


The dynamic and task-dependent representational transformation between the motor and sensory systems during speech production

Wenjia Zhang , Yiling Liu , Xuefei Wang & Xing Tian


To cite this article: Wenjia Zhang , Yiling Liu , Xuefei Wang & Xing Tian (2020): The dynamic and task-dependent representational transformation between the motor and sensory systems during speech production, Cognitive Neuroscience, DOI: [10.1080/17588928.2020.1792868](https://doi.org/10.1080/17588928.2020.1792868)

To link to this article: <https://doi.org/10.1080/17588928.2020.1792868>

 View supplementary material 

 Published online: 28 Jul 2020.

 Submit your article to this journal 

 Article views: 48

 View related articles 

 View Crossmark data 



The dynamic and task-dependent representational transformation between the motor and sensory systems during speech production

Wenjia Zhang^{a,b,c}, Yiling Liu^d, Xuefei Wang^e and Xing Tian^{a,b,c}

^aDivision of Arts and Sciences, New York University Shanghai, Shanghai, China; ^bShanghai Key Laboratory of Brain Functional Genomics (Ministry of Education), School of Psychology and Cognitive Science, East China Normal University, Shanghai, China; ^cNYU-ECNU Institute of Brain and Cognitive Science, New York University Shanghai, Shanghai, China; ^dDepartment of Educational Sciences, Tianjin Normal University, Tianjin, China; ^eDepartment of Computer Science, Fudan University, Shanghai, China

ABSTRACT

The motor and sensory systems work collaboratively to fulfill cognitive tasks, such as speech. For example, it has been hypothesized that neural signals generated in the motor system can transfer directly to the sensory system along a neural pathway (termed as motor-to-sensory transformation). Previous studies have demonstrated that the motor-to-sensory transformation is crucial for speech production. However, it is still unclear how neural representation dynamically evolves among distinct neural systems and how such representational transformation depends on task demand and the degrees of motor involvement. Using three speech tasks – overt articulation, silent articulation, and imagined articulation, the present fMRI study systematically investigated the representational formats and their dynamics in the motor-to-sensory transformation. Frontal-parietal-temporal neural pathways were observed in all three speech tasks in univariate analyses. The extent of the motor-to-sensory transformation network differed when the degrees of motor engagement varied among tasks. The representational similarity analysis (RSA) revealed that articulatory and acoustic information was represented in motor and auditory regions, respectively, in all three tasks. Moreover, articulatory information was cross-represented in the somatosensory and auditory regions in overt and silent articulation tasks. These results provided evidence for the dynamics and task-dependent transformation between representational formats in the motor-to-sensory transformation.

ARTICLE HISTORY

Received 3 February 2020
Revised 24 May 2020
Published online 28 July 2020

KEYWORDS

Speech production; internal forward model; efference copy/corollary discharge; motor-to-sensory transformation; sensorimotor integration; prediction; phonetic features; RSA

1. Introduction

Speech production is one of the most complex actions. Yet we can speak efficiently and effortlessly in daily life. One of the proposed mechanisms for controlling speech production is based on the motor-to-sensory transformation – the planned motor commands can transmit internally and activate sensory regions that respond to external feedback (Wolpert & Ghahramani, 2000). The motor-to-sensory transformation (c.f. efference copy/corollary discharge or internal forward model) provides a functional computation in which motor plans lead to predictions of the sensory consequences via motor-to-sensory neural projections (Parrell et al., 2017; Whitford et al., 2017). Specifically, in speech, the sensory consequences of speaking can be predicted by the motor system activation (hereafter also referred as motor-based prediction). This prediction is compared with feedback for error detection and correction in speech motor control (Guenther et al., 2006; Hickok, 2012;

Houde & Nagarajan, 2011; Tian et al., 2018; Tian & Poeppel, 2010, 2012, 2013, 2015; Tian et al., 2016).

Previous studies have proposed how this motor-to-sensory transformation mechanism facilitates the speech production in detail (Guenther et al., 2006; Hickok, 2012; Houde & Nagarajan, 2011; Tian & Poeppel, 2010, 2012). Specifically, a copy of motor signals from the frontal cortices (termed as *efference copy* or *corollary discharge*) is sent to the parietal and temporal areas to internally induce the somatosensory and auditory representations of the speech targets. The motor action can be updated according to the interaction among the estimated sensorimotor state, predicted sensory consequences, and the actual feedback (Guenther et al., 2006; Hickok, 2012; Houde & Nagarajan, 2011; Tian & Poeppel, 2010). Therefore, this motor-based prediction stream theoretically involves the dynamics of the representational format of speech (from the motor system to the sensory system) (Tian & Poeppel, 2013; Tian et al., 2016).

Previous neuroimaging studies demonstrate converging neural networks of sensory and motor systems that mediate different types of speech production. During overt articulation (*OA*), the temporal, parietal, and frontal lobes, including the insular cortex, are commonly observed for processing auditory, somatosensory, and motor information that are crucial for controlling speech production (Adank, 2012; Price, 2012). Comparing to *OA*, in silent articulation (*SA*), similar motor processes have been observed in the inferior frontal gyrus (Huang et al., 2002). Moreover, similar somatosensory and auditory representations have been suggested by the activation in the inferior parietal and superior temporal cortices during *SA* (Anumanchipalli et al., 2019; Cogan et al., 2014). For the imagined articulation (*IA*), similar frontal, parietal, and temporal networks as observed in *OA* have been observed, including the inferior frontal gyrus, insular, inferior parietal cortex, and superior temporal gyrus and sulcus (Hickok et al., 2003; Tian et al., 2016). Therefore, all types of speech production tasks activate common neural networks that are at least partially overlapped among tasks and indicate the common sensorimotor processes.

