Cerebral Cortex, September 2018;28: 3241–3254

doi: 10.1093/cercor/bhx195
Advance Access Publication Date: 28 August 2017
Original Article

ORIGINAL ARTICLE

Task-General and Acoustic-Invariant Neural Representation of Speech Categories in the Human Brain

Gangyi Feng[1,2,3,], Zhenzhong Gan[4], Suiping Wang[4,5], Patrick C. M. Wong[1,2] and Bharath Chandrasekaran[3,6,7,8,9]

[1]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China, [2]Brain and Mind Institute, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China, [3]Department of Communication Sciences & Disorders, Moody College of Communication, The University of Texas at Austin, 2504A Whitis Avenue (A1100), Austin, TX 78712, USA, [4]Center for the Study of Applied Psychology and School of Psychology, South China Normal University, Guangzhou 510631, China, [5]Guangdong Provincial Key Laboratory of Mental Health and Cognitive Science, South China Normal University, Guangzhou 510631, China, [6]Department of Psychology, The University of Texas at Austin, 108 E. Dean Keeton Stop A8000, Austin, TX 78712, USA, [7]Department of Linguistics, The University of Texas at Austin, 305 E. 23rd Street STOP B5100, Austin, TX 78712, USA, [8]Institute for Mental Health Research, College of Liberal Arts, The University of Texas at Austin, 305 E. 23rd St. Stop E9000, Austin, TX 78712, USA and [9]The Institute for Neuroscience, The University of Texas at Austin, 1 University Station Stop C7000, Austin, TX 78712, USA

Address correspondence to Bharath Chandrasekaran, The University of Texas at Austin, 2504A Whitis Avenue (A1100), Austin, TX 78712, USA. Email: bchandra@utexas.edu or Suiping Wang, Center for the Study of Applied Psychology and School of Psychology, South China Normal University, Guangzhou 510631, China. E-mail: wangsuiping@m.scnu.edu.cn

Gangyi Feng was a Postdoctoral fellow at The University of Texas at Austin when this work was completed

Abstract

A significant neural challenge in speech perception includes extracting discrete phonetic categories from continuous and multidimensional signals despite varying task demands and surface-acoustic variability. While neural representations of speech categories have been previously identified in frontal and posterior temporal-parietal regions, the task dependency and dimensional specificity of these neural representations are still unclear. Here, we asked native Mandarin participants to listen to speech syllables carrying 4 distinct lexical tone categories across passive listening, repetition, and categorization tasks while they underwent functional magnetic resonance imaging (fMRI). We used searchlight classification and representational similarity analysis (RSA) to identify the dimensional structure underlying neural representation across tasks and surface-acoustic properties. Searchlight classification analyses revealed significant "cross-task" lexical tone decoding within the bilateral superior temporal gyrus (STG) and left inferior parietal lobule (LIPL). RSA revealed that the LIPL and LSTG, in contrast to the RSTG, relate to 2 critical dimensions (pitch height, pitch direction) underlying tone perception. Outside this core representational network, we found greater activation in the inferior frontal and parietal regions for stimuli that are more perceptually similar during tone categorization. Our findings reveal the specific characteristics of fronto-tempo-parietal regions that support speech representation and categorization processing.

Key words: Mandarin tone categories, multivariate pattern analysis, speech perception, task-general neural representation

Introduction

A major goal of auditory neuroscience is to understand how behaviorally relevant information in conspecific sounds are extracted, represented, and mapped to meaningful constructs in the brain (Scott and Johnsrude 2003; Griffiths and Warren 2004; Hickok and Poeppel 2007; Hickok 2009). In speech perception, key acoustic features are extracted from continuous speech signals and mapped to behavioral-relevant equivalent classes, that is, categories. During speech perception, various task demands and surface-acoustic variability (e.g., talker variability) are salient factors that impact the organization of neural activity patterns that relate to the speech category (Chang et al. 2010; Chevillet et al. 2013; Bonte et al. 2014; Arsenault and Buchsbaum 2016; Cheung et al. 2016). Extracting discrete categories despite various task demands and talker variability, while a significant challenge in neural computation, is one that is critical for speech perception. Our goal in this study is to identify the brain regions that represent speech category information irrespective of task demands and talker variability and to assess the representational structures underlying the categorical representations.

Previous studies examining the neural representation of speech categories have provided important insights, but findings from these studies are somewhat inconsistent. Prior studies using functional magnetic resonance imaging (fMRI) and multivariate data analysis methods have revealed neural representations of speech categories from multivoxel patterns within the human superior temporal gyrus (STG) (Formisano et al. 2008; Boets et al. 2013; Chevillet et al. 2013; Bonte et al. 2014; Du et al. 2014; Arsenault and Buchsbaum 2015; Correia et al. 2015; Evans and Davis 2015). These results are consistent with studies using Electrocorticographic recordings (Chang et al. 2010; Mesgarani et al. 2014; Cheung et al. 2016). Further, several other studies have revealed that speech category information can also be decoded from activity patterns in sensorimotor areas and prefrontal regions that constitute the dorsal auditory stream (Lee et al. 2012; Du et al. 2014; Correia et al. 2015; Evans and Davis 2015; Cheung et al. 2016). An emerging view is that speech category representations are broadly distributed in the fronto-auditory network, including both ventral and dorsal streams (Hickok and Poeppel 2007; Leonard and Chang 2014; Poeppel 2014). Such widely distributed neural representation may allow for robust speech processing irrespective of the task demand variability (Bonte et al. 2014; Alho et al. 2016), talker variability (Evans and Davis 2015), and speech signal quality (e.g., adverse listening conditions) (Du et al. 2014).

Primarily, at least 2 potential neural mechanisms underlying speech categorization have been discussed in the literature. From an emergent perspective, the neural representation of speech categories is an emergent property of distributed neural dynamics that is "task-dependent". During speech perception, different task constraints interact with auditory stimulus processes to give rise to distinct neural processes, each associated with a task-specific neural activation pattern across frontal and auditory regions (Bonte et al. 2014; Arsenault and Buchsbaum 2016). For example, when participants are discriminating vowels relative to talker information, there is enhanced representation of vowels in the bilateral STG (Bonte et al. 2014). Similarly, syllable information decoded from activation patterns of the human frontal areas are more prominent during production task comparing to a passive perception task (Arsenault and Buchsbaum 2016). Thus, it is possible that categorical speech perception emerges from task-dependent distributed activation patterns, in which a functionally localized brain region *does not* encode the same categorical speech information *across* task demands. In comparison, a functional specialization perspective argues that some core brain regions are specialized to represent category information *across* task demands. Although task demands may modulate the extent of brain activation patterns in a task-specific manner, the neural representation of abstract speech categories is largely resistant to changes in task demand and surface-acoustic variability (Grieser and Kuhl 1989; Kuhl 1991).

In this study, we assessed task dependence in the neural representation of speech categories by examining category representation across different tasks. In regions revealing speech category distinctions across tasks, we further evaluated the representational structure (i.e., speech dimensional specificity) of encoding. Specifically, we probed the neural coding of Mandarin lexical tone categories across 3 different tasks by using fMRI combined with both a multivariate pattern classification (MVPC) approach (Haynes and Rees 2006; Tong and Pratte 2012) and RSA (Kriegeskorte et al. 2008; Kriegeskorte and Kievit 2013). In Mandarin, 4 linguistically relevant tone categories are primarily distinguished by dynamic pitch patterns that change a word's meaning, similar to consonants and vowels. For example, the syllable /ma/ in conjunction with a high-level tone (Tone 1) means "mother", while in conjunction with a low-dipping tone (Tone 3) means "horse". The 4 tones are phonetically described by 2 distinct dimensions: pitch height and direction (Tone 1: high-level, Tone 2: low-rising, Tone 3: low-dipping, and Tone 4: high-falling). Prior behavioral work has demonstrated that pitch height and direction are critical language-universal dimensions underlying variability in tone perception (Gandour and Harshman 1978; Chandrasekaran et al. 2007b; Francis et al. 2008). Further, cross-language studies comparing tone language and non-tone language speakers revealed that language experience does not modulate pitch height, but that the relative weighting of pitch direction is language-dependent (Gandour and Harshman 1978). In particular, native speakers of tone languages weight pitch direction more than non-tone language speakers, presumably because the pitch direction dimension is more resistant to talker variability.

In the present study, native speakers of Mandarin Chinese were instructed to listen to Mandarin tone categories in various syllabic contexts (e.g., /ba3/ and /ma1/) produced by different talkers in 3 different tasks within the scanner. These tasks included: (1) passive listening, (2) silent repetition and (3) tone categorization using a button box. These tasks were selected because cognitive components related to tone category processing are argued to vary across tasks significantly (e.g., less in silent repetition and more in tone categorization). Here, we first conducted univariate activation-based analyses on each of these tasks to identify brain networks underlying task-related processing of speech information. Second, we conducted a "cross-task" decoding approach in combination with a searchlight algorithm (Kriegeskorte et al. 2006; Fairhall and Caramazza 2013; Simanova et al. 2014) to determine the brain areas revealing above-chance classification of tone categories. Third, RSA (Kriegeskorte et al. 2008) was employed to reveal the representational structure in task-invariant core regions. Specifically, we constructed 3 theory-driven dissimilarity models to uncover the extent to which different regions within the fronto-auditory system represent different feature dimensions related to tone processing as well as how brain regions combine these dimensions to form abstract speech categories. Our results demonstrate that support for core regions that disambiguate speech categories across task demands and irrespective of surface-acoustic variability. These core regions were identified in the STG and inferior parietal lobule (IPL). Notably, frontal regions did not yield speech category-related information. As a post hoc analysis to assess the role of frontal regions identified in prior studies (Myers 2007; Myers et al. 2009; Lee et al. 2012), we conducted an additional voxel-wise parametric modulation analysis to reveal the relationship between perceptual categorical confusability and brain activation during tone categorization. This additional analysis revealed neural activations in frontal and parietal regions when a participant specifically makes an overt tone category decision and validated a specific hypothesis that these areas are involved in the mediating competition between exemplars.

