1 mm gap, Acceleration factor = 2]. T1-weighted high-resolution structural images were acquired using a magnetization prepared rapid acquisition gradient echo (MPRAGE) sequence (192 slices, TR = 2.0 sec, TE = 2.3 ms, flip angle = 8 deg, voxel size = 0.94 × 0.94 × 1 mm^3).

For the sound-to-category training experiment, MRI data were acquired using a Siemens 3T Magnetom Skyra MRI system with a 32-channel head coil at the Biomedical Imaging Center at The University of Texas at Austin. Functional images were obtained using a gradient-echo multi-band EPI pulse sequence (flip angle = 60 deg; TR = 1.8 s; TE = 30 ms; FOV = 250 × 250 mm^2; in-plane resolution = 2 × 2 mm^2; 36 axial slices; slice thickness = 2 mm; distance factor = 50%) using GRAPPA with an acceleration factor of 2. Whole-brain T1-weighted structural images were obtained via MPRAGE sequence (176 slices, TR = 2.53 sec; TE =

3.37 ms; FOV = 250 × 250 mm^2; 256 × 256 matrix; voxel size = 1 × 1 × 1 mm^3; distance factor = 0%).

Behavioral data modeling

Estimation of learning outcome and speed.—The learning outcome is defined as the average tone identification accuracy in the last three blocks. There are three considerations for this learning outcome definition. Firstly, at the group level, learning performance in the last three blocks was relatively stable compared to the first three blocks. That is, tone identification performance was not significantly improved in the last three blocks (Ps > 0.05), which suggests that the last three blocks may be a relatively stable learning phase. Secondly, individual differences in learning outcomes are based on the fact that the amount of training was the same across learners. Therefore, we selected the same number of training blocks for each learner. Thirdly, the division of two training phases (i.e., the first and the last three blocks) ensures that there is enough number of trails for the brain estimation of stimulus items for the multivariate analyses. Based on the above considerations, the last three blocks were defined as the "late phase" of training, and the first three blocks were defined as the "early phase." It is worthy to note that this training phase definition mainly refers to the amount of training that the learners received instead of the proficiency level in a certain phase achieved by individual learners.

To model the non-native tone learners' learning curves properly for the estimation of learning speed, we used four functions (i.e., hyperbolic, logarithmic, power, and linear regression functions) to fit each subject's block-by-block category identification accuracies, separately (Figure 1B). The goodness of fits (GOF) of the modeling with each function were calculated first by estimated the root mean square error (rmse) between a fitting line and the actual learning curve for each learner. The GOF of the three curvilinear functions were then compared with that of the linear function at the group level to examine whether the curvilinear functions are better in capturing the learning progression than that of the linear function. We found that only the GOF of the power function (i.e., $Y = aX^b$; X = training block and Y = tone identification accuracy) was significantly better than the linear function ($t_{(52)} = -4.16$, $P < 0.001$). Parameters a and b from the power function are both associated with the learning progression (i.e., the parameter a represents the steepness of the fitting curve like the slope parameter in the linear function, while the parameter b represents the changes in learning gain based on the same amount of training between training blocks; also see individual fitting curves with the power function in Figure S2, Supplementary Materials). Therefore, we combined the two parameters by multiplying them to represent the learning speed (i.e., LS = $a \times b$; see Figure 1C). The learning speed was significantly correlated with the learning outcome ($r = 0.91$, $P < 0.001$) but it was not significantly correlated with the first block categorization performance ($r = 0.216$, $P = 0.120$). The two learning measures (i.e., learning outcome and speed) share 82% variances. That is, there are around 18% non-overlapped variances that are unique to each learning measure. We hypothesize that these non-overlapped variances may be predicted by different neural sources. Therefore, we used both measures as learning success indices for predictive modeling.

Categorization response-pattern modeling with behavioral representational similarity analysis (bRSA).—We estimated learners' behavioral representational structure during training by using bRSA to model their behavioral confusion/response patterns. The bRSA reveals the model fits (i.e., Spearman's correlations) between predefined representational models (i.e., representational dissimilarity matrices [RDMs]) and the response confusion matrices for each block. We created six RDMs to examine what type of dimensions/information emerge or change following training. These RDMs are 20-by-20 dissimilarity matrices with four tones and five syllables, including dissimilarities between all pairs of tonal contrasts. The dissimilarities were calculated based on different acoustic and non-acoustic information, including native neural activation patterns (Native nRDM), binary tone-category labels (CAT), fundamental frequency (F0) height, F0 slope, and syllable identity (see Figure 1D). The native nRDM was constructed based on the neural activation-pattern dissimilarities between each pair of sound items derived from the native Mandarin speakers within a predefined speech/auditory-perception-related brain mask (see Figure S3, Supplementary Materials). This mask was generated from a meta-analysis in Neurosynth.org by searching the topic dataset with keywords "speech", "auditory", and "perception." The dataset consists of 400 topics extracted with linear discriminant analysis (LDA) from the abstracts of all articles in the Neurosynth database as of July 2018. This automatic meta-analysis included 269 studies (Topic 180) with a list of highly related topic words, including "auditory", "perception", "speech", "non-speech", "sound", "processing", and so on. We used this independent brain mask to avoid any ROI-selection bias. This brain mask was only used for creating native nRDM for bRSA. Including this native nRDM model for the bRSA was to estimate to what extent the learners' behavioral response patterns were similar to the native neural representation patterns. The F0 height RDM was constructed by calculating the acoustic distance between each pair of sounds based on their mean F0 estimates. The F0 slope RDM was constructed by calculating the distance between each pair of sounds based on their F0 slopes (i.e., F0 height changes over time). For the multidimensional (MD) model, we created a two-dimensional space with the F0 height and slope dimensions. Each dimension was normalized before calculating the distance. The Euclid distance between each sound pair within this two-dimensional space was computed and converted into a distance matrix (see Feng et al., 2018a for the detailed RDM construction procedure). We then normalized these RDMs by scaling between 0 (low dissimilarity, i.e., close in the distance) and 1 (high dissimilarity, i.e., far from each other in the distance). The binary tone-category RDM was constructed based on combinations of the four category labels (i.e., 0 for the same category, 1 for different categories). The syllable RDM was constructed based on the identity of the five syllables (i.e., 0 for the same syllable, 1 for different syllables). These six RDMs were correlated with learners' response confusion matrices in a block-by-block manner. Learners' response confusion matrices were created based on their categorization responses. If two sounds had an identical response, then this pair was coded as 0 in the confusion matrix; otherwise coded as 1. Using this procedure, we created two confusion matrices in each block (one for each talker) for each learner. The two matrices were then averaged for each block. Finally, we calculated the Spearman's correlations (i.e., model fits) between each RDM and confusion matrices. We also examined the relationships between the RDM model fits and learning outcome and speed across subjects to see which RDM explains most of the inter-individual variance in learning success.

Neuroimaging data analysis

Preprocessing for multivariate pattern analyses (MVPA).—All MRI data were preprocessed using SPM12 (Wellcome Department of Imaging Neuroscience, London, UK; www.fil.ion.ucl.ac.uk/spm/). Briefly, functional images were head-movement corrected by coregistering each image with the mean image. The high-resolution structure image was coregistered to the mean functional image for each subject. The normalization transformation parameters were then estimated using a segmentation-normalization procedure with the co-registered structure image and used to normalize the functional images to the Montreal Neurological Institute (MNI) space for group-level statistical analyses. To model single-trial brain activation responses for MVPA, the realigned functional images in the native space were fed into the subject-level GLM analysis with the least-squares single (LSS) approach (Mumford et al., 2014; Mumford et al., 2012). Specifically, for the tone-category training dataset, a design matrix was constructed with a regressor of interest for each trial during sound or feedback presentation; a regressor of non-interest consisted of other events (i.e., feedback or sound presentation for the current trial, and stimulus and feedback presentations for the other trials), six head movement regressors and a session mean regressor for each training block individually. Therefore, 480 subject-level GLM models (240 models for sound presentation and another 240 models for feedback) were constructed and estimated for each subject for the training experiment. Similarly, for the native tone-categorization experiment, 200 or 240 subject-level GLM models (for the sound presentation events) were constructed and estimated. The *t*-statistic brain maps were calculated for each trial and further used for MVPA (Misaki et al., 2010).

MVPA

Inter-subject neural representational similarity (IS-NRS) analysis.—We calculated three types of MVPA measures for learning-success prediction, including IS-NRS, model-based representational similarity analysis (RSA) measures, and neural feedback sensitivity. The three types of measures were considered as predictive features for predictive modeling (see Figure S4A for overall data processing pipeline). We quantified the degree of nativeness in neural representational structure for the learners by measuring the IS-NRS between each non-native learner and each of the native-Mandarin speakers for each anatomical defined region (see Figure 2A for a graphical illustration of the analysis procedure). Higher IS-NRS indicates greater similarity in the neural representations of the speech sound pairs relative to the native listeners. The IS-NRS is a derivative of the RSA, enabling us to evaluate similarity in neural representational structures (i.e., nRDMs) between subjects within the same stimulus space instead of in the subjects' voxel space (Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008). In the IS-NRS calculation pipeline, nRDMs were first generated within each subject and compared between subjects from the two groups (Figure 2A). Since both groups of participants heard the same sets of sound stimuli, their neural representations were compared in the same space. This two-step dissimilarity-similarity analysis approach can capture the similarity in representational structure between datasets from different scanners, populations, imaging modalities, and even species while ensuring that the representational similarity effect that is not due to the differences between datasets in these variables.

To calculate the IS-NRS, we extracted activation patterns from 94 regions of interest (ROIs) for both the learners and the native listeners based on an anatomical-defined atlas (i.e., Anatomical Automatic Labeling 2 [AAL2]) (Rolls et al., 2015). Cerebellum regions were removed because the cerebellum was not covered for some of the learners' data. Since two groups of datasets differ slightly in imaging parameters (e.g., number of voxels and voxel size, etc.) and there are large inter-individual differences in brain anatomy, the ROI-based approach ensures that the neural representational structures were compared in the same anatomical-defined areas between the two group of subjects. The AAL2 atlas in the MNI space was projected back to the native space for each subject, and the activation patterns when listening to the stimuli were extracted for each ROI (Figure 2A). Then, the nRDMs were calculated based on the activation patterns for each ROI and subject (note that dimensionality reduction was additionally conducted before the nRDM calculation to evaluate the representational dimensionality underlying successful learning; refer to the Representational dimensionality evaluation section). Calculating nRDMs ensures that different subjects' neural patterns were converted onto the same stimulus representational space. In this space, the distance (i.e., dissimilarity) between each pair of sounds (or phonetic contrasts) can be quantified based on their activation patterns and can be compared with other subjects or model-based RDMs. For each ROI, we calculated the IS-NRS for each learner-native speaker pair with Spearman's ranking correlation based on two vectorized nRDMs. The variances of hand-response RDM (left vs. right hand) were controlled for using the partial correlation approach to rule out the potential confounding of hand-response pattern similarity between the two groups. Therefore, each learner has 33 IS-NRSs (33 native listeners) for each ROI. These IS-NRSs were then averaged for the same learner in each ROI. The resulting IS-NRS data (i.e., a learner-by-ROI matrix) were used as predictive features to train and validate prediction models for the prediction of individual learners' behavioral learning outcome and speed.