Furthermore, different types of speech production tasks may induce different neural representations in the common sensorimotor networks. Previous studies mainly combined the *OA* with *IA* tasks to examine the motor-to-sensory transformation (Tian & Poeppel, 2013; Tian et al., 2016). It is assumed that both overt and covert speech share the motor-to-sensory transformation but are different in the involvement of primary motor cortices for action execution. That is, *OA* provided a reference of cortex activation involved in all aspects of speech production, whereas *IA* was hypothesized to implement the motor-based prediction using the motor-to-sensory transformation without the contamination of overt action and feedback (Tian & Poeppel, 2013; Tian et al., 2016). In addition to the *OA* and *IA* tasks, previous studies also used the *SA* task (same articulatory movement without producing sounds) to examine this transformation. These studies showed that the motor-based prediction mechanism is more involved in *SA* than *IA* task (Okada et al., 2017; Oppenheim & Dell, 2008, 2010). That is, the strength of representation along the motor-to-sensory transformation pathway would be graded according to the detailed articulatory movement.

Therefore, it is crucial to provide empirical evidence for the dynamics of representational format in the motor-to-sensory transformation and examine whether the dynamics was also task-dependant. In detail, the first and foremost question is the correspondence between the representational format and the transformation stage. Specifically, is the transformation of representation format

a strictly serial process that updates from the motor to sensory domains without overlapping? Second, how do the motor and sensory representational formats change among tasks that involve graded articulatory motor actions? Specifically, could the engagement of overt articulatory movement in *OA* and *SA* tasks strengthen the corresponding representation?

Previous studies mainly combined univariate analyses with overt and covert speech tasks to map out motor-to-sensory transformation pathways (Kleber et al., 2017; Tian & Poeppel, 2013, 2015). Note that univariate activation is quantified by measuring relative differences between conditions, and the results depend on what conditions are being compared. Therefore, the univariate analysis can reveal cortical regions that mediate cognitive functions. However, it is not optimal to examine the detailed representational formats. An advanced analysis method – representation similar analysis (RSA) – is more suitable to examine the neural representation of speech features (Evans & Davis, 2015). By taking advantage of systematic variance distributed across voxels like other multi-voxel pattern analysis (MVPA) methods, RSA makes it possible to examine representational content such as phonetic features (Feng et al., 2017; Mur et al., 2009). In detail, we can particularly create distinct articulatory and acoustic Representation Dissimilar Matrices (RDMs) to probe the detailed neural representation in the motor-to-sensory transformation (Carey & Mcgettigan, 2016; Carey et al., 2017).

The present study used three speech tasks (*OA*, *SA*, and *IA*) and combined the univariate analysis with RSA to systematically investigate the representational formats in the motor-to-sensory transformation. Univariate analysis was first used to reveal the regions engaged in the motor and sensory systems during speech production. More importantly, RSA was used to systematically investigate the representational transformation of phonetic formats. Specifically, we used 16 consonant-vowel syllables and constructed theoretical RDMs that reflected articulatory and acoustic information, respectively. The correlations between theoretical RDMs and the neural pattern RDMs were obtained to examine the representational format of speech features in the motor-to-sensory transformation among different speech tasks. Because the common motor and sensory neural networks mediate different types of speech production (Price, 2012), we first predicted that the motor-to-sensory transformation was available in all three speech tasks. Articulatory and acoustic information should be selectively represented in the motor and sensory regions among all three tasks. Furthermore, previous theoretical and empirical works have indicated a common representation linking between sensory and motor systems (Assaneo et al., 2019; Cogan et al., 2014; Hickok, 2012; Zhen

et al., 2019). Moreover, the engagement of overt articulatory movement strengthens the representations from phonological to phonetic levels (Oppenheim & Dell, 2008, 2010). Therefore, we further predicted that the OA and SA tasks could strengthen and extend the articulatory information to parietal somatosensory or even temporal auditory regions in the motor-to-sensory transformation pathway (Carey et al., 2017).

2. Methods

2.1. Participants

Nineteen university students (9 males, age range: 19–26, mean: 22.8, SD: 2.68) participated in the fMRI experiment. Their native languages were all Mandarin Chinese. They were all right-handed with normal or corrected to normal vision. All the participants reported no history of speech or language disorders. This study was approved by the New York University Shanghai Institutional Review Board (IRB). Written informed consent was obtained from all participants who received monetary incentives for their participation.

2.2. Stimuli and task

We combined each of the eight consonants (/b/, /p/, /d/, /t/, /w/, /f/, /z/, and /s/) with each of the two vowels (/a/ and /u/) to construct 16 Chinese consonant-vowel syllables (/ba/, /pa/, /da/, /ta/, /wa/, /fa/, /za/, /sa/, /bu/, /pu/, /du/, /tu/, /wu/, /fu/, /zu/ and /su/) that were used in this study. All syllables are common conversational speech in Mandarin Chinese. We adopted three tasks in this study: OA, SA, and IA. For the OA task, participants were asked to overtly articulate each syllable. For the SA task, they were asked to articulate each syllable without phonating. For the IA task, they were asked to imagine speaking each syllable without overtly articulating. Note that participants were required to generate the articulatory movements in the SA task. However, such movements were strongly discouraged in the IA task. Participants were asked to articulate with the first tone (high) in Chinese for all tasks.

2.3. Procedure

Each participant completed three functional runs, with each run contained one task. As shown in Figure 1, each trial began with a visual cue displayed at the center of the screen for 1000 ms. Different pictures of the visual cue indicated the task and the syllable labels beneath the picture informed the content of the task. A 1200 ms blank, a 600 ms blue circle, and 600 ms blank were then

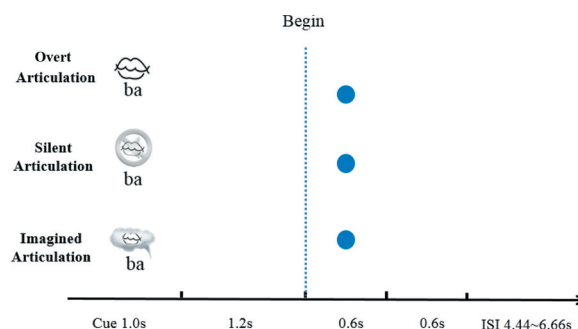


Figure 1. Experimental procedure. The visual cue and syllable label indicated a specific task and syllable content to perform. Participants began the task time-locked to the onset of the blue circle.

presented sequentially. Participants were asked to begin the task time-locked to the onset of the blue circle. They were asked to articulate the syllable only once for each trial but not explicitly instructed to hold the syllable for a specific amount of time to ensure the natural of articulation. The inter-trial interval (ITI) was randomly chosen from 4440 to 6660 ms (2 to 3 TRs, see MRI scanning protocol for details), temporally jittered by 148 ms increments. Therefore, the average duration of trials was 8950 ms (1000 ms + 2400 ms + 5550 ms).