Materials and Methods

Participants

A total of 30 right-handed native speakers of Mandarin participated in the MRI experiment (13 male; age = 23.1 ± 2.2 [mean ± SD] years). They reported normal hearing ability, had normal or corrected to normal vision, and were without neurological impairment, as confirmed by self-report, questionnaires, and interviews. All participants were native Mandarin speakers, and they had scores higher than second-class upper-level on the Putonghua Proficiency Test, indicating high-level proficiency in the ability to speak in Mandarin. We excluded 2 participants from further analysis due to poor task performance (accuracy <70%, *N* = 1) in the tone categorization task or due to excessive head movements (>2 mm in any directions, *N* = 1). Before the experiment, all participants signed written informed consent forms approved by the Institutional Review Boards of South China Normal University and The Chinese University of Hong Kong.

Stimulus Construction

Natural exemplars (*N* = 80) of the 4 Mandarin tones were produced by 4 native Mandarin speakers (originally from Beijing; 2 female) in the context of 5 monosyllabic Mandarin syllables (/bu/, /di/, /lu/, /ma/, /mi/). These tones were characterized by fundamental frequency (F0) height and slope variations, such as high-flat (tone 1), low-rising (tone 2), high-falling (tone 3), low-dipping (tone 4). The stimuli were recorded using 16-bit quantization and a 44.1-kHz sampling rate in a sound-isolated booth. The stimuli were normalized for RMS amplitude of 70 dB and duration of 442 ms (Perrachione et al. 2011). Independent native speakers (*N* = 5) correctly identified the 4 tone categories (>95%) and scored the stimuli as highly natural.

Tasks and Procedure

The fMRI experiment consisted of 3 different task sessions during scanning, including a passive listening task, a silent repetition task, and a tone categorization task (see Fig. 1). In the passive listening task, the participants were instructed to listen to the speech sounds without making any behavioral response. In the silent repetition task, participants were instructed to pronounce the item that they just heard covertly. Covert repetitions were utilized to minimize head movements. To make sure the participants internally produced the speech sounds during scanning, they were asked to produce sounds overtly first (1 min) and did the same task covertly for practice purposes before the fMRI experiment. In the tone categorization task, participants were required to judge which tone category they just heard by pressing a “1”, “2”, “3”, or “4” button, which corresponded to their left index, middle and right index, middle finger respectively, counterbalanced across participants. The participants also briefly practiced before scanning to establish the category-response mapping. Since the goal was to

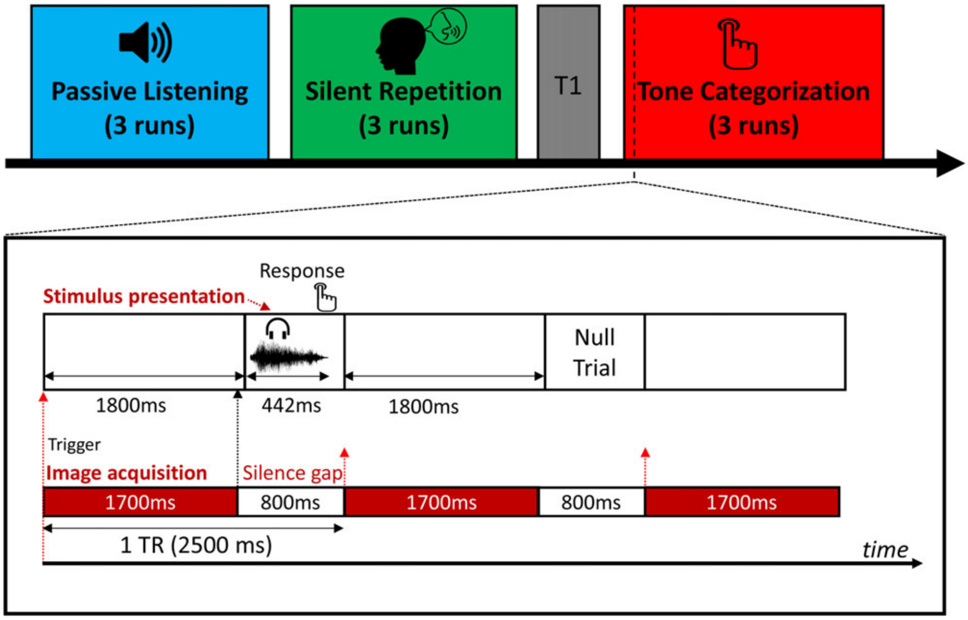

Figure 1. MRI scanning and stimulus presentation procedures. Spare sampling with 800-ms silent gaps was employed. Sound stimuli were presented during the silent gap. Button responses were required in the tone categorization task. No overt response was required for other 2 tasks.

assess tone representations across task, the specific task order (passive listening, silent repetition, and tone categorization) was crucial. To avoid interference from the preceding task (Stevens et al. 2010; Tomasi et al. 2014; Tung et al. 2013), we used a fixed task order, in which the passive listening task was performed first, and the tone categorization task was always performed last. We asked participants to perform the tone categorization task last as this task is most related to tone categorization and could potentially exert a greater influence on the other tasks (e.g., paying more attention to the tonal pattern instead of other speech information), if the tone categorization task was performed before the silent repetition task or the passive listening task.

Each task consisted of 3 runs. In each run, the stimuli were presented on a screen using an MRI-compatible LCD projector. Stimulus presentation and data collection were controlled by E-Prime (Psychology Software Tools; version 2.0). The stimulus presentation schema is described in Figure 1. We employed a spare-sampling sequence with an 800-ms silence gap between each imaging acquisition to reduce the interference of scanner noise on neural activity related to speech categories. Therefore, each stimulus was presented within each silence gap after each imaging acquisition scan (Fig. 1, lower panel). We designed an E-Prime program to receive continued trigger signals from the MRI scanner so that the onset of each trial was synchronized with the onset of each image acquisition. Also, to minimize the forward masking effect induced by scanning noise, we added a 100-ms silence period before the presentation of each stimulus. We presented each of the 80 stimuli once in a random order for each run. Therefore, there were 60 sound trials per each tone category were presented (240 sound trials total) for each task context. To better estimate the hemodynamic response to each item, we randomly added 20 null trials (i.e., silence, duration = 5 s) between sound trials as jittered intertrial intervals in each run. Therefore, each run consisted of 100 trials lasting about 4.8 minutes. We recorded each participant's response and reaction time (RT) in each trial during the tone categorization task.

MRI Data Acquisition

MRI data were acquired using a Siemens 3 T Tim Trio MRI system with a 12-channel head coil at the South China Normal University. Functional images were recorded using a T2*-weighted gradient echo-planar imaging (EPI) pulse sequence [repetition time (TR) = 2500 ms with 800-ms silence gap, TE = 30 ms, flip angle = 90°, 31 slices, field of view = 224 × 224 mm^2, in-plane resolution = 3.5 × 3.5 mm^2, 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^3).

MRI Data Preprocessing

All imaging data were preprocessed using SPM8 (Wellcome Department of Imaging Neuroscience; www.fil.ion.ucl.ac.uk/spm/). For the voxel-wise univariate activation analyses, the preprocessing procedure included correction for head movement, coregistration between structural and EPI images, normalization to a standard T1 template in the Montreal Neurological Institute (MNI) space by using segmentation-normalization procedure. The normalized images were then resampled to 2 × 2 × 2 mm^3 voxel size and underwent smoothing with a Gaussian kernel of 6-mm full width at half maximum. The preprocessing steps for the multivariate pattern analyses (both classification and representational similarity analyses) only included head movement correction and coregistration between EPI and T1-weighted images.

Univariate Activation-Based Analysis

To identify brain regions that activated in each task context, we performed subject-level analysis by using the general linear model (GLM). The design matrix of each task (passive listening, silent repetition, and tone categorization) was constructed and modeled separately. Within each task, a regressor of interest corresponding to the onset of the sound presentation was convolved with the canonical hemodynamic response function. We removed low-frequency drifts by using a temporal high-pass filter (cutoff at 128 s) and used the AR1 correction for autocorrelation. Six head movement parameters and the session mean were also added into each design matrix as nuisance regressors. The standard gray matter volume created from the segmentation step for each participant was used as an inclusive mask to restrict voxels of interest. In the group-level analysis, we used a random-effect GLM model. In each task, we used a one-sample *t*-test to identify brain areas that were activated during stimulus presentation. Brain maps were first thresholded at voxel-wise P = 0.005, and all reported brain areas have been corrected P = 0.05 at the cluster-level using the family-wise error (FWE) as implemented in the SPM package.

Multivariate pattern analysis

To investigate the local representation of tone category information and to further reveal how different acoustic dimensions of tone were encoded in the human brain, we first used MVPC in combination with a whole-brain searchlight procedure (Kriegeskorte et al. 2006) that selected local sphere voxels for training and testing. Further, we used a RSA to investigate how the brain areas identified from the MVPC above represented different dimensions of tone category information. Detailed methodological steps consisting of the construction of the fMRI feature space (fMRI feature construction and extraction), as well as the cross-task cross-validation (CV) procedure and RSA model construction, are described below.