Representational dimensionality evaluation.—To evaluate the representational dimensionality underlying successful learning (i.e., how many dimensions underlying learners' representational structure explains individual differences in learning success), we additionally used principal component analysis (PCA) with the single vector decomposition algorithm to decompose learners' sound-induced activation patterns into principal components (PCs) before IS-NRS calculation. We used a different number of PCs (maximal 20 PCs due to 20 sound items) to re-calculate the distance (i.e., Euclidean distance) between each sound pair and to construct the nRDMs for the learners. The IS-NRSs were calculated based on the dimension-constrained nRDM (see the procedure in Figure 5A and the overall processing pipeline in Figure S4, Supplementary Materials). Finally, a subject-by-ROI-by-PC IS-NRS matrix was obtained for the predictive modeling analysis. By using this dimension decomposition approach with predictive modeling, we can assess the dimensionality underlying individual learning success and parametrically examine the relationship between representational dimensionality and individual differences in learning success.

Model-based representational similarity analysis (RSA) and searchlight approach.—To compare the predictive power of the IS-NRS and other representational

measures, we also calculated the representational similarities between learners' nRDMs and five predefined model-based RDMs. The five RDMs were derived based on the acoustic properties and phonetic category labels of the stimuli (i.e., CAT, MD, F0 height, F0 slope, and Syllable; see Figure 2C). These model-based representational similarities were then entered into the predictive modeling to estimate to what extent these representational measures could be predictors of learning success. We further examined whether the predictive power of IS-NRS outperforms these model-based RSA measures. Cross-validation with bootstrapping and permutation procedures was used to determine the statistical significance and stability of the predictive models (see the next section for details). We also conducted model-based RSA with the searchlight approach (Kriegeskorte et al., 2006) for the five model RDMs to examine how the learners' neural representations of those stimuli-related information change following training. The same searchlight RSA was also conducted for native speakers for comparison. This searchlight approach has been described extensively in previous studies (Feng et al., 2021; Feng, Gan, et al., 2018; Feng et al., 2019). We briefly described the approach here. This searchlight RSA analysis was conducted across the whole brain. In each searchlight sphere (radius = 3 voxels), an nRDM was generated and then correlated with each of the five model RDMs with Spearman's correlation. The correlation value was then normalized with the Fisher *Z*-Transformation. This z value was then mapped back to the center voxel of the sphere. This RSA was conducted for each voxel to generate representational maps for each learner. We conducted this searchlight RSA for the early and late blocks separately. For the group-level analysis, the individual RSA maps in the learners' native space were first normalized to standard MNI space and then fed to a one-sample *t*-test against chance.

Neural sensitivity to feedback valence.—To examine the extent to which individual differences in neural sensitivity to corrective feedback relates to learning success and learners' nativeness in neural representations (i.e., IS-NRS), we estimated the ROI-based feedback-type classification accuracy and used it as a predictor to predict individual learning outcomes and speed as well as the IS-NRS. We operationally defined neural sensitivity to feedback valence as the feedback-type classification accuracy (correct vs. incorrect) based on the single-trial brain activation patterns. In each ROI, we used an LDA classifier (Chang & Lin, 2011) with a leave-one-block-out cross-validation (CV) procedure to classify individual trials' feedback types (correct or incorrect). Missing trials (i.e., trials for which there was no response; 7.1% on average across non-native learners) were removed before the classification analysis. We conducted the classification analysis separately for the early and late phases of training. Two learners were removed from the analysis because they achieved 100% accuracy in one of the last three blocks. The ratio of correct and incorrect feedback trials was varied across learners and training phases. To avoid this inherent imbalance, we used a balanced leave-one-block-out partition procedure. This procedure randomly selected the same number of correct and incorrect feedback trials for both training and testing so that each feedback type occurs equally often in the training and testing chunks. Higher classification accuracy indicates higher neural sensitivity to the feedback-valence of learners' brains. The ROI-based classification accuracies were used as predictive features to predict learning outcomes and speed as well as the IS-NRS. If the predictive power is

significantly higher than chance, the neural feedback sensitivity plays a critical role in learning success and the emergence of native-similar neural representations.

Predictive modeling analysis

To determine whether the neural measures significantly predict learning outcomes and speed as well as the nativeness of the emerging neural representations, we used multiple linear regression and linear support vector regression (SVR) as prediction algorithms in combination with a 10-fold CV procedure to train and validate prediction models. The neural measures (i.e., IS-NRS, five model-based representational measures, and neural feedback sensitivity index) obtained from all ROIs were used as predictive features, separately. Neural measures from all subjects were combined into an S-by-F matrix where S is the number of subjects, and F is the number of features (i.e., ROI). We used a nested 10-fold CV procedure for feature selection, dimension reduction, model construction, and estimation (see Figure 2B and Figure S4B for graphical illustrations). This CV procedure avoids obtaining overfit models with a large number of noisy features and ensures testing the models with unseen data points (i.e., generalization ability) (Feng, Ingvalson, et al., 2018). The nested CV procedure consisted of two levels of nesting (inner and outer) for feature selection, dimension reduction, and model validation. At the inner level, we employed the linear Pearson correlation analysis to remove irrelevant features based on the training sets (Pereira et al., 2009; Smialowski et al., 2010), where only features (e.g., IS-NRS in the superior temporal gyrus) showing significant correlations with learning outcomes or speed were selected. To avoid selecting features that were related to learners' first block tone categorization performance instead of speech category learning success (i.e., speed and outcome), we controlled for the inter-individual variance of the categorization accuracy in the first block in the feature selection step. Therefore, the predictive powers of the models reflect how well those selected ROIs in predicting learners' learning efficacy. Different feature selection thresholds (i.e., $P = 0.01$ and 10% of total features) were tested to assess the consistency and stability of the predictive performance. To further reduce the dimensionality of the predictors, we conducted the principal component analysis for the selected features and select the relevant principal components ($P < 0.05$) for further model training. The feature selection and dimension reduction procedures were conducted only on the training set, which was independent of the outer-level model testing (Figure 2B). That is, 90% (i.e., 9-fold) of the data were used for feature selection, dimension reduction, and model training while the hold-out 10% were for testing, repeating 10 times (i.e., 10-fold CV). The linear SVR algorithm with default parameters (i.e., C = 1, Gamma = 1/number of features) was also used to access the multivariate predictive power of the predictors. We used functions from a MATLAB package LIBSVM (Chang & Lin, 2011) in combination with in-house scripts to conduct the predictive modeling analysis. We examined the predictive power of a given neural measure by calculating the Pearson's correlation between the predicted and observed scores ($rval_{[\text{predicted, observed}]}$). The predictive modeling analysis was conducted separately for the early and late phases of training.

The statistical significance of the prediction was evaluated using a non-parametric permutation procedure. To test whether the predictive power of each model occurred by chance, we used a non-parametric permutation procedure to generate a null distribution of

the predictive power by fully shuffling the predictive features and learning performance across learners for each CV. Note that each feature and learning performance was permuted independently to generate a fully randomized data matrix, and the 10-fold CV procedure was conducted based on the randomized dataset. This data randomization and CV procedure were repeated 10,000 times, and the 95th percentile points of each distribution were used as the critical values for a one-tailed *t*-test against the null hypothesis with $P = 0.05$. To test the stability of the prediction, we used a bootstrapping procedure by randomly dividing all the learners into ten folds and conducted the 10-fold CV. Each CV prediction would be slightly different because the composition of the training and testing subjects were different for each iteration. We repeated this bootstrapping procedure 10,000 times. We identified the most contributing regions by comparing each region's correlation values derived from the feature selection procedure with its corresponding permutation-based correlation distribution. These regional permutation-based *p* values were corrected with the false discovery rate (FDR) approach.

Results

Behavioral results

Tone categorization performance for the native Mandarin speakers was close to ceiling (accuracy = 97.3 ± 2.66 % [mean ± SD], reaction time [RT] = 927.98 ± 109.26 ms). In the tone-category training fMRI experiment, English-speaking participants learned to categorize Mandarin tones significantly above-chance following training (first block: the mean accuracy across the participants was 22%, range = 0–45%, $SD = 9\%$; chance level = 25%; first block vs. chance: $t_{(52)} = -2.38$, $P = 0.021$; the final block: the mean accuracy was 47%, range = 13–100%, $SD = 26\%$; final block vs. chance: $t_{(52)} = 6.27$, $P < 0.001$; see Figure 1B for the group and individual learning curves). The category identification accuracy significantly increased over blocks (the first vs. final block paired *t*-test: $t_{(52)} = 7.69$, $P < 0.001$). Similarly, the mean accuracy of the late phase of training (i.e., the last three blocks) was significantly higher than in the early phase ($t_{(52)} = 6.41$, $P < 0.001$). The learning outcome was operationally defined as the mean accuracy in the late phase. Learning speed was operationally defined as the model fitting parameters for individuals' learning curves with a power function (Figure 1C). Learning speed was not significantly correlated with the accuracy in the first block ($r = 0.216$, $P = 0.120$) whereas learning outcome was significantly correlated with the accuracy in the first block ($r = 0.45$, $P < 0.001$). These results demonstrated that compared to the outcome, learning speed may be more related to learners' sound-to-category learning gains instead of the first block accuracy. Because the learning speed and outcome are two critical indices reflecting learning efficacy, we used both for the predictive modeling analyses while controlling for the inter-individual variance of block 1 accuracy.

The behavioral representational similarity analysis (bRSA, see Figure 1D for graphical analysis procedure) showed that RSA model fits significantly increased over blocks for the native nRDM (repeated measures ANOVA; main effect of block: $F_{(5, 260)} = 10.42$, $P < 0.001$) and other tone-category-related RDMs, including CAT ($F_{(5, 260)} = 20.24$, $P < 0.001$), MD ($F_{(5, 260)} = 12.78$, $P < 0.001$), F0 height ($F_{(5, 260)} = 13.43$, $P < 0.001$), and F0 slope

($F_{(5, 260)}$ = 5.94, P < 0.001). However, the RSA model fits of the Syllable RDM decreased over blocks ($F_{(5, 260)}$ = 9.24, P < 0.001; Figure 1E). Moreover, individual differences of the model fits were significantly correlated with the individual differences of learning outcome (Native nRDM: r = 0.91; CAT: r = 0.95; F0 height: r = 0.81; F0 slope: r = 0.83; MD: r = 0.89; Syllable: r = −0.73; Ps < 0.001; see Figure 1F for a representative scatter plot) and speed (Native nRDM: r = 0.80; CAT: r = 0.85; F0 height: r = 0.72; F0 slope: r = 0.74; MD: r = 0.78; Syllable: r = −0.61; Ps < 0.001). These modeling results indicate that native-similar categorization response patterns emerged for the learners following training. The response patterns were highly related to the individual differences in pitch encoding and learning success.