Each syllable was presented three times in each task (Yee et al., 2010). Each task also included six resting trials that were 8500 ms in length and visually cued with the word 'rest'. Therefore, each functional run contained 54 trials. The stimuli presentation of each run was preceded by 10 s and then followed by 20 s of passive viewing of an asterisk to enhance the estimation of baseline signals. Therefore, the total duration of each functional run was 510.6 s. The order of the tasks was presented with the Latin Square Design across participants to control the task order effect. The order of trials in each task was randomized once and then presented in the same order to each participant.

Before the fMRI experiment, each participant was asked to familiarize themselves with all the syllables and conduct a practice session. The procedure of practice session was the same as the fMRI experiment except that each syllable was presented only once. Participants were asked to focus on the timing of tasks as well as differences among tasks (articulation in OA without much head movement, mouth movement without making sounds in SA, and generate vividness of imagery in IA). Feedback was provided if needed. The practice lasted for about 15 min on average. After the fMRI experiment, we recorded the production of the 16 syllables by each participant in a sound-isolated booth. This is for the construction of participant-specific acoustic RDMS (see RDM construction part for details).

2.4. MRI scanning

Scanning was performed with a Siemens Trio Tim 3 T system in East China Normal University. Functional data were acquired using a gradient-echo, echo-planar pulse (EPI) sequence (TR = 2220 ms; TE = 30 ms; 38 slices; $3 \times 3 \times 3 \text{ mm}^3$ voxel size with 0.6 mm interslice gap). We rotated the scanning orientation counter-clockwise about 30° from the AC-PC line to maximize the coverage. Each functional run lasted 230 TRs ($510.6/2.22 = 230$). High resolution T1-weighted anatomical images were collected before the functional scan from each participant. Specifically, these images were acquired with a magnetization-prepared rapid acquisition gradient echo (MP-RAGE) sequence and sagittal slice orientation (176 slices, TR = 1900 ms, TE = 2.53 ms, FOV = 256 mm \times 256 mm, flip angle = 9° , voxel size = $1 \times 1 \times 1 \text{ mm}^3$, duration = 4 min 26 s).

2.5. MRI preprocessing and univariate analysis

MRI data were analyzed using SPM12 (<http://www.fil.ion.ucl.ac.uk/spm/>). For each participant, all functional images were corrected for head-motion and realigned to the first functional image. The mean functional images were coregistered with the structural image and then segmented. The parameters obtained in the segmentation step were used to normalize the functional images onto the Montreal Neurological Institute (MNI) space. The resulting normalization functional images were smoothed with a Gaussian kernel of 6-mm full width at half maximum.

The functional images were further analyzed using the general linear model (GLM) at the first level for each task, respectively. Specifically, a standard hemodynamic response function (HRF) model was fitted to the

data to estimate the voxel-wise statistical parameters (beta values in SPM). In each model, the first regressor corresponded to the presentation of all stimuli. An additional six regressors of no interest representing the motion parameters were entered into the GLM. We used a high-pass filter with a time constant of 128 s to reduce the influence of low-frequency noise. Parameter estimates for the events of interest were obtained, and statistical maps were created.

For the group-level analysis, we first combined the contrast maps across the participants and performed a voxel-wise t-test against the intrinsic baseline for each task. This was to identify brain regions activated for each task, and ensure that our activation maps were consistent with previous speech articulation studies. Furthermore, we performed voxel-wise t-tests for $OA > SA$, $OA > IA$, and $SA > IA$ contrasts to identify brain regions showing the task effect. All reported whole-brain analyses used voxels significant at 0.005 with an FWE correction at the cluster level with a critical cluster level of 0.05. We reported the resultant significance as t-value maps in MNI space.

2.6. RDM construction

We developed the articulator and acoustic (speech spectra and formants) theoretical model RDMs respectively (Figure 2). They are sensitive to different kinds of phonetic information based on the 16 syllables. We constructed the articulator RDM based on the number of same features in articulatory dimensions (the number of articulatory features that each pair of syllables share). According to previous literature (Correia et al., 2015), there are three articulatory dimensions in consonants: articulation manner, articulation place, and voicing. The

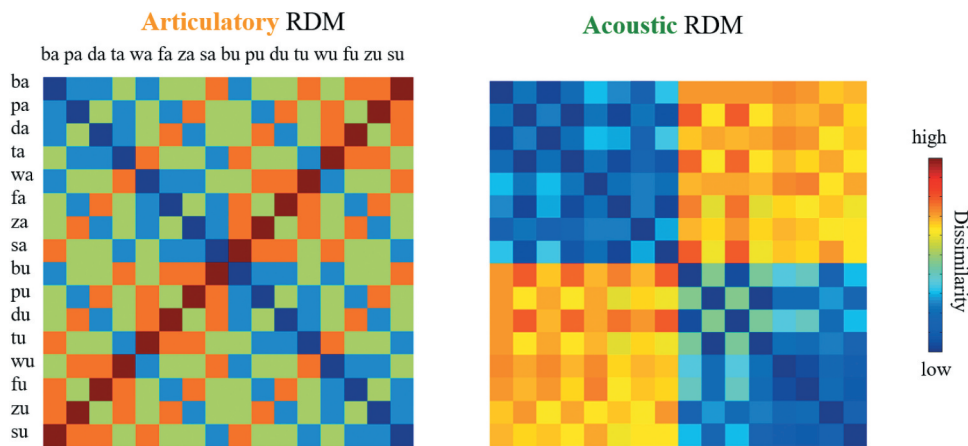


Figure 2. The RDMs for the RSA. The articulatory RDM was constructed based on the articulatory features. The acoustic RDM was constructed based on the PSD information. The syllables of the horizontal and vertical direction of RDMs are */ba/, /pa/, /da/, /ta/, /wa/, /fa/, /za/, /sa/, /bu/, /pu/, /du/, /tu/, /wu/, /fu/, /zu/* and */su/*.