MVPC Analysis

We conducted MVPA on data following realignment, but without normalization or smoothing, generating statistical maps for each participant. The resulting unsmoothed data for each participant in their native space were analyzed using the GLM with individual regressors for each item (e.g., /bu1/, collapsed across 4 talkers and 3 repetitions, 12 trials) that was used to calculate single-item *t*-statistic maps for each task. In addition to the stimulus regressors, 6 head movement regressors and a session mean regressor for each run were also included in the design matrix. The *t*-statistic maps were further used for multivariate analysis. We chose the *T*-statistic because it combines the effect size weighted by error variance so that it is not influenced by highly variable item estimates (Misaki et al. 2010). We used the searchlight algorithm (Kriegeskorte et al. 2006) to investigate neural representations of Mandarin tone categories by using a linear support vector machine (SVM) classifier as implemented in the LIBSVM toolbox (Chang and Lin 2011) and CoSMoMVPA toolbox (Oosterhof et al. 2016). Classifiers were trained and tested with each subject's data. At each voxel, sound induced activation values (*t*-values) for each item within a spherical searchlight (3-voxel-radius sphere, average contains 123 voxels) were extracted in each task. Therefore, in each spherical searchlight, $V \times I \times T$ value matrix was constructed, where V referred to voxel, I referred to item and T referred to task (i.e., $123 \times 20 \times 3$). This matrix was input to an SVM classifier for training and testing.

We operationally define "task-general" neural representation of speech categories as representations that emerge from multivoxel activation patterns across 3 different tasks. To investigate the extent to which category-related information can be decoded from brain activation patterns across tasks, we employed a leave-one-task-out CV procedure, wherein the classifier was trained on data set from 2 tasks and subsequently tested on the remained task data set. We repeated this procedure 3 times. Thus, only the tone category information common across tasks was informative to the classifier. Finally, mean classification accuracy was calculated and mapped back to the voxel at the center of each searchlight sphere. We conducted the same procedure across all voxels in the brain and generated classification accuracy brain maps for each participant. For comparison purposes, we also used the same searchlight procedure to classify syllables (5 classes), consonants (4 classes), and vowels (3 classes), respectively.

To investigate the "surface-acoustics-invariant" neural representation of speech category, we conducted cross-talker and cross-exemplar whole-brain searchlight classification analyses. In the cross-talker CV classification analysis, we divided all the data into 2-fold based on different talkers (female vs. male) irrespective of the task. To achieve this, we constructed another first-level GLM analysis in which we modeled male and female talkers' item separately. Therefore, there were 40 regressors in the design matrix (20 items by 2 talkers) for each task. We trained an SVM classifier by using the male talker's items and tested the classifier by using the female talker's items, and vice versa. Thus, only the tone (or syllable identity) information common across talkers was informative to the classifier. In the second cross-exemplar CV classification analysis, we divided all the data into k folds irrespective of the task, in which $k = 5$ (based on syllable) if we classified tone category, whereas $k = 4$ (based on tone) if we classified syllable identity. We subsequently conducted a whole-brain searchlight classification analysis with the k-fold CV. Therefore, only the tone (or syllable identity) information common across exemplars was informative to the classifier. The 2 CV classification analyses were conducted independently of task information to the classifier. Thus, if this approach identifies comparable brain regions, this would provide converging evidence of task-general and surface-acoustic-invariant neural representation of speech categories.

Additionally, we also conducted whole-brain searchlight classifications for tone category and syllable identity within each task to reveal task-specific neural representations of speech categories. During fMRI scanning, there were 3 runs for each task. Here, we constructed new first-level GLM models for each task, in which we modeled each item individually while collapsing the same item across talkers for each run. Therefore, each run consists of 20 items. We used a leave-one-run-out CV procedure to conduct the classification analysis for each task separately.

For the whole-brain group-level analysis, the classification accuracy map for each subject was first normalized to MNI space using the parameters estimated from the segmentation step and then entered into a one-sample t-test. All group statistical maps from multivariate analyses were thresholded at voxel-vise uncorrected $P < 0.005$, with cluster-level FWE-corrected $P < 0.05$. In addition to the whole-brain MVPC, we also conducted an ROI-based MVPC analysis on each predefined brain region to reveal brain representation differences between tone categories and other speech information (i.e., syllable identity, consonant, and vowel). All ROI brain masks were first defined and constructed in the standardized MNI space and then projected those masks back to the native space for each participant. We extracted brain activation patterns in the native subject space and further conducted classification analyses.

Representational Similarity Analysis

We used RSA (Kriegeskorte et al. 2008; Kriegeskorte and Kievit 2013) to delineate the relationship between neural representation similarity and stimulus-derived perceptual similarity for those regions that yielded significantly above-chance tone classification. Three hypothesized representational dissimilarity matrices (DSMs) were created according to unidimensional fundamental frequency (F0) height (pitch height), F0 slope (pitch direction), and multidimensional (F0 height plus F0 slope) respectively (see Fig. 4 for graphical illustration) (Maddox et al. 2014; Chandrasekaran et al. 2015; Yi et al. 2016). To create the pitch height model, we calculated the distance between each pair of items according to their F0 height. We constructed the pitch direction model in the same way but instead in accordance with each pair's F0 slope. For the multidimensional model, we first created a 2-dimensional space according to both F0 height and F0 slope dimensions (Fig. 4A). The distance between each pair in this 2-dimensional space was computed and converted into a distance matrix (i.e., dissimilarity matrix). We then normalized these distance DSMs by scaling between 0 (low dissimilarity, i.e., close in the distance) and 1 (high dissimilarity, i.e., far from each other in the distance). A binary tone category model was also constructed based on combinations of the 4 tone categories (i.e., 0 for the same category, 1 for different category). We observed a certain degree of correlation between these 4 DSMs. The Spearman's rank correlation between pitch height and pitch direction model was $rho = -0.17$ ($P = 0.02$), while $rho = 0.71$ ($P < 0.001$) between the multidimensional and tone category models. The correlation between the pitch height and multidimensional models was $rho = 0.43$ ($P < 0.001$), and $rho = 0.51$ ($P < 0.001$) between the pitch height and tone category models. The correlation between the pitch direction and multidimensional models was $rho = 0.79$ ($P < 0.001$), and $rho = 0.46$ ($P < 0.001$) between the pitch direction and tone category models.

Voxel activation values (*t-statistic* values) of each item within each ROI mask were extracted to calculate dissimilarity (using 1—Pearson's correlation) between each pair of items for creating a neural dissimilarity matrix. This neural DSM then correlated with each theoretical model DSM by using Spearman's rank correlation. In additional RSA analyses, to investigate unique contribution of each theoretical model DSM, fMRI DSM was correlated with each theoretical model DSM while controlling for effect of other theoretical model DSMs by using partial Spearman's rank correlation.

Additional Data Analyses: Voxel-Wise Parametric Modulation Analysis.

We examined the relationship between the tone confusability and brain activity in the categorization task by using an item-by-item parametric modulation analysis (Buchel et al. 1996, 1998). As a complementary analysis, this analysis highlights regions involved in the active tone categorical perception beyond the brain representation of tone categories. Here, we aimed to identify brain regions that would be more activated when an item is more perceptually confused with other items in tone categories. Because, the behavioral performance of the tone categorization task was the ceiling for the native Mandarin participants, we quantified the tone confusability by measuring the inverse acoustic distance between within-category and between-category items (see Fig. 6A). Here, we computed a perceptual tone confusion index (*CI*) according to the multidimensional model for each item by using the following equation:

$$CI\,(i) = 1 - \left| \frac{1}{N} \sum_{i \neq j} BD_{ij} - \frac{1}{n} \sum_{i \neq k} WD_{ik} \right|$$

Here, n denotes the number of items that belong to the same tone category as item i (i.e., within-category items), and N denotes the number of items that belong to different categories (i.e., between-category items). BD_{ij} refers to the between-category distance (BD) between items i and j, while WD_{ik} refers to the within-category distance (WD) between items i and k. Thus, $CI(i)$ is the summarized tone confusion score of item i (see Fig. 6A for details). Then, these confusion values were z-transformed before entering into further analysis.

In the subject-level analysis, the CI for each item was used as a parametric modulation weight in the design matrix. Also, we included the button press regressor, "−1" for the left-hand button press and "1" for the right-hand button press, in the design matrix to regress out button press response related effects. Head motion parameters and the session mean were also modeled separately as nuisance regressors. In the group-level analysis, a one-sample *t*-test was used to define significant voxels. Thus, the result would reveal which regions show a monotonic modulation in activity as a function of tone confusion.

Results

Behavioral Performance

The native Chinese participants achieved nearly perfect tone categorization performance (mean accuracy = 96.6%, SD = 3.6; mean RT = 1055.8 ms, SD = 151.3) in the tone categorization task. Also, to determine the relationship between RT and confusion index (CI), we conducted a correlation analysis between these 2 variables across items for each participant. At the group-level, we found that RT-CI correlation was significantly higher than null hypothesis ($t_{(27)} = 4.19$, $P < 0.001$), which suggests that items that were more perceptually confusable with other exemplars required a longer time to respond.

