Brain & Language 274 (2026) 105702

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l

Suprasegmental aspects of phonetic feature representation in human cortex: An fMRI investigation of Cantonese lexical tones

Ran Tao [a,1,*], Kaile Zhang [a,1], Yan Feng [b], Yi Weng [a], Gang Peng [a,*]

[a] Research Centre for Language, Cognition, and Neuroscience, Department of Language Science and Technology, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong Special Administrative Region of China
[b] School of Foreign Studies, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu Province, China

ARTICLE INFO

Keywords:
Lexical tone
Speech perception
Phonetic feature
Cantonese
fMRI
Representational similarity analysis

ABSTRACT

This study investigated the neural basis of lexical tone representation in Cantonese, a complex tone language that contrasts pitch height and slope to convey lexical meaning. We used sparse-sampling fMRI to measure brain activity from native Cantonese speakers performing three tasks involving tonal syllables: passive listening, silent repetition, and word identification. Behavioral performance with high identification rates confirmed effective stimulus processing. Group-level activation and multivariate pattern analyses revealed a distributed bilateral network encompassing the bilateral precentral gyri (PrCG), right superior frontal gyrus (RSFG), bilateral superior temporal gyri (STG), left inferior parietal sulcus (LIPS), and bilateral lingual gyri (BiLG), which reliably encoded tone categories. Using dissimilarity matrices constructed from tonal features and neural activation patterns, representational similarity analysis (RSA) showed bilateral STG encoding pitch height and LIPS processing pitch slope. The frontal regions, LIPS, and BiLG contribute to holistic tone processing. This contrasts with the temporal-parietal network identified in previous Mandarin studies, suggesting that Cantonese tones invoke a bilateral and more extended brain network. The inter-subject RSA results revealed significant brain-behavioral correlations in the frontal and parietal regions, suggesting that these regions are closely associated with tone categorization performance. Other regions showed non-significant correlations, indicating their involvement in tone processing but not directly predicting behavioral performance. Together, these findings enhance our understanding of the neural mechanisms underlying tone perception in complex tonal languages and highlight the intricate role of bilateral cortical networks supporting the representation of complex suprasegmental phonetic features.

1. Introduction

Lexical tones, a fundamental feature of tone languages spoken by over half the world's population (Yip, 2002), employ pitch variations to distinguish word meanings (Wang, 1967), adding a layer of complexity to language processing (Li et al., 2021). Despite extensive research on the neural basis of consonant and vowel processing (Mesgarani et al., 2014), studies on lexical tone systems are relatively scarce. Research on Mandarin Chinese, a tonal language with four tones, has revealed the neural basis of tone perception, involving regions such as the bilateral superior temporal gyrus (STG) and left precentral gyrus (LPrCG) (Liang & Du, 2018). The left auditory cortical regions show enhanced activation when processing linguistically relevant tonal patterns, whereas the right auditory cortical regions are involved in processing pitch information, regardless of linguistic relevance (Zatorre & Gandour, 2008). However, these findings from specific tonal systems raise questions about the neural architecture supporting more complex tonal languages and the specific contributions of different brain regions.

Neuroimaging studies have consistently demonstrated bilateral hemispheric involvement in lexical tone processing, although the degree of lateralization and the regions recruited may vary across tasks (Kwok et al., 2015; Liu et al., 2006). The dual-stream model of speech processing (Hickok, 2022; Hickok & Poeppel, 2007) provides a useful framework for understanding the neural underpinnings of lexical tone perception. According to this model, the ventral stream primarily supports sound-to-meaning mapping, whereas the dorsal stream mediates

* Corresponding authors.
E-mail addresses: ran.tao@polyu.edu.hk (R. Tao), kaile-keller.zhang@polyu.edu.hk (K. Zhang), yanny.feng@njust.edu.cn (Y. Feng), yi.weng@polyu.edu.hk (Y. Weng), gang.peng@polyu.edu.hk (G. Peng).
[1] Both authors contributed equally to this article.

https://doi.org/10.1016/j.bandl.2025.105702
Received 25 June 2025; Received in revised form 15 December 2025; Accepted 22 December 2025
Available online 31 December 2025

sensorimotor integration. Importantly, these two streams exhibit distinct patterns of hemispheric asymmetry. The ventral stream shows more bilateral organization, while the dorsal stream demonstrates stronger left-hemisphere dominance. Consistent with this account, some studies suggested left-hemisphere dominance for the linguistic processing of lexical tones (Shuai & Gong, 2014). Using a lexical tone perception task in the reading modality, researchers found that the left inferior frontal gyrus (LIFG), right middle temporal gyrus (RMTG), and bilateral superior temporal gyri (STG) are responsible for the perception of linguistic pitch (Kwok et al., 2016). In contrast, a previous *meta*-analysis revealed that the bilateral inferior prefrontal regions, bilateral superior temporal regions, and right caudate are significantly activated during auditory processing of lexical tones in tonal languages (Kwok et al., 2017). While these findings confirm the bilateral involvement of lexical tone, they also point to additional complexities. Specifically, the hemispheric asymmetry observed in frontal and temporal cortices during tone perception may reflect the differential engagement of ventral and dorsal streams, depending on task demands and the particular acoustic features being processed (Hickok, 2022).

The above neuroimaging findings primarily emerged from studies involving Mandarin speakers. Only a few studies have examined other East Asian tone languages, such as Thai (Gandour et al., 1998, 2000) and Cantonese (Zhang et al., 2016, 2017). Compared with non-tonal language speakers, Thai speakers showed activation in the left frontal operculum when discriminating between Thai tones (Gandour et al., 1998).This functional specialization was also observed when comparing Thai and Mandarin speakers, suggesting language-specific-experience effects (Gandour et al., 2000). An fMRI study on Cantonese lexical tone perception showed that the bilateral superior temporal gyrus (STG) was activated when listeners heard lexical tone changes, even though they were instructed to detect changes in talker identity during the task (Zhang et al., 2016). The functional role of the STG in Cantonese lexical tone processing was further supported by evidence showing that Cantonese-speaking individuals with amusia exhibited reduced and atypical activation in the right STG during lexical tone processing (Zhang et al., 2017). Although evidence from non-Mandarin tone languages remains limited, researchers generally agree that there is a shared neural basis for lexical tone processing among speakers of tone languages that is largely absent in non-tone language speakers (Kwok et al., 2017; Liang & Du, 2018).

However, it is important to recognize that extending the observed patterns from Mandarin to other tone languages may be limited because these languages can rely on different phonetic features. In Mandarin, tone processing primarily depends on pitch slope, while Cantonese relies more on pitch height (Gandour, 1983). Cantonese has a complex tonal system with six contrastive tones in open syllables (Peng, 2006), which include level tones with similar slopes but different heights and tones with similar heights but different slopes, creating a more nuanced system for investigating the neural representation of both pitch features. For example, the syllable /ji/ means “doctor” with a high-level tone, “meaning” with a mid-level tone, and “two” with a low-level tone, which is difficult to distinguish even for native speakers as the differences are primarily in pitch height (Zhang et al., 2013). The larger number of categories in Cantonese may necessitate finer-grained neural representation to support accurate perceptual identification and discrimination. The richly contrasted tonal inventory of Cantonese thus offers a valuable framework for testing whether the neural representation patterns identified in Mandarin can be generalized or instead require refinement when applied to a more acoustically complex tone system such as Cantonese.

While earlier studies primarily contrasted linguistic units with and without tonal information, more recent investigations have explored the specific acoustic features that constitute lexical tones and how these features might explain the observed lateralization patterns (Feng et al., 2018; Li et al., 2021). Lexical tones can be described using phonetic features such as pitch height (mean fundamental frequency, F0) and pitch slope (F0 curve gradient) (Peng, 2006). These acoustic parameters may be processed differently across regions and hemispheres, potentially accounting for the mixed findings regarding lateralization. Recent neuroimaging studies have provided insights into the feature-level processing of lexical tones. Feng et al. (2018) examined the neural representation of Mandarin Chinese tones, discovering that pitch height and slope are processed in distinct cortical regions. In their study, multivariate pattern analysis (MVPA) revealed task-general lexical tone decoding within the bilateral STG and left inferior parietal lobule (LIPL). Representational similarity analysis (RSA) further indicated that multivoxel patterns in the right STG are predominantly sensitive to pitch height, while those in the left STG and LIPL represent a combination of pitch height and slope. Complementing this evidence, a high-density electrocorticography (ECoG) study characterized the temporal dynamics of tone processing on the cortical surface and revealed that the superior temporal gyrus rapidly encodes pitch movements, with different neural populations responding selectively to specific tone features (Li et al., 2021).