8 consonants we selected orthogonally cover all the 3 articulatory dimensions: stop (/b/,/p/,/d/,/t/) and fricative (/w/,/f/,/z/,/s/) for manner of articulation; bilabial/labiodental (/b/,/p/,/w/,/f/) and alveolar (/d/,/t/,/z/,/s/) for place of articulation; and voiced (/b/,/d/,/w/,/z/) and unvoiced (/p/,/t/,/f/,/s/) for voicing. Furthermore, the feature of the vowels (/a/and/u/) is similar in articulation manner and voicing dimensions but different in the articulation place dimension. We constructed the articulatory RDM as follows. Firstly, we made pairs of all 16 syllables to yield 120 ($16 \times (16-1)/2$) syllable pairs. For each pair of syllables, we calculated the number of the same features in all three articulatory dimensions. We calculated this value for the consonants and vowels separately and subsequently added them to represent the overall articulatory information of syllable. These values were normalized to a range between 0 and 1 using the min-max method. The normalized scores were subtracted from 1 to represent the articulatory dissimilarity of each pair.

We constructed the acoustic RDMs based on the power spectral density (PSD) information of syllables. We used the recorded sound files of syllables from each participant to construct the participant-specific acoustic RDM. Specifically, we first extracted the samples of each syllable in Matlab. We then derived the PSD matrix of each syllable, using a Goertzel discrete Fourier transform spectrogram algorithm (range 0.1–5000 Hz; 260 window length; 0.1 Hz increment) (see Carey et al., 2017, for previous use of the same settings). Each PSD matrix was then cross-correlated over all possible pairs of syllables (yielding a 16×16 matrices). The correlation values were subtracted from 1 to create the acoustic RDM.

2.7. Regional-level RSA for each task

We were mainly interested in the neural representation of phonetic features in specific cortices. Therefore, we used ROI-based RSA for the following two reasons (Feng et al., 2017; Hjortkjaer et al., 2017). First, we can directly examine the neural representation of specific ROIs in the motor-to-sensory transformation stream. Second, ROI analysis avoids the problem of multiple comparisons correction that may penalize the relatively weak effect (Poldrack, 2007).

According to previous studies, articulatory preparation and execution mainly localizes in the anterior insula (INS) (Baldo et al., 2011) and the inferior frontal gyrus (IFG) (Rampinini et al., 2017). Estimating the somatosensory consequences localizes in the angular gyrus (AG) (in the temporal-parietal junction) (Rogalsky et al., 2015). Speech perception localizes in superior temporal gyrus (STG) (Bonte et al., 2014; Mesgarani et al., 2014). Therefore, we created ROIs in these brain regions for this analysis. Specifically, we used the Harvard-Oxford Atlas to define the independent anatomical ROI (probability > 0.3). We manually extracted the anterior part of the insula region in the Harvard-Oxford Atlas. In all, we defined the following 6 ROIs in each hemisphere: anterior INS (aINS), IFG-pars triangularis (IFG-PT), IFG-pars opercularis (IFG-PO), AG, posterior STG (pSTG) and anterior STG (aSTG) (see Figure 3). We showed the voxel number of each ROI in Supplementary Table 1.

We used the CoSMo RSA toolbox (Oosterhof et al., 2016) and customized Matlab functions for this analysis. We conducted RSA on functional images following realignment and normalization but without smoothing. We

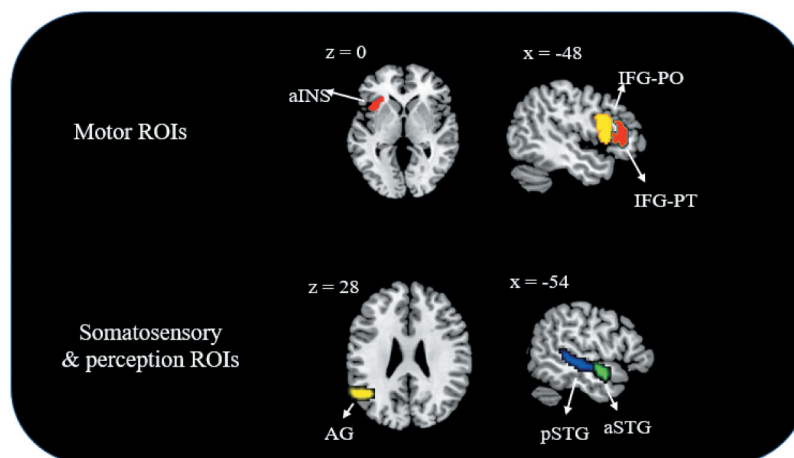


Figure 3. The ROIs used for Regional-level RSA analysis. Note that symmetrical ROIs in both hemispheres were defined for analysis. Only ROIs in the left hemisphere were illustrated here. We divided the ROIs into motor ROIs (aINS, IFG-PT and IFG-PO) and somatosensory & auditory ROIs (AG, pSTG, and aSTG). Abbreviations: aINS: anterior insula; IFG-PT: inferior frontal gyrus-pars triangularis; IFG-PO: inferior frontal gyrus-pars opercularis; AG: angular gyrus; pSTG: posterior superior temporal gyrus; aSTG: anterior superior temporal gyrus.

used the GLM with individual regressors for each syllable and further calculated the single-syllable t-statistic maps for each task. The ROI images were resliced to the same resolution as that of the t-statistic images. For each ROI, the t-statistic value was extracted to calculate the dissimilarity of the neural activation pattern between each pair of syllables (using 1-Pearson's correlation). This created the neural activation pattern RDM. We then directly computed the correlation coefficient between the neural RDM with the theoretical model RDM (Feng et al., 2017). The resulting correlation values were then Fisher's r -to- z transformed. At the group level, we used one-sample t-tests to assess whether the averaged correlation value of specific ROI was significantly higher than 0 (Zhao et al., 2016).