Univariate Brain Activation

The univariate GLM analysis revealed robust common activation in the auditory cortex (bilateral Heschl's gyrus and STG) induced by the sound stimulus compared with baseline across tasks (Fig. 2A). We also observed different activation patterns in different tasks. In the passive listening task, besides the bilateral auditory cortex, the dorsal superior frontal gyrus adjacent to supplementary motor areas (SMA) and the superior parietal cortex were also activated, while the bilateral precentral gyrus, SMA, bilateral putamen, and cerebellum were found to be activated in the silent repetition task. Furthermore, during the tone categorization task, we observed bilateral STG as well as attention- and motor-related region activations, which consisted of the bilateral middle frontal gyrus, precentral gyrus, bilateral inferior and superior parietal regions, as well as SMA and bilateral putamen activation.

Cross-Task Searchlight Classification of Speech Category

The overall cross-task tone category and syllable identity classification maps, attained by the whole-brain searchlight multi-voxel pattern classification, was determined separately. We found 3 regions associated with significantly above-chance tone category classification performance (see Supplementary Fig. S6 for searchlight classification results with different CV procedures), including the left anterior superior temporal gyrus (LaSTG, peak MNI coordinates: $x = -52$, $y = -18$, $z = 2$), right superior temporal gyrus (RSTG, peak MNI coordinates: $x = 62$,

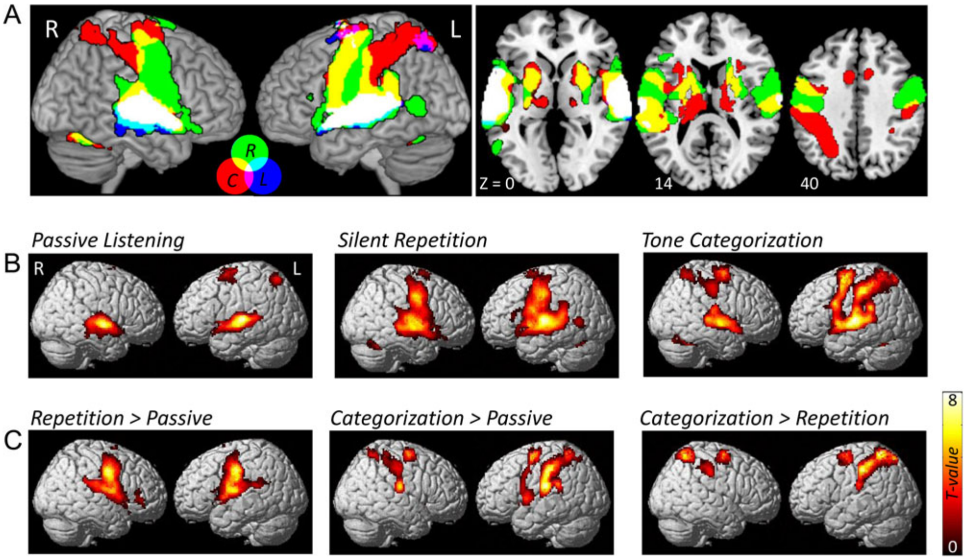

Figure 2. Brain activation maps of sound versus baseline for each task. (A) Overlap brain map of sound activations across the 3 tasks. Label abbreviation: *L*, Passive Listening; *R*, Silent Repetition; *C*, Tone Categorization. (B) Sound versus baseline activation maps for each task. (C) Differences in brain activation between each pair of tasks.

y = −24, z = 12) and LIPL (peak MNI coordinates: x = −32, y = −36, z = 40). In contrast, only bilateral STG were found to be significant in classifying syllable identity (Fig. 3A, see Supplementary Fig. S3 for slice view of the results). The determination of the neural representation of tone categories is also different from that of other segmental information (e.g., consonants and vowels). We found that the right STG was significantly above-chance and most salient for both consonant and vowel classification (see Supplementary Fig. S1), which was similar to the syllable identity classification searchlight pattern. Although the task-specific neural representation of tone category and syllable identity are not the focus of the present study, we conducted within-task searchlight classification for each task separately for completeness. Supplementary Figure S2 showed the task-specific brain representation of tone categories and syllable identity, respectively.

To further delineate the differences between tone and syllable classification in brain geometry, we conducted additional ROI-based classification analyses to compare classification performance between the anterior and posterior LSTG, and between left and right STG. The upper panel of the Figure 3B shows a sliced view of the 2 subregions of the LSTG that were derived from the searchlight tone and syllable classification maps, respectively. For visualization purposes, we extracted all voxels' classification performance (*t*-values) within an anatomically defined LSTG mask derived from the atlas of Automated Anatomical Labeling (Tzourio-Mazoyer et al., 2002) for separate tone and syllable classification. This quantitative analysis showed that tone-decoding ability was gradually increased from the posterior to the anterior of STG, while the opposite pattern was observed for syllable-decoding (Fig. 3B, lower panel). To further access the anterior–posterior segregation in speech information representation, we examined syllable classification performance in the tone-decoding mask (LaSTG) and examined tone classification performance in the syllable-decoding mask (LpSTG). Note that the ROI selection is independent of the classification analysis we conducted to avoid double dipping bias. We found that tone classification performance was not significantly above-chance for the syllable-decoding LpSTG ($t_{(27)}$ = 1.13, P = 0.27). Similarly, syllable identity classification performance was not significantly above-chance for the tone-decoding LaSTG ($t_{(27)}$ = 0.59, P = 0.56). These results suggest that there is an anterior–posterior segregation in the representation of tone category and syllable identity information.

In the second analysis, we compared the classification performance between the left and right STG for tone category and syllable identity separately. First, 2 STG masks (LSTG and RSTG) were created by using a conjunction analysis to identify the common brain regions that were activated in sound versus baseline across tasks (Fig. 2A, brain areas in white color). Then, the activation patterns of the left and right STG for each item were extracted and fed into the cross-task classification procedure separately to generate classification accuracy for each subject. We conducted a 2-by-2 repeated measure ANOVA (hemisphere-by-classification type). We found a significant interaction effect (F = 9.49, P = 0.0045), and a main effect of classification type (F = 26.67, P < 0.001). However, the main effect of hemisphere was not significant (F = 0.79, P = 0.38). In addition, We found apparent left laterality of the STG for tone decoding, while quantitative rather than qualitative, with weaker effects being evident in the right hemispheric regions (left STG: $t_{(27)}$ = 3.92, P < 0.001, right STG: $t_{(27)}$ = 1.88, P = 0.07; left versus right STG: $t_{(27)}$ = 1.29, P = 0.209). In contrast, right laterality of STG both quantitatively and qualitatively was more evident for syllable-decoding (left: $t_{(27)}$ = 0.51, P = 0.61; right: $t_{(27)}$ = 3.16, P = 0.003; left versus right STG: $t_{(27)}$ = 2.62, P = 0.01). Altogether,

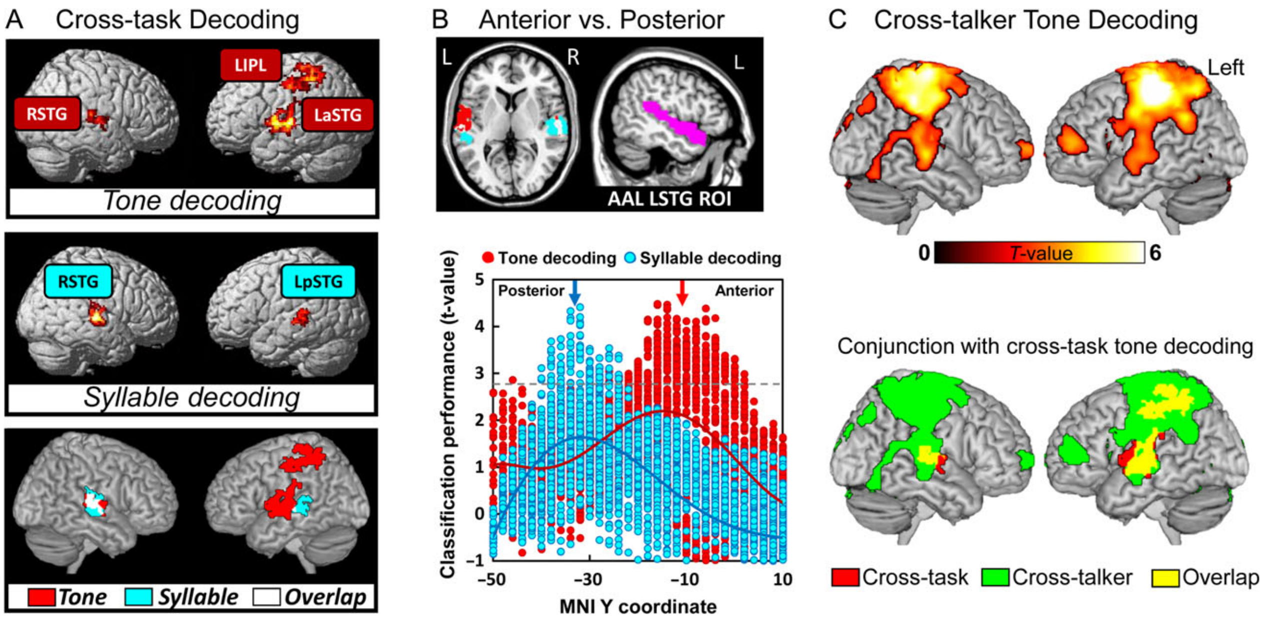

Figure 3. Brain-based classification of speech categories. (*A*) Searchlight brain classification maps of cross-task tone decoding and syllable-decoding. Voxel-level P < 0.005, cluster-level FWE-corrected P < 0.05. The left-side brain is right hemisphere and the right-side brain is left hemisphere. Upper panel, the whole-brain searchlight results of tone classification; middle panel: the searchlight results of syllable identity classification; lower panel, conjunction map of tone and syllable classification. (*B*) Anterior–posterior gradient of tone category and syllable identity representation in the LaSTG. Upper panel left, slice view of anterior–posterior separation from whole-brain searchlight analyses; Upper panel right, anatomical defined LSTG mask derived from the atlas of Automated Anatomical Labeling (AAL) (Tzourio-Mazoyer et al., 2002); Bottom panel, Dissociation between tone category and syllable identity classification performance in anatomical location (anterior vs. posterior STG). The left anterior STG was more sensitive to tone category than the left posterior STG. (*C*) Cross-talker whole-brain searchlight classification results (top panel) and conjuction with cross-task tone decoding results (buttom panel).

these results reveal the topological and hemispheric differences in neural representation of tone category and syllable identity.