Building on previous research that established distinct neural representations for pitch height and slope in Mandarin (Feng et al., 2018), the current study investigated the neural correlates of lexical tone processing in Cantonese to advance our understanding of suprasegmental phonetic representation. The primary research objective was to identify task-general neural patterns shared across diverse tone-related–tasks, thereby revealing common cortical mechanisms that support tone perception, covert articulation, and categorical decision-making in Cantonese. The research questions were summarized as follows: (1) Do Cantonese lexical tones, characterized by both pitch height and slope, recruit a more extensive and potentially bilateral neural network for tone processing than previously reported for Mandarin? (2) How are different pitch features (e.g., height, slope, and their combination) represented in the brain, as revealed by representational similarity analysis? (3) How do these neural representations relate to individual differences in tone categorization performance?

We hypothesize that Cantonese lexical tone perception may recruit a more extensive bilateral processing network compared to Mandarin, which may extend beyond the classic temporal-parietal network to other cortical regions, reflecting the increased complexity of tonal categories and contrasts. The brain regions identified in this network may represent tone features differently and may also relate to individual differences in the behavioral tone categorization performance. To comprehensively characterize the neural representations of lexical tones in Cantonese, the present study employed three complementary analytical approaches. First, a univariate General Linear Model (GLM) analysis was used to identify brain regions showing overall activation during the various tasks. Second, using MVPA and RSA, we examined how tonal features are represented in the brain and whether they engage distinct or additional neural circuits beyond those identified in previous Mandarin studies. Third, we employed inter-subject representational similarity analysis (IS-RSA) to link neural representations to behavioral outcomes, identifying cortical nodes where neural similarity patterns predict perceptual decision-making patterns across individuals.

2. Material and methods

2.1. Participants

A total of 42 participants (age [mean ± SD] = 23.2 ± 3.86, 20 females) were recruited from the neighborhood of The Hong Kong Polytechnic University (PolyU). All participants were native Cantonese speakers who had also acquired Mandarin (another tonal language) and English (a non-tonal language). The reported age of acquisition (AoA) for Mandarin ranged from 0 to 12 years (mean = 4.7 ± 2.43 years), while Mandarin proficiency averaged 4.6 ± 0.93 on a 7-point self-rating scale (range = 2–6). For English, AoA was 3.5 ± 1.47 years (range = 0–7) and proficiency was 4.8 ± 0.84 (range = 2–6). All participants were

right-handed, as confirmed using the Edinburgh handedness inventory (Oldfield, 1971). They reported normal hearing ability, normal or corrected-to-normal vision, and no neurological impairment. All participants signed written informed consent forms before the experimental sessions. The study was approved by the Institutional Review Board of PolyU.

2.2. Stimuli

The stimuli consisted of three Cantonese monosyllables, /fan/, /fu/, and /ji/, which respectively contained a low front, high back, and high front vowel. These were selected to cover the major corners of the Cantonese vowel space and to minimize phonetic confounds, and also were used in previous studies on tonal perception (Zhang et al., 2018). The syllable /se/ was included to familiarize the listeners with each speaker's pitch range and served as a context stimulus functioning as a pitch normalization cue: prior presentation of tones on /se/ established a perceptual reference for pitch height and slope, reducing potential confusion among level tones and facilitating talker normalization during subsequent tone processing (Tao et al., 2021; Zhang et al., 2013; Zhang et al., 2023).

Each syllable had six Cantonese tones: three level tones (T1: high-level, T3: middle-level, and T6: low-level) and three contour tones (T2: high-rising, T4: low-falling, and T5: low-rising). These tones are distinguished by systematic variations in the fundamental frequency (F0) height and slope patterns. Two native Hong Kong Cantonese speakers (one male and one female) produced all the stimuli, yielding 48 unique tokens (four syllables, six tones, and two speakers). Recordings were conducted in a soundproof room using a high-fidelity microphone with 16-bit quantization and a 44.1-kHz sampling rate. All stimuli were digitally processed to achieve a uniform root-mean-square (RMS) amplitude of 70 dB and a duration of 450 ms using Praat software (Boersma, 2001).

2.3. Tasks and procedure

The fMRI study comprised three distinct task sessions conducted during scanning: passive listening (PAS), silent repetition (REP), and word identification (WID) (Fig. 1). In the PAS task, the participants passively listened to speech sounds without making any overt responses. In the REP task, participants silently repeated each heard item covertly to minimize their head movements. To ensure that the participants internally produced the speech sounds during the scanning, they first practiced producing the sounds overtly before transitioning to silent repetition after several practice trials. In the WID task, participants identified which of the six words they heard by pressing the corresponding button, with the left button (pressed with the left thumb) indicating that they heard a word displayed on the left side of the visual cue and the right button (pressed with the right thumb) indicating that they heard a word on the right side of the cue. For each trial, six prespecified word options were presented, representing the six tones of the same base syllable (see Fig. 1, top right corner for an example), and participants practiced and established category-response mapping before scanning.

To prevent interference from preceding tasks (Feng et al., 2018; Stevens et al., 2010; Tomasi et al., 2014; Tung et al., 2013), a fixed task order was used, with the PAS task performed first and the WID task performed as the last task. The WID task was performed last because it is mostly related to tone categorization and could potentially have more influence on other tasks (e.g., paying more attention to the tonal pattern rather than other speech information) if it was performed before the PAS or REP (Feng et al., 2018).

Each task comprised four runs, during which stimuli were presented on a screen using an MRI-compatible LCD projector controlled by E-Prime (version 3.0) for stimulus presentation and data collection. The stimulus presentation sequence is illustrated in Fig. 1. To minimize the effect of scanner noise on neural activity related to speech categories, we employed a sparse-sampling sequence with a 700-ms silence gap between each imaging acquisition. Each stimulus was presented within this gap after each image acquisition scan, as shown in the lower panel of Fig. 1. We designed an E-Prime program to ensure that sounds began 100 ms after the silent scanning period, thereby reducing the forward-masking effects induced by scanner noise.

Stimulus presentations were grouped by gender within each run and counterbalanced among the runs. For each speaker, the base syllable /se/ served as context and was presented in the sequence of T1 to T6. Presenting a context is necessary for Cantonese tone perception because of its ambiguity, as shown in previous research (Tao et al., 2021; Zhang et al., 2023). Following the context, the 18 monosyllabic /fan/, /fu/, and /ji/ served as target stimuli and were presented randomly. The stimuli for each speaker were presented once in each run and repeated in four runs, resulting in 24 target trials per tone category for each task. To improve the estimation of the hemodynamic response for each item, we randomly added 12 null trials (i.e., 3-second silence) as jittered intertrial

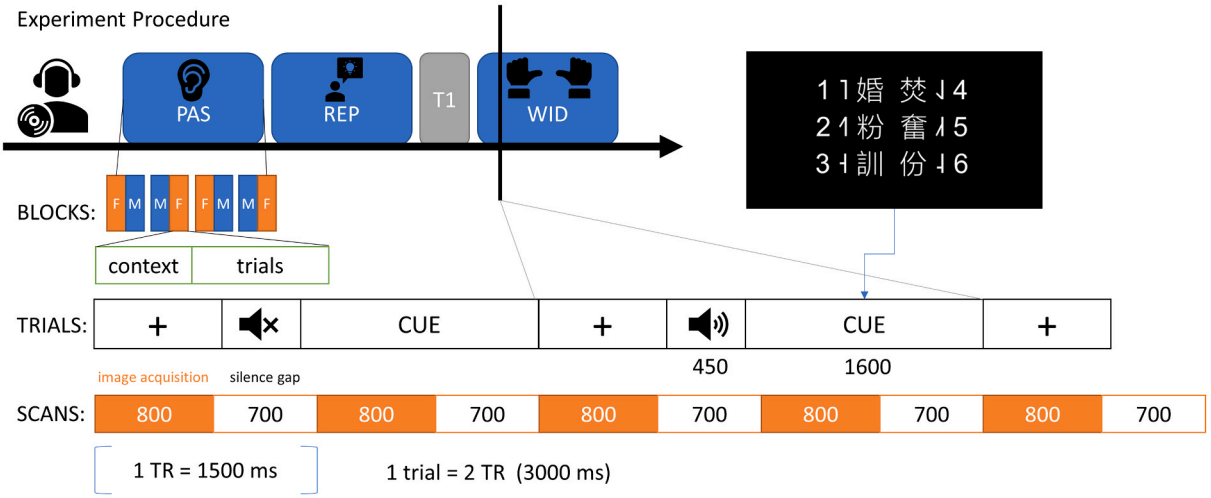

Fig. 1. Experimental procedure. Tasks included passive listening (PAS, no overt response), silent repetition (REP, covert repetition), and word identification (WID, button response). The sequence shows task blocks, trial structure (2 TRs = 3000 ms per trial, with 800 ms acquisition and 700 ms gap cycles), and stimulus presentation timing to reduce scanner noise masking. Sparse sampling with 700 ms silent gaps separated image acquisitions. Sound stimuli (Cantonese monosyllables with tones) were presented during these gaps.

intervals in each run. Each run consisted of 12 context stimuli and 48 experimental trials (target + null) lasting approximately 3 min, and we recorded each participant's responses during the WID task.