The degree of nativeness in neural representational structure predicts learning success

We employed inter-subject neural representational similarity analysis to measure the degree of nativeness in neural representational structure (i.e., IS-NRS) for individual learners, as compared to a group of native Mandarin-speaking listeners (see Figure 2 for the IS-NRS calculation procedure). Significant similarities with native listeners in neural representational structure emerged at the late phase of training in the bilateral superior temporal gyrus (STG) and right precentral gyrus (R.PreCG) (Figure 4B). Similar to the emerging native-similar neural representations, the learner's neural representations of tone categories and pitch-related information emerged in the late phase of training, demonstrated using the searchlight-based RSA with predefined category and pitch-related RDMs (CAT, MD, F0 height, and slope RDMs; Figure S5; Also see Figure S6 for IS-NRS comparisons between learners and native speakers). Comparing the whole-brain searchlight model-based RSA brain maps between the native listeners and the learners for the tone-category-related RDMs, we found that the searchlight RSA patterns of the learners in the late phase was approaching the patterns of the native listeners, although the RSA correlations were less robust in extent and yielded lower intensity. In contrast, the syllable-related information was less and less represented in the brain following training (Figure S5). Altogether, these results indicate that the learners enhanced the neural representations of learning/task-relevant tone-category-related information while decreased or suppressed the representations of the learning/task-irrelevant segmental units (e.g., consonants and vowels).

We used IS-NRS as an indicator of learners' nativeness in neural representational structure. The ROI-based IS-NRS and other model RSA measures were used as predictive features for learning-success prediction analyses (see Figure S4 for the overall analysis pipeline). We used cross-validation (CV) and non-parametric permutation procedures with 10,000 iterations to determine the statistical significance of each predictive model (see Figure 2B for the CV procedure). We also employed the bootstrapping procedure to evaluate the reliability of the prediction models. We found that the IS-NRS in the late phase of training was significantly predictive of learning outcome (permutation test: P = 0.004) and speed (permutation test: P = 0.006; see Figure 3A&B for the predictive powers), whereas the predictive powers were at chance levels for both outcome (P = 0.582) and speed (P = 0.915) predictions in the early phase (blue-color distributions in Figure 3B). We conducted additional prediction analyses with fine-tune distinction between different phases of training. To increase the signal-to-noise ratios of the activation estimation for individual stimulus

items, we combined data from two consecutive blocks. Therefore, the whole training session was divided into five parts (i.e., blocks 1–2, 2–3, 3–4, 4–5, and 5–6). We re-calculated the IS-NRSs for these blocks and re-conducted the learning-outcome and -speed prediction analyses. The results were shown in Figure S7 (Supplementary Materials). We found that the prediction powers increased as a function of training blocks. Only the IS-NRSs at the last three blocks (i.e., blocks 4–5 and blocks 5–6) were predictive of learning success. These additional results were consistent with the above prediction results showing that the IS-NRSs at the initial phase of training were not predictive of the learning speed and outcome. Altogether, these results demonstrated that the degree of the nativeness of the neural representational structure in the late training sessions was tightly related to individual differences in learning efficacy.

To further compare the predictive power of IS-NRS with other model-based representational measures, we conducted the same predictive modeling with other model-based RSA measures as predictors. Four tone-category-related RDMs (i.e., CAT, MD, F0 height, and F0 slope) and one segmental-unit-related RDM (i.e., Syllable) were used to generate RSA representational measures for all ROIs (Figure S4A). With the predictive modeling, we found that the IS-NRS yielded the highest predictive power (median $r_{[\text{predicted, observed}]}$ = 0.510, P = 0.004 for outcome prediction; median $r_{[\text{predicted, observed}]}$ = 0.412, P = 0.006 for speed prediction). Three of the tone-category-related RDMs also yielded predictive powers significantly better than chance (**CAT**: outcome prediction: $r_{[\text{predicted, observed}]}$ = 0.430, P = 0.013, SD = 0.097; speed prediction: $r_{[\text{predicted, observed}]}$ = 0.416, P = 0.012, SD = 0.077; **MD**: outcome prediction: $r_{[\text{predicted, observed}]}$ = 0.379, P = 0.014, SD = 0.082; speed prediction: $r_{[\text{predicted, observed}]}$ = 0.391, P = 0.013, SD = 0.074; **F0 height**: outcome prediction: $r_{[\text{predicted, observed}]}$ = 0.304, P = 0.025, SD = 0.072; speed prediction: $r_{[\text{predicted, observed}]}$ = 0.353, P = 0.017, SD = 0.083). However, the F0 slope and Syllable models did not show significant better-than-chance predictive powers (**F0 slope**: outcome prediction: $r_{[\text{predicted, observed}]}$ = 0.036, P = 0.388, SD = 0.119; speed prediction: $r_{[\text{predicted, observed}]}$ = 0.083, P = 0.273, SD = 0.118; **Syl**: outcome prediction: $r_{[\text{predicted, observed}]}$ = −0.053, P = 0.649, SD = 0.120; speed prediction: $r_{[\text{predicted, observed}]}$ = 0.151, P = 0.182, SD = 0.118).

To further confirm that the predictive power of the IS-NRS was not due to sharing the same segmental information (i.e., consonants and vowels) between learners and native listeners, we recalculated the IS-NRS while additionally controlling for the Syl model. We confirmed that the resulting predictive power remained significant (outcome prediction: $r_{[\text{predicted, observed}]}$ = 0.510, P = 0.003; speed prediction: $r_{[\text{predicted, observed}]}$ = 0.403, P = 0.008). We also examined to what extent the predictive power of the IS-NRSs was due to the joint variances of F0 height and slope representations by controlling for the variance of the two RDMs. We found that the predictive power of the IS-NRS was diminished (outcome prediction: $r_{[\text{predicted, observed}]}$ = −0.056, P = 0.669; speed prediction: $r_{[\text{predicted, observed}]}$ = −0.081, P = 0.732) when controlled for the variances of both RDMs. These results indicate that the representational models derived from native listeners' neural patterns and the resulting IS-NRSs outperformed other representational measures in differentiating successful from less-successful learners.

To examine whether the ROIs with positive or negative correlation patterns contributed equally to the predictive performance, we reran the predictive modeling with two feature selection (FS) procedures to disentangle the effects of the two types of ROIs. In one FS, we only selected ROIs that showed positive correlations between the IS-NRS and learning performance in the training sets to build predictive models, while in another FS, we only selected the ROIs with negative correlations. We found that those models with only positive-correlation ROIs were able to significantly predict learning success while the predictive powers with negative-correlation ROIs were at chance. These results indicate that more native-similar neural representations for the learners are associated with higher learning efficacy.

To identify brain regions that significantly contributed to the prediction models with IS-NRS, we estimated the statistical significance of each region using a non-parametric permutation-based approach. In the predictive modeling, we generated a bootstrapping-based correlation distribution and a permutation-based null distribution (10,000 iterations) for each region based on the training sets (i.e., 90% of randomly-selected learners, $n = 48$). The median of the bootstrapping distribution for a given region was compared with the 95th percentile of its corresponding null distribution to determine statistical significance. Multiple comparison correction was conducted based on the false discovery rate (FDR) approach. We found that a speech-related brain network, including the triangular part of left inferior frontal gyrus (L.IFGtri), left inferior parietal lobule (L.IPL), left supramarginal gyrus (L.SMG), bilateral superior temporal gyrus (STG), left middle temporal gyrus (MTG), right angular gyrus (R.AG), and right precentral gyrus (R.PreCG) showed significant contribution to the speed-prediction modeling in the late phase (Figure 4A, right panel). Similarly, L.IFGtri, L.STG, and R.PreCG contributed significantly to the outcome prediction (Figure 4A, left panel). Additional searchlight IS-NRS analyses within the bilateral STG were conducted to identify which STG subregions contributing to individual differences in learning success. The IS-NRSs were significantly correlated with learning outcomes primarily in the middle and anterior portions of the STG (see Figure S8, Supplementary Materials). Taken together, these results indicate that learners with greater IS-NRS (i.e., more nativeness in neural representations) in the fronto-temporoparietal speech perception network are more successful in learning to categorize novel speech categories.

To further examine whether individual differences in the neural representations at the initial phase of training relate to the individual differences in the neural representations at the late phase, we conducted additional prediction analyses with six neural representational measures (i.e., IS-NRS [native-similar representations], CAT [tone category representations], F0 height [pitch height representations], F0 slope [pitch direction representations], MD [multidimensional pitch representations], and Syl [syllable representations]) as predictive features derived from the first two blocks to predict these representational measures at the last two blocks. To increase the signal-to-noise ratio of the neural representational dissimilarity matrices (nRDMs), we combined the data from blocks 1 and 2 as well as blocks 5 and 6, respectively. Across the whole brain (94 ROIs), we did not find any region shown a significant prediction effect after correction (i.e., FDR $q = 0.05$). This finding suggests that the initial neural representations may change significantly following training,

with successful learners' representation having native-similar representations relative to less successful learners.

Multidimensionality in learners' neural representations contributes to the learning success

To further reveal the nature of the dimensionality of the emerging native-similar neural representational structure underlying successful learning, we used principal component analysis with the singular value decomposition (SVD) algorithm to decompose learners' brain activation patterns of the stimuli into independent principal components (PCs) and recalculated the IS-NRS with PC-constrained nRDM for predictive modeling (see Figure 5A for graphical analysis procedure). This procedure allows us to assess how many dimensions of the learners' representations underlie individual differences in learning success. We found that representation dimensionality significantly modulated the predictive powers for both learning-speed and -outcome predictions. Predictive power increased as the dimensionality increased. Importantly, predictive powers reached a plateau with approximately five PCs (speed prediction: median $r = 0.52$, $P = 0.001$; outcome prediction: $r = 0.42$, $P = 0.005$; 1PC's vs. 5PCs' predictive power: $Ps < 0.001$; Figure 5B). We also conducted the same prediction analysis with PCA for six predefined speech-perception-related regions and found that the number of PCs for the maximum predictive powers were varied across regions (ranged from two to nine PCs; see Figure S9 in Supplementary Materials). These results indicate that successful learners use a multidimensional but also cost-efficient neural representational mechanism (i.e., a moderate number of dimensions) to encode the newly-acquired speech categories. A more straightforward demonstration is shown in Figure 5C. We extracted the IS-NRSs (controlled for both hand-response and Syllable RDMs) from the significant contributing regions (i.e., L.IFGtri, L.STG, and R.PreCG) and compared them between the successful and less-successful learners across different numbers of PCs. Two groups of learners were created based on the median split of the learning outcome (successful: $n = 26$, $M = 65.0\%$; less successful: $n = 25$, $M = 24.4\%$; two of them in the median line were removed). We found that successful learners showed greater IS-NRS than those of the less-successful learners (group-by-dimensionality ANOVA; main effect of the group: $F_{(1, 49)} = 16.33$, $P < 0.001$) but only in the late phase; while the group differences were significantly increased as the dimensionality increased and reached the maximum at 5–6 PCs, which was evidenced by a significant group-by-dimensionality interaction effect ($F_{(5, 245)} = 4.84$, $P < 0.001$).