Cross-Exemplar Searchlight Classification of Speech Category

We conducted additional ROI-based and whole-brain searchlight classification analyses to address the extent to which the "core representation" of tone categories is surface-acoustic-invariant to talker and syllable information. Here, we define surface-acoustic-invariant tone category representation as tone category information that is shared across exemplars (i.e., talkers or syllables). Therefore, we constructed another CV procedure that trained classifiers from some exemplars and tested the classifier on the remaining exemplar. First, we conducted an ROI-based cross-talker CV procedure to classify tone categories. To achieve this, we constructed another first-level GLM analysis in which we modeled male and female talker separately. Therefore, in each task, there were 40 regressors in the design matrix (20 items, 2 talkers). We trained the SVM classifier by using the male talker's items and tested the classifier by using the female talker's item, and vice versa. We found that the cross-talker tone classification was significantly above-chance for all the 3 regions (LaSTG, $t_{(27)} = 2.59$, $P = 0.014$; RSTG, $t_{(27)} = 3.51$, $P = 0.002$; LIPL, $t_{(27)} = 4.61$, $P < 0.001$). We also conducted a whole-brain searchlight cross-talker tone classification analysis to confirm the ROI analysis results. Figure 3C shows the cross-talker whole-brain searchlight tone classification results. We found that there were largely overlapping brain areas between the cross-task and cross-talker classification brain maps (Fig. 3C, bottom panel; also see Supplementary Fig. S4A). These findings suggest that tone category representation is talker-invariant for the 3 brain areas that were identified by the cross-task CV procedure.

Second, we used a cross-syllable CV procedure to confirm our findings further. We divided all the data into 5 folds according to the syllable identity. We trained the SVM classifier by using items from 4 syllables and tested the classifier on the remaining syllable. This procedure was repeated 5 times, and the average accuracy was computed. A ROI-based cross-syllable tone classification analysis revealed that the classification performance were significantly better than chance for all the 3 regions (LaSTG, $t_{(27)} = 2.98$, $P = 0.006$; RSTG, $t_{(27)} = 2.45$, $P = 0.021$; LIPL, $t_{(27)} = 5.30$, $P < 0.001$). A whole-brain searchlight analysis further confirmed this finding (see Supplementary Fig. S4B). In summary, both cross-talker and cross-syllable tone classification analyses confirmed that tone category information is shared across exemplars.

RSA Results

The searchlight classification analysis allowed us to identify brain regions with a high overall sensitivity to tone categories but did not enable us to assess the detailed relationship between items or between categories based on the multivoxel patterns. By using RSA, we further delineated what aspect of speech information content was encoded in the multivoxel pattern of activity in the 3 identified regions (i.e., LaSTG, RSTG, and LIPL). Pitch height, pitch direction, and multidimensional (height plus direction) dissimilarity matrices (see Fig. 4B) were derived from the speech signals, representing different sources of stimulus information.

We found that these regions were sensitive to different dimensions of speech signal. First, the neural DSM of all 3 regions were significantly correlated with the pitch height model (Fig. 5B; LaSTG: $t_{(27)} = 4.85$, $P < 0.001$; RSTG: $t_{(27)} = 4.42$, $P < 0.001$; LIPL: $t_{(27)} = 6.51$, $P < 0.001$). Such effects were still statistically significant even though the contribution of the pitch direction model was controlled using a partial correlation approach (RSTG: $t_{(27)} = 4.38$, $P < 0.001$; LaSTG: $t_{(27)} = 5.06$, $P < 0.001$; LIPL: $t_{(27)} = 6.84$, $P < 0.001$; see Figure 5B bar graphs with label D).

Second, the pitch direction model was not significantly associated with neural DSM in the RSTG ($t_{(27)} = -0.15$, $P = 0.88$; controlling for the pitch height model: $t_{(27)} = 0.19$, $P = 0.85$), but the pitch direction model was marginally significantly related to the neural DSM in the LaSTG ($t_{(27)} = 1.86$, $P = 0.07$; controlling for the pitch height model: LaSTG: $t_{(27)} = 2.11$, $P = 0.04$). We found the most robust effect in the LIPL even when the variance of pitch height model was controlled (LIPL: $t_{(27)} = 7.05$, $P < 0.001$).

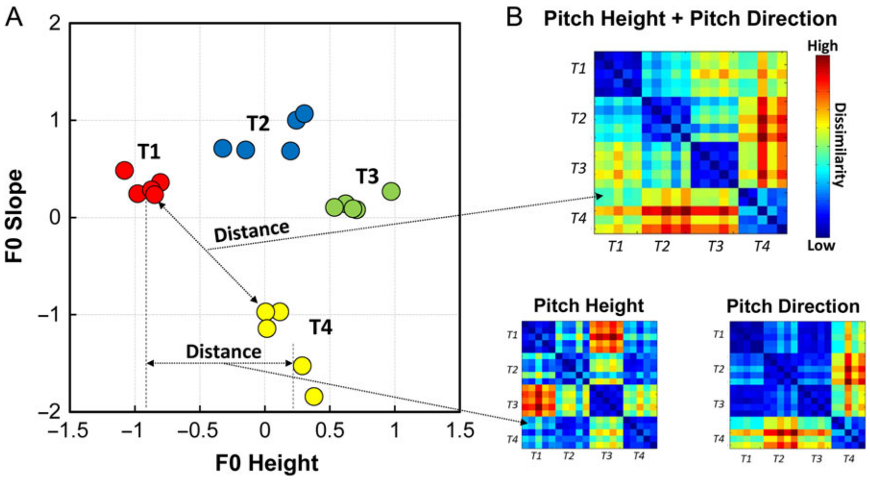

Figure 4. Illustration of constructing stimulus-derived dissimilarity matrices (DSM). (A) Scatterplot of all items (averaged and normalized across talkers) from the fMRI experiment is showed in a 2-dimension (F0 height, a correlate of pitch height; F0 slope, represents pitch direction) space. (B) Three stimulus-derived DSMs were constructed by calculating Euclidean distance between each pair of items based on unidimensional F0 height, F0 slope, and the 2-dimensional space, separately. The DSMs were then scaled to between 0 and 1, in which blue color squares indicate high similarity while warm color squares indicate high dissimilarity. T in the figure refer to tone.

