Published in final edited form as:
Neuroimage. 2021 January 01; 224: 117410. doi:10.1016/j.neuroimage.2020.117410.

A distributed dynamic brain network mediates linguistic tone representation and categorization

Gangyi Feng[a,b,*], **Zhenzhong Gan**[c], **Fernando Llanos**[d], **Danting Meng**[c], **Suiping Wang**[c,e], **Patrick C.M. Wong**[a,b], **Bharath Chandrasekaran**[d,*]

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

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

[c]Center for the Study of Applied Psychology and School of Psychology, South China Normal University, Guangzhou 510631, China

[d]Department of Communication Science and Disorders, School of Health and Rehabilitation Sciences, University of Pittsburgh, Pittsburgh, PA 15260, United States

[e]Guangdong Provincial Key Laboratory of Mental Health and Cognitive Science, South China Normal University, Guangzhou 510631, China

Abstract

Successful categorization requires listeners to represent the incoming sensory information, resolve the "blooming, buzzing confusion" inherent to noisy sensory signals, and leverage the accumulated evidence towards making a decision. Despite decades of intense debate, the neural systems underlying speech categorization remain unresolved. Here we assessed the neural representation and categorization of lexical tones by native Mandarin speakers ($N = 31$) across a range of acoustic and contextual variabilities (talkers, perceptual saliences, and stimuluscontexts) using functional magnetic imaging (fMRI) and an evidence accumulation model of decision-making. Univariate activation and multivariate pattern analyses reveal that the acoustic-variability-tolerant representations of tone category are observed within the middle portion of the left superior temporal gyrus (STG). Activation patterns in the frontal and parietal regions also contained category-relevant information that was differentially sensitive to various forms of variability. The robustness of neural representations of tone category in a distributed fronto-temporoparietal network is associated with trial-by-trial decision-making parameters. These findings support a

[*]Corresponding authors. g.feng@cuhk.edu.hk (G. Feng), b.chandra@pitt.edu (B. Chandrasekaran).

CRediT authorship contribution statement
Gangyi Feng: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Zhenzhong Gan:** Investigation. **Fernando Llanos:** Software, Resources, Writing review & editing. **Danting Meng:** Investigation. **Suiping Wang:** Resources, Supervision, Project administration. **Patrick C.M. Wong:** Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Bharath Chandrasekaran:** Conceptualization, Resources, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition.

Supplementary materials
Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117410.

hybrid model involving a representational core within the STG that operates dynamically within an extensive frontoparietal network to support the representation and categorization of linguistic pitch patterns.

Keywords

Categorization decision; Lexical tones; Neural decoding; Neural representation; Perceptual constancy; Speech categorization

1. Introduction

Categorization involves mapping continuous and highly variable sensory information into discrete behavioral classes. Speech perception is construed as a particularly challenging categorization problem (James, 1890; Diehl et al., 2004; Holt and Lotto, 2010; Liberman et al., 1967). Neural systems mediating speech perception are tasked with mapping continuous, variable, and noisy acoustic signals to discrete long-term stored representations of speech categories. The neural mechanisms underlying the representation and categorization of speech categories are unresolved and will be the focus of this study. We examine the neural regions that robustly represent speech category information in the face of various forms of sensory-perceptual variability. We then assess the extent to which the robustness of neural category representation relates to the trial-by-trial fluctuations of categorization decision, modeled by a single-trial evidence accumulation model. Our goal is to provide an integrative account of how the human brain extracts relevant speech category information from noisy and variable acoustic signals and accumulates this information in making efficient category decisions.

Auditory associative regions within the middle STG are hypothesized to be functionally-specialized in representing familiar auditory categories, including speech (Arsenault and Buchsbaum, 2015; Chevillet et al., 2013; Desai et al., 2008; Feng et al., 2018; Formisano et al., 2008; Liebenthal et al., 2010; Xin et al., 2019; Yi et al., 2019). While most neural models attribute a key role for the STG in speech categorization, there is increasing evidence that speech category information is also present in the frontal and parietal regions (Cheung et al., 2016; Correia et al., 2015; Du et al., 2014; Evans and Davis, 2015; Myers et al., 2009; Raizada and Poldrack, 2007). An emerging hypothesis is that the STG may be a functionally specialized ‘core’ region that subserves speech representation and categorization in optimal listening environments; under less supportive (e.g., noisy or greater perceptual confusability) listening conditions, achieving perceptual constancy and categorization involve frontoparietal networks in addition to the STG (Alain et al., 2018; Alavash et al., 2019; Feng et al., 2018).

Current models have mostly focused on the neural substrates underlying the perception of segmental units (consonants and vowels) in speech (e.g., Du et al., 2014; Yi et al., 2019). In the current study, we assess the neural mechanisms underlying the representation and categorization of a critical speech feature in tone languages —lexical pitch patterns. In tone languages, pitch contours play a similar role as consonants and vowels in altering lexical or word meaning (Yip, 2002). Extracting pitch contours from the incoming speech stream and mapping key pitch features to tone categories are critical for speech communication in a tone language. Prior neuroimaging studies have shown that lexical tone perception involves activation of a network across inferiorfrontal, precentral, inferior-parietal, and temporal cortices (Feng et al., 2018; Liang and Du, 2018; Si et al., 2017; Zatorre and Gandour, 2008). Regions within the bilateral STG and left inferior parietal lobule (IPL) have been proposed to be functionally specialized in the multidimensional representations of lexical tone categories (Feng et al., 2018). Inferior frontal regions contained limited information regarding tone categories but were more active for tones that are perceptually confusable (Feng et al., 2018). To our knowledge, no study has systematically examined the robustness of lexical-tone categories in the STG and frontoparietal regions across various forms of acoustic and perceptual variability. Further, the role of the neural representations within the distributed frontoparietal-STG regions in mediating lexical tone-category decision is unresolved.

Here we conducted functional magnetic resonance imaging (fMRI) while native Mandarin speakers listened to various acoustic exemplars of Mandarin lexical tones and categorized them based on their pitch patterns. The stimuli were designed to introduce various forms of acoustic variability: talker (male vs. female fundamental frequencies), a range of pitch saliency (from noisy to robust pitch patterns), and stimulus context (speech vs. non-speech exemplars). In the speech conditions, the pitch contours were produced by male and female talkers; in the non-speech conditions, the pitch contours were iterative ripple noise (IRN) analogs of the pitch contours produced by the talkers. IRN stimuli have been extensively used in auditory neuroscience research focused on pitch processing. These stimuli afforded us a unique opportunity in parametrically varying pitch perceptual salience. Increasing iteration steps systematically increases the temporal regularity that is generated by broadband noise. Higher iteration steps correspond to systematic increases in pitch saliency and robustness in the representation of lexical pitch contours at the level of the auditory sub-cortex (Krishnan et al., 2010). The continuum of IRN steps (2–32) allowed us to examine the robustness of neural representations of tone categories in the face of varying pitch saliency. Importantly, while IRN stimuli faithfully retain pitch, they are less confounded by variability associated with waveform periodicity and segmental cues inherent to speech signals, thus providing us with a well-controlled non-speech context, devoid of lexical-semantic confounds. In addition to examining univariate activations across the various conditions, we used multivariate representational similarity analysis (RSA) to examine the variability-tolerant core neural representation of tone category by controlling for variance across acoustic (talker variability and pitch salience) and contextual (speech vs. non-speech) factors. We predict that core regions mediating tone perceptual constancy would be highly sensitive to between-category changes while tolerant of different forms of within-category variability. We further examined to what extent the robustness of neural representations of tone category was influenced by the different forms of variability using machine-learning classification (MVPC) with various cross-validation procedures.

Finally, we examined to what extent the neural representations contribute to the categorization decision by examining the association between trial-by-trial robustness of neural category representations and decision-making components that are involved in the accumulation of sensory evidence in support of categorization decisions. Previous studies have demonstrated that a distributed frontoparietal network subserves efficient decision-making across different modalities and tasks (Busemeyer et al., 2019; Frank et al., 2015; Mulder et al., 2014; van Maanen et al., 2011). Most of these studies investigated the association between univariate activations (i.e., blood-oxygen-level-dependent [BOLD] responses) and decision-making parameters across trials or subjects. While some studies revealed positive correlations, others have found negative correlations even for the same decision parameter (Heekeren et al., 2004; Ho et al., 2009; Mulder et al., 2014; Noppeney et al., 2010). This inconsistency may be due to the differences in task used and the nature of the task-induced BOLD responses between studies. In contrast to prior studies, here we used a multivariate neural category index (NCI) that reflects the amount of category information represented in the neural patterns at the single-trial level to assess the extent to which the robustness of the representations relates to categorization decision. We test a hypothesis that the robustness of neural representations of tone category, as reflected by the NCI metric contributes to online categorization decisions. To estimate latent decision components, we employed a widely adopted, parsimonious response-choice model: the linear ballistic accumulator (LBA) (Brown and Heathcote, 2008). LBA provides a simple computational framework to tease apart different cognitive processes underlying decision-making. These cognitive processes are modeled with a fixed set of parameters (e.g., starting point and accumulation rate) that can serve as an explainable latent middle ground between the observed behavioral data (i.e., accuracy and reaction time) and its underlying neuronal processing (Forstmann et al., 2011; Mulder et al., 2014). We assessed the relationship between decision-making parameters and the NCI to examine how the neural representations relate to the latent categorization decision components. Using converging behavioral, computational modeling, and neuroimaging approaches, we examine the extent to which tone-category representation and decision are achieved by a functionally-specialized representational ‘core’ within the STG and a distributed frontoparietal network.

2. Material and methods

2.1. Participants

Native speakers of Mandarin ($N = 31$, 14 males; right-handed; age = 20.9 ± 2.3 [mean ± SD] years) were recruited from the neighboring communities of South China Normal University to participate in this study. We selected candidates who were originally from the middle and north of China because dialects in those areas (e.g., Beijing, Hebei, Liaoning, Tianjin, etc.) are very close to standard Mandarin. These dialects have the same tonal inventory as the Mandarin. We excluded candidates who can speak Cantonese, Teochew, and other southern Min dialects. All participants demonstrated high proficiency in spoken standard Mandarin (higher than second-class upper-level on a standardized spoken Mandarin proficiency test []). Participants reported no neurological or hearing-related impairment and had normal or corrected to normal vision. All participants signed written informed consent approved by the Institutional Review Board of School of Psychology at South China Normal University and The Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee. They were monetarily compensated for their participation.

2.2. Stimulus construction

Two native Mandarin speakers (one female) produced speech sounds of three tones (i.e., T1: high-flat tone; T2: low-rising tone; T4: high-falling tone) in the syllable context of /di/ (see Fig. 1A for spectrograms of the sample stimuli). We used the three Mandarin tones to reduce the listener fatigue and confusion between the dipping tone (i.e., T3) and the other tones (Xie et al., 2018). The stimuli were recorded using 16-bit quantization and a 44.1-kHz sampling rate in a sound-treated booth. The speech stimuli were normalized to the same duration (300 ms) and intensity (72 dB). The non-speech iterated ripple noise (IRN) homologs of the three Mandarin tones were then synthesized with broadband noise using a delay-and-add procedure described by Swaminathan et al. (2008). The fundamental frequency (F0) of the IRN stimuli mimicked natural citation-form productions of the three tones (Swaminathan et al., 2008). The procedure is briefly described as follows. A polynomial was interpolated across five F0 values, equally spaced in time, of the estimated contour derived from the citation-form speech production. The resulting polynomial was used to create IRNs of the corresponding speech utterance with a different number of iterations (i.e., 2, 4, 8, 16, and 32). An increasing number of iterations is associated with increased pitch perceptual salience. Therefore, stimuli of six conditions (i.e., Speech, IRN 2, 4, 8, 16, and 32) were constructed for the fMRI experiment where each condition consists of a female and a male version of the sounds.