Neural sensitivity to feedback-valence in the frontostriatal system contributes to individual differences in learning success and in the nativeness of neural representations

To evaluate the extent to which the neural sensitivity to feedback valence is a driving factor of the behavioral learning success (i.e., outcome and speed) and the degree of nativeness of neural representational structure (i.e., IS-NRS), we used the multivariate feedback-type classification accuracy as a neural feedback-sensitivity index to predict the learning performance and IS-NRS. Higher feedback-type classification accuracy indicates more sensitivity to feedback valance (i.e., more robust feedback-valance representations) in the brain. At the group level, with a cross-validation procedure that strictly balancing the number of correct and incorrect feedback trials, we found that widespread brain regions showed significantly above-chance classification accuracy for both early and late phases of

training (Figure 6A), including cortical and sub-cortical striatal areas. Note that, quantitatively, the classification accuracies in the late phase were slightly higher than in the early phase, especially in the frontostriatal regions. The most significant feedback-sensitive regions across training phases were within the frontostriatal network, which is consistent with previous findings derived by univariate activation analysis that used contrasts of correct vs. incorrect feedback (Feng et al., 2019; Yi et al., 2016). Importantly, the neural feedback sensitivity in the late phase significantly predicted learners' behavioral learning outcome (median $r_{[\text{predicted, observed}]} = 0.54$, $P = 0.003$; permutation test) and speed (median $r_{[\text{predicted, observed}]} = 0.60$, $P = 0.002$) (Figure 6B). In contrast, the predictions with feedback classification accuracies in the early phase were not significantly better than chance (Ps > 0.05). Furthermore, the neural feedback sensitivity in the late phase significantly predicted the IS-NRSs (median $r_{[\text{predicted, observed}]} = 0.40$, $P = 0.011$) of the L.IFGtri, bilateral STG, and R.PreCG (IS-NRSs collapsed across these regions; see Figure 6B), where these regions showed significant predictive powers of learning success as well as the emergent native-similar neural representations in the late phase relative to the early phase.

The feedback-sensitive regions that significantly contribute to the learning predictions were identified in the frontostriatal network (Figure 6C, outcome prediction regions; also see Figure S10 for regions significantly contributing to the speed and IS-NRS predictions), which indicates that the neural sensitivity of feedback valence at the late phase of training within this network is a neuromarkers of tone-category learning successful. The most contributing regions in predicting the IS-NRSs were also within the frontostriatal network, including the L.IFGtri, left caudate, right angular gyrus, right IFGorb, right middle frontal gyrus, and right posterior cingulate cortex (permutation-based FDR-corrected $q < 0.05$). These results demonstrate that the frontostriatal network plays important role in facilitating the formation of native-similar neural representations.

Discussion

We employed a novel inter-subject neural representational similarity analysis and rigorous predictive modeling approach to examine the neural underpinnings of individual differences in non-native speech category learning success. We demonstrate that native-similar neural representational structure emerges during training and the degree of nativeness in neural representational structure in the left inferior frontal gyrus (IFG), left superior temporal gyrus (STG), and right precentral gyrus (PreCG) is robustly predictive of behavioral learning success. The emerging native-similar neural representations in successful learners are multidimensional and economical in encoding pitch-related phonetic/phonological category information. Further, individual differences in neural sensitivity to feedback valence within the frontostriatal network are highly predictive of individual differences in learning success and of the degree of nativeness of the emerging representations. These findings provide new insights into the neural representational mechanism underlying successful non-native speech category acquisition and the role of feedback in mediating individual differences in learning success.

The nativeness in neural representational structure predicts sound-to-category learning success

It has been previously demonstrated that task-general and acoustic-invariant neural representations of Mandarin tone categories for native listeners are evidenced in the superior temporal areas and inferior parietal lobule using multivariate pattern classification (MVPC) (Feng, Gan, et al., 2018). While MVPC reveals category-level representations, this analytic method cannot capture the fine representational structures underlying the neural activation patterns. Here we used native listeners' neural representational dissimilarity matrices (RDMs) as a native representation model to estimate learners' neural representational structure and to assess the extent to which native-similar representations emerged during learning at the group level and in relation to individual differences in learning success. At the group level, the native-similar neural representational structure emerged in the late phase of training, similar to the emerging neural representations of tone categories and multidimensional pitch information (see Figure S5 in Supplementary Materials), which suggests that feedback-based training protocol could not only facilitate adult learners forming task-relevant categorical representations (Chandrasekaran, Koslov, et al., 2014; Feng et al., 2019) but also resulted in a representational structure that was increasingly similar to native listeners within just hundreds of training trials. This finding is consistent with the previous observation that neural representations of tone category emerge following training (Feng et al., 2019), and further reveals the native-similar nature of the representational structure underlying successful sound-to-category acquisition.

In addition to the emerging native-similar representations, a key finding is that greater neural similarity in representational structure between learners and native listeners (i.e., IS-NRS) predicts better learning performance. This finding suggests that IS-NRS is a robust neural representational indicator of sound-to-category learning success. The prediction results are validated by the rigorous predictive modeling approach with cross-validation, bootstrapping, and permutation procedures and are not explained by the contextual factors, response similarity, or individual differences in tone identification performance in the first block of training. The left IFG, STG, and right PreCG are crucial brain regions that reliably contributed to the learning-success prediction. These findings demonstrate that more successful learners reveal greater similarity to native listeners in their neural representations of Mandarin tone categories, even though the mechanisms underlying how the category representations are acquired might be fundamentally different (e.g., unsupervised vs. feedback-based learning) (Hernandez et al., 2005; Lim et al., 2019; MacWhinney, 1998).

The native listeners' neural representational dissimilarity structure serves as an excellent tone-contrast model to quantitatively evaluate the degree of nativeness of neural representational patterns for the learners. The native listeners' dissimilarity structure is also better in differentiating successful from less successful learners comparing to other representational models (i.e., CAT, MD, etc.). The IS-NRS prediction model yielded the best predictive powers in predicting how fast and how well learning could be achieved among other predefined pitch-related category models, reflecting on the predictive accuracy as well as the reliability of the predictive models revealed by the bootstrapping procedure (Figure 3C). Previous studies have documented several neural indicators of speech learning success

(Deng et al., 2016; Golestani & Zatorre, 2009; Liu & Holt, 2011; Myers & Swan, 2012; Sheppard et al., 2012; Wong et al., 2011; Wong & Ettlinger, 2011; Wong et al., 2007; Zhang et al., 2009). However, these studies have largely focused on pre-training neural measures to predict learning outcomes or on examining the group-level neural changes in response to training. Here, we categorically focus on the neural representational dynamics during the process of learning and how neural plasticity contributes to individual differences in learning. Our results provide key insights into how successful learners form multidimensional and economical representations as a function of training with a goal of more efficient categorization.

The dimensionality of the emerging native-similar neural representational patterns

Theoretical models in L2 acquisition, largely in the domains of grammar and syntax, posit that the representational structure in L2 learners may be shallow and inefficient (Clahsen & Felser, 2006a, 2018). However, in term of non-native speech category learning, our results, demonstrate that the emerging neural representations of newly acquired speech categories for successful learners are not only significantly similar to those of native listeners but also multidimensional and cost-efficient, where the speech categories are encoded in a neural representational space with a moderate number of dimensions. Mathematically, a high-dimensional representational space provides flexibility in encoding different categories but may come with a greater cost in terms of neural resources. In contrast, a low-dimensional space expends fewer resources but may not be capable of robustly differentiating behaviorally relevant categories. An optimal learning-induced representational mechanism would need to balance these two competing factors—maximizing behaviorally-relevant information in the signal with minimal resources to encode information (Gervain & Geffen, 2019; Tang et al., 2019).

Using the single vector decomposition approach, we decomposed the neural patterns of speech sounds into independent dimensions and reconstructed the representational spaces parametrically with different numbers of dimensions to assess the relationship between dimensionality and predictive power as well as to estimate how many dimensions are needed to differentiate successful learners from less successful learners. Our results show that predictive powers change as a non-linear function of dimensionality, which reflects an interaction between learning success and dimensionality. Successful learners' neural representations show increasingly native-similar as the number of dimensions increase, whereas less successful learners did not show such a relationship. Importantly, learning-success predictions did not increase linearly with the number of dimensions increase (i.e., close to plateau at around five dimensions). This result suggests that the emerging neural representations in successful learners are cost-efficient, in which activation patterns encode the new categories with a limited number of dimensions that can maximally differentiate them, similar to native listeners (Gandour, 1983; Gandour & Harshman, 1978). Using other representational models' RSA measures as predictors, we further demonstrate that multidimensional pitch information are critical constituents of the emerging native-similar neural representations for successful learners. Consistent with previous findings in native listeners (Chandrasekaran, Gandour, et al., 2007; Gandour & Harshman, 1978), we posited that pitch height and direction (i.e., contour) are important category-defining components

that represent tone-category distinctions in successful learners. Although prior behavioral studies have shown that other dimensions may also differentiate tones (Gandour & Harshman, 1978), we found that when we controlled for the variance of both F0 height and slope, the prediction is diminished. These results suggest that pitch height and direction are two critical components underlying both the native listeners' and successful learners' neural representations, in line with our original hypothesis.

The brain areas that significantly contributed to the learning-success prediction are within a large speech network involved in encoding pitch information for both native listeners and successful learners. These include the left IFG, left IPL, bilateral STG, left SMG, left MTG, right AG, and PreCG. Intriguingly, these brain areas encode the two pitch components differently for learners at the group-level. The bilateral STG, PostCG, PreCG, and the left IFG are dominated by F0 height, whereas much fewer regions are dominated by representing F0 slope (see Figure S5, Supplementary Materials). However, for native listeners, the above regions encode the multidimensional pitch information of the categorical representation. It is important to note that sound-to-category training only involved 240 trials. The mean accuracy for even the successful learners ($n = 26$) in the last training block ($M = 70\%$) is therefore far from perfect that the native speakers performed ($M = 97\%$). Therefore, the greater dominance of pitch height in learners relative to native listeners may because the learners are still novices. In line with a recent study demonstrating changes at early auditory processing stages with extensive multi-day sound-to-category training (Reetzke et al., 2018), we posit that a more extended training phase may yield better neural alignment of dimensional structure between native listeners and successful learners.

Neural sensitivity to feedback valence drives learning success and the emergence of native-similar neural representations

Our results demonstrate the significant similarity between native listeners and successful non-native learners in how tone-category-related information is represented in the brain. It is important to note that this significant neural similarity emerges following a relatively short period of sound-to-category training, which fundamentally differs from the mechanisms underlying category acquisition during infancy. Acquiring speech categories in adulthood is argued to require greater supervision and recent models (e.g., dual learning systems model) have highlighted the role of multiple corticostriatal systems in mediating adult speech learning (Chandrasekaran, Koslov, et al., 2014; Chandrasekaran, Yi, et al., 2014; Maddox & Chandrasekaran, 2014). Here, we provide supporting evidence from the perspective of individual differences in learning success that the neural sensitivity to feedback valence in the frontostriatal system is highly predictive of both behavioral learning success and the emerging native-similar neural representations. We posit that learners are reliant on feedback processing to update the internal representation, which could guide the formation of correct sound-to-category representations and efficient categorization behaviors. Individual differences in feedback processing and sensitivity are presumably critical factors associated with individual differences in learning outcomes. A previous study has identified the putamen, a core region in the striatum, dynamically coupling with the representational areas in the left STG when learners processed corrective feedbacks (Feng et al., 2019). In expanding this finding with a novel prediction analytic method, we find that individual

differences in neural feedback sensitivity in a more extended cortico-striatal network, including the striatum as well as lateral and medial frontal, precentral gyrus, inferior parietal cortex, and hippocampus areas robustly contributed to the prediction of learning success and the degree of native-similar representations. These findings suggest that feedback sensitivities in the two putative category learning systems (i.e., reflective and reflexive systems) are critical neural sources mediating individual differences in speech category learning success, at least during the transition learning stage (from novice to experienced phase).