3. Results

3.1. Univariate analysis results

To obtain the main effects of each speech task and reveal the core network that mediates motor-to-sensory transformation, univariate analysis was conducted. The univariate analyses of the three tasks showed similar activation. As shown in Figure 4(a), activation was observed in the INS, IFG, AG, and STG regions under both the OA and SA tasks. Under the IA task, activation was observed in the bilateral IFG, left AG, STG, and middle temporal gyrus (MTG). Apart from the

classic frontal-parietal-temporal cortex, we also found activation in the cerebellum in all three tasks.

In the paired comparison, the OA > SA contrast revealed greater activation in the bilateral STG and superior frontal gyrus (SFG). For both the OA > IA and SA > IA contrasts, we found greater activation in bilateral posterior IFG, INS, and STG. Moreover, we also found greater activation in the cerebellum for the SA > IA contrast (See Figure 4(c)).

3.2. RSA results

To investigate the detailed representation in the motor-to-sensory transformation network, RSA was conducted in anatomically defined ROIs for each speech task. According to the distinct systems in motor-to-sensory transformation, we divided these ROIs into two groups – motor ROIs (aINS, IFG-PT, and IFG-PO) and somatosensory & auditory ROIs (AG, pSTG, and aSTG). The overall RSA results showed clear patterns. That is, 1) articulatory and acoustic information was represented in motor and somatosensory & auditory regions, respectively; and 2) the information flow between motor and perceptual systems showed a gradient transformation as the articulatory information was also represented in the somatosensory and sensory regions in the OA and SA tasks.

Specifically, for the articulatory RDM, in the motor ROIs that included bilateral aINS, IFG-PT and IFG-PO (left group of bars in the left column of Fig. 5), significant results were observed in bilateral aINS, left IFG-PT and left IFG-PO in the

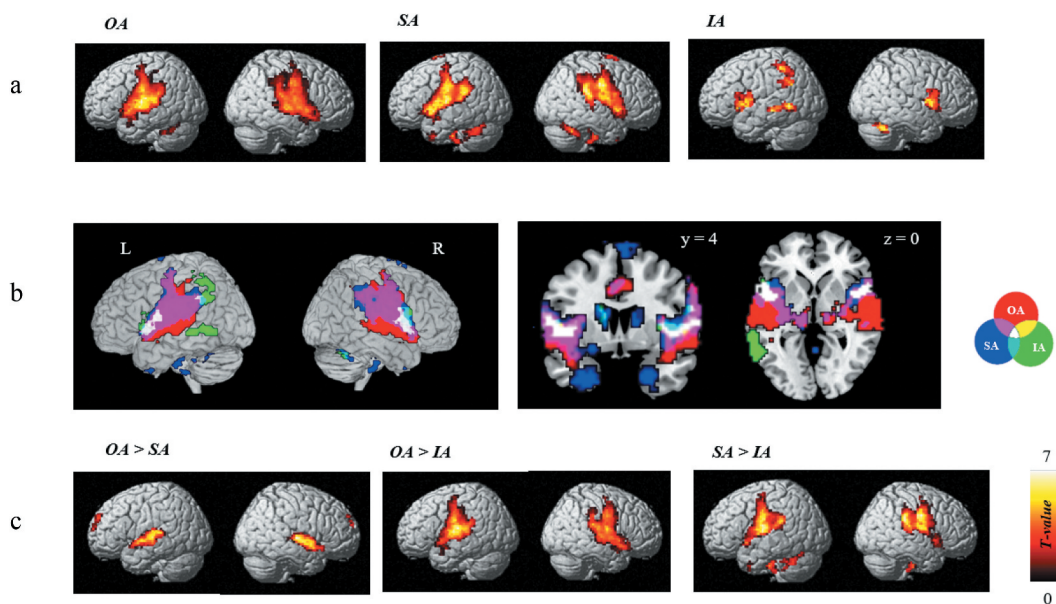


Figure 4. Brain activation maps of univariate analysis results in three tasks. (a) Activation maps for each task against the baseline. (b) Overlap maps of significant brain regions that were activated in three tasks. Left, the activation regions were superimposed on lateral surface rendering templates. Right, the activation regions were superimposed on coronal and axial templates. (c) Activation maps of differences between a pair of tasks. L: left hemisphere; R: right hemisphere; OA: overt articulation; SA: silent articulation; IA: imagined articulation.

OA task. We also found significant results in the left aINS and IFG-PO in the SA task and in left aINS, IFG-PT and IFG-PO in the IA task. In the somatosensory & auditory ROIs that included bilateral AG, pSTG and aSTG (right group of bars in the left column of Fig. 5), significant results were obtained in the left AG, bilateral pSTG and right aSTG in the OA task. We also found significant results in right pSTG and aSTG in the SA task. In the IA task, we did not find any significant results for the articulatory RDM in somatosensory or auditory ROIs.

For the acoustic RDM, in the somatosensory & auditory ROIs that included bilateral AG, pSTG and aSTG (right group of bars in the right column of Fig. 5), significant results were observed in left AG and bilateral pSTG in the OA task. We also found significant results in bilateral AG, right pSTG and right aSTG in the SA task and significant result in left aSTG in the IA task. However, we did not find any significant results for the acoustic RDM in motor area ROIs (left group of bars in the right column of Fig. 5). Detailed statistics of regional-level RSA results

with the articulatory and acoustic RDM were summarized in Supplementary Table 1 and 2 respectively.

4. Discussion

The present study investigated the dynamics and task dependence of neural representations in the motor-to-sensory transformation during speech production. The univariate analyses revealed the frontal-parietal-temporal neural network in all three speech tasks. The extent of this motor-to-sensory transformation network showed a monotonic decrease from OA, SA to IA task. These results were consistent with previous studies (Okada et al., 2017; Tian et al., 2016) and provided further evidence for the motor-based prediction mechanism. More importantly, the RSA results revealed that articulatory and acoustic information was represented in motor and auditory regions, respectively, in all three tasks. Furthermore, articulatory information was cross-represented in the somatosensory and auditory regions