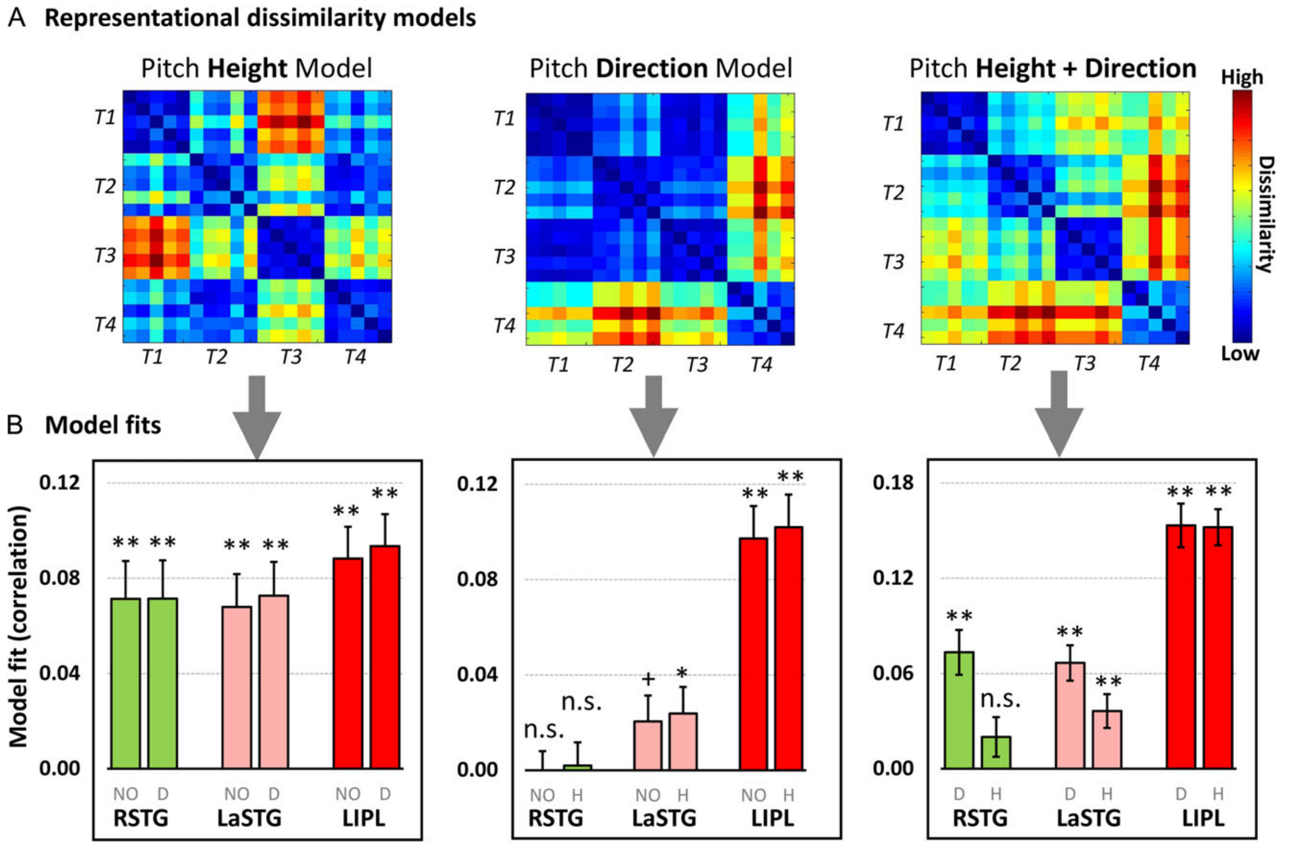

Figure 5. RSA on the 3 regions that were identified by the searchlight cross-task tone classification analysis. (A) Three hypothesized dissimilarity matrices, which were constructed by calculating the distance between each pair based on pitch height, direction and both height and direction dimensions (see Fig. 4 for details); (B) Model fits (Spearman rank correlations) between hypothesized models and fMRI dissimilarity matrices were showed. The gray color labels under each bar represent different model fit methods: NO, not controlled the variance of any model; D, controlled the variance of the pitch direction model; H, controlled the variance of the pitch height models. **, $P < 0.01$; *, $P < 0.05$; $^{+}P < 0.1$; n.s., not signification. Error bars denote standard error of the mean.

Thirdly, the multidimensional model (i.e., pitch height plus direction model) was significantly correlated with neural DSM for all 3 regions (RSTG: $t_{(27)} = 3.47$, $P < 0.001$; LaSTG: $t_{(27)} = 4.76$, $P < 0.001$; LIPL: $t_{(27)} = 9.83$, $P < 0.001$). However, this effect was diminished in RSTG when the variance of pitch height model was controlled ($t_{(27)} = 1.68$, $P = 0.1$), which indicated that the RSTG was dominantly associated with the representation of pitch height information. In contrast, this multidimensional model was still significantly correlated with neural DSM in both the LaSTG (controlling for pitch height model: $t_{(27)} = 3.14$, $P = 0.004$; controlling for pitch direction model: $t_{(27)} = 4.75$, $P < 0.001$) and the LIPL (controlling for pitch height: $t_{(27)} = 8.22$, $P < 0.001$; controlling for pitch direction: $t_{(27)} = 9.19$, $P < 0.001$) even either pitch height or direction model was controlled (Fig. 5, right panel). These results suggested that the neural activity pattern in the LaSTG and LIPL could be best characterized by combining multidimensional speech information (pitch height and direction).

Voxel-wise Parametric Modulation Analysis of Tone Category Confusion

Besides bilateral auditory cortices, previous studies have shown that activity in the prefrontal cortex can differentiate speech categories (between- vs. within-category) during speech perception (Myers et al. 2009; Myers and Swan 2012; Chevillet et al. 2013; Alho et al. 2016). Using an adaptation paradigm, activations in the prefrontal cortices are sensitive to changes between phonetic categories but insensitive to surface-acoustic changes within a category. One potential explanation for this finding is that the prefrontal cortices (e.g., inferior frontal gyrus) relate to decisional/cognitive processes rather than categorical speech representation. As an ad hoc analysis, we tested the extent to which prefrontal cortex activity is increasingly engaged for more confusable tone category items. We performed a trial-by-trial voxel-wise parametric modulation analysis looking for effects of perceptual tone category confusability. Our metric of tone confusability was derived from the multidimensional pitch height plus direction model (see Fig. 6A and "Materials and Methods" section).

We found that increased tone confusability was associated with increased activity in the left inferior frontal gyrus (LIFG, peak MNI coordination: $x = -40$, $y = 8$, $z = 24$) and LpIPL (peak MNI coordination: $x = -30$, $y = -62$, $z = 36$) (Fig. 6B). In addition to tone category information being reflected in the multivoxel activation pattern in the RSTG, LaSTG, and LIPL, local activity in fronto-parietal regions are sensitive to the extent of perceptual similarity with other between-category items, therefore requiring additional cognitive/decisional processes to differentiate category information.

Discussion

We assessed task-general and talker-invariant neural representation of native speech categories using MVPC analysis and RSA. We examined neural responses to stimulus presentation

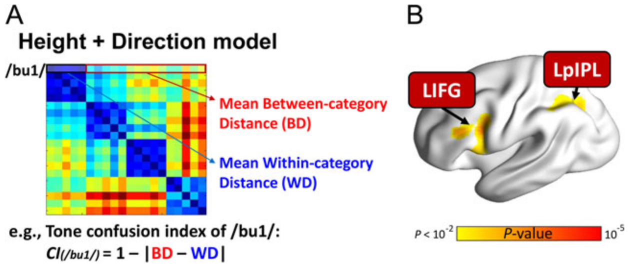
A
Height + Direction model
/bu1/
Mean Between-category Distance (BD)
Mean Within-category Distance (WD)
e.g., Tone confusion index of /bu1/:
CI(/bu1/) = 1 – |BD – WD|
B
LIFG
LpIPL
P < 10-2
P-value
10-5

Figure 6. Left inferior frontal and parietal regions were more activated when an item is more perceptually confused with other items from other tone categories compared with items from its category during tone categorization. (A) Tone confusion index was calculated for each item based on the multidimensional pitch height plus direction model. (B) Brain activation in the LIFG and LpIPL were significantly positively correlated with confusion scores. LIFG, left inferior frontal gyrus; LpIPL, left posterior inferior parietal lobule. Voxel-wise $P < 0.005$; cluster-size FWE-corrected $P < 0.05$. This figure is visualized on the standard rendered cortex surface. The threshold was set to voxel-wise $P = 0.01$ for visualization purpose.

in different task constraints: passive listening, repetition, and categorization. Activation in response to speech stimuli was task-dependent: all tasks elicited the bilateral STG; when repetition was the goal, there were greater fronto-motor-related activities; when categorization was the task goal, there was more activity in the bilateral parietal and premotor regions, compared with passive listening (Fig. 2). By employing MVPC, we found 3 core regions within a temporoparietal network, including the RSTG, LaSTG, and LIPL, which yielded a significantly above-chance classification of native Mandarin tone categories. The classification was significantly above chance, common across tasks and irrespective of surface features, which supports the functional specialization perspective (i.e., core neural representation). The brain representation of tone categories, revealed by cross-task decoding procedures is qualitatively differing from that of in syllable, consonant, and vowel. Moreover, by using RSA, we found region-specific representational differences: multivoxel patterns in the RSTG is predominantly sensitive to unidimensional pitch height information, while that in the LaSTG and LIPL represent multidimensional information related to the combination of pitch height and direction. In an additional analysis, we found that tone confusability derived from the multidimensional representational model is related to the engagement of the fronto-parietal cognitive control network during tone categorization. Specifically, we found that the left inferior frontal and parietal regions are more engaged for items that are more behaviorally confusable (validated by slower RT despite high accuracies) with other items. Altogether, our results not only identify core brain regions that represent task-general and talker-invariant speech category information but also reveal the fine-grained representational structure underlying the neural representation of tone categories. Outside these core regions, a fronto-parietal network is engaged when categorization is particularly challenging due to high confusability.

Task-General and Acoustic-Invariant Neural Representation of Native Speech Categories

In this study, we identified the neural representation of speech categories across 3 different task contexts and under various acoustic realizations (i.e., talkers and syllables). Previous studies have revealed a broad brain network involved in the representation of speech information that encompasses the anterior and posterior language areas. This widespread brain network is associated with the mapping of the acoustic signal to the internal representation of speech categories (Formisano et al. 2008; Chang et al. 2010; Lee et al. 2012; Myers and Swan 2012; Du et al. 2014; Mesgarani et al. 2014). However, prior studies have also revealed that neural representations of speech categories are highly task-dependent. Since a majority of studies have employed a single task to examine speech representation, the extent to which brain areas within the broad language network actually reflects the stored representations of speech categories is debatable. Previous findings on both visual and speech domains have shown that task demand modulates the neural representation of external signals. For example, Bonte et al. (2014) showed that the activity pattern in the bilateral STG contributing to speech category representations is sensitive to whether the participants performed a talker or a phoneme identification task on the same set of stimuli. Moreover, recruitment of frontal and motor regions in support of building speech categories representation is also shown to be related to the task context. Arsenault and Buchsbaum (2016) revealed that activation pattern in the fronto-motor cortex is related to speech classification only in a speech production task, but not in a passive speech processing task. Activity pattern in the bilateral STG, but not in fronto-motor region, can be used to decoded different dimension of speech features in a sex discrimination task (Arsenault and Buchsbaum 2015). This evidence together with findings from other domains suggested that different task contexts may trigger different cognitive processes and strategies, which would potentially warp the neural representation of specific information (Kok et al. 2012; Cukur et al. 2013). Here, we showed that bilateral STG and LIPL are core regions that represent linguistically relevant categories across task demands and irrespective of surface-acoustic variation. These findings reveal a functional specialized and focal brain network underlying forming native speech category representation.