2.4. MRI data acquisition

Brain imaging was performed using a Siemens Prisma 3 T MRI Scanner equipped with a 32-channel head coil. Functional images were acquired using a T2*-weighted gradient echo-planar imaging (EPI) sequence with simultaneous multislice (SMS) acceleration. A sparse sampling paradigm was implemented to ensure stimulus presentation during quiet periods. The key acquisition parameters were: repetition time (TR) = 1500 ms including 700 ms silence gap, echo time (TE) = 38 ms, flip angle = 52°, SMS acceleration factor = 8, field of view (FoV) = 208 × 208 mm^2, 72 interleaved slices, voxel size = 2.0 × 2.0 × 2.0 mm^3, and total acquisition time (TA) for each task run = 3:02 min. The 700–ms silent gap was selected because each speech stimulus lasted for 450 ms, and the gap needed to exceed this duration so that the entire sound could occur within the noise–free period. The silent gap allowed a 100–ms delay before sound onset to minimize forward masking from the preceding scan and accommodate the natural rise/fall of the waveform. Thus, a 700–ms gap guaranteed complete stimulus delivery during the scanning silence. High-resolution structural images were acquired using a T1-weighted magnetization prepared rapid acquisition gradient echo (MPRAGE) sequence with the following parameters: TR = 3.2 ms, TE = 1.37 ms, flip angle = 8°, field of view = 260 mm, 128 sagittal slices, voxel size = 1.6 × 1.6 × 1.6 mm^3, GRAPPA acceleration factor = 3, and acquisition time = 3:22 min.

2.5. Data analysis

2.5.1. MRI data preprocessing

All imaging data were preprocessed using SPM12 (Wellcome Department of Imaging Neuroscience; https://www.fil.ion.ucl.ac.uk/spm/). For voxel-wise univariate activation analysis, the preprocessing procedure involved correcting for head movement, registering the structural and EPI images, and normalizing the images to the standard T1 template (i.e., the default MNI152 template from the Montreal Neurological Institution as implemented in SPM12) using a segmentation-normalization procedure. The normalized images were then resampled to a voxel size of 2 × 2 × 2 mm^3 and smoothed using a Gaussian kernel with a 6-mm full width at half maximum.

The preprocessing steps for the multivariate pattern analysis (both classification and RSA) only included correcting for head movement and registering EPI and T1-weighted images. Normalization and smoothing were performed after the first-level analysis to preserve the fine-grained spatial patterns in individuals.

2.5.2. Univariate activation analysis

We conducted a subject-level analysis using a general linear model (GLM) to identify the brain regions activated in each task. For each task (PAS, REP, and WID), we constructed a design matrix and modeled neural activity separately. We convolved a regressor of interest corresponding to the onset of sound presentation with a canonical hemodynamic response function for each task. To remove low-frequency drifts, we applied a temporal high-pass filter (cutoff at 128 s) and used the AR1 correction for autocorrelation. We also added six head-movement parameters and the session mean as nuisance regressors to the design matrix. We used the standard gray matter volume created from the segmentation step for each participant as an inclusive mask to restrict the voxels of interest. In the group-level analysis, we used a random-effects GLM. For each task, we conducted a one-sample *t*-test to identify the brain areas activated during stimulus presentation. We thresholded the brain maps at voxel-wise p = 0.001, and all reported brain areas were corrected for multiple comparisons using family wise error (FWE p = 0.05) as implemented in the SPM package.

2.5.3. Multivariate pattern analysis

This study aimed to elucidate how the human brain processes tone-category information and encodes distinct acoustic features of lexical tones. To achieve this goal, we implemented a combination of whole-brain multivariate pattern classification and searchlight-based representational analyses to characterize the distributed neural representations of tone categories (Haxby et al., 2014). Within this analytical framework, we define task-general neural representations as multivoxel activation patterns that remain consistent across the three functional tasks (PAS, REP, WID) reflecting the shared perceptual encoding of tonal information independent of task-specific cognitive demands. In contrast, task-specific representations correspond to neural patterns that vary with the additional cognitive or motor requirements of each task (e.g., covert articulatory rehearsal during REP, lexical judgment during WID). Based on prior evidence linking tone processing to bilateral temporal and parietal cortices, we anticipated convergent multivoxel patterns in these regions representing task-general tone encoding. However, given the greater acoustic and linguistic complexity of Cantonese tones, we further expected a more extensive cortical network to be involved. To localize such representations, we employed a voxel-wise searchlight analysis for classifier training and testing, followed by RSA to examine the neural representation of phonetic features in the identified brain areas. The subsequent subsections detail the procedures for constructing the fMRI feature space, performing cross-task cross-validation, and implementing the RSA (section 2.5.4).

We performed MVPA on the realigned data without normalization or smoothing and generated statistical maps for each participant. These unsmoothed data were analyzed using a GLM with individual regressors for each item (e.g., /ji1/, accounting for eight trials from two talkers and four runs), which were used to calculate single-item t-statistic maps for each task. In addition to the stimulus regressors, six head-movement regressors and one session-mean regressor were included in the design matrix for each run. T-statistic maps were used for the multivariate analysis because they combine the effect size weighted by the error variance, thus minimizing the influence of the highly variable item estimates.

The searchlight algorithm was employed to investigate the neural representations of Cantonese tone categories using a linear support vector machine (SVM) classifier implemented in the decoding toolbox (TDT) (Hebart et al., 2015). Classifiers were trained and tested with each subject's data, and at each voxel, sound-induced activation values (t-values) for each item within a spherical searchlight (3-voxel-radius sphere, averaging 123 voxels) were extracted for each task. Consequently, a V × I × T value matrix was constructed for each spherical searchlight, where V represents the voxel, I represents the item, and T represents the task (i.e., 123 × 18 × 3). This matrix was input into an SVM classifier for training and testing.

Based on previous research, we operationally defined “task-general” neural representations of speech categories as representations that result from multivoxel activation patterns across three different tasks (Feng et al., 2018). To investigate how much category-related information could be decoded from brain activation patterns across tasks, we employed a leave-one-task-out cross-validation (CV) procedure. In this procedure, the classifier was trained on data from the two tasks and subsequently tested on the remaining task data. This process was repeated three times, meaning that only the tone category information common across all three tasks was informative to the classifier. Finally, we calculated the mean classification accuracy and mapped it back to the voxel at the center of each searchlight sphere. We repeated this process across all voxels in the brain of each participant to generate brain maps with classification accuracy.

For whole-brain group-level analysis, each subject’s classification accuracy map was normalized to the standard space using the parameters obtained from the segmentation step, followed by a one-sample *t*-test on the group level. Subsequently, group statistical maps from the multivariate analyses were thresholded at voxel-wise uncorrected $p <$

0.005, with cluster-level FWE-corrected $p < 0.05$. Thresholded regions from the MVPA were selected as regions of interest (ROIs) in the following analyses. The resulting clusters were binarized and used as ROI masks for subsequent RSA analyses, ensuring that RSA was applied to brain regions that exhibited significant multivariate tone classification effects at the group level.

2.5.4. Representation similarity analysis

We employed RSA to examine the neural representation of phonetic features by investigating the relationship between neural representation similarity and stimulus-derived perceptual similarity for regions that showed significantly above-chance tone classifications (Kriegeskorte, 2008). To accomplish this, we created three models of stimulus-derived perceptual dissimilarity matrices (DSMs) based on the unidimensional F0 height (pitch height model), F0 slope (pitch slope model), and F0 vector (pitch vector model, integrating F0 height and F0 slope as orthogonal components, see Fig. 2 for a graphical illustration), as well as the neural DSM (Feng et al., 2018; Maddox et al., 2014; Yi et al., 2016).

The stimulus-derived models were first created for each speaker and then averaged across speakers. To construct the pitch height model, we calculated the distance between each pair of items based on their normalized F0 height. We built the pitch slope model in the same manner, but instead used each pair's normalized F0 slope. For the multidimensional vector model, we constructed a two–dimensional acoustic space in which the horizontal and vertical axes represented the normalized F0 height and F0 slope values, respectively, averaged across talkers. We then computed the Euclidean distance between each pair of tones within this two–dimensional feature space and averaged these distances across speakers, converting the results into a dissimilarity matrix. The vector model captured the integrated two–dimensional structure of pitch information (height + slope) within each tone token, allowing us to test whether certain brain regions encode multidimensional pitch patterns rather than individual parameters. This approach follows prior work indicating that tone perception often depends on the conjoint variation of height and contour rather than on either factor alone. Finally, all DSMs were normalized to a 0–1 range (0 = similarity; 1 = high dissimilarity) to facilitate cross–model comparison and visualization.

To create a neural dissimilarity matrix, voxel activation values (t-statistic values) for each item within each regional ROI mask were extracted and used to calculate the dissimilarity (using 1 − Pearson's correlation) between each pair of items. The neural DSM was then correlated with each model of stimulus-derived perceptual DSMs using Spearman rank correlation. In addition, to assess the unique contribution of each model, the neural DSM was correlated with each model while controlling for the effect of other models using partial Spearman's rank correlation.