The neural sensitivity to feedback valence is prominent in the late training phase relative to the early phase. Similarly, the representation/learning-success predictions (based on feedback-valence sensitivity) are more powerful for the late relative to the early training phase. We posit that trial-by-trial corrective feedback information facilitates rapidly updating learners' internal representations to enhance categorization success. More successful and faster learners likely leverage the feedback better, leading to the more native-similar multidimensional representations of the acquired speech categories. During sound-to-category learning, interactions between the striatum, auditory cortex, and frontoparietal regions might enable the integration of perceptual representation and feedback valence, mediating the shift from novice to skilled behavioral performance (Reetzke et al., 2018).

Learning non-native novel phonemic contrasts is a key step towards acquiring new words in a foreign language. Previous studies have demonstrated that both learning non-native phonemic contrasts and learning new words rely on the feedback/reward-sensitive striatal regions (Feng et al., 2019; Lim et al., 2019; Ripolles et al., 2016; Ripolles et al., 2014) and interactions within corticocortical and corticostriatal networks (Li et al., 2014; Lopez-Barroso et al., 2013; Shtyrov, 2012). The striatal activations are associated with domain-general reward processes, where a reward signal (e.g., gaining money or receiving feedback) may facilitate the formation of new memories in general (Adcock et al., 2006; Wolosin et al., 2012) and drive the acquisition of different language components (beyond phonetic/phonological learning). The interaction of the striatum and cortical regions have been proposed to be a neural driving force for the formation of cortical representations in language learning (Feng et al., 2019; Ripolles et al., 2014). Here, we further demonstrate that individual differences in learning success and the robustness of the emerging native-similar neural representations are both associated with the feedback sensitivity in the corticostriatal network. This corticostriatal interaction mechanism may not be restricted to the learning of novel non-native phonemic contrasts but also could be used in other aspects of language learning, e.g., learning new words and grammar. Further studies need to be conducted across different domains of language learning to directly address this question.

To what extent can our results generalize to typical language learning contexts? Prior studies have trained participants on a sound-to-meaning training paradigm that involves learning novel words as well as tone categories (Deng et al., 2016; Wong et al., 2007). Similar to the current study, large individual differences underlie learning performance in this learning context as well. Interestingly, during the initial word learning, learners often make lexical errors, but by the end of the training, most errors are in disambiguating tonal categories. Indeed, a prior study demonstrated that learning success in such a paradigm may be driven

by poorer representations of tone category information in subcortical auditory regions (Chandrasekaran et al., 2012). Thus, representational plasticity may underlie individual differences in learning to map pitch information irrespective of learning context. However, it is also possible that in a more ecological word learning context, there would be a need for greater coupling between the lexical-semantic network, superior temporal gyrus, as well as the reward-related corticostriatal pathways. In this context, individual differences in learning success may depend on emerging representations of tone categories as well as lexical-semantic representations.

Conclusion

Using the multivariate inter-subject representational similarity analysis and predictive modeling approaches, we deconstructed the neural sources of inter-individual differences in learning success during the process of learning to map non-native speech sounds into discrete categories. Successful learners can build robust and detailed speech representations that are similar to those in native listeners. The greater similarity between non-native learners and native listeners in neural representations of tone-category-related pitch information is associated with more rapid learning and better learning outcomes. Neural representations in successful learners are encoded in a cost-efficient manner: the representational space is multidimensional but with a limited number of dimensions that maximize the categorization of newly acquired speech categories. The emerging native-similar representations in more successful learners are associated with neural sensitivity to feedback valence in a distributed frontostriatal network. We provide new evidence and insights into the neural mechanisms underlying the successful acquisition of non-native speech categories in adulthood and into the scaffolding for the development of individualized speech training protocols that maximize learning outcomes with effective feedback.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments and Funding sources

This work was supported by grants from the General Research Fund (Ref. No. 14619518 to G.F.) of the Research Grants Council of Hong Kong and the National Institute on Deafness and other Communication Disorders of the National Institutes of Health under Award No. (R01DC013315 to B.C.). We thank Casey Roark for her helpful comments and language editing for the early version of the manuscript.

References

Abutalebi J (2008). Neural aspects of second language representation and language control. Acta Psychologica, 128(3), 466–478. DOI: 10.1016/j.actpsy.2008.03.014 [PubMed: 18479667]

Adcock RA, Thangavel A, Whitfield-Gabrieli S, Knutson B, & Gabrieli JD (2006). Reward-motivated learning: mesolimbic activation precedes memory formation. Neuron, 50(3), 507–517. DOI: 10.1016/j.neuron.2006.03.036 [PubMed: 16675403]

Best CT (1995). A direct realist view of cross-language speech perception. In Strange W (Ed.), Speech perception and linguistic experience (pp. 171–206). York Press.

Birdsong D (2018). Plasticity, Variability and Age in Second Language Acquisition and Bilingualism. Frontiers in Psychology, 9, 81. DOI: 10.3389/fpsyg.2018.00081 [PubMed: 29593590]

Birn RM, Cox RW, & Bandettini PA (2002). Detection versus estimation in event-related fMRI: choosing the optimal stimulus timing. Neuroimage, 15(1), 252–264. DOI: 10.1006/nimg.2001.0964 [PubMed: 11771993]

Bley-Vroman R (1990). The logical problem of foreign language learning. Linguistic Analysis, 20(1–2), 1–49.

Bley-Vroman R (2009). The Evolving Context of the Fundamental Difference Hypothesis. Studies in Second Language Acquisition, 31(2), 175–198. DOI: 10.1017/S0272263109090275

Chandrasekaran B, Gandour JT, & Krishnan A (2007). Neuroplasticity in the processing of pitch dimensions: a multidimensional scaling analysis of the mismatch negativity. Restorative Neurology and Neuroscience, 25(3–4), 195–210. https://www.ncbi.nlm.nih.gov/pubmed/17942999 [PubMed: 17942999]

Chandrasekaran B, Koslov SR, & Maddox WT (2014). Toward a dual-learning systems model of speech category learning. Frontiers in Psychology, 5, 825. DOI: 10.3389/fpsyg.2014.00825 [PubMed: 25132827]

Chandrasekaran B, Kraus N, & Wong PC (2012). Human inferior colliculus activity relates to individual differences in spoken language learning. Journal of Neurophysiology, 107(5), 1325–1336. DOI: 10.1152/jn.00923.2011 [PubMed: 22131377]

Chandrasekaran B, Krishnan A, & Gandour JT (2007). Mismatch negativity to pitch contours is influenced by language experience. Brain Research, 1128(1), 148–156. DOI: 10.1016/j.brainres.2006.10.064 [PubMed: 17125749]

Chandrasekaran B, Sampath PD, & Wong PC (2010). Individual variability in cue-weighting and lexical tone learning. Journal of the Acoustical Society of America, 128(1), 456–465. DOI: 10.1121/1.3445785

Chandrasekaran B, Yi HG, & Maddox WT (2014). Dual-learning systems during speech category learning. Psychon Bull Rev, 21(2), 488–495. DOI: 10.3758/s13423-013-0501-5 [PubMed: 24002965]

Chang CC, & Lin CJ (2011). LIBSVM: A Library for Support Vector Machines. Acm Transactions on Intelligent Systems and Technology, 2(3), 27:21–27:27. DOI: 10.1145/1961189.1961199

Chee MW, Caplan D, Soon CS, Sriram N, Tan EW, Thiel T, & Weekes B (1999). Processing of visually presented sentences in Mandarin and English studied with fMRI. Neuron, 23(1), 127–137. DOI: 10.1016/s0896-6273(00)80759-x [PubMed: 10402199]

Chen J, Leong YC, Honey CJ, Yong CH, Norman KA, & Hasson U (2017). Shared memories reveal shared structure in neural activity across individuals. Nature Neuroscience, 20(1), 115–125. DOI: 10.1038/nn.4450 [PubMed: 27918531]

Cheour M, Ceponiene R, Lehtokoski A, Luuk A, Allik J, Alho K, & Naatanen R (1998). Development of language-specific phoneme representations in the infant brain. Nature Neuroscience, 1(5), 351–353. DOI: 10.1038/1561 [PubMed: 10196522]

Clahsen H, & Felser C (2006a). Continuity and shallow structures in language processing. Applied Psycholinguistics, 27(1), 107–126. DOI: 10.1017/S0142716406060206

Clahsen H, & Felser C (2006b). How native-like is non-native language processing? Trends in Cognitive Sciences, 10(12), 564–570. DOI: 10.1016/j.tics.2006.10.002 [PubMed: 17071131]

Clahsen H, & Felser C (2018). Some Notes on the Shallow Structure Hypothesis. Studies in Second Language Acquisition, 40(3), 693–706. DOI: 10.1017/S0272263117000250

Dale AM (1999). Optimal experimental design for event-related fMRI. Human Brain Mapping, 8(2–3), 109–114. DOI: 10.1002/(SICI)1097-0193(1999)8:2/3<109::AID-HBM7>3.0.CO;2-W [PubMed: 10524601]

Deng Z, Chandrasekaran B, Wang S, & Wong PC (2016). Resting-state low-frequency fluctuations reflect individual differences in spoken language learning. Cortex, 76, 63–78. DOI: 10.1016/j.cortex.2015.11.020 [PubMed: 26866283]

Diedrichsen J, & Kriegeskorte N (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLoS

Computational Biology, 13(4), e1005508. DOI: 10.1371/journal.pcbi.1005508 [PubMed: 28437426]

Ellis R (2004). Individual Differences in Second Language Learning. The handbook of applied linguistics, 525.