2.3. Experimental procedure

We conducted functional magnetic resonance imaging (fMRI) while participants performed a tone categorization task. Participants were instructed to listen to sounds and categorize them into one of the three categories based on their pitch patterns by pressing a “1”, “2”, or “4” button, which corresponded to their left middle, index, and right middle fingers (or right middle, index, and left middle fingers), respectively. Therefore, fingers and hands were not only counterbalanced across participants but also within each participant (i.e., the pair of tone categories 1 and 2 shared the same hand but differed in fingers, while the pair of tone categories 1 and 4 shared the same finger but differed in hands). Participants practiced before scanning to establish the category-response mapping. Auditory stimuli were presented and controlled using E-Prime (Psychology Software Tools, Inc.; version 2.0). The stimulus presentation schema and fMRI acquisition procedure are described in Fig. 1B. To reduce the interference of scanner noise on auditory perception, a customized sparse-sampling fMRI sequence was employed, during which stimuli were presented within an 800-ms silent interval between each of the imaging acquisitions (Feng et al., 2018). The onset of each trial was synchronized to the onset of each image acquisition to ensure the stimuli were presented during the silence gap. To minimize the forward masking effect induced by the scanner noise, we presented a sound stimulus 100 ms after each of the 1700-ms imaging acquisition. The stimulus set consists of a speech and five non-speech IRN conditions (i.e., Speech, IRN 2, 4, 8, 16, and 32). Each condition reflects F0 contours (original and modeled) from male and female talkers, resulting in 36 unique stimuli items (six conditions, two talkers [male and female], and three tones). In each block (i.e., an fMRI run or session), we presented the 36 stimuli four times in random order controlled by E-prime. Participants categorized these stimuli across six blocks resulting in 864 trials per subject in total. Therefore, there were 24 repetitions per sound item during the fMRI experiment. To efficiently estimate the hemodynamic response to each stimulus item, 40 null trials (i.e., silence trial) were randomly inserted between sound trials as jittered intertrial intervals in each block. The above randomization procedure ensures that the stimulus presentation order was different between blocks and participants, while the silent trials were randomly distributed within each block. For each trial, the participant's categorization response and reaction time (RT) were recorded.

2.4. MRI data acquisition

All MRI data were acquired using a Siemens 3-Tesla Tim Trio MRI system with a 12-channel head coil in the Brain Imaging Center at South China Normal University. Functional MRI images were acquired with the T2*-weighted gradient echo-planar imaging pulse sequence using the following parameters: repetition time (TR) = 2500 ms with 800-ms silence gap, TE = 30 ms, flip angle = 90°, 31 slices, field of view = 224 × 224 mm, in-plane resolution = 3.5 × 3.5 mm, slice thickness = 3.5 mm with 1.1 mm gap. T1-weighted high-resolution structural images were acquired using a magnetization prepared rapid acquisition gradient echo sequence (176 slices, TR = 1900 ms, TE = 2.53 ms, flip angle = 9°, voxel size = 1 × 1 × 1 mm).

2.5. MRI data preprocessing

MRI data were preprocessed using SPM12 (Wellcome Department of Imaging Neuroscience; www.fil.ion.ucl.ac.uk/spm/). For univariate activation analysis, raw functional images were corrected for head movement using a least-squares approach and a six-parameters (rigid body) spatial transformation (Friston et al., 1995). A two pass procedure was used to spatially register all the images to the mean of the images after the first realignment (i.e., the register to mean approach). The slice-time correction was not implemented. The high-resolution T1 image was then co-registered with the mean functional image (i.e., reference image) using the Normalized Mutual Information algorithm (Separation = [4 2]; Histogram Smoothing = [7 7]) (Studholme et al., 1999). The co-registered T1 image was processed with the unified segmentation procedure (Ashburner and Friston, 2005). The deformation fields estimated in the segmentation procedure were used for normalization by converting the realigned functional images in native space to the Montreal Neurological Institute (MNI) space. The normalized functional images were resampled to 3 mm^3 voxel size and smoothed with a Gaussian kernel of 6-mm full width at half maximum. For the multivariate pattern analysis, the preprocessing steps for the functional images included head movement correction and co-registration but without normalization and smoothing.

2.6. Univariate activation analysis

Univariate voxel-wise activation analysis was conducted with the general linear model (GLM) to identify brain regions that are commonly responsive to tone categorization across talkers (female vs. male), pitch salience (different IRN iterations), and stimulus contexts (speech vs. non-speech) and to examine brain responses that are modulated by the three factors. For the subject-level analysis, a GLM with a design matrix including 12 sound regressors of interest (i.e., two talkers by six stimulus conditions [Speech, IRN, 2, 4, 8, 16, and 32]) was constructed for each participant. Trials of incorrect or missing responses were put into a controlled regressor. The silent trials were considered as a part of the model baseline and not modeled explicitly in the GLM. These regressors corresponding to the onset of each trial were convolved with the canonical hemodynamic response function. Low-frequency drifts were removed by a temporal high-pass filter (cutoff at 128 s). The AR(1) approach was used for autocorrelation correction. The six parameters derived from head movement correction and the session mean were also added into the GLM model as nuisance regressors. The gray-matter image generated from the segmentation step was converted as a binary inclusive mask for each participant to restrict voxels of interest.

For the group-level analysis, an overall categorization-related brain activation map (i.e., all sounds vs. baseline) and four conjunctive brain maps were generated. The conjunctive maps were computed by using a conjunction analysis procedure to identify regions that demonstrated significant overlapping activations across talkers (talker-general regions), pitch salience (salience-general regions), stimulus contexts (context-general regions), and all forms of variability (variability-general regions). For example, to identify the talker-general regions, brain maps of female talker's items and male talker's items were generated separately with voxel-level $P = 0.001$. Overlapping regions that survived with a cluster-level threshold (family-wise error [FWE] rate = 0.05) were considered as talker-general categorization regions. In addition to the conjoint activations across these factors, activations modulated by the talker variability (male vs. female talkers), pitch salience, and stimulus context (speech vs. non-speech) were examined separately. Changes in activation as a function of pitch salience were assessed using parametric modulation analysis (described in the next section). Contrast images from the subject-level analysis were entered into the group-level one-sample *t*-test or repeated measures one-way analysis of variance (ANOVA). Brain maps were initially thresholded at voxel-level $P = 0.001$, and all reported brain areas were corrected at the cluster-level $P = 0.05$ using the FWE rate approach implemented in the SPM package.

2.7. Parametric modulation analysis in pitch salience

The relationship between the pitch salience (i.e., IRN iteration steps) and the level of brain activation was examined using the trial-by-trial parametric modulation analysis (Buchel et al., 1996). We aimed to identify brain regions that systematically vary with pitch salience. To this end, in the subject-level GLM analysis, we constructed a design matrix with a parametric modulation regressor in which the weights were coded as a linear function (i.e., −2, −1, 0, 1, 2) for IRN 2, 4, 8, 16, and 32 trials, respectively. Quadratic function (i.e., mean-centered [2, 4, 8, 16, 32]) was used in a separate parametric analysis. A speech condition regressor, a button-press regressor, and six head movement parameters were included as nuisance regressors in the design matrix to control for the related effects. At the group-level analysis, a one-sample *t*-test was used to identify significant voxels that parametrically vary with pitch salience.

2.8. Multivariate pattern analysis (MVPA)

Firstly, to identify the core neural representations of tone category across different forms of variability (i.e., talker, pitch salience, and stimulus context), we used representational similarity analysis (RSA) (Kriegeskorte and Kievit, 2013; Kriegeskorte et al., 2008) while controlling for the variance of acoustic, perceptual, and contextual variabilities. Secondly, to examine to what extent these factors influence the robustness of the representation, multivariate pattern classification (MVPC) analysis was used in the combination of different cross-validation (CV) procedures. The searchlight-based (Kriegeskorte et al., 2006) RSA and MVPC were conducted. Detailed methodological steps consisting of the construction of the RSA models, feature space (feature construction and extraction), and CV procedures are described in the next section.

2.9. Representational similarity analysis (RSA)

The RSA analysis was used to reveal the variability-tolerant neural representations of tone category by controlling for the variance of other acoustic and non-acoustic variables (i.e., variations in pitch height, pitch direction, pitch salience, talker, stimulus context, and motor response). Seven stimulus-derived and behavior-derived representational dissimilarity matrices (RDMs) were constructed according to discrete tone category, fundamental frequency (F0) height (i.e., pitch height), F0 slope (i.e., pitch direction), talker (i.e., female and male), stimulus context (i.e., speech and non-speech stimuli), stimulus condition (i.e., IRN 2, 4, 8, 16, 32 and speech conditions), and motor response (i.e., hands and fingers for button presses), respectively (see Fig. S1 in Supplementary Materials for graphical illustrations). The construction procedure of the RDMs has been described in previous studies (Feng et al., 2018). The procedure was briefly described below. The tone-category RDM was constructed based on the combinations of the three tones (i.e., 0 s for sound pairs of the same tone, 1 s for sound pairs of different tones). The stimulus-context, stimulus-condition, and talker RDMs were constructed with the same procedure. To construct the F0-height RDM, the standardized Euclidean distance was calculated between each pair of sounds according to their mean F0. The F0 slope was accessed by estimating a simple regression slope based on the F0 time-varying pattern of each sound using a linear fitting function in Mathlab (R2016a). The F0-slope RDM was created by calculating the differences in F0-slope estimates between each pair of sounds. The two F0 RDMs were normalized by scaling the distance or difference values between 0 (low dissimilarity, i.e., close in the distance) and 1 (high dissimilarity, i.e., far from each other in the distance). To create the motor-response RDM, we first constructed two vectors representing hands and fingers used for responses. We coded 1 s for the sounds using the left hand while coded 2 s for the sounds using the right hand. The same procedure was used to generate the finger vector (i.e., 1 s for the middle finger and 2 s for the index finger). We created the motor-response RDM by estimating the weighted Euclidean distances between each pair of sounds based on the hand (weighted score = 1) and finger (weighted score = 0.5) vectors using the 'pdist' function. The unweighted motor-response RDM was also created for comparison. A certain degree of correlation between these RDMs was found. Spearman's rank correlations between the tone-category and F0-slope RDMs ($r = 0.80$), between the tone-category and (weighted) motor-response RDMs ($r = 0.83$), and between the F0-height and talker RDMs ($r = 0.87$) were relatively higher compared to that of other pairs of RDMs (see Fig. S1 for more details).