2.5.5. Inter-subject representational similarity analysis for tone categorization performance

To identify correlations between brain activity and tone categorization performance, we performed a series of ROI-based Inter-subject Representational Similarity Analyses (IS-RSA) using data from our WID task. Unlike the direct correlation between individual behavioral scores and neural activation, the IS-RSA evaluates the similarity of representational structures across individuals. By constructing inter-subject

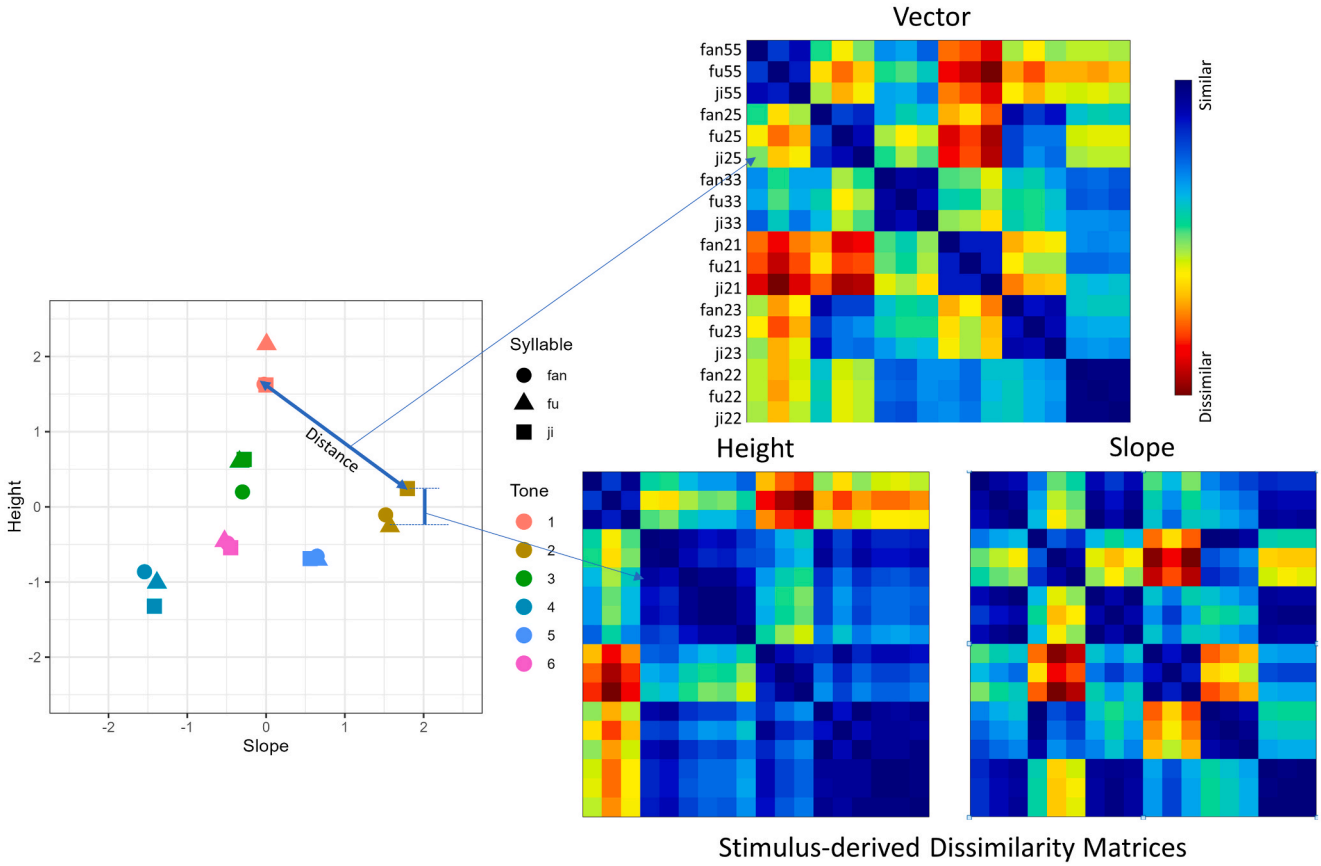

Fig. 2. Construction of stimulus-derived dissimilarity matrices (DSM). On the left, a scatterplot of all items (averaged between male and female speaker produced counterparts) is presented in two-dimensional space (normalized F0 height on the y axis and normalized F0 slope on the x axis). On the right, three DSMs were constructed by calculating the Euclidean distance between each pair of items based on the unidimensional F0 height, F0 slope, and 2-dimensional space, separately. The DSMs were then scaled to a range of 0 to 1, where blue squares indicate high similarity and warm squares indicate high dissimilarity. The DSMs for male speakers were constructed in the same way and averaged with the female-speaker DSMs for the RSA.

correlation matrices for both tone categorization and neural RDMs, this approach captures the shared variance in the geometry of behavioral and neural representational spaces rather than absolute performance levels (Gu et al., 2024). Thus, higher brain–behavioral similarity in IS-RSA indicates that participants who behaviorally organize tone categories in a comparable manner also exhibit analogous neural representational topographies, even if their overall accuracy differs. Therefore, this method provides a measure of the common representational architecture beyond the scope of conventional univariate correlations. The ROIs in this analysis were the same as those in the RSA.

For each participant, we first calculated the mean accuracy in the WID task across the four blocks for each of the six tone categories (Tones 1 to 6). This yielded six accuracy values for each participant. We then constructed a Representational Dissimilarity Matrix (RDM) for each participant by comparing the differences between pairs of tone categories. Subsequently, we generated an Inter-subject Correlation (ISC) matrix of tone categorization performance by performing pairwise correlations of the RDMs between subjects.

Similarly, for the ROI-based brain ISC matrices, we first extracted neural responses from the ROIs by averaging the T-values across all voxels with an ROI for each tone category, as derived from univariate activation analysis. Thus, each participant had six T-values corresponding to six tone categories. We then calculated the neural RDM for each participant by comparing the differences between pairs of tone categories. Using these neural RDMs, we constructed a pairwise brain ISC matrix for each ROI by calculating the pairwise correlations of neural RDMs among participants in the WID task.

Spearman's correlation coefficients were calculated between the ISC matrix of each ROI and the tone categorization performance of the ISC matrix. This allowed us to assess the relationship between neural and behavioral patterns across participants. To determine the statistical significance of brain-behavioral representational similarity, we used a non-parametric permutation test. We randomly permuted the subject labels of the brain ISC matrix (i.e., rows and columns) 10,000 times to generate a null distribution of surrogate Spearman correlation coefficients. We then compared the observed Spearman's correlation coefficient with this null distribution to obtain a p-value for each ROI. To correct for multiple comparisons, we applied the Bonferroni correction across all ROIs derived from the MVPA. This approach allowed us to identify the brain regions most closely coupled with individuals' behavioral performance in tone categorization, providing insights into the neural mechanisms underlying tone perception.

2.6. Brain visualization

The surface-view brain results and ROIs were produced using the BrainNet Viewer (RRID: SCR_009446; https://www.nitrc.org/projects/bnv/, (Xia et al., 2013) with the "Interpolated" map algorithm on a standard MNI space brain surface (e.g., BrainMesh_ICBM152_smoothed). Multi-slice view brain results were obtained using MRIcroGL (RRID: SCR_002403; https://www.nitrc.org/projects/mricrogl).

3. Results

3.1. Behavioral performance

The native Cantonese participants demonstrated good performance in categorizing tones [mean accuracy = 85.1 %, SD = 8.8 %; mean reaction time (RT) = 868.8 ms, SD = 91.8] in the WID task. Moreover, participants' tone category processing abilities were evaluated in a post-fMRI behavioral test (mean accuracy = 86.6 %, SD = 9.4 %; RT = 1201.5 ms, SD = 252.5), where they were required to press six response keys for each of the tone categories. There was a strong correlation between fMRI performance and post–fMRI behavioral test (based on accuracy scores; $r = 0.899$, $p < 0.001$). Because the reaction time requirements differed across sessions (time–limited in scanner vs. self–paced post–fMRI), accuracy was therefore selected as a consistent and theoretically meaningful index of tone categorization proficiency. The high correlation between fMRI and post-fMRI behavioral performance indicated that the participants exerted maximum effort in the fMRI environment.

3.2. Univariate brain activation

Univariate GLM analysis showed distinct activation patterns were observed for each task (Fig. 3). In the PAS task, the precentral gyrus, left superior and middle frontal gyri, and bilateral auditory cortices were activated. In the REP task, the left precentral gyrus, left superior frontal gyrus, left superior parietal lobule, and left inferior occipital gyrus were activated. During the WID task, we observed bilateral STG activation extending to the attention- and motor-related regions, including the bilateral middle frontal gyrus, precentral gyrus, bilateral inferior and superior parietal regions, SMA, bilateral putamen, and bilateral occipital gyrus. Strong common activation was found in the auditory cortex, including both Heschl's gyrus and the STG, in response to the sound stimulus compared with the baseline across all tasks (Fig. 3).