Feng G, Chen HC, Zhu Z, He Y, & Wang S (2015). Dynamic brain architectures in local brain activity and functional network efficiency associate with efficient reading in bilinguals. Neuroimage, 119(0), 103–118. DOI: 10.1016/j.neuroimage.2015.05.100 [PubMed: 26095088]

Feng G, Gan Z, Llanos F, Meng D, Wang S, Wong PCM, & Chandrasekaran B (2021). A distributed dynamic brain network mediates linguistic tone representation and categorization. Neuroimage, 224, 117410. DOI: 10.1016/j.neuroimage.2020.117410 [PubMed: 33011415]

Feng G, Gan Z, Wang S, Wong PCM, & Chandrasekaran B (2018). Task-General and Acoustic-Invariant Neural Representation of Speech Categories in the Human Brain. Cerebral Cortex, 28(9), 3241–3254. DOI: 10.1093/cercor/bhx195 [PubMed: 28968658]

Feng G, Ingvalson EM, Grieco-Calub TM, Roberts MY, Ryan ME, Birmingham P, Burrowes D, Young NM, & Wong PCM (2018). Neural preservation underlies speech improvement from auditory deprivation in young cochlear implant recipients. Proceedings of the National Academy of Sciences of the United States of America, 115(5), E1022–E1031. DOI: 10.1073/pnas.1717603115 [PubMed: 29339512]

Feng G, Yi HG, & Chandrasekaran B (2019). The Role of the Human Auditory Corticostriatal Network in Speech Learning. Cerebral Cortex, 29(10), 4077–4089. DOI: 10.1093/cercor/bhy289 [PubMed: 30535138]

Gabrieli JDE, Ghosh SS, & Whitfield-Gabrieli S (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron, 85(1), 11–26. DOI: 10.1016/j.neuron.2014.10.047 [PubMed: 25569345]

Gandour J (1983). Tone Perception in Far Eastern-Languages. Journal of Phonetics, 11(2), 149–175. DOI: 10.1016/S0095-4470(19)30813-7

Gandour JT, & Harshman RA (1978). Crosslanguage differences in tone perception: a multidimensional scaling investigation. Language and Speech, 21(1), 1–33. DOI: 10.1177/002383097802100101 [PubMed: 692240]

Garcia-Lazaro JA, Ahmed B, & Schnupp JW (2011). Emergence of tuning to natural stimulus statistics along the central auditory pathway. PloS One, 6(8), e22584. DOI: 10.1371/journal.pone.0022584 [PubMed: 21850231]

Gervain J, & Geffen MN (2019). Efficient Neural Coding in Auditory and Speech Perception. Trends in Neurosciences, 42(1), 56–65. DOI: 10.1016/j.tins.2018.09.004, [PubMed: 30297085]

Golestani N, & Zatorre RJ (2009). Individual differences in the acquisition of second language phonology. Brain and Language, 109(2–3), 55–67. DOI: 10.1016/j.bandl.2008.01.005 [PubMed: 18295875]

Hartshorne JK, Tenenbaum JB, & Pinker S (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. Cognition, 177, 263–277. DOI: 10.1016/j.cognition.2018.04.007 [PubMed: 29729947]

Hernandez A, Li P, & MacWhinney B (2005). The emergence of competing modules in bilingualism. Trends in Cognitive Sciences, 9(5), 220–225. DOI: 10.1016/j.tics.2005.03.003 [PubMed: 15866148]

Kidd E, & Donnelly S (2020). Individual Differences in First Language Acquisition. Annual Review of Linguistics, Vol 6, 6, 319–340. DOI: 10.1146/annurev-linguistics-011619-030326

Kidd E, Donnelly S, & Christiansen MH (2018). Individual Differences in Language Acquisition and Processing. Trends in Cognitive Sciences, 22(2), 154–169. DOI: 10.1016/j.tics.2017.11.006 [PubMed: 29277256]

Kriegeskorte N, Goebel R, & Bandettini P (2006). Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America, 103(10), 3863–3868. DOI: 10.1073/pnas.0600244103, [PubMed: 16537458]

Kriegeskorte N, & Kievit RA (2013). Representational geometry: integrating cognition, computation, and the brain. Trends in Cognitive Sciences, 17(8), 401–412. DOI: 10.1016/j.tics.2013.06.007 [PubMed: 23876494]

Kriegeskorte N, Mur M, & Bandettini P (2008). Representational similarity analysis–connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience, 2.

Kuhl PK (2004). Early language acquisition: cracking the speech code. Nature Reviews Neuroscience, 5(11), 831–843. DOI: 10.1038/nrn1533 [PubMed: 15496861]

Kuhl PK (2010). Brain mechanisms in early language acquisition. Neuron, 67(5), 713–727. DOI: 10.1016/j.neuron.2010.08.038 [PubMed: 20826304]

Kuhl PK, Conboy BT, Coffey-Corina S, Padden D, Rivera-Gaxiola M, & Nelson T (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 363(1493), 979–1000. DOI: 10.1098/rstb.2007.2154 [PubMed: 17846016]

Li P, Legault J, & Litcofsky KA (2014). Neuroplasticity as a function of second language learning: anatomical changes in the human brain. Cortex, 58, 301–324. DOI: 10.1016/j.cortex.2014.05.001 [PubMed: 24996640]

Lim SJ, Fiez JA, & Holt LL (2019). Role of the striatum in incidental learning of sound categories. Proceedings of the National Academy of Sciences of the United States of America, 116(10), 4671–4680. DOI: 10.1073/pnas.1811992116 [PubMed: 30782817]

Liu R, & Holt LL (2011). Neural changes associated with nonspeech auditory category learning parallel those of speech category acquisition. Journal of Cognitive Neuroscience, 23(3), 683–698. DOI: 10.1162/jocn.2009.21392 [PubMed: 19929331]

Liu TT, Frank LR, Wong EC, & Buxton RB (2001). Detection power, estimation efficiency, and predictability in event-related fMRI. Neuroimage, 13(4), 759–773. DOI: 10.1006/nimg.2000.0728 [PubMed: 11305903]

Lively SE, Logan JS, & Pisoni DB (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. Journal of the Acoustical Society of America, 94(3 Pt 1), 1242–1255. DOI: 10.1121/1.408177

Lopez-Barroso D, Catani M, Ripolles P, Dell'Acqua F, Rodriguez-Fornells A, & de Diego-Balaguer R (2013). Word learning is mediated by the left arcuate fasciculus. Proceedings of the National Academy of Sciences of the United States of America, 110(32), 13168–13173. DOI: 10.1073/pnas.1301696110 [PubMed: 23884655]

MacWhinney B (1998). Models of the emergence of language. Annual Review of Psychology, 49, 199–227. DOI: 10.1146/annurev.psych.49.1.199

Maddox WT, & Chandrasekaran B (2014). Tests of a Dual-systems Model of Speech Category Learning. Biling (Camb Engl), 17(4), 709–728. DOI: 10.1017/S1366728913000783 [PubMed: 25264426]

Misaki M, Kim Y, Bandettini PA, & Kriegeskorte N (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage, 53(1), 103–118. DOI: 10.1016/j.neuroimage.2010.05.051 [PubMed: 20580933]

Mumford JA, Davis T, & Poldrack RA (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. Neuroimage, 103, 130–138. DOI: 10.1016/j.neuroimage.2014.09.026 [PubMed: 25241907]

Mumford JA, Turner BO, Ashby FG, & Poldrack RA (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage, 59(3), 2636–2643. DOI: 10.1016/j.neuroimage.2011.08.076 [PubMed: 21924359]

Myers EB (2014). Emergence of category-level sensitivities in non-native speech sound learning. Frontiers in Neuroscience, 8, 238. DOI: 10.3389/fnins.2014.00238 [PubMed: 25152708]

Myers EB, & Swan K (2012). Effects of category learning on neural sensitivity to non-native phonetic categories. Journal of Cognitive Neuroscience, 24(8), 1695–1708. DOI: 10.1162/jocn_a_00243 [PubMed: 22621261]

Nakahara H, Zhang LI, & Merzenich MM (2004). Specialization of primary auditory cortex processing by sound exposure in the "critical period". Proceedings of the National Academy of Sciences of the United States of America, 101(18), 7170–7174. DOI: 10.1073/pnas.0401196101 [PubMed: 15118079]

Perani D, & Abutalebi J (2005). The neural basis of first and second language processing. Current Opinion in Neurobiology, 15(2), 202–206. DOI: 10.1016/j.conb.2005.03.007 [PubMed: 15831403]

Perani D, Dehaene S, Grassi F, Cohen L, Cappa SF, Dupoux E, Fazio F, & Mehler J (1996). Brain processing of native and foreign languages. Neuroreport, 7(15–17), 2439–2444. DOI: 10.1097/00001756-199611040-00007 [PubMed: 8981399]

Pereira F, Mitchell T, & Botvinick M (2009). Machine learning classifiers and fMRI: a tutorial overview. Neuroimage, 45(1 Suppl), S199–209. DOI: 10.1016/j.neuroimage.2008.11.007 [PubMed: 19070668]

Perrachione TK, Lee J, Ha LY, & Wong PC (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. Journal of the Acoustical Society of America, 130(1), 461–472. DOI: 10.1121/1.3593366

Reetzke R, Xie Z, Llanos F, & Chandrasekaran B (2018). Tracing the Trajectory of Sensory Plasticity across Different Stages of Speech Learning in Adulthood. Current Biology, 28(9), 1419–1427 e1414. DOI: 10.1016/j.cub.2018.03.026 [PubMed: 29681473]

Reid A, Burnham D, Kasisopa B, Reilly R, Attina V, Rattanasone NX, & Best CT (2015). Perceptual assimilation of lexical tone: the roles of language experience and visual information. Atten Percept Psychophys, 77(2), 571–591. DOI: 10.3758/s13414-014-0791-3 [PubMed: 25465395]

Ripolles P, Marco-Pallares J, Alicart H, Tempelmann C, Rodriguez-Fornells A, & Noesselt T (2016). Intrinsic monitoring of learning success facilitates memory encoding via the activation of the SN/VTA-Hippocampal loop. Elife, 5. DOI: 10.7554/eLife.17441

Ripolles P, Marco-Pallares J, Hielscher U, Mestres-Misse A, Tempelmann C, Heinze HJ, Rodriguez-Fornells A, & Noesselt T (2014). The role of reward in word learning and its implications for language acquisition. Current Biology, 24(21), 2606–2611. DOI: 10.1016/j.cub.2014.09.044 [PubMed: 25447993]

Rolls ET, Joliot M, & Tzourio-Mazoyer N (2015). Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. Neuroimage, 122, 1–5. DOI: 10.1016/j.neuroimage.2015.07.075 [PubMed: 26241684]

Rosenberg MD, Finn ES, Scheinost D, Papademetris X, Shen X, Constable RT, & Chun MM (2016). A neuromarker of sustained attention from whole-brain functional connectivity. Nature Neuroscience, 19(1), 165–171. DOI: 10.1038/nn.4179 [PubMed: 26595653]

Sheppard JP, Wang JP, & Wong PC (2012). Large-scale cortical network properties predict future sound-to-word learning success. Journal of Cognitive Neuroscience, 24(5), 1087–1103. DOI: 10.1162/jocn_a_00210 [PubMed: 22360625]

Shtyrov Y (2012). Neural bases of rapid word learning. Neuroscientist, 18(4), 312–319. DOI: 10.1177/1073858411420299 [PubMed: 22020546]

Smialowski P, Frishman D, & Kramer S (2010). Pitfalls of supervised feature selection. Bioinformatics, 26(3), 440–443. DOI: 10.1093/bioinformatics/btp621 [PubMed: 19880370]

So CK, & Best CT (2010). Cross-language perception of non-native tonal contrasts: effects of native phonological and phonetic influences. Language and Speech, 53(Pt 2), 273–293. DOI: 10.1177/0023830909357156 [PubMed: 20583732]