The searchlight-based RSA analyses were conducted for each subject on the functional images following realignment but without normalization or smoothing. Firstly, sound-induced item-based statistical brain activation maps were generated with the GLM approach for each subject for the RSA. The unsmoothed images in the subject's native space were analyzed using a GLM with individual regressors for each item (i.e., collapsed the same item across repetitions within a block) to calculate single-item *t*-statistic maps for each block, while other items were pooled into a regressor of non-interest. This Least Squares Single (LSS) approach was designed to model brain activities for each item while controlling for the variance of other covariant items in the same block (Mumford et al., 2014, 2012). Specifically, for each sound item, a design matrix was constructed with a regressor of interest for that item; a regressor of non-interest consisted of other items, six head movement regressors, and a session mean regressor for each block individually. Therefore, 216 subject-level GLM models (36 unique items per block; six blocks in total) were constructed and estimated for each subject. The *t*-statistic image was calculated for each item by contrasting the item regressor with the baseline and further used for the RSA and MVPC analyses. The *t*-statistic was used because it combines the effect size weighted by error variance; therefore *t*-statistic is less affected by highly variable item estimates than that of the beta estimation (Misaki et al., 2010).

Secondly, the searchlight algorithm (Kriegeskorte et al., 2006) implemented in the CoSMoMVPA toolbox (Oosterhof et al., 2016) was used to identify brain areas where their neural RDMs were correlated with the tone-category RDM. The searchlight analysis was restricted to broad brain regions of interest using a mask (see Fig. S2 for the brain mask, Supplementary Materials) generated by a meta-analysis from Neurosynth.org (http://neurosynth.org/). We aimed to include all possible brain regions that relate to auditory and speech perception. Thus, we searched the Neurosynth topic dataset with keywords "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, categorization, and so on. The resulting "auditory-perception" brain map includes distributed fronto-temporoparietal regions, consisting of the bilateral inferior frontal gyrus (IFG), insula, middle frontal gyrus (MFG), precentral gyrus (preCG), inferior parietal lobule (IPL), superior temporal gyrus (STG), superior temporal sulcus (STS), middle temporal gyrus (MTG), and supplementary motor areas (SMA). Activation clusters less than 80 voxels (corresponding to the average searchlight sphere) in the brain mask (3 mm^3 voxel size) were removed for the searchlight-based analyses.

For the searchlight-based RSA, at each voxel, sound-induced activation patterns (i.e., *t* statistic values) within each spherical searchlight (three-voxel-radius sphere) were extracted for all items to generate a neural representational dissimilarity matrix (nRDM) by calculating the dissimilarity (i.e., 1 - Pearson's correlation) between each pair of items (i.e., a 36-by-36 matrix). Different spherical sizes (e.g., four-voxel-radius sphere) were also tested to ensure the RSA results were not significantly different between the size chosen. The tone-category RDM was then correlated with the nRDM (both RDMs were first vectorized) for each spherical searchlight by using Spearman's rank correlation. The correlation values were normalized using Fisher's *r*-to-*z* transformation.

For the partial RSA analysis, the unique contribution of the tone-category RDM in explaining the variance of the nRDM was examined by correlating the tone-category RDM with the nRDMs while controlling for the variances of other RDMs (i.e., F0 slope, F0 height, talker, stimulus type, stimulus condition, and motor response) using the partial Spearman's rank correlation. This analysis has been previously used to separate different contributing factors of the neural representations (Cichy et al., 2019; Feng et al., 2018; Kriegeskorte and Kievit, 2013). All the RDMs were first vectorized using the 'squareform' function. The rank partial correlation coefficients were calculated between the tone-category RDM and nRDMs while the other vectorized RDMs were controlled using the "partialcorr" function in MATLAB (R2016a). This partial RSA approach can reveal the neural representations of tone categories that are linearly independent of the experimenter-induced acoustic and contextual variabilities. For the group-level analysis, the individual RSA maps in the native space were first normalized to MNI space and then fed to a one-sample *t*-test against chance.

2.10. Multivariate pattern classification (MVPC) analysis

The searchlight-based MVPC of tone category was employed with different CV procedures to examine variability-tolerant neural representations of tone category and access the extent to which acoustic, perceptual, and contextual variables modulate the robustness of the neural representations of category information. We operationally define the variability-tolerant neural representation of tone category as significantly above-chance tone classification performances that emerge from multivoxel activation patterns across repetitions (i.e., cross-block), acoustic variants (i.e., cross-talker and cross-IRN), and stimulus contexts (i.e., cross-stimulus-type). Therefore, four leave-one-X-out (X denotes block, talker, IRN step, and stimulus context, respectively) CV procedures were used to establish classifier generalizability and estimate the robustness of variability-tolerant neural representation of tone category. The leave-one-block-out (i.e., cross-block) CV procedure was used to gain the overall effect of the neural representations of tone category, testing generalizability of the trained classifier across item repetitions (i.e., overall tone decoding performance). The leave-one-talker-out (i.e., cross-talker) CV procedure was used to identify talker-general neural representations. Classifiers were trained on items from a male talker and subsequently tested on the items from a female talker, and vice versa. The leave-one-IRN-step-out (i.e., cross-IRN) CV procedure was used to identify the neural representations that are tolerant of variability in perceptual pitch salience. The final approach was the leave-one-stimulus-context-out (i.e., cross-stimulus-type) CV procedure, wherein the classifier was trained on data from the speech items and subsequently tested on the non-speech (IRN) items, and vice versa. Thus, only the tone-category information general across item variants (i.e., talker, pitch salience, or stimulus context, respectively) was informative to the classifier. For more rigorous MVPC analyses, we combined the cross-block CV procedure with each of the other three CV procedures (e.g., "cross-block&talker" CV: the trials for classifier training and testing are differed both in block and talker) to further demonstrate the extent to which these acoustic and contextual variabilities modulate the robustness of neural representations of tone-category respectively.

For the searchlight MVPC analysis, a $V \times I \times B$ matrix was generated for each spherical searchlight, where V referred to the number of voxels, I referred to the number of items and B referred to the number of blocks (e.g., 80 × 36 × 6). This matrix was entered into a linear support vector machine (SVM) classifier implemented in the LIBSVM toolbox (Chang and Lin, 2011) for training and testing with the CV procedures, respectively. The mean classification accuracy was calculated and mapped back to the voxel at the center of each sphere. The same classification procedure was conducted across all voxels of interest and generated classification accuracy maps for each subject. The classification accuracy maps were contrasted with chance accuracy maps that were generated by the same MVPC analysis with a permutation procedure, in which the tone category labels were shuffled independently for each subject. For the group-level analysis, the resulting MVPC maps were first normalized to the MNI space and then fed into a one-sample t-test. All group statistical maps from the multivariate analyses were thresholded at the voxel-level $P = 0.001$, with cluster-level FWE-corrected $P = 0.05$.

2.11. Model-based estimation of categorization decision variables

The Linear Ballistic Accumulator (LBA) model (Brown and Heathcote, 2008; Donkin et al., 2011) was employed to examine decision-making components underlying tone categorization that is not directly observable with raw reaction times (RTs) or accuracies and to examine whether the robustness of the neural representations of tone category relates to the trial-by-trial fluctuations of the decision parameters.

For each categorization, the LBA model defines three processing components (see Fig. 5A for schematic illustration), including sensory (yellow), decision (white), and motor (gray) processes. The stimulus initially undergoes sensory processing and the tone-category information (i.e., evidence) is accumulated toward one of the possible stimulus-response mappings until the response threshold (b) is reached. In the LBA modeling, the decision (reaction) time of a choice is modeled as an evidence accumulation process. The accumulation process terminates when the amount of evidence accumulated in support of given a choice (e.g., the low-rising tone) reaches the response threshold before other competing choices (e.g., the high-falling tone). The LBA has four parameters that are linked to categorization processes: the initial amount of evidence in support of a choice selection (i.e., starting point [SP]); the amount of evidence needed from the SP to reach the response threshold (i.e., response caution [b-SP]); the rate at which evidence accumulates (i.e., accumulation or drift rate [AR]); and the processing time t_0 spent in non-decisional processing (e.g., sensory encoding and motor response). The SP and AR characterize two critical aspects of the decision-making process reflecting the amount of category information and information accumulation processing, which we hypothesized to be associated with the multivoxel representations of tone category information (Fig. 5A). Therefore, we focused on the two decision components in the current study, corresponding to the parameter A and v in the model. The LBA model assumes the SP of each accumulator has a uniform distribution between 0 and an upper limit A. The ARs are normally distributed across trials with a mean v and variance sv. In Fig. 5A, the two red lines denote the evidence accumulation process at different rates. The red dash line denotes a higher AR than the solid line, while they have different SP (green lines). A higher SP (the green dash line in Fig. 5A) is typically associated with more initial evidence for a particular choice selection and a more confident -less conservative- decision strategy. An increase in SP could result in more (correct) choices are made for the biased alternative and faster responses. In contrast, a lower SP (the green solid line) results in that accumulators are required to accumulate more evidence for a decision, which could require relatively longer decision times and lower AR. If SP stays constant, higher ARs are typically associated with more accurate responses, as a greater amount of evidence informs decisions (Brown and Heathcote, 2008; Mulder et al., 2014). Therefore, a higher AR reflects more efficient in evidence accumulation for decision-making.

We fit an LBA model to each subject's behavioral data (both RTs and accuracy) using the R library glba (Ingmar Visser, 2015). Trials with no response or RTs shorter than 200 ms or RT higher than 2.5 *SD* of the mean were removed from the modeling because these trials are not likely related to decisional processes. The mean removal rate was 8.90% (*SD* = 6.13) of total trials across subjects. We estimated the four LBA parameters (*A, v, b*, and *t0*) with the maximum likelihood estimation approach (van Maanen et al., 2011). The parameters were free to vary across the six stimulus conditions and the two talkers (with initial values of the parameters: $v = 0.5$, $A = 0.1$, $b = 0.3$, and $t0 = 0.1$). The *sv* of the accumulation rate was fixed to 0.2 to improve model fitting. To assess the goodness-of-fit of the model, we calculated the optimized log-likelihood and correlations between the model-predicted (i.e., simulated) behavioral performance and the observed behavioral performances. The model-predicted and the observed behavioral data were highly correlated for both ACC ($r = 0.97$, $P < 0.001$) and RT ($r = 0.91$, $P < 0.001$) across variability instances (Fig. 2B). These results indicate that the models accounted for a large proportion of the variance of interest. In addition, to test the model generalization ability across speech and non-speech items, we fit LBA models with speech and non-speech data separately and compared the model-predicted RTs of the non-speech model with the observed RTs of the speech condition, and vice versa. Significant correlations between predicted and observed RTs across contexts would suggest good generalizability of the estimated model to novel stimulus contexts. We confirmed a robust generalizability of the models across contexts (mean $r = 0.82$, $P < 0.001$).

In addition to the standard LBA modeling described above, we used a maximum likelihood approach to (re)estimate the LBA parameters at the single-trial level (single-trial LBA, STLBA) (van Maanen et al., 2011) for further neural-behavioral correlation analysis. Our goal here was to examine to what extent the robustness of multivoxel representations of tone-category information was associated with categorization decision processes. For the estimation of single-trial LBA parameters, the subject-level LBA parameters were used as initial input parameters to evaluate the parameter values of v_i and a_i (*i* denotes trial *i)* for each trial following the criteria and equations described in van Maanen et al. (2011). We used the single-trial SP (a_i) and AR (v_i) as two decision parameters of interest for the neural-behavioral correlation analysis described in the next section.