3.3. Multivariate pattern of classifying tone categories

We identified seven regions that demonstrated significantly above-average performance in tone category classification, including the left superior temporal gyrus (LSTG, peak coordinate: X = -56, Y = -14, Z = 4), right superior temporal gyrus (RSTG, X = 58, Y = -2, Z = 0), left inferior parietal sulcus (LIPS, X = -30, Y = -50, Z = 42), left pre-central gyrus (LPrCG, X = -46, Y = 6, Z = 34), right pre-central gyrus (RPrCG, X = 30, Y = -4, Z = 54), right superior frontal gyrus (RSFG, X = 18, Y = 52, Z = 30), and bilateral lingual gyrus (BiLG, X = 18, Y = -82, Z = -8) (Fig. 4C). Interestingly, the bilateral STG and LIPS are very close to the regions reported in previous studies of Mandarin tone categorization (Feng et al., 2018).

3.4. RSA results

Searchlight classification analysis enabled us to identify brain regions that exhibited high sensitivity to tone categories. However, it does not provide information on the intricate relationships between items or categories based on multivoxel patterns. Further analysis using RSA revealed specific aspects of speech information encoded in the activity patterns of seven identified regions (LSTG, RSTG, LIPS, LPrCG, RPrCG, RSFG, and BiLG, as shown in Fig. 4C).

First, the neural DSM of these regions, including the RSTG, LSTG, LIPS, and BiLG, was significantly correlated with the pitch height model. Furthermore, this correlation remained statistically significant even when the contribution of the pitch slope model was controlled using a partial correlation approach, except for LIPS. Second, the pitch slope model exhibited a significant association with the neural DSM in the LIPS, and this association was robust, as the effect persisted even when the variance of the pitch height model was controlled. Third, the vector model (pitch height plus slope model) demonstrated significant correlations with the neural DSM in all regions except the RSFG. However, when the variance of the pitch height model was controlled, the effect diminished in the bilateral STG and LPrCG, confirming that the bilateral STG mainly represented pitch height information. In contrast, the vector model remained significantly correlated with the neural DSM in the LIPS, RPrCG, and BiLG, even when either the pitch height or slope model was controlled (Fig. 4B, bottom panel). These results suggest that neural activity patterns in the LIPS, RPrCG, and BiLG can be best characterized by combining multidimensional speech information (i.e., pitch height and slope).

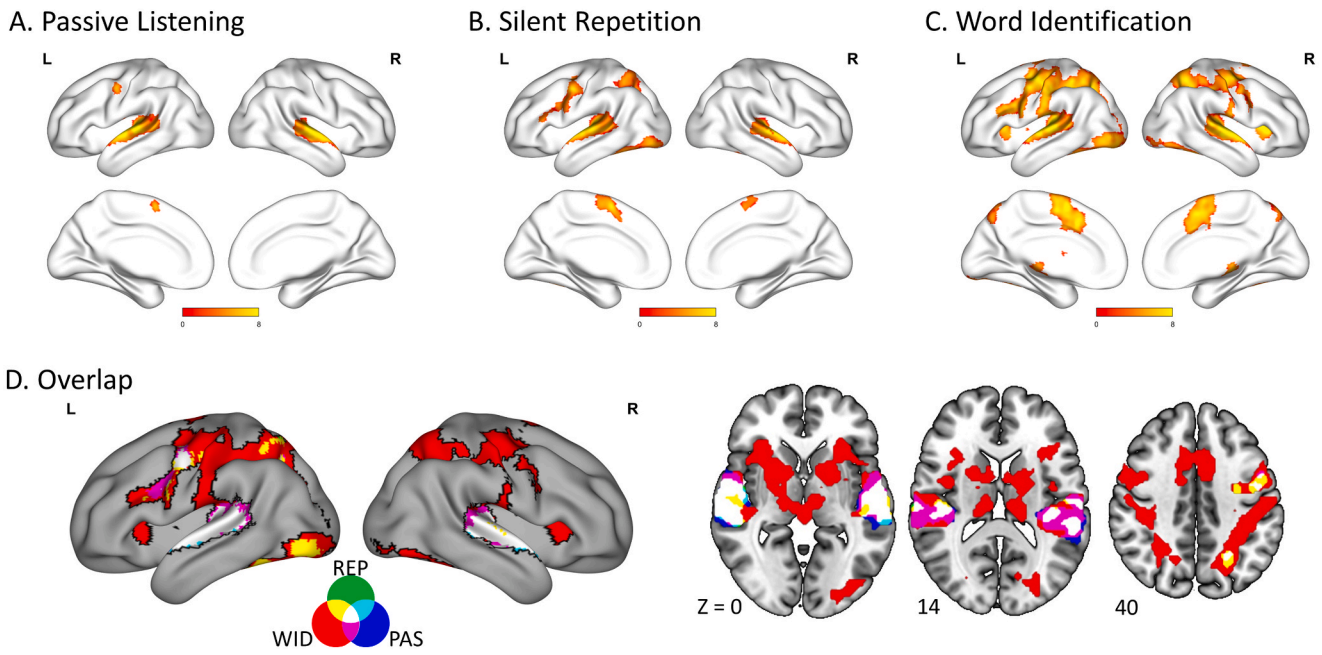

Fig. 3. Brain activation maps of speech sound versus baseline for each task: A. Passive Listening, B. Silent Repetition, and C. Word Identification. D. Overlap brain map of speech sound activation across the three tasks. Label abbreviation: PAS: passive listening task; REP: silent repetition task; WID: word identification task.

3.5. IS-RSA results

The ROI-based IS-RSA results demonstrated positive correlations across all seven ROIs, ranging from $r = 0.016$ to $r = 0.081$. Three ROIs exhibited significant brain-behavioral similarity correlations: LIPS ($r = 0.074$, $p < 0.001$, adjusted $p = 0.005$), LPrCG ($r = 0.081$, $p < 0.001$, adjusted $p = 0.002$), and RSFG ($r = 0.080$, $p = 0.004$, adjusted $p = 0.030$). However, the other four ROIs did not show significant brain-behavioral similarity correlations, such as the LSTG ($r = 0.062$, $p = 0.027$, adjusted $p = 0.188$), RSTG ($r = 0.074$, $p = 0.015$, adjusted $p = 0.102$), RPrCG ($r = 0.032$, $p = 0.100$, adjusted $p = 0.694$), and BiLG ($r = 0.016$, $p = 0.190$, adjusted $p = 1$). These results suggest that brain regions, such as the LIPS, LPrCG, and RSFG, are closely associated with behavioral performance in tone categorization. Conversely, other regions may be involved in tone category representation but do not show a direct correlation with behavioral performance. This may indicate that these regions contribute to the overall neural network for tone processing but are not the primary drivers of individual differences in categorization performance.

4. Discussion

In this study, we investigated the neural correlates of Cantonese lexical tone representation using fMRI during three tone processing tasks: PAS, REP, and WID. Our findings extend previous research on simpler tonal languages by revealing a bilateral and extensive brain network involved in tone processing, including the bilateral temporal regions (e.g., STG), left parietal region (IPS), bilateral frontal regions (e.g., LPrCG, RPrCG, and RSFG), and bilateral occipital regions (e.g., BiLG). The RSA indicated that pitch height information was predominantly processed in the bilateral STG and LG, whereas pitch slope information engaged only the left IPS. Additionally, the frontal regions, particularly the bilateral pre-central regions (LPrCG and RPrCG), have been implicated in holistic tone processing. Beyond their specialized roles in processing specific tonal features, both LIPS and BiLG also exhibited a strong representation of holistic pitch information. The IS-RSA further suggests that the LIPS, LPrCG, and RSFG are most closely associated with individual tone categorization performance, whereas other regions may contribute to tone categorization but are not the primary drivers of individual differences.

4.1. Extensive brain regions contribute to Cantonese lexical tone categorization

Previous research has consistently reported the crucial role of the bilateral STG in phonetic information processing (Kwok et al., 2017; Liang & Du, 2018). An ECoG study further demonstrated that bilateral STG activation represents information on pitch height and slope in native Mandarin speakers (Li et al., 2021). This representation appears to be language universal, as Native English speakers also exhibited a similar pattern, although native Mandarin speakers' neural representation was more sensitive because of their language experience. ECoG provides fine–grained temporal information on tone–related neural responses, however, its cortical coverage is inherently limited. Utilizing an approach similar to that described in the current article, Feng et al. (2018) identified three cortical regions responsible for representing Mandarin lexical tones, including the bilateral STG and the left IPL. The bilateral STG contributed to pitch height representation, whereas the left IPL contributed to both pitch height and slope representation. It is evident that, although Mandarin tones vary significantly in pitch slope, pitch height information remains critical for differentiating lexical tones (Feng et al., 2018; Li et al., 2021).