Here, we also observed that different task demands modulated the extent of category-based information, reflected by differences in classification performance between tasks (see Supplementary Fig. S2 for within-task searchlight classification results). We found that the tone categorization task-induced higher tone classification performance in comparison to passive listening and repetition tasks. In the tone categorization task, participants are required to focus on the tone categories and plan a motor response (button-press) to indicate (tone) category information. This task only requires the participants focus on the tone information of the stimulus while ignoring syllabic/talker variation and response as accurate as possible. In contrast, during the silent repetition task, participants need to (silently) repeat what they just heard, which requires multiple cognitive-linguistic processes, including focusing on syllabic (and tone) information, and covert articulatory planning. Therefore, the amount of tone category information encoding in the activation pattern would be significantly different (much more in the tone categorization task compare to the repetition task). Nevertheless, our data show that even in tasks that do not require a primary focus on tone information (silent repetition and passive listening), we can still decode tone category information using cross-task, CV approach, which supports the functional specialization to tone information in these regions.

Fine-Grained Acoustic Structure in the Neural Representation of Tone Categories

Extensive prior work has shown that multiple dimensions underlie perception of linguistic tones (Gandour 1978; Gandour and Harshman 1978; Chandrasekaran et al. 2007a; Francis et al. 2008).

Multidimensional scaling studies have demonstrated that 2 critical pitch-related dimensions are used to disambiguate tone categories across languages: pitch height and pitch direction (Gandour and Harshman 1978; Chandrasekaran et al. 2010). Although pitch height is the dominant dimension for tone differentiation across languages (Gandour and Harshman 1978), native speakers of tone languages do not weight this dimension more than non-native speakers (Chandrasekaran et al. 2007a). In contrast, native speakers demonstrate a more robust neural encoding of dynamic pitch information, as revealed by electrophysiological studies (Krishnan et al. 2005; Chandrasekaran et al. 2009), and attend more to pitch direction as a dimension, relative to non-native speakers (Gandour 1983). Integrating pitch height and direction within a representational space allows listeners to extract tone categories with less interference from surface features (talker and syllabic variability). Indeed, when non-native speakers of tone languages are trained to categorize tones, successful learners tend to switch from focusing on unidimensional (height) cues to using more multidimensional strategies (pitch height + direction) (Chandrasekaran et al. 2016; Yi et al. 2016). Here, we employed RSA to assess the representational structure of these core regions that elicited task-general and talker-invariant tone classification. Neural similarity structures of the core regions correlate with the dominant pitch height model, but the LaSTG and LIPL regions were more sensitive to multidimensional pitch cues compared with RSTG. Pitch height information was prominent in bilateral STG, which is consistent with previous activation findings that the bilateral STG represents static pitch patterns (Warren et al. 2003; Hall and Plack 2009). Critically, both the LaSTG and LIPL are sensitive to multidimensional features, even when the variance of pitch height, the dominant dimension, is controlled. This finding is consistent with prior work that has demonstrated that activation of the left STG/STS is related to speech signal processing (Mazoyer et al. 1993; Scott et al. 2000; Narain 2003; Spitsyna et al. 2006). Also, the LSTG shows increased activation after a short-term lexical tone categorization training (Wang et al. 2003) and a sound-to-meaning lexical tone training task (Wong et al. 2007) for non-native speakers. Our findings, extending this previous observation, indicate that the LaSTG is related to activation patterns that encode multiple dimensions related to category information.

Moreover, we found that multivoxel patterns in the LIPL encoded information related to multidimensional (pitch height + direction) speech information even after the variance of pitch height, pitch direction, or tone category models was controlled. This finding suggested that the LIPL is critical in representing higher-order abstract speech information as a result of integrating different dimensions of speech information compared with functions of bilateral STG. Prior work has argued that the left IPL may be critical in computing the relative weighting of multidimensional cues in a task-specific manner (Scharinger et al. 2015), and may be a part of a domain-general network that relates to flexibility in cue utilization (Geng and Mangun 2009). Additionally, Du et al. (2014) have found that multivoxel pattern activity of the LIPL exhibited robust phoneme categorization. In the present study, we observed that activation strength in the IPL is particularly strong in the categorization task, which requires the very specific goal-directed use of pitch cues. However, in the cross-task MVPC analyses, only information related to tone category (irrespective of surface structure and task context) was captured by the classifier and generalize to the items in another task. Moreover, we have shown convergent evidence supporting LIPL representing tone category and multidimensional information by using searchlight classification and ROI-based RSA respectively. We posit that the representational structure within the LIPL contains multidimensional acoustic features that do not vary by task or surface features. However, the cues may be differentially-weighted in a goal-directed manner during the overt act of categorization.

Neural Representation of Tone Categories Differs from that of Syllable Information

Our results revealed topographical and hemispheric differences in the neural representation of tone category and syllable identity, as well as segmental information (consonants and vowels). The neural representation of tone category information was localized to the RSTG, LaSTG, and LIPL, while the neural representation of syllable identity was localized to the bilateral STG (dominant on the right hemisphere). Within the left STG, searchlight classification analysis further revealed a more anterior brain representation for tone categories, compared with that of syllable identity.

Moreover, compared with syllable representation, tone representation was more dominant in the left hemisphere. Previous studies have found evidence that lateralized activation pattern of speech processing is dependent on the linguistic relevance (native vs. non-native language) (Wang et al. 2004; Xu et al. 2006; Zatorre and Gandour 2008) and categorical nature of the stimulus (Liebenthal et al. 2005). However, since both tone and syllable information can signal word meaning in tonal languages, the topographical and hemispheric differences cannot be accounted for by differences in linguistic relevance. Another possibility is that native Chinese speakers may weight tones differently compared with other segmental information (e.g., consonants and vowels) due to differences in information value in constraining word meaning (each tone is associated with more words than consonants and vowels) and acoustic properties (dynamic vs. static). In line with this hypothesis, previous behavioral studies have found that tone is relatively more vulnerable to interference compared with other speech dimensions by using the Garner interference paradigm (Tong et al. 2008). Similarly, using a priming paradigm, researchers have shown that the syllable itself triggered implicit priming effects, whereas tone-alone prime did not (Chen et al. 2002). Underlying neural representation differences in both hemisphere and topography between tone and syllable identity may drive these behavioral effects. Further studies are required for delineating the direct relationship between the neural representation of these speech cues and behavioral consequences.

The Role of Fronto-parietal Network in Tone Categorization

Prior work has pointed towards a role for frontal regions in speech representation (Myers et al. 2009; Lee et al. 2012; Alho et al. 2016). However, our cross-task MVPC and RSA analyses did not reveal frontal areas in the core network. As an ad hoc analysis, we examined the extent to which frontal regions are driven by decisional processes that mediate competition between confusable categories (Blumstein et al. 2005; Myers 2007). Note that even though behavioral performance on the categorization task is close to the ceiling, categorization is nonetheless computationally challenging given the considerable acoustic overlap between tone categories. Moreover, we found a significantly positive correlation between the tone confusion scores and RT across items (reported in the “Results” section). Given that native listeners are excellent at categorizing tone patterns despite the variability, the goal of the parametric modulation analysis was to evaluate the brain mechanisms that assist in the resolution of category confusions (based on the acoustic-feature space, Fig. 4) in contrast to the encoding the similarity between exemplars (of a category) that is revealed by multivariate pattern analyses. Here, we found that the inferior frontal and parietal regions engage more for items that are more confusable in tone category during tone identification task. These results showed that, in addition to bilateral STG and LIPL representing different aspects of tone category information, the fronto-parietal network has greater involvement in the discrimination of more confusable items.

Consistent with our findings, brain lesion studies have shown that damage to fronto-parietal areas in the left hemisphere is related to deficits in tasks that require the discrimination or identification of speech syllables (Blumstein et al. 1977; Caplan et al. 1995). Moreover, previous imaging studies have found that frontal activation is related to the speech category by using different priming paradigms. For example, Myers et al. (2009) found that LIFG activation was insensitive to subtle within-category acoustic changes but sensitive to between-category phonetic changes by using an adaptation paradigm. In the additional parametric analysis, we used a single-item regressor to reveal the extent to which activation strength in LIFG and LpIPL are associated with tone confusability (between-category vs. within-category). We also confirmed that the multivoxel activation pattern in the LIFG and LpIPL were not significantly related to any theoretical similarity model across tasks. These convergent findings altogether imply that the LIFG and LpIPL were associated with task-related cognitive processes that may assist in decisional processes underlying categorization. However, future studies are required to reveal the functionality of these brain areas in speech perception.

Conclusion

In conclusion, we found functionally specialized core regions representing speech categories across task demands and irrespective of talker variability. We further revealed a fine-graded representational structure for each of those brain areas, demonstrating a brain system with a graded hierarchical organization representing speech information from unidimensional to multidimensional structure. These results emphasize a critical role of bilateral STG and LIPL in representing task-general and talker-invariant speech categories and providing further evidence for the role of the left inferior frontal and parietal regions in supporting the precise distinction of speech items during active categorization.

Supplementary Material

Supplementary data are available at *Cerebral Cortex* online.

Funding

National Institute on Deafness and other Communication Disorders of the National Institutes of Health under Award No. (R01DC013315 834 to B.C.), Key Project of National Social Science Foundation of China (15AZD048), and Key Project of National Natural Science Foundation of Guangdong Province, China (2014A030311016).

Notes

Conflict of Interest: None declared.