2.12. Trial-by-trial model-based neural-behavioral correlation analysis

We examined to what extent the robustness of the neural representations of tone category relates to decision components by calculating the inter-trial correlations between a neural category index (NCI) and the two decision parameters (i.e., SP and AR). The NCI is a multivariate neural index measuring the robustness of tone category representation. We quantified the NCI by estimating the difference in neural-pattern dissimilarity between within-category and between-category trial pairs (see Fig. 5B for a graphical illustration). Similar measures have been created and used in prior studies (e.g., phoneme or tone-category selectivity) (Du et al., 2014; Feng et al., 2018). The NCI reflects the neural sensitivity in differentiating tone categories based on their activation patterns. To calculate single-trial NCIs, we modeled single-trial brain activities by using the LSS approach described above. Trials with no response and RTs lower than 200 ms or RTs higher than 2.5 *SD* of the mean were removed from the LSS estimation and the neural-behavioral correlation analysis. We constructed an nRDM using single-trial activation patterns for each block and searchlight sphere individually. We then computed the NCI for each trial according to the following equation:

$$NCI_i = \frac{1}{N}\sum_{i \neq j} BCD_{ij} - \frac{1}{n}\sum_{i \neq k} WCD_{ik}$$

The n denotes the number of trials that belong to the same category as trial i (i.e., within-category trials), and the N denotes the number of trials that belong to different categories (i.e., between-category trials). BCD_{ij} refers to the between-category neural pattern dissimilarity between trial i and j, while $\mathrm{WCD}_{\mathrm{ik}}$ refers to the within-category neural pattern dissimilarity between trial i and k. Therefore, NCI_i is a weighted summarized neural representation value of tone category for the trial i, where a higher NCI reflects a greater tone-category selectivity. We calculated the NCI for each trial and z-transformed the NCI values for each block separately. The NCIs derived from a specific brain area were then correlated with the single-trial decision parameters across blocks for each subject. The NCI-decision-parameters correlation analysis was conducted with the searchlight approach to identify regions that their local representations contribute significantly to categorization decision processes. Therefore, we used the same brain mask as the MVPA to restrict the searchlight areas. Finally, a group-level t-test was used to test whether the neural-behavioral correlations were significantly higher than chance.

3. Results

3.1. Behavioral categorization performance

Participants were highly accurate in identifying native Mandarin tone categories across talkers, pitch salience, and stimulus contexts (mean accuracy [ACC] = 96.5%) (Fig. 1C). Behavioral accuracies were impressively resistant to the various forms of variability (talker, iteration steps, stimulus types). We did not find a significant difference between female and male talker items (ACC: $t_{(30)} = 1.23$, $P = 0.228$; reaction time [RT]: $t_{(30)} = 1.59$, $P = 0.122$). Examining effect of stimulus context, we did not find any significant difference in ACC or RT between the speech and non-speech condition (collapsed across IRNs) (ACC: $t_{(30)} = 0.27$, $P = 0.783$; RT: $t_{(30)} = 0.93$, $P = 0.361$). Similarly, we did not find a main effect of pitch salience in terms of ACC ($F_{(4,30)} = 0.75$, $P = 0.554$). However, we found a significant main effect of pitch salience in RT ($F_{(4,30)} = 6.96$, $P < 0.001$). Planned *posthoc* pairwise comparisons showed that participants were slower in responding to the IRN2 condition relative to the other IRN conditions (Ps < 0.05), while the comparisons for other pairs of IRN conditions in RT were not significant (all Ps > 0.1). We further conducted linear mixed-effects modeling with the talker (female vs. male) and stimulus conditions (i.e., IRN2, 4, 8, 16, 32 and speech) as two fixed-effect variables and the subject, subject-by-talker, and subject-by-condition as random-effects variables (the coefficients were set to be summed to 0) to validate the above main effects and examine the interaction effects between the factors. We found a significant main effect of stimulus condition in RT ($F_{(5,1080)} = 3.19$, $P = 0.007$) but not in ACC ($F_{(5,1080)} = 0.357$, $P = 0.878$). We did not find any significant main effect of talker for ACC ($F_{(1,1080)} = 0.672$, $P = 0.412$) or RT ($F_{(1,1080)} = 2.09$, $P = 0.148$). No talker-by-stimulus condition interaction effect was found. In summary, the talker variability, pitch salience (iteration step), and stimulus context did not significantly modulate the tone categorization accuracy, but subtle effects were evidenced in the RT.

3.2. The categorization decision processing is similar across talkers, pitch salience, and stimulus contexts

We applied the LBA model (Brown and Heathcote, 2008; Donkin et al., 2011) to examine the extent to which categorization decision components (e.g., starting point [SP] and accumulation rate [AR]) are modulated by variations in talker, pitch salience and stimulus context. LBA enables us to isolate the decision components from the non-decisional sensory-motor processes. For the raw RT responses, we found that the RT distributions were substantially similar across talkers, IRN steps, and stimulus contexts (Fig. 2A). The model accounted for a substantial proportion of the behavioral response variance, which evidenced by a robust model predictive performance, shown by strong inter-individual correlations between the model-predicted behavioral performance (i.e., RT and ACC) and the observed performance across talkers, IRN steps, and stimulus contexts (Fig. 2B; mean $r = 0.91$ for RT and $r = 0.97$ for ACC). Note that the goodness-of-fit of the model fittings (i.e., prediction) were similar across takers, IRN steps, and stimulus contexts (shown in Fig. 2B, regression lines with different colors and lightnesses). A linear mixed-effects modeling analysis further revealed that the SP and AR, two critical model parameters reflecting two decision components, were not significantly modulated by talkers, IRN steps, or stimulus contexts (Ps > 0.1) (Fig. 2C&D). No significant main effect of the variability factors was found for the other two model parameters (i.e., response threshold and non-decision time) (Fig. 2E&F). Note that there was a positive correlation between SP and AR across subjects and stimulus conditions ($r = 0.46$, $P < 0.001$).

3.3. Variability-tolerant multivariate neural representations of tone category

We used a multivariate pattern analysis approach, RSA, to identify brain regions representing the tone-category information that was tolerant of different forms of inter-item variability. We calculated the partial correlations between the tone-category representational dissimilarity matrix (RDM) to the neural RDMs (nRDMs) derived from the activation patterns while controlling for the variance of all other relevant factors (i.e., RDMs). We constructed a binary tone-category RDM and another six RDMs according to the distances or differences in fundamental frequency (F0) slope, mean F0 height, stimuli type or context (speech vs. non-speech), stimulus condiction (Speech, IRN 2, 4, 8, 16, and 32), talker, and motor response (see Fig. S1 for the RDM graphs). With the standard RSA, we found that the tone-category RDM model was significantly correlated with the local nRDMs in a distributed fronto-temporoparietal network, including the left inferior frontal gyrus (IFG), left precentral gyrus (preCG), bilateral inferior parietal lobule (IPL) and bilateral STG/STS (Fig. 3A). This RSA brain pattern is visually dominant in the left STG and IPL. With the partial RSA approach, we identified significant partial correlations of tone-category RDM in the left middle STG/STS (LmSTG/STS; Peak MNI coordinates: $x = -51$, $y = -19$, $z = 2$; peak t-value = 4.84; cluster size = 93 voxels) exclusively when the variances of the other six RDMs were controlled for (Fig. 3B). These results demonstrate a critical role for the LmSTG/STS in representing tone-category information irrespective of all potential sources of experimental-induced inter-item variability. Meanwhile, these results also indicate that different forms of variability modulated the robustness of neural representations of tone-category prominently in the bilateral IFG, preCG, IPL, and right STG. Further control analyses demonstrated that the neural representations of tone categories in these fronto-temporoparietal regions were primarily driven by acoustic and contextual factors instead of motor response differences (see Supplementary control analyses and results section and Fig. S4; Supplementary Materials).

3.4. Neural representations of tone category are differentially sensitive to various forms of variability

We used MVPC with different cross-validation (CV) procedures to examine the extent to which the fronto-temporoparietal tone-category representations identified by the RSA were modulated by the acoustic and contextual factors. Tone categories were classified (T1 vs. T2 vs. T4) significantly above chance from local activation patterns in the bilateral STG, IFG, IPL, and the left preCG with leave-one-block-out (i.e., “cross-block”) CV procedure (Fig. 4A). Brain regions demonstrating greater tone differentiation were relatively dominant in the left hemisphere compared to the right counterparts. This result demonstrates the overall tone-decoding brain pattern, which is similar to the RSA correlation pattern with the tone-category RDM shown in Fig. 3A. To identify neural representations of tone categories that were general across talkers, we conducted another searchlight classification analysis with the cross-talker CV procedure, where the support vector classification model was trained and tested on different talker items. Thus, tone-category information shared across talkers were captured by this CV procedure and reflects on the classification accuracy. We found that the talker-general (or talker-invariant) tone-decoding maps were similar to the one identified by the cross-block CV approach but with less robust in extent, especially in the bilateral IFG and preCG (Fig. 4B; also see Fig. S6B for results using the “cross-block&talker” CV procedure). To examine the neural representations that were resistant to the variability of pitch salience, we conducted a searchlight classification analysis with the cross-IRN CV procedure for the IRN items exclusively. We found that the searchlight tone-decoding brain pattern was similar to the one derived from the cross-talker CV approach in the temporoparietal regions but less robust in the bilateral frontal areas (e.g., the IFG) (Fig. 4C; also see Fig. S6C for similar RSA patterns using the “cross-block&IRN” CV procedure). Finally, we identified significantly above-chance tone-category information only in the bilateral middle portion of STG (mSTG) and the LIPL with the cross-stimulus-type CV procedure (Fig. 4D; also see Fig. S6D for similar RSA patterns using the “cross-block&stimulus-type" CV procedure). The only similarity between speech and IRN stimuli is restricted to the commonalities in time-varying F0 (pitch) patterns. Therefore, these cross-stimulus-type tone-decoding results indicate that the local activation patterns in the bilateral mSTG and LIPL represent neural ensembles that are finely tuned to representing dynamic pitch patterns. In summary, these results are not only converged with the partial RSA results but also demonstrate that the neural representations of tone category, especially in the bilateral frontal regions are differentially sensitive to different types of variability.

With the univariate voxel-wise conjunction activation analysis, we revealed common categorization-related activations in the inferior frontal, precentral, inferior parietal, and temporal areas across talkers (Fig. 4F, talker conj. map), which was similar to the overall categorization activations (Fig. 4E). Similar brain activation patterns, but with less extent in frontal and precentral areas, were found to be tolerant of variability in pitch salience (Fig. 4G, Pitch-salience conj. map). Importantly, we identified brain activations that were tolerant of all the three forms of variability (Fig. 4H, common activations across all conditions), including bilateral STG, left IPL, and bilateral supplementary motor areas (SMA). In addition, we found that the levels of univariate activation across fronto-temporoparietal regions were differentially modulated by the three factors. No region showed significant talker modulation effect (i.e., female vs. male; Fig. S3A, Supplementary Materials). In contrast, significant stimulus-context modulation effects were found in the bilateral fronto-temporoparietal areas (see Fig. S3B). The speech items elicited greater activations compared to the non-speech items in the bilateral IFG, STG, posterior middle temporal gyrus (pMTG), left IPL, anterior cingulate cortex, and the left insula. For the pitch-salience modulation effect, parametric modulation analysis with the linear modulation function revealed two brain areas showed significant effects (Fig. S3C). The left precentral gyrus (LpreCG) and the left anterior inferior parietal lobe (LIPLa) exhibited a linear effect of reduced activations with increasing IRN steps (Fig. S3D). We did not find any region shown a significant quadratic modulation effect.