Our MVPA results expand on these previous findings by revealing a more extensive cortical network that effectively differentiates tone categories across various tasks in Cantonese. This network not only includes previously reported temporal and parietal regions (e.g., bilateral STG and LIPS), but also encompasses the frontal and occipital regions (e.g., LPrCG, RPrCG, RSFG, and BiLG). While the univariate activation results from the passive listening task showed a spatially restricted pattern centered in the auditory cortices, the multivariate searchlight analysis captured distributed representational patterns that differentiate tone categories even when mean activation amplitude remains stable. This indicates that MVPA and univariate approaches offer complementary perspectives: the former reveals how tonal information is encoded in distributed activation patterns, whereas the latter highlights where activity intensity increases most strongly during auditory processing.

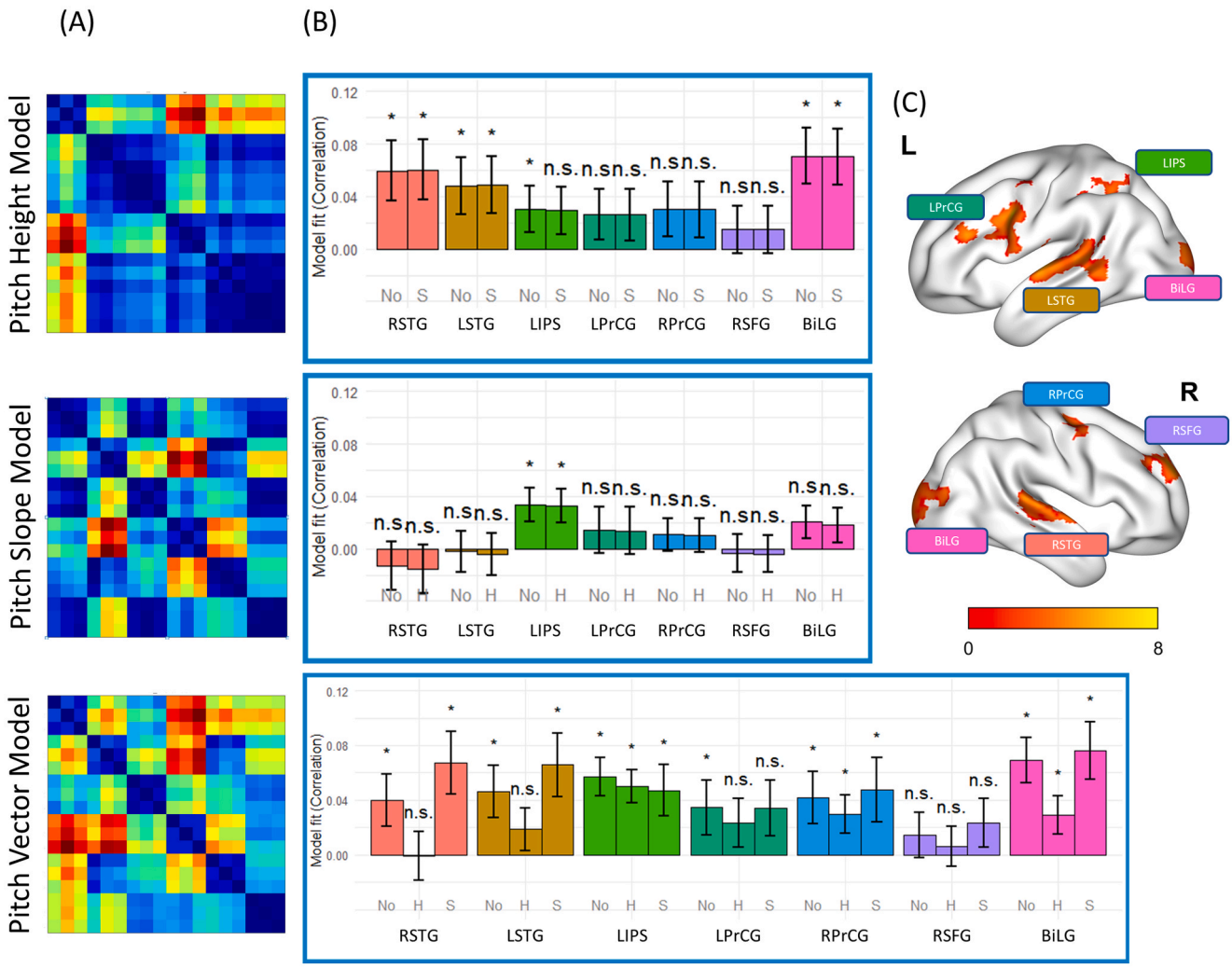

Fig. 4. RSA of regions identified using searchlight cross-task tone classification analysis. (A) Three models of stimulus-derived perceptual dissimilarity matrices (DSMs) were constructed by calculating the distance between each pair of tonal syllables based on pitch height, slope, and vector. (B) Model fits (Spearman rank correlations) between the stimulus-derived perceptual DSMs and neural DSMs. The gray labels under each bar represent different model fit methods: No, not controlling the variance of any model; S, controlling the variance of the pitch slope model; and H, controlling the variance of the pitch height models. The error bars denote the standard error of the mean. *, $p < 0.05$; n.s., not significant. (C) The brain regions identified in the multivariate pattern analysis of tone category classification, including the left pre-central gyrus (LPrCG), left superior temporal gyrus (LSTG), left inferior parietal sulcus (LIPS), and bilateral Lingual Gyrus (BiLG) shown on the top panel, and right superior frontal gyrus (RSFG), right pre-central gyrus (RPrCG), right superior temporal gyrus (RSTG), and BiLG shown on the bottom panel.

Such complementarity suggests that the more extensive network revealed by MVPA reflects additional representational sensitivity rather than inconsistency between analytic methods.

This broader cortical involvement supports our hypothesis that Cantonese lexical tone representation engages partially distinct neural mechanisms, potentially reflecting greater acoustic and linguistic complexity than Mandarin. In contrast to Mandarin, which comprises four tones, Cantonese has six lexical tones, and their differentiation is governed not only by pitch slope but also by pitch height (Peng, 2006). This added acoustic complexity may require more extensive recruitment of the temporal-parietal network. Compared with the more discretely distributed Mandarin tones, Cantonese tones can share similar pitch heights but differ in slope (e.g., T2 and T3; T5 and T6), or share similar slope but differ in height (e.g., T1, T3, and T6). Such overlap increases perceptual ambiguity and the need for finer-grained categorization, thereby placing greater demands on perceptual decision-making processes and potentially requiring additional recruitment of fronto-parietal regions. Taken together, the larger acoustic and phonological complexity of Cantonese tones likely imposes greater perceptual and categorization demands, consistent with the involvement of a more extensive cortical network for tone processing and lexical tone categorization.

An alternative explanation for the extensive cortical involvement is the task design. Compared with previous research, our paradigm incorporated visual cues in which six Chinese characters corresponding to base syllables were displayed. These visual stimuli were presented consistently across the three tasks to ensure identical perceptual input, with differences arising only from the task instructions and cognitive demands. Visual cues are necessary for native Cantonese speakers to perform tone identification, as they acquire lexical tones through natural language exposure rather than through auxiliary phonetic symbols (e.g., tone numbers) commonly used in Mandarin learning. Thus, Cantonese participants needed to categorize tones by selecting the corresponding characters, a process that may engage additional visual–phonological and cross-modal integration regions. This task configuration ensured a fair comparison across the tone categories. As the visual inputs are identical across tone categories, the processing of visual stimuli per se may not directly contribute to the observed

extensive cortical network that effectively differentiates tone categories. Nonetheless, the presentation of visual cues may have triggered additional processing and contributed to the recruitment of an extended neural network for tone feature processing. This effect may stem from the strong association between tone processing and orthographic learning in Cantonese speakers. Future research that omits visual elements could further test this hypothesis and help disentangle the contributions of auditory–phonological and visual-orthographic processing to tone representation.

Finally, we should also consider the potential influence of participants' multilingual background. All participants were native Cantonese speakers who had also acquired Mandarin (another tonal language) and English (a non-tonal language). This multilingual experience may have modulated neural mechanisms during tone processing, particularly with the involvement of visual cues. For example, when native Cantonese speakers were presented with Chinese characters corresponding to tonal syllables, they might need to inhibit lexical and phonological associations derived from their Mandarin experience to correctly categories Cantonese tones and identify Cantonese words (Yang et al., 2025). Moreover, bilingual or multilingual exposure has been shown to enhance attentional control and cross-linguistic phonological interaction mechanisms, which could partially explain the engagement of additional frontal regions observed in our MVPA results. Future research should explicitly examine how different degrees of multilingual proficiency modulate tone perception and neural representation in native Cantonese speakers.