Tang E, Mattar MG, Giusti C, Lydon-Staley DM, Thompson-Schill SL, & Bassett DS (2019). Effective learning is accompanied by high-dimensional and efficient representations of neural activity. Nature Neuroscience, 22(6), 1000–1009. DOI: 10.1038/s41593-019-0400-9 [PubMed: 31110323]

Ullman MT (2006). The declarative/procedural model and the shallow structure hypothesis. Applied Psycholinguistics, 27(1), 97–105. DOI: 10.1017/S014271640606019x

Vallabha GK, McClelland JL, Pons F, Werker JF, & Amano S (2007). Unsupervised learning of vowel categories from infant-directed speech. Proceedings of the National Academy of Sciences of the United States of America, 104(33), 13273–13278. DOI: 10.1073/pnas.0705369104 [PubMed: 17664424]

Wolosin SM, Zeithamova D, & Preston AR (2012). Reward modulation of hippocampal subfield activation during successful associative encoding and retrieval. Journal of Cognitive Neuroscience, 24(7), 1532–1547. DOI: 10.1162/jocn_a_00237 [PubMed: 22524296]

Wong FC, Chandrasekaran B, Garibaldi K, & Wong PC (2011). White matter anisotropy in the ventral language pathway predicts sound-to-word learning success. Journal of Neuroscience, 31(24), 8780–8785. DOI: 10.1523/JNEUROSCI.0999-11.2011 [PubMed: 21677162]

Wong PC, & Ettlinger M (2011). Predictors of spoken language learning. Journal of Communication Disorders, 44(5), 564–567. DOI: 10.1016/j.jcomdis.2011.04.003 [PubMed: 21601868]

Wong PC, Perrachione TK, & Parrish TB (2007). Neural characteristics of successful and less successful speech and word learning in adults. Human Brain Mapping, 28(10), 995–1006. DOI: 10.1002/hbm.20330 [PubMed: 17133399]

Wong PCM, & Perrachione TK (2007). Learning pitch patterns in lexical identification by native English-speaking adults. Applied Psycholinguistics, 28(4), 565–585. DOI: 10.1017/S0142716407070312

Yi HG, Maddox WT, Mumford JA, & Chandrasekaran B (2016). The Role of Corticostriatal Systems in Speech Category Learning. Cerebral Cortex, 26(4), 1409–1420. DOI: 10.1093/cercor/bhu236 [PubMed: 25331600]

Yip M (2012). Tone. Cambridge University Press. 10.1017/cbo9781139164559

Zhang Y, Kuhl PK, Imada T, Iverson P, Pruitt J, Stevens EB, Kawakatsu M, Tohkura Y, & Nemoto I (2009). Neural signatures of phonetic learning in adulthood: a magnetoencephalography study. Neuroimage, 46(1), 226–240. DOI: 10.1016/j.neuroimage.2009.01.028 [PubMed: 19457395]

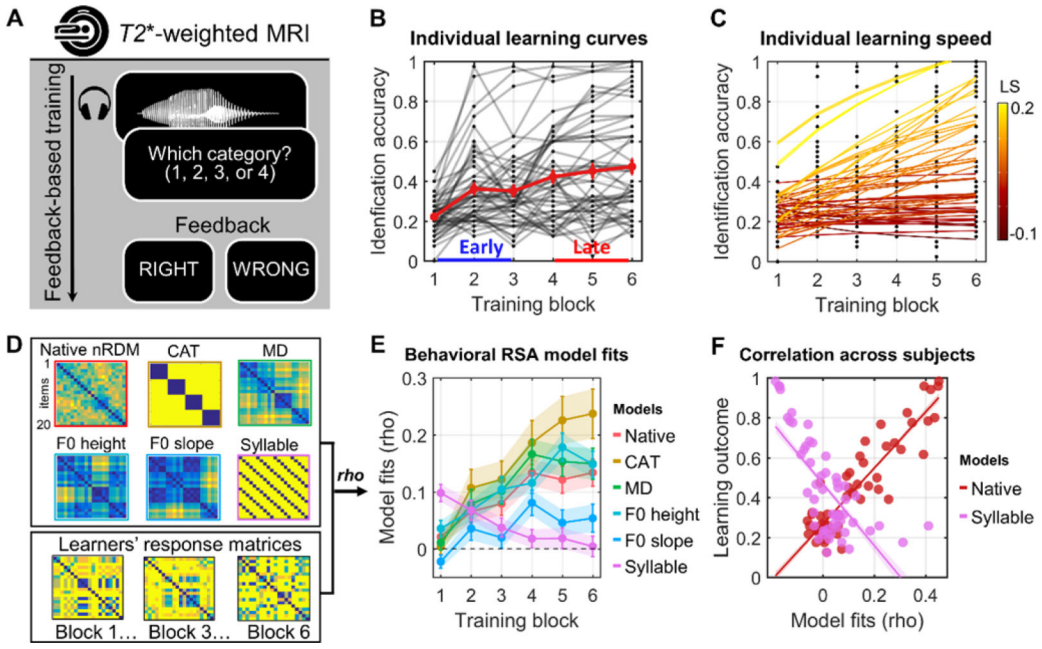

Figure 1.

behavioral tone-category training procedure, learning performance, and response-pattern modeling. **A**, feedback-based sound-to-category training procedure was used during MRI scanning for the learners. The native Mandarin listeners performed the same tone categorization task but without feedback (see Figure S1 for the experimental procedure). **B**, line graphs showing the group-level and individual learning curves across six training blocks. Early, the early phase of training; Late, the late phase. **C**, learning speed (LS) was estimated by fitting each learner's block-by-block accuracies with a power function. See Methods for the detailed learning curve modeling procedure. **D**, six predefined representational dissimilarity matrices (RDMs) were constructed for modeling learners' categorization response patterns using the behavioral representational similarity analysis (bRSA): Native nRDM = native listeners' neural RDM derived from a predefined brain mask; CAT = binary tone-category RDM; MD = multidimensional pitch RDM; F0 height = pitch height RDM; F0 slope = pitch direction RDM; Syllable = binary syllable-identity RDM; See Methods for the detailed RDM construction procedure. **E**, the bRSA reveals that native-similar tone-category-related information emerges following training, whereas task-unrelated segmental information decreases. Error bar: s.e.m. **F**, the model fits of the native nRDM (also other tone-category-related RDMs) are highly correlated with the learning outcome (red dots) and speed (not shown). In contrast, an inverse relationship was found between the Syllable model fits and learning outcome (pink dots).

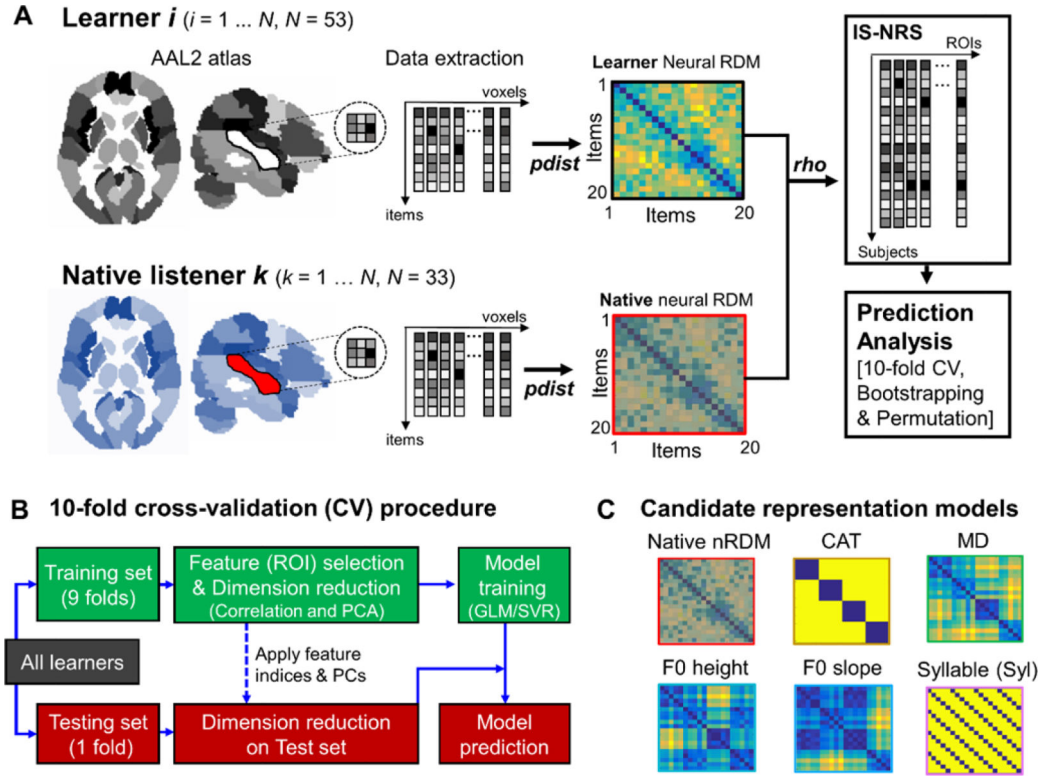

Figure 2.

inter-subject neural representational similarity (IS-NRS) analysis, predictive modeling procedure, and candidate representation models for RSA and prediction analysis. **A**, graphical illustration of the calculation of IS-NRS. Neural activation patterns were extracted from predefined ROIs based on the AAL2 atlas for both learners and native listeners. The neural RDMs (nRDMs) were calculated separately for each group. Each learner's nRDM was compared with every native listener's nRDM for each ROI. The IS-NRSs were then generated and used as predictive features to predict learning success (i.e., outcome and speed). **B**, the 10-fold cross-validation (CV) procedure for model construction and validation. All learners were split into 10 folds where 90% of the learners' data were used to train a GLM/SVR model and then the trained model was used to predict the left-over 10% of the learners and repeated 10 times. Predictive powers were estimated by calculating correlations between the predicted and observed learning performance. Permutation and bootstrapping procedures were used to determine the statistical significance and stability of the predictive powers. See Figure S4 for the overall data analysis pipeline. **C**, candidate representational models for the generation of neural representational predictive features. To compare the predictive powers of IS-NRS with those of other representational measures, the same prediction analyses were conducted with the predictors derived from the other five RDMs (i.e., CAT, MD, F0 height, F0 slope, and Syl).