3.5. The neural representations of tone category in a distributed fronto-temporoparietal network contribute to categorization decision

To investigate to what extent the neural representations of tone category contribute to categorization decisions, we correlated the neural representations (i.e., NCI) with the two LBA parameters (i.e., SP and AR) separately. That is, we examined how trial-by-trial fluctuations in the robustness of neural representations of tone category related to the fluctuations in categorization decision. We conducted the NCI-SP and NCI-AR correlation analyses with the searchlight approach (see Fig. 5B). For the NCI-SP correlation, we found significant positive correlations (voxel-level $P < 0.001$, FWE-corrected $P < 0.05$) in the bilateral IFG (including the orbital and opercular part of IFG), left preCG, bilateral IPL, and posterior and anterior STG (Fig. 5C; also see Fig. S5A for the NCI-SP correlation results after controlling for the motor responses; Supplementary Materials), which indicate that more robust tone-category representations in these regions were associated with more initial evidence (information) supporting choice selection during online categorization decision. This NCI-SP network was partially overlapped with the variability-tolerant LmSTG/STS in the superior temporal cortex with a less conservative threshold (voxel-level $P = 0.005$; 31

voxels overlapped). Similarly, we identified a distributed fronto-temporoparietal network that showed significant positive NCI-AR correlations, including the left IFG (opercular), right IFG (orbital, triangular, and opercular parts), left preCG, bilateral IPL, and bilateral anterior-mid STG/STS (Fig. 5D; also see Fig. S5B for the NCI-AR correlation results after controlling for the motor responses; Supplementary Materials). This NCI-AR network was also partially overlapped with the variability-tolerant LmSTG/STS in the temporal cortex with a less conservative threshold (voxel-level $P = 0.005$; 76 voxels overlapped). We did not find any region that showed significant negative NCI-SP or NCI-AR correlation. These findings together demonstrated that more robust tone-category representations in the fronto-temporoparietal regions are associated with higher efficiency in categorization decisions.

4. Discussion

We assessed the neural systems underlying representation and categorization of a critical feature in tone languages: lexical pitch patterns. Our behavioral results demonstrate that native listeners are highly adept at tone categorization. The ability to categorize linguistically-relevant pitches is highly tolerant of various forms of variability. Neural activation patterns within the left mid-STG region are highly sensitive to tone-category changes while resistant to acoustic and contextual variabilities. A distributed frontoparietal representational network is differentially and dynamically involved in the face of different forms of variability. On a single-trial level, the robustness of neural category representations within the distributed fronto-temporoparietal network is strongly associated with the efficiency of categorization decision processes. Our findings point to a specialized, variability-resistant representational core within the left STG and a distributed frontoparietal network that dynamically supports efficient tone-category representation and decision in native listeners.

4.1. Variability-tolerant representation of lexical tones in the left middle STG

Multivariate pattern analyses revealed that local activation patterns in the LmSTG/STS represent tone categories with high tolerance to different forms of variability. The partial RSA enables us to identify the core tone-category representation areas by controlling for the variance of all other acoustic and non-acoustic factors. Importantly, this approach revealed that only the local activation patterns in the LmSTG/STS were highly sensitive to between-category changes while tolerant of within-category inter-item variabilities induced by acoustic and contextual factors. This key property of perceptual constancy reflecting in the activation patterns demonstrates that the LmSTG/STS plays a key role as a core region in mediating tone perceptual constancy. The current findings are consistent with previous findings that the middle STG is a critical node along the ventral auditory stream involved in representing abstract category information (Desai et al., 2008; Liebenthal et al., 2010) that are tolerant of acoustic variabilities (Feng et al., 2018). Response properties in the middle-anterior STG show a greater propensity towards sustained neuronal responses rather than onset sensitivity (Hamilton et al., 2018). This functional specialization in the middle-anterior STG may be crucial for representing longer-duration suprasegmental information compared to segmental units representing in more posterior sections (Feng et al., 2018). These findings suggest that the local activation patterns in the LmSTG/STS may be important in representing abstract-level native lexical tone categories while resistant to various forms of variability induced by the current experiment. It is worth noting that other acoustic factors (e.g., nonlinear changes in F0) and contextual factors (e.g., task demands) unaccounted for in the present setting may mediate the robustness of the tone-category representations in the LmSTG/STS. Further studies are needed to assess the extent to which these factors interact with each other in mediating speech perception and the formation of the neural representations.

In the current study, we generated well-controlled non-speech analogs of lexical tone-category-like pitch patterns and manipulated variations in talker, pitch salience, and stimulus context within the same experiment. While the neural representations in the LmSTG/STS are tolerant of these variations, the robustness of representations in the dorsal fronto-motor stream is differentially modulated by the three factors. Among the three types of variability, stimulus context variability (speech vs. non-speech stimuli) modulated most the robustness of the neural representations of tone category in the bilateral frontal and precentral regions revealed by both univariate activation and multivariate classification analyses. The left inferior frontal and precentral areas, bilateral primary auditory regions and STG vary in their sensitivity to speech relative to non-speech stimuli (e.g., Binder 2001). Our univariate activation results corroborate these findings (see Fig. S3B, Supplementary Materials). However, prior neural and behavioral studies have also shown that processing of non-speech is similar to that of speech stimuli when they share critical temporal acoustic properties (Leech et al., 2009; Miller et al., 1976; Pisoni, 1977; Stevens and Klatt, 1974). Consistent with these findings, our univariate conjunction and multivariate classification analyses demonstrated that the left mSTG/STS, IPL, and right STG robustly represented tone categories irrespective of stimulus context. These findings suggest that these temporoparietal regions are highly sensitive to changes to pitch features that distinguish lexical-relevant tone categories while representations in the dorsal fronto-motor regions are more vulnerable to various forms of variability.

4.2. Multivariate neural representations of tone category dynamically contribute to categorization decision

One of the aims of this study is to examine the extent to which the multivariate neural representations of tone category contribute to categorization decision processes. In our decision modeling, the starting point (SP) reflects the amount of initial evidence for choice selection after the sensory-processing component, while accumulation rate (AR) reflects the efficiency in the evidence accumulation process for decision-making. Higher SP is typically associated with more prior information related to the task-relevant decision, which results in more confident (less conservative) choice selection strategies (Brown and Heathcote, 2008). We found significantly positive correlations between the neural category index (NCI) and SP across inferior frontal, parietal, and temporal regions (dominant in the bilateral frontoparietal network). These NCI-SP associations remained after controlling for the motor-response differences between categories within participants (Fig. S5A). This result suggests that the robustness of tone-category representations in the network is associated with the amount of initial category information for choice selection in category decision instead of motor processing. Previous behavioral decision-making studies have demonstrated that when the task becomes more challenging and subjects are instructed to produce fast responses, reaction time is typically reduced at the expense of accuracy (i.e., speed-accuracy tradeoffs) (Brown and Heathcote, 2008; Forstmann et al., 2010). In this case, lower SP typically results in less confident and incorrect choice behavior because of less prior evidence informing decisions. When the task is not difficult or subjects are not instructed to prioritize speed over accuracy, higher SP is typically associated with more confident -less conservative- decisions. Incorporating these prior behavioral findings, the strong positive NCI-SP correlations in the fronto-temporoparietal regions suggest a potential link between the more robust neural representations and more confident tone categorization decision-making processes. Moreover, the NCI-SP associations were mainly found in the bilateral IFG, left preCG, bilateral IPL, posterior and anterior STG, with minimal overlap with the variability-tolerant LmSTG/STS region. These results altogether suggest that the fronto-temporoparietal NCI-SP correlation regions may play an important role in utilizing the amount of category representations for categorization decision-making, complementing the role of the ‘core’ LmSTG/STS. The dynamic interplay between the core and the extended frontoparietal regions may be crucial for efficient tone perception and categorization decision-making under high-variability conditions.

In addition, we observed strong NCI-accumulation rate (AR) associations in the fronto-temporoparietal areas, especially the bilateral temporoparietal regions, which indicates the multi-function of these regions in tone-category representation and categorization decision. The NCI-AR associations indicate that more robust neural representations of tone category were associated with higher rates in evidence accumulation (i.e., higher efficiency) for categorization decisions. These results support our prediction that trial-by-trial fluctuations in the robustness of neural representations of tone category contribute to the fluctuations in the efficiency of evidence accumulation for categorization decisions. We posit that trial-by-trial acoustic and contextual variations of sounds and/or other factors such as attention and consciousness dynamically modulate the robustness of the neural representations which impacts both initial evidence (i.e., SP) and the efficiency of evidence accumulation for perceptual decision. Across different decision-related cognitive models, evidence accumulation is a common and important latent process for decision-making (Mulder et al., 2014). Decisions are made based on the amount of accumulating perceptual evidence. Higher AR is usually associated with more robust evidence, which results in more accurate and faster responses. More robust neural representations of tone category may provide sharper category distinction, which could more efficiently guide the selection of the correct category response while inhibiting or rejecting incorrect responses. Importantly, brain areas demonstrating significant NCI-AR correlations included a portion of the core category representation in the left mSTG/STS as well as a distributed frontoparietal network involving the bilateral ventrolateral prefrontal cortex (VLPFC) (e.g., IFG), preCG, IPL, anterior and posterior STG. Prior studies in animal models have demonstrated a high degree of category selectivity within the VLPFC (Rauschecker and Scott, 2009; Romanski et al., 2005). The VLPFC, preCG, IPL, and STG have been identified as important nodes of a distributed speech-motor system and sensorimotor interface (Hickok and Poeppel, 2007; Liebenthal et al., 2013). These systems are hypothesized to constrain speech perception by generating internal models involving an articulatory code (IFG) and matching the internal models with the incoming auditory signals via a sensorimotor interface (IPL) (Du et al., 2014). Accumulating linguistically relevant category-related representations in these regions may further facilitate optimal speech categorization decisions.

It is worth noting that there are overlaps in the frontoparietal regions between the NCI-SP and NCI-AR correlations, which may suggest that the SP and AR have similar sources in neural representations in tone categorization decision. Conceptually, the LBA is an evidence-accumulation-based decision-making model where the SP and AR parameters are linked with each other. The SP refers to the initial amount of evidence and the AR refers to the rate at which evidence accumulates. In the model, they are two intercorrelated parameters that both contribute to evidence accumulation. A lower SP results in that accumulators are required to accumulate more evidence, which could require relatively longer decision times and lower AR (see Fig. 5A). The correlations between AR and SP have been commonly found in previous studies (see Mulder et al., 2014 for a review). Consistent with the model description and previous observations, we found that the AR and SP parameters are significantly correlated with each other ($r = 0.46$, $P < 0.001$). Therefore, partially overlaps in neural category representations were expected due to the theoretical interrelation and the statistical correlation between AR and SP parameters. This is also consistent with a previous meta-analysis that the univariate activation correlations of AR and SP have common regions in the bilateral frontal and parietal regions (Mulder et al., 2014). Further studies need to examine whether the neural correlates of SP and AR change differently according to the demand and nature of decision tasks.