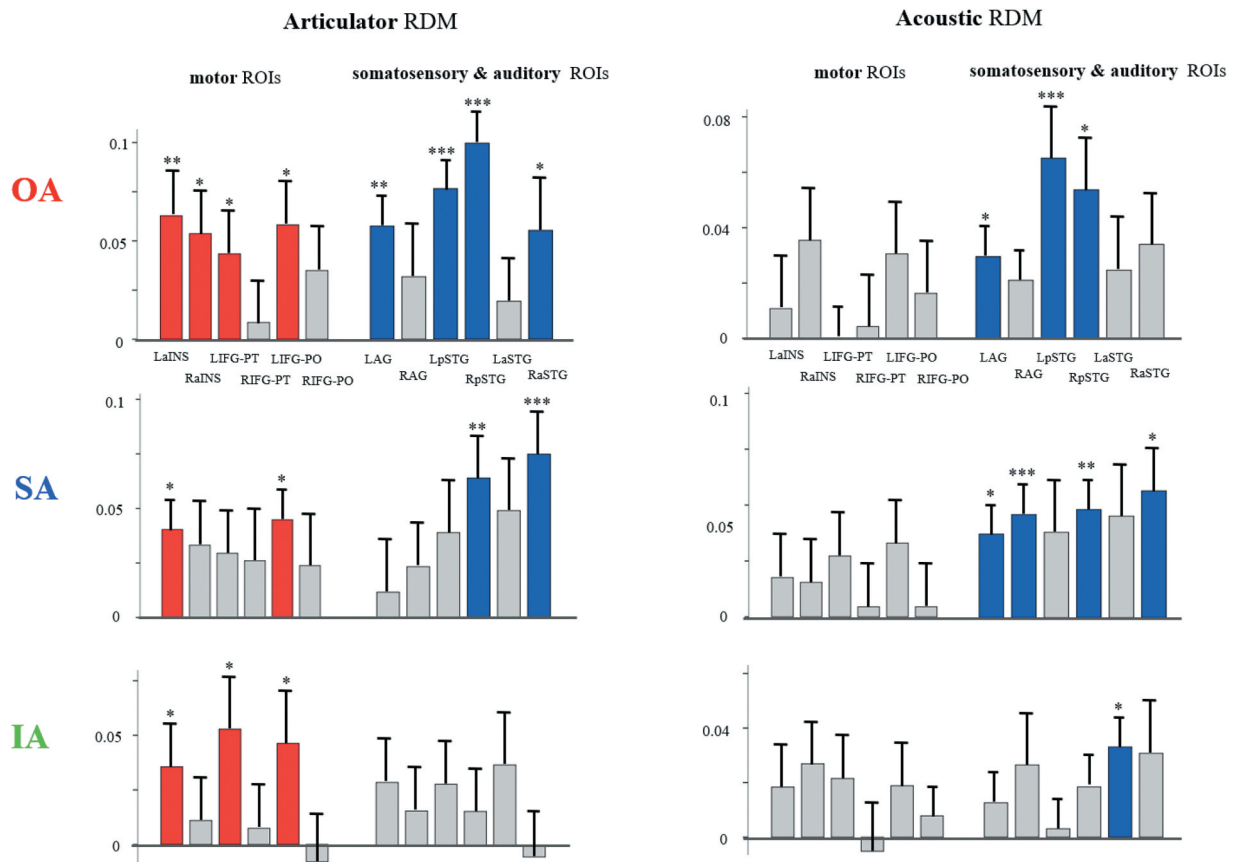


Figure 5. Regional-level RSA results for each theoretical RDM and each task. ROIs are divided into motor ROIs (aINS, IFG-PT and IFG-PO) and somatosensory & auditory ROIs (AG, pSTG and aSTG). The colors of the bars represent the significance of results – red for significant results in the motor ROIs and blue for significant results in the somatosensory & auditory ROIs, whereas gray for no significant results. Error bars denote standard error of the mean. Abbreviations: aINS: anterior insula; IFG-PT: Inferior Frontal Gyrus-pars triangularis; IFG-PO: inferior frontal gyrus-pars opercularis; AG: angular gyrus; pSTG: posterior superior temporal gyrus; aSTG: anterior superior temporal gyrus. OA: overt articulation; SA: silent articulation; IA: imagined articulation. ***, $P < 0.005$; **, $P < 0.01$; *, $P < 0.05$.

in *OA* and *SA* tasks. These RSA results further suggested that the representational format of phonetic features is dynamic and task-dependent in the motor-to-sensory transformation.

The univariate analysis results showed that the frontal-parietal-temporal neural network was activated in all three speech tasks (Fig. 4). This provided evidence for the motor-to-sensory transformation and its neural pathway in speech production. That is, motor simulation involves the motor cortex such as *INS* and *IFG* to induce the efference copy. The efference copy is sent to the parietal and temporal areas and internally induces the corresponding sensory neural representations. Note we found activation of the *IA* > baseline contrast in the motor cortex (*INS*, *IFG*), somatosensory cortex (*AG*), and auditory representation cortex (*STG*, *MTG*). These results demonstrated that imagery without actual articulator movements was sufficient to form the motor-to-sensory transformation stream (Tian & Poeppel, 2012; Tian et al., 2016). In other words, the realization of articulated movements is not necessary for the formation of this motor-based prediction mechanism. However, note bilateral posterior *IFG*, *AG*, and *STG* were more activated in the *SA* > *IA* contrast, and bilateral *STG* and *SFG* were more activated in the *OA* > *SA* contrast. These results suggested that articulation-related movements, especially for actual speech articulation, could strengthen the neural activity in the motor-to-sensory transformation (Okada et al., 2017).

An important innovation of the present study was that we used RSA to examine the neural representations of phonetic features in the motor-to-sensory transformation. The RSA results first showed a clear tendency that articulatory information was more represented in motor regions, and acoustic information was more represented in somatosensory and sensory regions. Specifically, in all three tasks, articulatory information was represented in motor areas such as *aINS* and *IFG*, whereas acoustic information was represented in somatosensory and sensory regions such as *AG* and *STG* (Fig. 5). These results provide a holistic picture of the dynamics of representational formats in the motor-to-sensory transformation. That is, the transformation of representation format is a relatively cascaded process. Articulatory information involves in the motor simulation stage, then articulatory information was updated to acoustic information after the motor signal is sent for the somatosensory estimation and auditory formation (Hickok, 2012; Houde & Nagarajan, 2011; Tian & Poeppel, 2010, 2012, 2013; Tian et al., 2016).