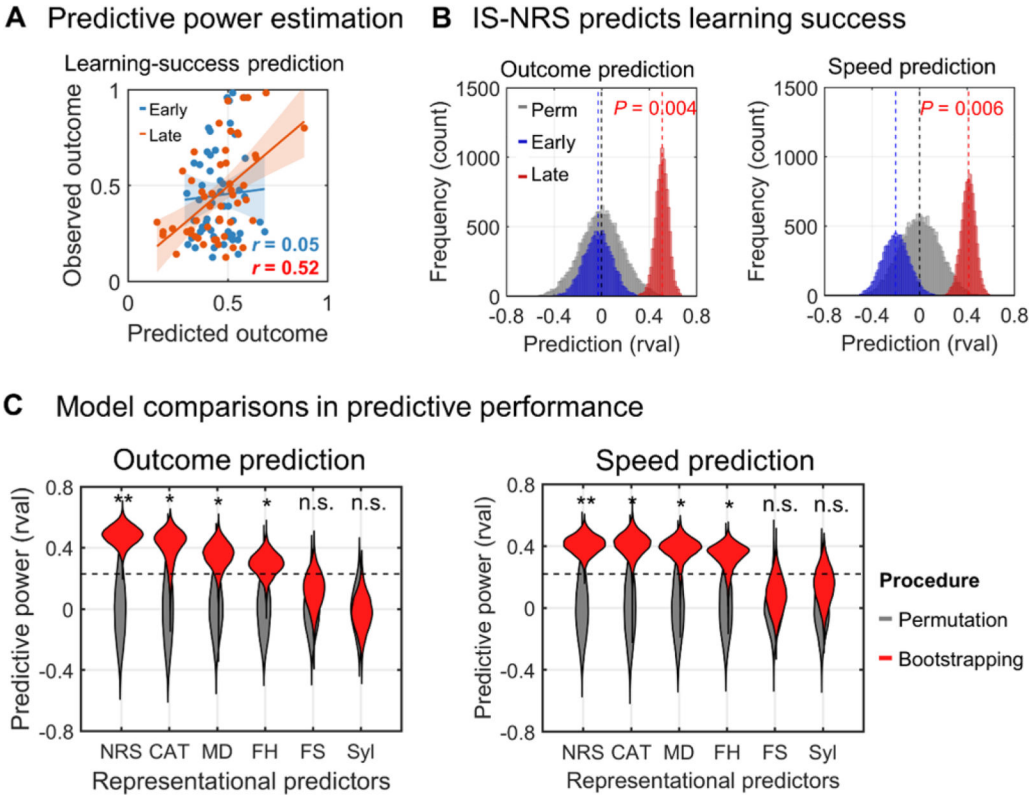

Figure 3.

predictive powers of the IS-NRS and other five model-based representational predictors. **A**, predictive powers were estimated based on the linear correlations between the predicted and observed learning scores. A representative scatter plot with linear fits showed strong predictive power in the late training phase instead of the early. **B**, IS-NRS predictive power distributions for the outcome and speed predictions for the early and late phases of training, respectively. Bootstrapping-based distributions were compared with the permutation-based (i.e., Prem) distributions to determine the statistical significance of the prediction models. Models only in the late phase revealed significant effects for both outcome and speed predictions. **C**, the IS-NRS showed more predictive power and prediction stability compared with the other five representational predictors. The dashed line indicates the 95th percentile of a permutation-based distribution. Representational predictors: NRS, native listeners' regional neural model, i.e., IS-NRS; CAT, tone-category model; FH, F0 height; FS, F0 slope; MD, multidimensional pitch model; Syl, syllable-identity model; permutation-based significance test: **, $P < 0.01$; *, $P < 0.05$; n.s., non-significant.

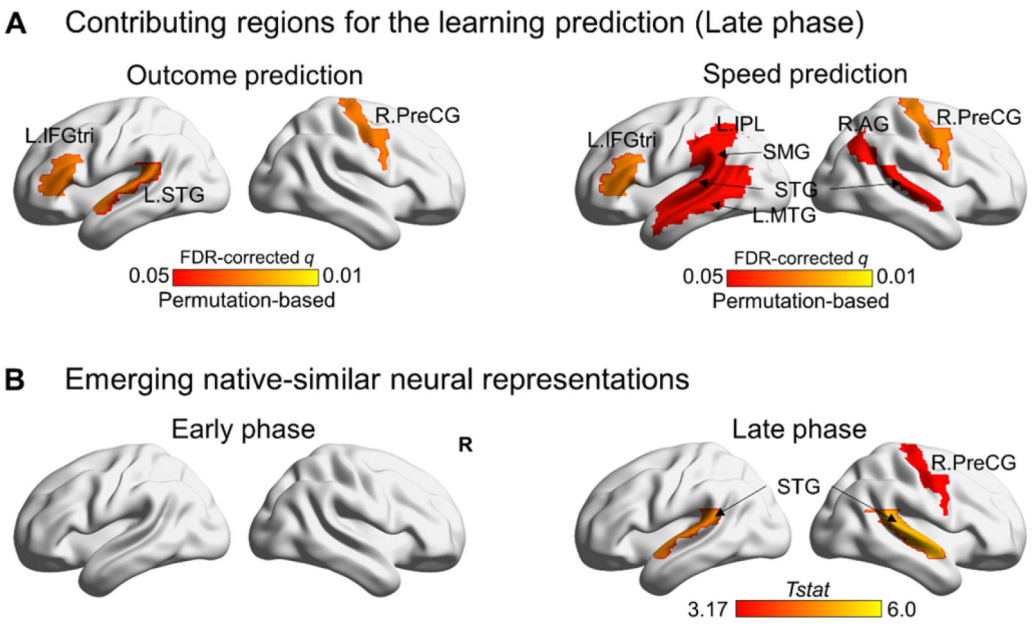

Figure 4.

The brain regions that significantly contributed to the predictive models and regions that showed significant emerging native-similar neural representations. **A**, regions significantly contributing to outcome and speed predictions in the late phase of training. Permutation-based FDR-corrected $q = 0.05$. ROI abbreviation: L.IFGtri, triangular part of left inferior frontal gyrus; L.IPL, left inferior parietal lobe; L.MTG, left middle temporal gyrus; STG, superior temporal gyrus; SMG, supramarginal gyrus; R.AG, right angular gyrus; R.PreCG, right precentral gyrus; L, left hemisphere; R, right hemisphere. **B**, brain regions that showed emerging native-similar neural representations in the late phase of training. FDR-corrected $q = 0.05$.

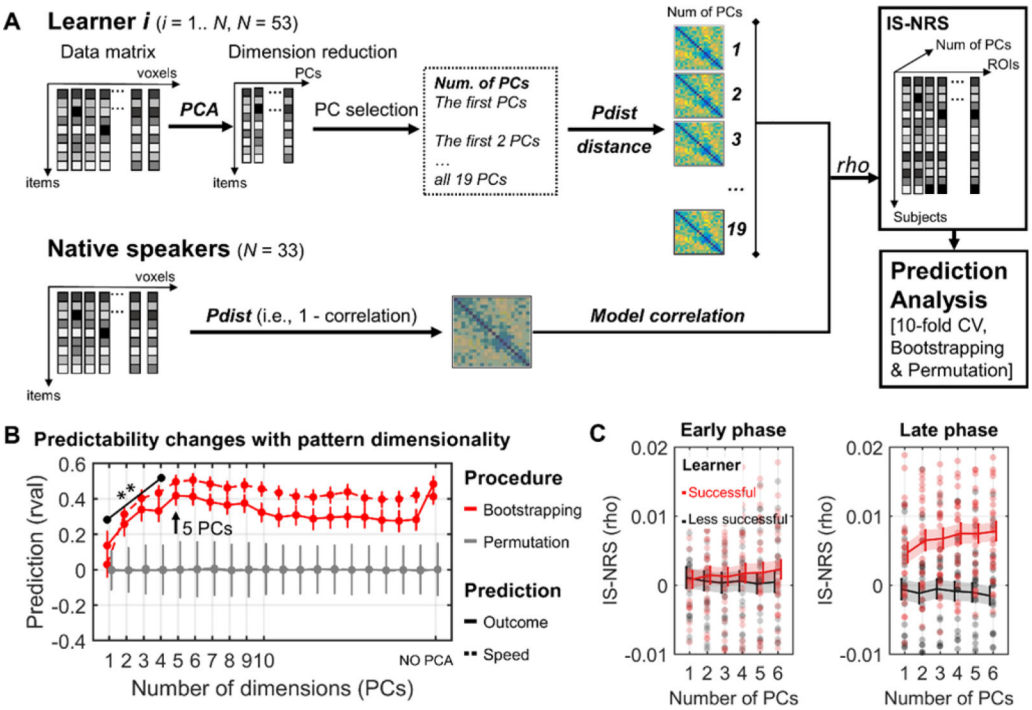

Figure 5.

moderate-to-high dimensionality of learners' native-similar neural representations best predicts individual learning success. **A**, the IS-NRS was recalculated with a dimensional decomposition procedure in which learners' activation patterns were decomposed into principal components (PCs). We constructed learners' nRDMs with different numbers of PCs (from 1 to *p*, *p* = number of PC). The non-native learners' dimension-constrained nRDMs were then correlated with native nRDM individually to calculate IS-NRS. These IS-NRSs were used to predict learning success with different numbers of PCs. **B**, predictive power reached a plateau with around five PCs (the black arrow). Predictive powers increased as a non-linear function of dimensionality. **, $P < 0.01$. **C**, group differences in IS-NRS across training phases. Learners were split into two groups, successful and less successful, based on the median of their outcomes. In the late phase of training, successful learners show more robust native-similar neural representations (i.e., IS-NRS) compared to less-successful learners. This group difference was more salient in the moderate-to-high dimensional space than that in the low-dimensional space.

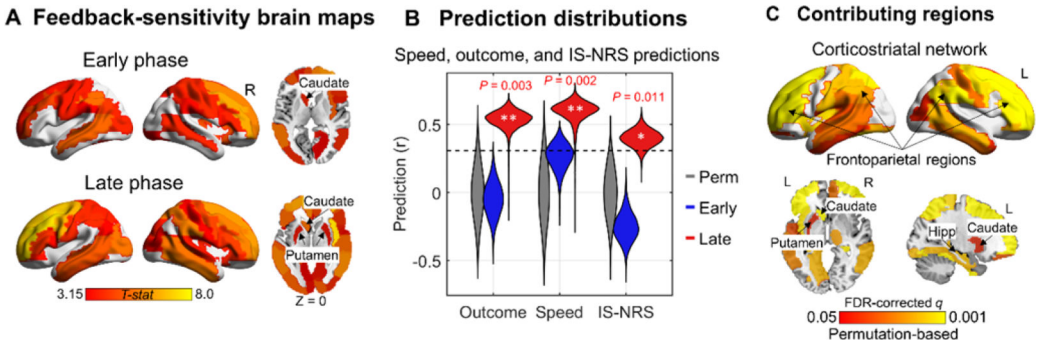

Figure 6.

Neural sensitivity to feedback valence predicts individual learning success and the degree of nativeness in neural representational structure in the late phase of training. **A**, Feedback-valence sensitivity brain maps for both early and late phases of training. Feedback-valence sensitivity was measured by the ROI-based multivariate feedback-type classification analysis. The group-level brain maps were thresholded at FDR-corrected $q = 0.05$. **B**, a violin graph shows the prediction distributions of predicting speed and outcome as well as the IS-NRS of L.IFGtri, bilateral STG, and R.PreCG. Neural feedback sensitivity in the late phase significantly predicts behavioral learning success and the robustness of learners' native-similar neural representations (i.e., IS-NRS). The dashed line indicates the 95th percentile of a permutation-based distribution. **C**, corticostriatal regions significantly contributed to the learning-outcome prediction mapping onto a brain template. The color bar indicates the significance (vs. permutation distributions) of an ROI in correlating the neural feedback sensitivity with the learning outcome, derived from the feature selection and permutation procedures. Permutation-based FDR-corrected $q = 0.05$. ROI abbreviation: Hipp, hippocampus; L, left hemisphere; R, right hemisphere.