4.3. Dynamic network mechanism subserves lexical-tone representation and categorization

We demonstrate that the human brain utilizes a flexible and dynamic mechanism to achieve perceptual constancy and efficient categorization of lexical tones. Previous studies show that the frontoparietal regions are associated with executive control processes, such as inhibition and selection (Braver, 2012; Braver and Barch, 2006). The LBA decision parameters have been associated with BOLD activities in this frontoparietal network (Mulder et al., 2014; van Maanen et al., 2011). Moreover, previous studies have demonstrated that the frontoparietal regions are dynamically involved when the participants categorize confusable lexical tones (Feng et al., 2018), undertake a difficult linguistic or non-linguistic task (Cocchi et al., 2013; Waskom et al., 2014), effortfully process a non-native language (Abutalebi and Green, 2008; Feng et al., 2015), or switch between tasks (Cole et al., 2013). Therefore, the differential representations and dynamical representation-decision relationships found in the frontoparietal network suggest that additional executive processes (e.g., inhibition and attention selection) may be recruited when facing challenging perception situations (e.g., in a noisy and confusable context).

Prior work has examined the neural circuitry underlying auditory categorization behavior under various degrees of variability (Arsenault and Buchsbaum, 2015; Binder et al., 2004; Bizley and Cohen, 2013; Noppeney et al., 2010; Tsunada et al., 2016; Xin et al., 2019; Yi et al., 2019). Variability arises due to perceptual noise induced by various factors (e.g., talker, listening context, and/or listeners' encoding of the sensory signal) (Mattys et al., 2012), which induce challenges to the auditory perceptual systems. While some studies advocate a critical role for sparse units in the STG in driving auditory decisions (Tsunada et al., 2016; Yi et al., 2019), others contextualize the STG within a more extensive fronto-temporoparietal network (Russ et al., 2007, 2008) in resolving variability and driving categorization. We posit that dynamic interactions between a representational core and an extended executive network are an effective neural mechanistic solution to increase accuracy and efficiency in categorization decisions under different forms of acoustic and perceptual variability. Consistent with this viewpoint, ensembles within the bilateral STG and inferior parietal cortex are tolerant of talker variability and contextual variables (i.e., segmental contexts and task demands) (Feng et al., 2018) but are susceptible to perceptual noise (Du et al., 2014, 2016). In contrast to temporal regions, representations of speech category within the dorsal fronto-motor network are more resistant to noise (Du et al., 2014, 2016) but may be sensitive to the talker and contextual variables (Feng et al., 2018, 2019; Myers et al., 2009; Salvata et al., 2012). Our findings in lexical-tone representation and categorization extended these previous findings and further demonstrated that the perceptual relevance of the fronto-temporoparietal representations of speech category is dynamically contributing to different levels of perceptual constancy and online categorization decision. Increased robustness of neural representations is dynamically associated with increased efficiency in categorization decisions.

It is worth considering the extent to which the dynamic representation mechanism identified in the fronto-temporoparietal network can be generalized to other scenarios or tasks. In the current study, participants performed an explicit categorization task that requires attention and executive-control-related decision processes. Listeners are unlikely to engage in this kind of explicit metalinguistic task during ecological speech processing. Nevertheless, in natural speech perception and daily communication, listeners often engage in various challenging listening situations (e.g., low signal-to-noise ratio [e.g., noisy environments], the high degree of talker variability [i.e., listening to multiple talkers], degraded speech, and ambiguous or complex linguistic contexts, etc.). Previous neuroimaging studies have shown that speech perception in challenging contexts additionally recruits a fronto-temporoparietal network while different regions within the network are differentially sensitive to various sources of variability as well as different experimental and task settings (Alain et al., 2018; Alho et al., 2016; Bonte et al., 2014; Du et al., 2014; Evans and Davis, 2015; Feng et al., 2018; Hickok and Poeppel, 2007). Therefore, the dynamic representational and decisional mechanisms found in the fronto-temporoparietal network in response to various forms of variability may be operational during speech processing. Explicit processing (e.g., categorization task) may enhance the engagement of this network. Future studies need to test the generalization of the current findings in ecologically valid conditions.

Conclusion

We show that native listeners are highly proficient at categorizing lexical tones in Mandarin Chinese. Efficient categorization is associated with the robustness of the representation in a distributed fronto-temporoparietal network that includes a core neural representation in the left mSTG/STS that is tolerant of variabilities in surface acoustic properties, perceptual salience, and linguistic context. Converging evidence from univariate activation analysis, multivariate pattern analysis, and computational modeling point to a variability-tolerant 'core' STG that dynamically operates within a widely-distributed fronto-temporoparietal network in mediating efficient tone categorization.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This work was supported by grants from the National Institute on Deafness and other Communication Disorders of the National Institutes of Health under Award No. (R01DC013315 834 to B.C.), the General Research Fund (Ref. No. 14619518 to G.F. and Ref. No. 34000118 to P.W.) by the Research Grants Council of Hong Kong, and Direct Grant for Research (Ref. No. 4051137 to G.F.) by the Chinese University of Hong Kong.

Declaration of Competing Interest

P.W. is a founder of a company in Hong Kong supported by a Hong Kong SAR government startup scheme for universities.

References

Abutalebi J, Green DW, 2008 Control mechanisms in bilingual language production: neural evidence from language switching studies. Lang. Cogn. Process. 23, 557–582.

Alain C, Du Y, Bernstein LJ, Barten T, Banai K, 2018 Listening under difficult conditions: an activation likelihood estimation meta-analysis. Hum. Brain Mapp. 39, 2695–2709. [PubMed: 29536592]

Alavash M, Tune S, Obleser J, 2019 Modular reconfiguration of an auditory control brain network supports adaptive listening behavior. Proc. Natl. Acad. Sci. U. S. A. 116, 660–669. [PubMed: 30587584]

Alho J, Green BM, May PJ, Sams M, Tiitinen H, Rauschecker JP, Jaaske-lainen IP, 2016 Early-latency categorical speech sound representations in the left inferior frontal gyrus. Neuroimage 129, 214–223. [PubMed: 26774614]

Arsenault JS, Buchsbaum BR, 2015 Distributed neural representations of phonological features during speech perception. J. Neurosci. 35, 634–642. [PubMed: 25589757]

Ashburner J, Friston KJ, 2005 Unified segmentation. Neuroimage 26, 839–851. [PubMed: 15955494]

Binder JR, Liebenthal E, Possing ET, Medler DA, Ward BD, 2004 Neural correlates of sensory and decision processes in auditory object identification. Nat. Neurosci. 7, 295–301. [PubMed: 14966525]

Bizley JK, Cohen YE, 2013 The what, where and how of auditory-object perception. Nat. Rev. Neurosci. 14, 693–707. [PubMed: 24052177]

Bonte M, Hausfeld L, Scharke W, Valente G, Formisano E, 2014 Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. J. Neurosci. 34, 4548–4557. [PubMed: 24672000]

Braver TS, 2012 The variable nature of cognitive control: a dual mechanisms framework. Trends Cogn. Sci. 16, 106–113. [PubMed: 22245618]

Braver TS, Barch DM, 2006 Extracting core components of cognitive control. Trends Cogn. Sci. 10, 529–532. [PubMed: 17071129]

Brown SD, Heathcote A, 2008 The simplest complete model of choice response time: linear ballistic accumulation. Cogn. Psychol. 57, 153–178. [PubMed: 18243170]

Buchel C, Wise RJ, Mummery CJ, Poline JB, Friston KJ, 1996 Nonlinear regression in parametric activation studies. Neuroimage 4, 60–66. [PubMed: 9345497]

Busemeyer JR, Gluth S, Rieskamp J, Turner B, 2019 Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. Trends Cogn. Sci. 23, 251–263. [PubMed: 30630672]

Chang C−C, Lin C−J, 2011 LIBSVM: a library for support vector machines. ACM T. Intel. Syst. Tec. 2 27, 21–27 27.

Cheung C, Hamiton LS, Johnson K, Chang EF, 2016 The auditory representation of speech sounds in human motor cortex. Elife 5, e12577. [PubMed: 26943778]

Chevillet MA, Jiang X, Rauschecker JP, Riesenhuber M, 2013 Automatic phoneme category selectivity in the dorsal auditory stream. J. Neurosci. 33, 5208–5215. [PubMed: 23516286]

Cichy RM, Kriegeskorte N, Jozwik KM, van den Bosch JJF, Charest I, 2019 The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. Neuroimage 194, 12–24. [PubMed: 30894333]

Cocchi L, Zalesky A, Fornito A, Mattingley JB, 2013 Dynamic cooperation and competition between brain systems during cognitive control. Trends Cogn. Sci. 17, 493–501. [PubMed: 24021711]

Cole MW, Reynolds JR, Power JD, Repovs G, Anticevic A, Braver TS, 2013 Multi-task connectivity reveals flexible hubs for adaptive task control. Nat. Neurosci. 16, 1348–1355. [PubMed: 23892552]

Correia JM, Jansma BM, Bonte M, 2015 Decoding articulatory features from fMRI responses in dorsal speech regions. J. Neurosci. 35, 15015–15025. [PubMed: 26558773]

Desai R, Liebenthal E, Waldron E, Binder JR, 2008 Left posterior temporal regions are sensitive to auditory categorization. J. Cogn. Neurosci. 20, 1174–1188. [PubMed: 18284339]

Diehl RL, Lotto AJ, Holt LL, 2004 Speech perception. Annu. Rev. Psychol. 55, 149–179. [PubMed: 14744213]

Donkin C, Brown S, Heathcote A, 2011 Drawing conclusions from choice response time models: a tutorial using the linear ballistic accumulator. J. Math. Psychol. 55, 140–151.

Du Y, Buchsbaum BR, Grady CL, Alain C, 2014 Noise differentially impacts phoneme representations in the auditory and speech motor systems. Proc. Natl. Acad. Sci. U. S. A. 111, 7126–7131. [PubMed: 24778251]

Du Y, Buchsbaum BR, Grady CL, Alain C, 2016 Increased activity in frontal motor cortex compensates impaired speech perception in older adults. Nat. Commun. 7, 12241. [PubMed: 27483187]

Evans S, Davis MH, 2015 Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. Cereb. Cortex 25, 4772–4788. [PubMed: 26157026]

Feng G, Chen H−C, 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, 103–118. [PubMed: 26095088]

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. Cereb. Cortex 28, 3241–3254. [PubMed: 28968658]

Feng G, Yi HG, Chandrasekaran B, 2019 The role of the human auditory corticostriatal network in speech learning. Cereb. Cortex 29, 4077–4089. [PubMed: 30535138]

Formisano E, De Martino F, Bonte M, Goebel R, 2008 “Who” is saying “what”? Brain-based decoding of human voice and speech. Science 322, 970–973. [PubMed: 18988858]

Forstmann BU, Anwander A, Schafer A, Neumann J, Brown S, Wagenmakers EJ, Bogacz R, Turner R, 2010 Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. Proc. Natl. Acad. Sci. U. S. A. 107, 15916–15920. [PubMed: 20733082]

Forstmann BU, Wagenmakers EJ, Eichele T, Brown S, Serences JT, 2011 Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? Trends Cogn. Sci. 15, 272–279. [PubMed: 21612972]

Frank MJ, Gagne C, Nyhus E, Masters S, Wiecki TV, Cavanagh JF, Badre D, 2015 fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. J. Neurosci. 35, 485–494. [PubMed: 25589744]

Friston KJ, Frith CD, Frackowiak RS, Turner R, 1995 Characterizing dynamic brain responses with fMRI: a multivariate approach. Neuroimage 2, 166–172. [PubMed: 9343599]

Hamilton LS, Edwards E, Chang EF, 2018 A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. Curr. Biol. 28, 1860–1871. [PubMed: 29861132]

Heekeren HR, Marrett S, Bandettini PA, Ungerleider LG, 2004 A general mechanism for perceptual decision-making in the human brain. Nature 431, 859–862. [PubMed: 15483614]

Hickok G, Poeppel D, 2007 The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402. [PubMed: 17431404]

Ho TC, Brown S, Serences JT, 2009 Domain general mechanisms of perceptual decision making in human cortex. J. Neurosci. 29, 8675–8687. [PubMed: 19587274]

Holt LL, Lotto AJ, 2010 Speech perception as categorization. Atten. Percept. Psychophys. 72, 1218–1227. [PubMed: 20601702]

James W, 1890 The Principles of Psychology. Henry Holt and Company, New York.