4.2. Additional brain regions are recruited for phonetic feature processing

Our feature-specific analyses provide novel insights into how the brain processes these complex tonal features. Our RSA results indicate that the pitch height model is associated with the bilateral STG and BiLG which has not been reported in previous Mandarin study (Feng et al., 2018). For the pitch slope model, we replicated previous findings that the left inferior parietal regions (LIPS) play a significant role. For the vector model, we observed strong associations with the frontal and occipital regions, in addition to the temporal and parietal regions that have been previously identified. Collectively, these findings support the hypothesis that brain regions may represent tone features differently and suggest a more complex cortical representation of Cantonese tone.

The methodological approach used in our study may explain some of these novel findings. Our experimental task design with Cantonese lexical tones provides a means to observe frontal and occipital associations with lexical tone representation. Unlike previous Mandarin studies that utilized direct tone category judgment tasks (Feng et al., 2018, 2021), our experiment did not require explicit learning of tone categories because native Cantonese speakers do not explicitly learn tone categories in formal educational settings (Chan & Leung, 2020). Nevertheless, native Cantonese speakers can effectively recognize lexical tone contrasts even without explicit knowledge of the tone system. For instance, while Mandarin learners are aware of four tones and a null tone, Cantonese learners do not usually possess explicit knowledge of six unchecked and three checked tones. However, native Cantonese speakers can still acquire proficiency in the language and differentiate tones in daily communication. By having native Cantonese speakers directly identify words with tonal syllables rather than tone categories, our task design more closely approximates natural language processing. This approach avoids examining tone-related processing using only acoustic analysis. The semantic representations of tonal syllables must be activated before participants can successfully judge the perceived word. This approach allows for a more holistic assessment of tonal processing, which may assimilate real-world language use.

One of our intriguing findings was the involvement of BiLG in tone processing. Through MVPA, we identified BiLG as being involved in tone category representation, and our RSA revealed its high correlation with pitch height and pitch vector models, even when controlling for other models. The vector model contributes additional insight by revealing cortical regions that represent complex acoustic–phonetic combinations of tone features rather than single attributes, complementing the height– and slope–specific patterns observed in STG and LIPS. This finding suggests that Cantonese tone perception may involve multimodal processing by integrating visual information with auditory processing. This visual representation is unlikely to stem from the visual cues of the corresponding characters, as identical cues were presented for the six tones of each syllable, ensuring that the classification of tones was not attributable to the visual word forms presented in the cues during tasks.

The involvement of BiLG could indicate cross-modal integration, wherein visual representations of pitch contours are utilized to facilitate auditory tone perception. Previous research observed the recruitment of the occipital lobe in congenitally and late-onset blind individuals, suggesting that the occipital lobe can be 'recycled' for auditory processing (Collignon et al., 2013). Even in sighted individuals, auditory attention can activate the peripheral regions of the visual cortex, particularly when attending to sound sources outside their visual field (Cate et al., 2009). This activation is enhanced under more difficult listening conditions, indicating its potential role in complementing task difficulties in speech perception. In the context of Cantonese tone processing, we propose that utilizing occipital regions to represent pitch variations may assist in the robust neural representation of complex tonal patterns, providing a supplementary mechanism for distinguishing between acoustically similar tones.

4.3. Fronto-parietal network contributes to individual differences in tone categorization performance

Beyond identifying the neural architecture supporting Cantonese tone processing, our study provides insights into the brain regions that drive individual differences in tone categorization abilities. Further IS-RSA results suggest that regions like the LIPS, LPrCG, and RSFG are most closely coupled with individuals' behavioral performance in tone categorization, emphasizing their crucial role in processing and categorizing lexical tones. The left inferior parietal region, known for its involvement in higher cognitive functions such as attention (Larson & Lee, 2014), working memory (Alain et al., 2008), and spatial processing (Behrmann et al., 2004; Silk et al., 2010), plays a significant role in managing the complex demands of processing pitch slopes and integrating this information with other auditory cues. In the context of phonetic processing, this 'spatial processing' does not refer to literal physical space but to the transformation of dynamic acoustic cues (e.g., pitch height and slope) into an internal representational space that supports comparison, categorization, and integration with other auditory dimensions. Through this mechanism, the LIPS helps manage the complex perceptual demands of tone processing by linking attention– and memory–based resources with multidimensional acoustic representations.

Similarly, the bilateral frontal regions, which are associated with executive functions (Costafreda et al., 2006; Snyder et al., 2007), including cognitive control, decision-making, and attentional processes, may be critical for the holistic processing, integrating sensory information, and planning of motor responses. This finding highlights the role of frontal network in the later stages of tone perception, likely reflecting sensorimotor integration and perceptual decision-making processes. The involvement of frontal regions also aligns with recent evidence suggesting that motor regions (e.g., LPrCG) contribute to speech perception through internal simulation mechanisms (Liang et al., 2023). The brain-behavior pattern coupling in frontal regions may reflect the perceptual demands imposed by Cantonese's large tonal inventory, which may require more extensive recruitment of sensorimotor and executive resources for accurate categorical identification and discrimination. Overall, our findings suggest that successful tone categorization in complex tonal systems like Cantonese appears to rely not only on accurate early sensory encoding, but also on higher-order

processes supporting attention, working memory, and decision-making, consistent with the functional profile of these fronto-parietal regions.

4.4. Limitations

Our findings suggest a potentially broader cortical involvement in Cantonese tone processing than that reported for Mandarin; however, this interpretation should be approached with caution. The present study did not include a direct within–subject comparison between Cantonese and Mandarin using matched stimuli, tasks, or analytic pipelines, and differences in task settings (e.g., presenting visual cues) may also contribute to variability across studies. Therefore, the observed extensiveness of cortical representation of phonetic features likely reflects both the higher tonal complexity of Cantonese and potential methodological factors rather than a definitive language–specific effect. Future research directly comparing bilingual Mandarin and Cantonese speakers' processing of both tonal systems under equivalent experimental conditions would provide valuable insights.

5. Conclusion

This study reveals a more extensive neural network for Cantonese lexical tone processing than that previously reported for other tonal languages, reflecting the increased complexity of the Cantonese tone system. Our findings demonstrate that pitch height information is processed in the bilateral STG and LG, whereas pitch slope processing engages the LIPS. Holistic tone processing is supported by the LIPS, RPrCG, and BiLG. The involvement of the BiLG suggests a novel multimodal integration mechanism in tone processing, potentially supporting the complex discrimination required for Cantonese. The fronto-parietal network, particularly the LIPS, LPrCG, and RSFG, has emerged as a critical predictor of individual differences in tone categorization ability, suggesting the importance of executive function and attention in successful tone processing. These findings advance our understanding of speech perception by demonstrating that complex tonal systems recruit additional neural resources beyond the classical language networks. Future research should investigate how this expanded network develops during tone language acquisition and whether it can be strategically engaged in enhancing tone language learning among non-native speakers.

CRediT authorship contribution statement

Ran Tao: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Kaile Zhang:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Yan Feng:** Writing – original draft, Methodology, Data curation, Conceptualization. **Yi Weng:** Writing – original draft, Methodology, Formal analysis, Data curation. **Gang Peng:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The work was substantially supported by a fellowship award from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU/RFS2122-5H01).

Data availability

Data will be made available on request.

References

Alain, C., He, Y., & Grady, C. (2008). The contribution of the inferior parietal lobe to auditory spatial working memory. *Journal of Cognitive Neuroscience, 20*(2), 285–295. https://doi.org/10.1162/jocn.2008.20014

Behrmann, M., Geng, J. J., & Shomstein, S. (2004). Parietal cortex and attention. *Current Opinion in Neurobiology, 14*(2), 212–217. https://doi.org/10.1016/j.conb.2004.03.012

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9), 341–345.

Cate, A. D., Herron, T. J., Yund, E. W., Stecker, G. C., Rinne, T., Kang, X., Petkov, C. I., Disbrow, E. A., & Woods, D. L. (2009). Auditory attention activates peripheral visual cortex. *PLoS One1, 4*(2), e4645.

Chan, R. K. W., & Leung, J. H. C. (2020). Why are lexical tones difficult to learn? *Studies in Second Language Acquisition, 42*(1), 33–59. https://doi.org/10.1017/S0272263119000482

Collignon, O., Dormal, G., Albouy, G., Vandewalle, G., Voss, P., Phillips, C., & Lepore, F. (2013). Impact of blindness onset on the functional organization and the connectivity of the occipital cortex. *Brain, 136*(9), 2769–2783. https://doi.org/10.1093/brain/awt176

Costafreda, S. G., Fu, C. H. Y., Lee, L., Everitt, B., Brammer, M. J., & David, A. S. (2006). A systematic review and quantitative appraisal of fMRI studies of verbal fluency: Role of the left inferior frontal gyrus. *Human Brain Mapping, 27*(10), 799–810. https://doi.org/10.1002/hbm.20221

Feng, G., Gan, Z., Llanos, F., Meng, D., Wang, S., Wong, P. C. M., & Chandrasekaran, B. (2021). A distributed dynamic brain network mediates linguistic tone representation and categorization. In. *NeuroImage, 224*. https://doi.org/10.1016/j.neuroimage.2020.117410

Feng, G., Gan, Z., Wang, S., Wong, P. C. M., & Chandrasekaran, B. (2018). Task-general and acoustic-invariant neural representation of speech categories in the human brain. *Cerebral Cortex, 28*(9), 3241–3254. https://doi.org/10.1093/cercor/bhx195

Gandour, J. (1983). Tone perception in Far Eastern languages.pdf. *Journal of Phonetics, 11*, 149–175.