The detailed transformation process was revealed in this study. In the RSA results, we observed that the articulatory information was represented in both motor

and somatosensory & auditory regions. Further, both the articulatory and acoustic information was represented in *LAG*, and bilateral *pSTG* in *OA* task (Figure 5). These results are consistent with previous studies that showed the representation of articulatory information (vocal tract or *F1–F2* 2D distance) in the auditory temporal lobe (Carey et al., 2017). This bimodal representation is one of the possible mechanisms that serve as an intermediate step that bridges the motor and sensory systems.

The observations of articulatory representations in the auditory regions may imply the target of speech production. Compared to the common consensus of an auditory target in speech production, the somatosensory outcome could be another target for the action of speech to fulfill (e.g. Tremblay et al., 2003). The articulatory representations in the somatosensory areas during *OA* and auditory areas during *OA* and *SA* are consistent with the possible somatosensory target of speech production. Our results of bimodal representations in the auditory areas via motor-to-sensory transformation provide possible solutions to reconcile the differences between the auditory and somatosensory targets of speech production.

During the dynamics of representational format, the RSA results also suggested the task-dependent representation of phonetic features in the motor-to-sensory transformation. Articulatory information was represented in somatosensory and sensory regions in the *OA* and *SA* tasks but not apparently in the *IA* task (Fig. 5). This task-dependent modulation on the motor-to-sensory transformation is consistent with the hypothesis that motor representation serves as a modulatory role in speech (Hickok et al., 2009; Stokes et al., 2019; Tian & Poeppel, 2012). Specifically, the theory states that motor activation is not a core representation of speech. The engagement of motor representation depends on the task demand. In our results, only the involvement of articulator movement in *OA* and *SA* induced the articulatory representations in the auditory areas, but not in *IA*. These results suggest that the imagery task selectively induces the core representation of speech, whereas the task demand of articulator movement in *OA* and *SA* may evoke the auxiliary representation of speech.

The motor-to-sensory transformation is a canonical computation that enables the interaction between production and perception. The observed dynamics of representational transformation and the task-dependent modulation are consistent with our recently proposed functional distinctions between different motor signals along the entire time course of speech production (Li et al., 2020). Specifically, we found that

the speech preparation of a specific target could presumably induce the *efferece copy* that selectively enhanced the processing of the auditory target. Whereas general action preparation without a speech target could presumably induce the *corollary discharge* that ubiquitously suppressed the processing of all speech sound, similar to the speech-induced suppression observed during the execution phase (e.g. Houde et al., 2002). IA can be considered as preparation without execution, which could selectively induce the *efferece copy*, link to the specific auditory representation and enhance the sensitivity of the auditory target (Ma & Tian, 2019; Tian & Poeppel, 2013; Tian et al., 2016). Whereas the execution in OA and SA causes articulatory representation in the auditory areas, which may provide additional sources for suppressing the processing of the auditory target, and result in a function of separating ex-afferent (external stimuli) and re-afferent (feedback) information.

Comparing the univariate analyses with RSA in this study, the results are mostly consistent with each other, as the most ROIs we defined were also activated in the univariate analysis. In the literature, the univariate analysis and multi-voxel pattern classification (MVPC) were mainly used on the same contrasts (Kriegeskorte et al., 2007), with the MVPC taking advantage of systematic variance distributed across voxels to increase the detection sensitivity (Jimura & Poldrack 2011). Note that the purposes of the univariate analysis and RSA in this study were essentially different. Specifically, we used the contrasts of one task against the baseline or another task in the univariate analysis. However, we conducted RSA within a specific task to examine the neural representation of phonetic features. Hence, these two methods are complementary to provide distinct characteristics about the strength, extent, and patterns of representation in the motor-to-sensory transformation network.

The important methodological implication of this study is the combination of conventional univariate analyses and RSA methods. Specifically, univariate analyses could examine the involvement of a region in specific cognitive activities. RSA could be used complementarily to indicate specific representational content in a region (Mur et al., 2009). Future studies can use RSA to examine the neural representational content in speech learning and speech-related disorders. For example, RSA could be used to examine the discrepancy of representational content between the noisy perceptual estimation and external feedback in stuttering (Tian & Poeppel, 2012) as well as the similarity between the motor-to-sensory transformation and speech goal (Liu & Tian, 2018).

In conclusion, the present study employed three speech tasks and fMRI to systematically examine the motor-to-

sensory transformation during speech production. The univariate and RSA results revealed the frontal-parietal-temporal network that mediates the motor-to-sensory transformation. Moreover, the RSA results further suggested a bimodal representation that subserves the transformation between motor and sensory systems, and this extension of motor representation into sensory regions is constrained by task and motor involvement. These consistent results provided evidence for the dynamics and task-dependent transformation between representational formats in the motor-to-sensory transformation.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was supported by the National Natural Science Foundation of China 31871131, Major Program of Science and Technology Commission of Shanghai Municipality (STCSM) 17JC1404104, and Program of Introducing Talents of Discipline to Universities, Base B16018.

ORCID

Xing Tian  <http://orcid.org/0000-0003-1629-6304>

References

- Adank, P. (2012). The neural bases of difficult speech comprehension and speech production: Two Activation Likelihood Estimation (ALE) meta-analyses. *Brain and Language*, 122(1), 42–54. <https://doi.org/10.1016/j.bandl.2012.04.014>
- Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753), 493–498. <https://doi.org/10.1038/s41586-019-1119-1>
- Assaneo, M. F., Ripolles, P., Orpella, J., Lin, W. M., De Diego-balaguer, R., & Poeppel, D. (2019). Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nature Neuroscience*, 22(4), 627–632. <https://doi.org/10.1038/s41593-019-0353-z>
- Baldo, J. V., Wilkins, D. P., Ogar, J., Willock, S., & Dronkers, N. F. (2011). Role of the precentral gyrus of the insula in complex articulation. *Cortex*, 47(7), 800–807. <https://doi.org/10.1016/j.cortex.2010.07.001>
- Bonte, M., Hausfeld, L., Scharke, W., Valente, G., & Formisano, E. (2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *Journal of Neuroscience*, 34(13), 4548–4557. <https://doi.org/10.1523/JNEUROSCI.4339-13.2014>
- Carey, D., & Mcgettigan, C. (2016). Magnetic resonance imaging of the brain and vocal tract: Applications to the study of speech production and language learning. *Neuropsychologia*, 98(2017), 201–211. <https://doi.org/10.1016/j.neuropsychologia.2016.06.003>