Kriegeskorte N, Goebel R, Bandettini P, 2006 Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868. [PubMed: 16537458]

Kriegeskorte N, Kievit RA, 2013 Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. 17, 401–412. [PubMed: 23876494]

Kriegeskorte N, Mur M, Bandettini P, 2008 Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 1–28. [PubMed: 18958245]

Krishnan A, Bidelman GM, Gandour JT, 2010 Neural representation of pitch salience in the human brainstem revealed by psychophysical and electrophysiological indices. Hear. Res. 268, 60–66. [PubMed: 20457239]

Leech R, Holt LL, Devlin JT, Dick F, 2009 Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. J. Neurosci. 29, 5234–5239. [PubMed: 19386919]

Liang B, Du Y, 2018 The functional neuroanatomy of lexical tone perception: an activation likelihood estimation meta-Analysis. Front. Neurosci. 12, 495. [PubMed: 30087589]

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M, 1967 Perception of the speech code. Psychol. Rev. 74, 431–461. [PubMed: 4170865]

Liebenthal E, Desai R, Ellingson MM, Ramachandran B, Desai A, Binder JR, 2010 Specialization along the left superior temporal sulcus for auditory categorization. Cereb. Cortex 20, 2958–2970. [PubMed: 20382643]

Liebenthal E, Sabri M, Beardsley SA, Mangalathu-Arumana J, Desai A, 2013 Neural dynamics of phonological processing in the dorsal auditory stream. J. Neurosci. 33, 15414–15424. [PubMed: 24068810]

Mattys SL, Davis MH, Bradlow AR, Scott SK, 2012 Speech recognition in adverse conditions: a review. Lang. Cogn. Process. 27, 953–978.

Miller JD, Wier CC, Pastore RE, Kelly WJ, Dooling RJ, 1976 Discrimination and labeling of noise – buzz sequences with varying noise-lead times: an example of categorical perception. J. Acoust. Soc. Am. 60, 410–417. [PubMed: 993463]

Misaki M, Kim Y, Bandettini PA, Kriegeskorte N, 2010 Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage 53, 103–118. [PubMed: 20580933]

Mulder MJ, van Maanen L, Forstmann BU, 2014 Perceptual decision neurosciences - a model-based review. Neuroscience 277, 872–884. [PubMed: 25080159]

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. [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, 2636–2643. [PubMed: 21924359]

Myers EB, Blumstein SE, Walsh E, Eliassen J, 2009 Inferior frontal regions underlie the perception of phonetic category invariance. Psychol. Sci. 20, 895–903. [PubMed: 19515116]

Noppeney U, Ostwald D, Werner S, 2010 Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. J. Neurosci. 30, 7434–7446. [PubMed: 20505110]

Oosterhof NN, Connolly AC, Haxby JV, 2016 CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU octave. Front. Neuroinform. 10, 1–27. [PubMed: 26834620]

Pisoni DB, 1977 Identification and discrimination of the relative onset time of two component tones: implications for voicing perception in stops. J. Acoust. Soc. Am. 61, 1352–1361. [PubMed: 881488]

Raizada RD, Poldrack RA, 2007 Selective amplification of stimulus differences during categorical processing of speech. Neuron 56, 726–740. [PubMed: 18031688]

Rauschecker JP, Scott SK, 2009 Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci. 12, 718–724. [PubMed: 19471271]

Romanski LM, Averbeck BB, Diltz M, 2005 Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. J. Neurophysiol. 93, 734–747. [PubMed: 15371495]

Russ BE, Lee YS, Cohen YE, 2007 Neural and behavioral correlates of auditory categorization. Hear. Res. 229, 204–212. [PubMed: 17208397]

Russ BE, Orr LE, Cohen YE, 2008 Prefrontal neurons predict choices during an auditory same-different task. Curr. Biol. 18, 1483–1488. [PubMed: 18818080]

Salvata C, Blumstein SE, Myers EB, 2012 Speaker Invariance for Phonetic Information: an fMRI Investigation. Lang. Cogn. Process. 27, 210–230. [PubMed: 23264714]

Si X, Zhou W, Hong B, 2017 Cooperative cortical network for categorical processing of Chinese lexical tone. Proc. Natl. Acad. Sci. U. S. A. 114, 12303–12308. [PubMed: 29087324]

Stevens KN, Klatt DH, 1974 Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Am. 55, 653–659. [PubMed: 4819867]

Studholme C, Hill DLG, Hawkes DJ, 1999 An overlap invariant entropy measure of 3D medical image alignment. Pattern Recogn 32, 71–86.

Swaminathan J, Krishnan A, Gandour JT, Xu Y, 2008 Applications of static and dynamic iterated rippled noise to evaluate pitch encoding in the human auditory brainstem. IEEE Trans. Biomed. Eng. 55, 281–287. [PubMed: 18232372]

Tsunada J, Liu AS, Gold JI, Cohen YE, 2016 Causal contribution of primate auditory cortex to auditory perceptual decision-making. Nat. Neurosci. 19, 135–142. [PubMed: 26656644]

van Maanen L, Brown SD, Eichele T, Wagenmakers EJ, Ho T, Serences J, Forstmann BU, 2011 Neural correlates of trial-to-trial fluctuations in response caution. J. Neurosci. 31, 17488–17495. [PubMed: 22131410]

Waskom ML, Kumaran D, Gordon AM, Rissman J, Wagner AD, 2014 Frontoparietal representations of task context support the flexible control of goal-directed cognition. J. Neurosci. 34, 10743–10755. [PubMed: 25100605]

Xie Z, Reetzke R, Chandrasekaran B, 2018 Taking attention away from the auditory modality: context-dependent effects on early sensory encoding of speech. Neuroscience 384, 64–75. [PubMed: 29802881]

Xin Y, Zhong L, Zhang Y, Zhou T, Pan J, Xu N. l., 2019 Sensory-to-category transformation via dynamic reorganization of ensemble structures in mouse auditory cortex. Neuron 103, 909–921.e906. [PubMed: 31296412]

Yi HG, Leonard MK, Chang EF, 2019 The encoding of speech sounds in the superior temporal gyrus. Neuron 102, 1096–1110. [PubMed: 31220442]

Yip M, 2002 Tone. Cambridge University Press, Cambridge.

Zatorre RJ, Gandour JT, 2008 Neural specializations for speech and pitch: moving beyond the dichotomies. Phil. Trans. R. Soc. B 363, 1087–1104. [PubMed: 17890188]

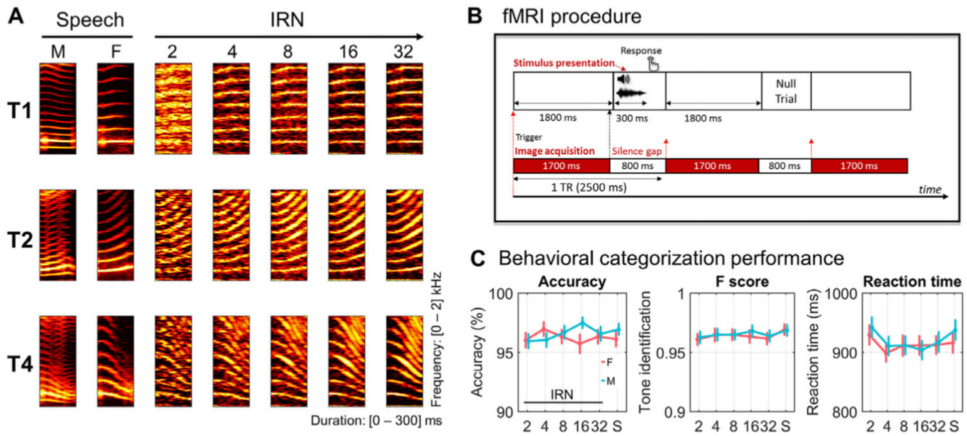

Fig. 1.

Stimuli, fMRI protocol, and behavioral categorization performance. **A**, the spectrogram of sample stimuli for the speech (S) sounds produced by a male (M) and a female (F) speaker and iterative ripple noise (IRN) homologs generated with five different iteration steps (i.e., IRN 2, 4, 8, 16, and 32, modeled with a female talker [male talker stimuli were not shown]). **B**, the customized fMRI sparse-sampling scanning procedure. The stimuli were presented in the silence gaps between imaging acquisitions to minimize the impact of scanner noise on perception/categorization. **C**, behavioral categorization accuracy, *F* score (a composite measure reflects tone identification sensitivity and specificity), and reaction times were displayed across talkers, IRNs, and stimulus contexts. S, speech condition.

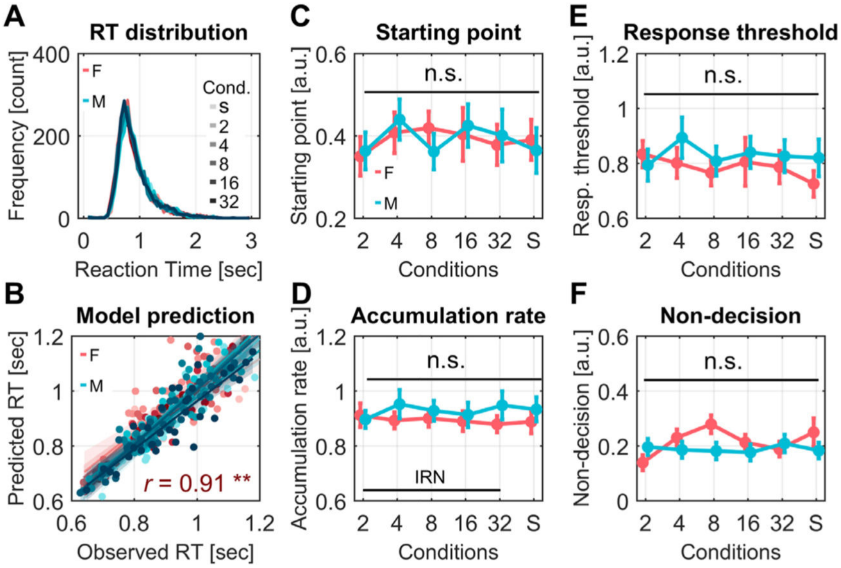

Fig. 2.