References

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

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

Arsenault JS, Buchsbaum BR. 2016. No evidence of somatotopic place of articulation feature mapping in motor cortex during passive speech perception. Psychon B Rev. 23:1231–1240.

Blumstein SE, Baker E, Goodglass H. 1977. Phonological factors in auditory comprehension in aphasia. Neuropsychologia. 15:19–30.

Blumstein SE, Myers EB, Rissman J. 2005. The perception of voice onset time: an fMRI investigation of phonetic category structure. J Cogn Neurosci. 17:1353–1366.

Boets B, Op de Beeck HP, Vandermosten M, Scott SK, Gillebert CR, Mantini D, Bulthé J, Sunaert S, Wouters J, Ghesquiere P. 2013. Intact but less accessible phonetic representations in adults with dyslexia. Science. 342:1251–1254.

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.

Buchel C, Holmes AP, Rees G, Friston KJ. 1998. Characterizing stimulus-response functions using nonlinear regressors in parametric fMRI experiments. Neuroimage. 8:140–148.

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

Caplan D, Gow D, Makris N. 1995. Analysis of lesions by MRI in stroke patients with acoustic-phonetic processing deficits. Neurology. 45:293–298.

Chandrasekaran B, Gandour JT, Krishnan A. 2007a. Neuroplasticity in the processing of pitch dimensions: a multidimensional scaling analysis of the mismatch negativity. Restor Neurol Neuros. 25:195–210.

Chandrasekaran B, Krishnan A, Gandour JT. 2007b. Mismatch negativity to pitch contours is influenced by language experience. Brain Res. 1128:148–156.

Chandrasekaran B, Krishnan A, Gandour JT. 2009. Relative influence of musical and linguistic experience on early cortical processing of pitch contours. Brain Lang. 108:1–9.

Chandrasekaran B, Sampath PD, Wong PC. 2010. Individual variability in cue-weighting and lexical tone learning. J Acoust Soc Am. 128:456–465.

Chandrasekaran B, Yi HG, Blanco NJ, McGeary JE, Maddox WT. 2015. Enhanced procedural learning of speech sound categories in a genetic variant of FOXP2. J Neurosci. 35:7808–7812.

Chandrasekaran B, Yi HG, Smayda KE, Maddox WT. 2016. Effect of explicit dimensional instruction on speech category learning. Atten Percept Psychophys. 78:566–582.

Chang CC, Lin CJ. 2011. LIBSVM: a library for Support Vector Machines. Acm T Intel Syst Tec. 2(27):21–27. 27.

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT. 2010. Categorical speech representation in human superior temporal gyrus. Nat Neurosci. 13:1428–1432.

Chen J-Y, Chen T-M, Dell GS. 2002. Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. J Men Lang. 46:751–781.

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

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

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

Cukur T, Nishimoto S, Huth AG, Gallant JL. 2013. Attention during natural vision warps semantic representation across the human brain. Nat Neurosci. 16:763–770.

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 USA. 111:7126–7131.

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.

Fairhall SL, Caramazza A. 2013. Brain regions that represent amodal conceptual knowledge. J Neurosci. 33:10552–10558.

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.

Francis AL, Ciocca V, Ma L, Fenn K. 2008. Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. J Phonetics. 36:268–294.

Gandour JT. 1978. Perceived dimensions of 13 tones: a multidimensional scaling investigation. Phonetica. 35:169–179.

Gandour JT. 1983. Tone perception in far eastern-languages. J Phonetics. 11:149–175.

Gandour JT, Harshman RA. 1978. Crosslanguage differences in tone perception: a multidimensional scaling investigation. Lang Speech. 21:1–33.

Geng JJ, Mangun GR. 2009. Anterior intraparietal sulcus is sensitive to bottom-up attention driven by stimulus salience. J Cogn Neurosci. 21:1584–1601.

Grieser D, Kuhl PK. 1989. Categorization of speech by infants: support for speech-sound prototypes. Dev Psychol. 25: 577–588.

Griffiths TD, Warren JD. 2004. What is an auditory object? Nat Rev Neurosci. 5:887–892.

Hall DA, Plack CJ. 2009. Pitch processing sites in the human auditory brain. Cereb Cortex. 19:576–585.

Haynes J-D, Rees G. 2006. Decoding mental states from brain activity in humans. Nat Rev Neurosci. 7:523–534.

Hickok G. 2009. The functional neuroanatomy of language. Phys Life Rev. 6:121–143.

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

Kok P, Jehee JF, de Lange FP. 2012. Less is more: expectation sharpens representations in the primary visual cortex. Neuron. 75:265–270.

Kriegeskorte N, Goebel R, Bandettini P. 2006. Information-based functional brain mapping. Proc Natl Acad Sci USA. 103: 3863–3868.

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

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

Krishnan A, Xu Y, Gandour J, Cariani P. 2005. Encoding of pitch in the human brainstem is sensitive to language experience. Cogn Brain Res. 25:161–168.

Kuhl PK. 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. Percept Psychophys. 50:93–107.

Lee YS, Turkeltaub P, Granger R, Raizada RD. 2012. Categorical speech processing in Broca's area: an fMRI study using multivariate pattern-based analysis. J Neurosci. 32:3942–3948.

Leonard MK, Chang EF. 2014. Dynamic speech representations in the human temporal lobe. Trends Cogn Sci. 18:472–479.

Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA. 2005. Neural substrates of phonemic perception. Cereb Cortex. 15:1621–1631.

Maddox WT, Chandrasekaran B, Smayda K, Yi HG, Koslov S, Beevers CG. 2014. Elevated depressive symptoms enhance reflexive but not reflective auditory category learning. Cortex. 58:186–198.

Mazoyer BM, Tzourio N, Frak V, Syrota A, Murayama N, Levrier O, Salamon G, Dehaene S, Cohen L, Mehler J. 1993. The cortical representation of speech. J Cogn Neurosci. 5:467–479.

Mesgarani N, Cheung C, Johnson K, Chang EF. 2014. Phonetic feature encoding in human superior temporal gyrus. Science. 343:1006–1010.

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

Myers EB. 2007. Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: an fMRI investigation. Neuropsychologia. 45:1463–1473.

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

Myers EB, Swan K. 2012. Effects of category learning on neural sensitivity to non-native phonetic categories. J Cogn Neurosci. 24:1695–1708.

Narain C. 2003. Defining a left-lateralized response specific to intelligible speech using fMRI. Cereb Cortex. 13:1362–1368.

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.

Perrachione TK, Lee J, Ha LYY, Wong PCM. 2011. Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. J Acoust Soc Am. 130:461–472.

Poeppel D. 2014. The neuroanatomic and neurophysiological infrastructure for speech and language. Curr Opin Neurobiol. 28:142–149.

Scharinger M, Henry MJ, Obleser J. 2015. Acoustic cue selection and discrimination under degradation: differential contributions of the inferior parietal and posterior temporal cortices. Neuroimage. 106:373–381.

Scott SK, Blank CC, Rosen S, Wise RJS. 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain. 123:2400–2406.

Scott SK, Johnsrude IS. 2003. The neuroanatomical and functional organization of speech perception. Trends Neurosci. 26:100–107.

Simanova I, Hagoort P, Oostenveld R, van Gerven MA. 2014. Modality-independent decoding of semantic information from the human brain. Cereb Cortex. 24:426–434.

Spitsyna G, Warren JE, Scott SK, Turkheimer FE, Wise RJS. 2006. Converging language streams in the human temporal lobe. J Neurosci. 26:7328–7336.

Stevens WD, Buckner RL, Schacter DL. 2010. Correlated low-frequency BOLD fluctuations in the resting human brain are modulated by recent experience in category-preferential visual regions. Cereb Cortex. 20:1997–2006.

Tomasi D, Wang R, Wang G-J, Volkow ND. 2014. Functional connectivity and brain activation: a synergistic approach. Cereb Cortex. 24:2619–2629.

Tong F, Pratte MS. 2012. Decoding patterns of human brain activity. Annu Rev Psychol. 63:483–509.

Tong YX, Francis AL, Gandour JT. 2008. Processing dependencies between segmental and suprasegmental features in Mandarin Chinese. Lang Cogn Proc. 23:689–708.

Tung K-C, Uh J, Mao D, Xu F, Xiao G, Lu H. 2013. Alterations in resting functional connectivity due to recent motor task. Neuroimage. 78:316–324.

Wang Y, Behne DM, Jongman A, Sereno JA. 2004. The role of linguistic experience in the hemispheric processing of lexical tone. Appl Psycholinguist. 25:449–466.

Wang Y, Sereno JA, Jongman A, Hirsch J. 2003. fMRI evidence for cortical modification during learning of Mandarin lexical tone. J Cogn Neurosci. 15:1019–1027.

Warren JD, Uppenkamp S, Patterson RD, Griffiths TD. 2003. Separating pitch chroma and pitch height in the human brain. Proc Natl Acad Sci USA. 100:10038–10042.

Wong PC, Perrachione TK, Parrish TB. 2007. Neural characteristics of successful and less successful speech and word learning in adults. Hum Brain Mapp. 28:995–1006.

Xu YS, Gandour J, Talavage T, Wong D, Dzemidzic M, Tong YX, Li XJ, Lowe M. 2006. Activation of the left planum temporale in pitch processing is shaped by language experience. Hum Brain Mapp. 27:173–183.

Yi HG, Maddox WT, Mumford JA, Chandrasekaran B. 2016. The role of corticostriatal systems in speech category learning. Cereb Cortex. 26:1409–1420.

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