Gandour, J., Wong, D., Hsieh, L., Weinzapfel, B., Lancker, D. V., & Hutchins, G. D. (2000). A crosslinguistic PET study of tone perception. *Journal of Cognitive Neuroscience, 12*(1), 207–222. https://doi.org/10.1162/089892900561841

Gandour, J., Wong, D., & Hutchins, G. (1998). Pitch processing in the human brain is influenced by language experience. *Neuroreport, 9*(9), 2115–2119. https://doi.org/10.1097/00001756-199806220-00038

Gu, C., Peng, Y., Nastase, S. A., Mayer, R. E., & Li, P. (2024). Onscreen presence of instructors in video lectures affects learners' neural synchrony and visual attention during multimedia learning. *Proceedings of the National Academy of Sciences, 121*(12), 2017. https://doi.org/10.1073/pnas.2309054121

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. In *Annual Review of Neuroscience* (Vol. 37, pp. 435–456). Annual Reviews Inc. 10.1146/annurev-neuro-062012-170325.

Hebart, M. N., Gärgen, K., & Haynes, J.-D. (2015). The Decoding Toolbox (TDT): A versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics, 8*(JAN). https://doi.org/10.3389/fninf.2014.00088

Hickok, G. (2022). The dual stream model of speech and language processing. In *Handbook of Clinical Neurology* (Vol. 185, pp. 57–69). Elsevier B.V. 10.1016/B978-0-12-823384-9.00003-7.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402. https://doi.org/10.1038/nrn2113

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*(NOV). https://doi.org/10.3389/neuro.06.004.2008

Kwok, V. P. Y., Dan, G., Yakpo, K., Matthews, S., Fox, P. T., Li, P., & Tan, L.-H. (2017). A meta-analytic study of the neural systems for auditory processing of lexical tones. *Frontiers in Human Neuroscience, 11*(July), 1–10. https://doi.org/10.3389/fnhum.2017.00375

Kwok, V. P. Y., Dan, G., Yakpo, K., Matthews, S., & Tan, L. H. (2016). Neural systems for auditory perception of lexical tones. *Journal of Neurolinguistics, 37*, 34–40. https://doi.org/10.1016/j.jneuroling.2015.08.003

Kwok, V. P. Y., Wang, T., Chen, S., Yakpo, K., Zhu, L., Fox, P. T., & Tan, L. H. (2015). Neural signatures of lexical tone reading. *Human Brain Mapping, 36*(1), 304–312. https://doi.org/10.1002/hbm.22629

Larson, E., & Lee, A. K. C. (2014). Switching auditory attention using spatial and non-spatial features recruits different cortical networks. *NeuroImage, 84*(1), 681–687. https://doi.org/10.1016/j.neuroimage.2013.09.061

Li, Y., Tang, C., Lu, J., Wu, J., & Chang, E. F. (2021). Human cortical encoding of pitch in tonal and non-tonal languages. *Nature Communications, 12*(1), 1161. https://doi.org/10.1038/s41467-021-21430-x

Liang, B., & Du, Y. (2018). The functional neuroanatomy of lexical tone perception: An activation likelihood estimation meta-analysis. *Frontiers in Neuroscience, 12*(JUL), 1–17. https://doi.org/10.3389/fnins.2018.00495

Liang, B., Li, Y., Zhao, W., & Du, Y. (2023). Bilateral human laryngeal motor cortex in perceptual decision of lexical tone and voicing of consonant. *Nature Communications, 14*(1). https://doi.org/10.1038/s41467-023-40445-0

Liu, L., Peng, D., Ding, G., Jin, Z., Zhang, L., Li, K., & Chen, C. (2006). Dissociation in the neural basis underlying Chinese tone and vowel production. *NeuroImage, 29*(2), 515–523. https://doi.org/10.1016/j.neuroimage.2005.07.046

Maddox, W. T., Chandrasekaran, B., Smayda, K., Yi, H. G., Koslov, S., & Beevers, C. G. (2014). Elevated depressive symptoms enhance reflexive but not reflective auditory category learning. *Cortex, 58*, 186–198. https://doi.org/10.1016/j.cortex.2014.06.013

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Peng, G. (2006). Temporal and tonal aspects of Chinese syllables: A corpus-based comparative study of mandarin and cantonese. *Journal of Chinese Linguistics, 34*(1), 134–154.

Shuai, L., & Gong, T. (2014). Temporal relation between top-down and bottom-up processing in lexical tone perception. *Frontiers in Behavioral Neuroscience, 8*(MAR), 1–16. https://doi.org/10.3389/fnbeh.2014.00097

Silk, T. J., Bellgrove, M. A., Wrafter, P., Mattingley, J. B., & Cunnington, R. (2010). Spatial working memory and spatial attention rely on common neural processes in the intraparietal sulcus. *NeuroImage, 53*(2), 718–724. https://doi.org/10.1016/j.neuroimage.2010.06.068

Snyder, H. R., Feigenson, K., & Thompson-Schill, S. L. (2007). Prefrontal cortical response to conflict during semantic and phonological tasks. *Journal of Cognitive Neuroscience, 19*(5), 761–775. https://doi.org/10.1162/jocn.2007.19.5.761

Stevens, W. D., Buckner, R. L., & Schacter, D. L. (2010). Correlated low-frequency BOLD fluctuations in the resting human brain are modulated by recent experience in category-preferential visual regions. *Cerebral Cortex, 20*(8), 1997–2006. https://doi.org/10.1093/cercor/bhp270

Tao, R., Zhang, K., & Peng, G. (2021). Music does not facilitate lexical tone normalization: A speech-specific perceptual process. *Frontiers in Psychology, 12* (October), 1–14. https://doi.org/10.3389/fpsyg.2021.717110

Tomasi, D., Wang, R., Wang, G.-J., & Volkow, N. D. (2014). Functional connectivity and brain activation: A synergistic approach. *Cerebral Cortex, 24*(10), 2619–2629. https://doi.org/10.1093/cercor/bht119

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. https://doi.org/10.1016/j.neuroimage.2013.04.006

Wang, S. (1967). Phonological features of tone. *International Journal of American Linguistics, 33*(2), 93–105. https://doi.org/10.1086/464946

Xia, M., Wang, J., & He, Y. (2013). BrainNet Viewer: A network visualization tool for human brain connectomics. *PLoS One, 8*(7), Article e68910.

Yang, S., Jiang, S., & Jiang, M. (2025). Brain potentials reveal activation and inhibition of the non-target L2 in bilinguals' L1 context. *International Journal of Bilingualism*. https://doi.org/10.1177/13670069251384055

Yi, H.-G., Maddox, W. T., Mumford, J. A., & Chandrasekaran, B. (2016). The role of corticostriatal systems in speech category learning. *Cerebral Cortex, 26*(4), 1409–1420. https://doi.org/10.1093/cercor/bhu236

Yip, M. (2002). Tone. In *Tone*. Cambridge University Press. 10.1017/CBO9781139164559.

Zatorre, R. J., & Gandour, J. T. (2008). Neural specializations for speech and pitch: Moving beyond the dichotomies. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 1087–1104. https://doi.org/10.1098/rstb.2007.2161

Zhang, C., Peng, G., Shao, J., & Wang, W. S. Y. (2017). Neural bases of congenital amusia in tonal language speakers. *Neuropsychologia, 97*(July 2016), 18–28. 10.1016/j.neuropsychologia.2017.01.033.

Zhang, C., Peng, G., & Wang, W. S. Y. (2013). Achieving constancy in spoken word identification: Time course of talker normalization. *Brain and Language, 126*(2), 193–202. https://doi.org/10.1016/j.bandl.2013.05.010

Zhang, C., Pugh, K. R., Mencl, W. E., Molfese, P. J., Frost, S. J., Magnuson, J. S., Peng, G., & Wang, W. S.-Y. (2016). Functionally integrated neural processing of linguistic and talker information: An event-related fMRI and ERP study. *NeuroImage, 124*, 536–549. https://doi.org/10.1016/j.neuroimage.2015.08.064

Zhang, K., Peng, G., Li, Y., Minett, J. W., & Wang, W. S. Y. (2018). The effect of speech variability on tonal language speakers' second language lexical tone learning. *Frontiers in Psychology, 9*, 1–13. https://doi.org/10.3389/fpsyg.2018.01982

Zhang, K., Tao, R., & Peng, G. (2023). The advantage of the music-enabled brain in accommodating lexical tone variabilities. *Brain and Language, 247*, Article 105348. https://doi.org/10.1016/j.bandl.2023.105348