Behavioral reaction-time (RT) distributions and LBA model parameters across talkers, pitch-salience conditions, and stimulus contexts. **A**, the RT distributions are highly overlapped across talkers and conditions. Talker: *F* = female; *M* = male. Conditions: *S* = speech condition; 2 – 32 = IRN steps. The distribution lightness denotes stimulus conditions. **B**. the goodness-of-fit of the LBA modeling was measured by calculating the inter-individual correlations between model-predicted and observed RTs across talkers and stimulus conditions. The mean Pearson correlation coefficient $r = 0.91$. The lightness denotes stimulus conditions (i.e., Speech, IRN 2, 4, 8, 16, and 32). Each dot in the scatterplot represents the mean reaction time of a subject in a condition. **, $P < 0.001$. **C&D**, the two model parameters of interest (i.e., starting point and accumulation rate) did not show significant difference across talkers or stimulus conditions (*P*s > 0.1). **E&F**, the other two related model parameters, response threshold and non-decision time did not show significant difference across talker or stimulus conditions (*P*s > 0.1).

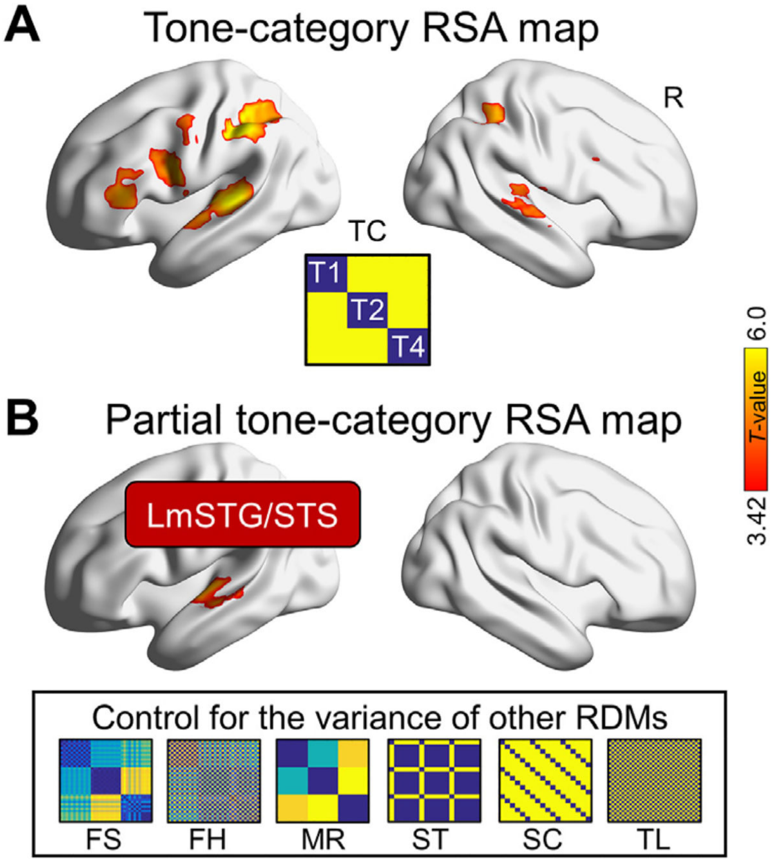
A
Tone-category RSA map
R
TC
T1
T2
T4
B
Partial tone-category RSA map
LmSTG/STS
6.0
T-value
3.42
Control for the variance of other RDMs
FS
FH
MR
ST
SC
TL

Fig. 3.

Neural representations of tone category revealed by the representational similarity analysis (RSA). **A**, the searchlight-based RSA maps showed significant correlations between the tone-category (TC) RDM and neural RDMs in the bilateral fronto-temporoparietal regions. **B**, the partial RSA revealed that the left middle portion of the STG/STS (LmSTG/STS) remained significant after controlling for the variance of other predefined RDMs. These control RDMs from left to right refer to F0 slope (FS), F0 height (FH), motor response (MR), stimulus type/context (ST), stimulus condition (SC), and talker (TL), respectively (see more details in Material and methods section and Fig. S1). All searchlight RSA maps were initially thresholded at voxel-level $P < 0.001$ and FWE-corrected $P < 0.05$.

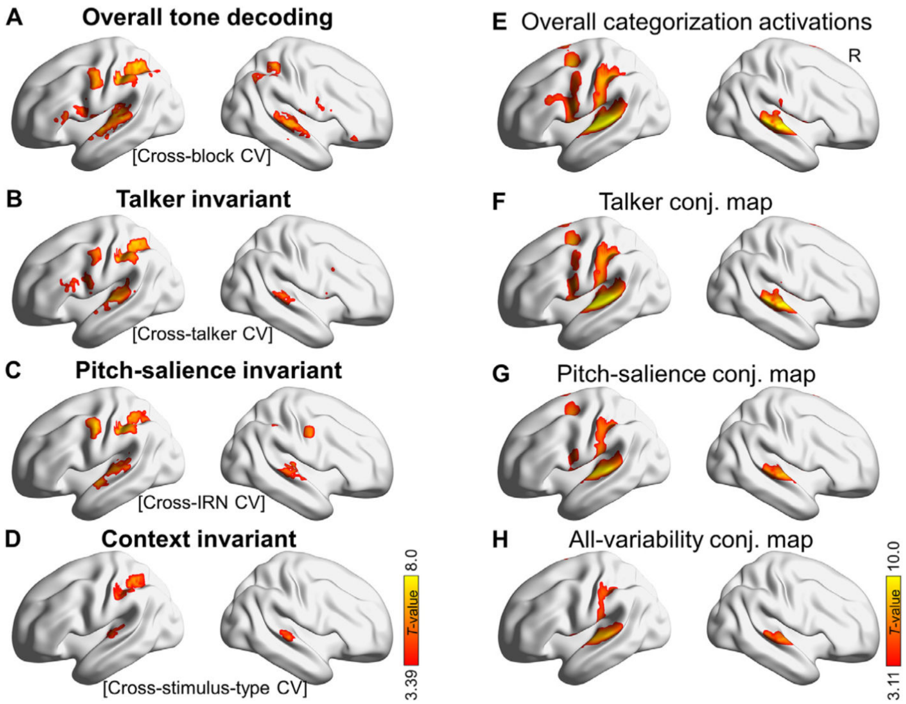

Fig. 4.

Multivariate pattern classification (MVPC) analysis (left panel A-D) with different cross-validation (CV) procedures and univariate activation analysis with different conjunction procedures (right panel E-H). **Left panels: A-D**, searchlight-based tone-category classification maps derived from four CV procedures. **A**, tone-classification (i.e., tone-decoding) maps derived from “cross-block” (i.e., leave-one-block-out) CV procedure, in which classifiers were trained and tested on the same sets of stimuli but from different blocks. **B&C**, tone-classification maps derived from the “cross-talker” (B) and “cross-IRN” (C) CV procedures (i.e., training and testing classifiers with data that are different in surface acoustic properties, i.e., talker and pitch salience). **D**, context-invariant tone-classification maps were generated by using “cross-stimulus-type” CV procedure (i.e., classifiers were trained with speech stimuli and tested with the IRN stimuli, and vice versa). **Right panels: E**, distributed fronto-temporoparietal regions were activated during tone categorization (relative to baseline). **F**, the talker-conjunction analysis revealed that the fronto-temporoparietal regions were commonly activated during categorization across talkers (i.e., talker-general areas). **G**, common categorization-related activations were found across IRN steps (i.e., pitch-salience-general areas). **H**, the overall conjunction analysis revealed that only the temporoparietal regions were commonly activated across all the three types of variants (i.e., all-variability-general regions). All brain maps were initially thresholded at voxel-level $P < 0.001$ and cluster-level FWE-corrected $P < 0.05$.

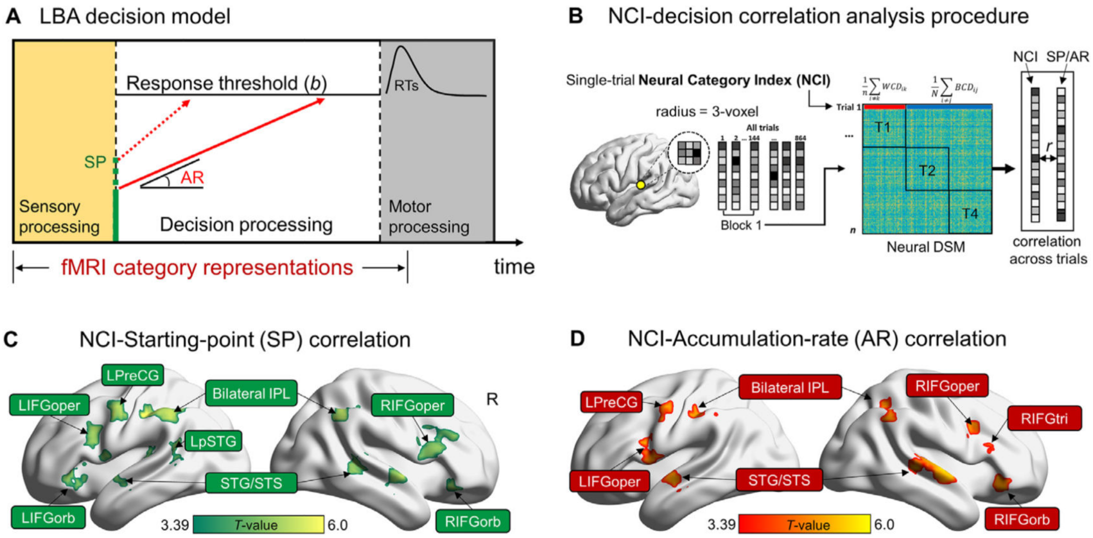

Fig. 5.

Graphical illustration of LBA components and parameters and trial-by-trial model-based behavioral-neural (i.e., NCI-decision) correlation analysis. **A**, LBA model components and parameters illustration. For each categorization, LBA defines three processing components, including sensory (yellow), decision (white), and motor (gray) processing. The stimulus initially undergoes sensory processing and the tone-category information (i.e., evidence) is then accumulated toward one of the possible stimulus-response mappings until the response threshold (*b*) is reached. The two red lines denote the accumulation process at different rates. The red dash line denotes a higher accumulation rate (AR) than the solid line, while they have different starting points (SP) (green lines). The fMRI category representations measured by the neural category index (NCI) are hypothesized relating to the amount of initial evidence and evidence accumulation process of decision-making (i.e., SP and AR). **B**, a graphical illustration of the trial-by-trial NCI-decision correlation analysis procedure. At each local searchlight sphere, the single-trial NCIs were calculated and then were correlated with the SP and AR separately across trials. **C**, searchlight-based NCI-SP correlation brain maps. Significant positive NCI-SP correlations were identified across the fronto-temporoparietal areas, which demonstrate that more robust neural category representations are associated with higher SP. **D**, searchlight-based NCI-AR correlation brain maps. A distributed bilateral fronto-temporoparietal network was also identified showing significant positive NCI-AR correlations, which indicate that more robust neural representations of tone category in these regions are associated with higher efficiency in evidence accumulation for decision. All brain maps were initially threshold at the voxel-level $P < 0.001$ and cluster-level FWE-corrected at $P < 0.05$.