- Carey, D., Miquel, M. E., Evans, B. G., Adank, P., & McGettigan, C. (2017). Vocal tract images reveal neural representations of sensorimotor transformation during speech imitation. *Cerebral Cortex*, 27(5), 3064–3079. <https://doi.org/10.1093/cercor/bhx056>
- Cogan, G. B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., & Pesaran, B. (2014). Sensory–motor transformations for speech occur bilaterally. *Nature*, 507(7490), 94–98. <https://doi.org/10.1038/nature12935>
- Correia, J. M., Jansma, B. M., & Bonte, M. (2015). Decoding Articulatory Features from fMRI Responses in Dorsal Speech Regions. *Journal of Neuroscience*, 35(45), 15015–15025. <https://doi.org/10.1523/JNEUROSCI.0977-15.2015>
- Evans, S., & Davis, M. H. (2015). Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis. *Cerebral Cortex*, 25(12), 4772–4788. <https://doi.org/10.1093/cercor/bhv136>
- Feng, G., Gan, Z., Wang, S., Wong, P. C. M., & Chandrasekaran, B. (2017). Task-general and acoustic-invariant neural representation of speech categories in the human brain. *Cerebral Cortex*, 28(9), 1–14. <https://doi.org/10.1093/cercor/bhx195>
- Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96(3), 280–301. <https://doi.org/10.1016/j.bandl.2005.06.001>
- Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, 13(2), 135–145. <https://doi.org/10.1038/nrn3158>
- Hickok, G., Buchsbaum, B., Humphries, C., & Muftuler, T. (2003). Auditory–motor interaction revealed by fMRI: Speech, music, and working memory in area Spt. *Journal of Cognitive Neuroscience*, 15(5), 673–682. <https://doi.org/10.1162/089892903322307393>
- Hickok, G., Holt, L. L., & Lotto, A. J. (2009). Response to Wilson: What does motor cortex contribute to speech perception? *Trends in Cognitive Sciences*, 13(8), 330–331. <https://doi.org/10.1016/j.tics.2009.05.002>
- Hjortkjaer, J., Kassuba, T., Madsen, K. H., Skov, M., & Siebner, H. R. (2017). Task-modulated cortical representations of natural sound source categories. *Cerebral Cortex*, 28(1), 1–12. <http://doi.org/10.1093/cercor/bhx263>
- Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, 5(2011), 82. <https://doi.org/10.3389/fnhum.2011.00082>
- Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: An MEG study. *Journal of Cognitive Neuroscience*, 14(8), 1125–1138. <https://doi.org/10.1162/089892902760807140>
- Huang, J., Carr, T. H., & Cao, Y. (2002). Comparing cortical activations for silent and overt speech using event-related fMRI. *Human Brain Mapping*, 15(1), 39–53. <https://doi.org/10.1002/hbm.1060>
- Jimura, K., & Poldrack, R. A. (2011). Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*, 50(4), 544–552. doi:10.1016/j.neuropsychologia.2011.11.007
- Kleber, B., Friberg, A., Zeitouni, A., & Zatorre, R. (2017). Experience-dependent modulation of right anterior insula and sensorimotor regions as a function of noise-masked auditory feedback in singers and nonsingers. *Neuroimage*, 147(2017), 97–110. <https://doi.org/10.1016/j.neuroimage.2016.11.059>
- Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, 104(51), 20600–20605. <https://doi.org/10.1073/pnas.0705654104>
- Li, S., Zhu, H., & Tian, X. (2020). Corollary Discharge Versus Efference Copy: Distinct neural signals in speech preparation differentially modulate auditory responses. *Cerebral Cortex*. <https://doi.org/10.1093/cercor/bhaa154>
- Liu, X., & Tian, X. (2018). The functional relations among motor-based prediction, sensory goals and feedback in learning non-native speech sounds: Evidence from adult Mandarin Chinese speakers with an auditory feedback masking paradigm. *Scientific Reports*, 8(1), 1–13. <https://doi.org/10.1038/s41598-017-17765-5>
- Ma, O., & Tian, X. (2019). Distinct mechanisms of imagery differentially influence speech perception. *eNeuro*, 6(5), 1–11. <https://doi.org/10.1523/ENEURO.0261-19.2019>
- 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>
- Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, 4(nsn044), 101–109. <https://doi.org/10.1093/scan/nsn044>
- Okada, K., Matchin, W., & Hickok, G. (2017). Neural evidence for predictive coding in auditory cortex during speech production. *Psychonomic Bulletin Review*, 25(1), 423–430. <https://doi.org/10.3758/s13423-017-1284-x>
- Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, 10(2016), 27. <https://doi.org/10.3389/fninf.2016.00027>
- Oppenheim, G. M., & Dell, G. S. (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition*, 106(1), 528–537. <https://doi.org/10.1016/j.cognition.2007.02.006>
- Oppenheim, G. M., & Dell, G. S. (2010). Motor movement matters: The flexible abstractness of inner speech. *Memory & Cognition*, 38(8), 1147–1160. <https://doi.org/10.3758/MC.38.8.1147>
- Parrell, B., Agnew, Z., Nagarajan, S., Houde, J., & Ivry, R. B. (2017). Impaired feedforward control and enhanced feedback control of speech in patients with cerebellar degeneration. *Journal of Neuroscience*, 37(38), 9249–9258. <https://doi.org/10.1523/JNEUROSCI.3363-16.2017>
- Poldrack, R. A. (2007). Region of interest analysis for fMRI. *Social Cognitive & Affective Neuroscience*, 2(1), 67–70. <https://doi.org/10.1093/scan/nsm006>
- Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, 62(2), 816–847. <https://doi.org/10.1016/j.neuroimage.2012.04.062>
- Rampinini, A. C., Handjaras, G., Leo, A., Cecchetti, L., Ricciardi, E., Marotta, G., & Pietrini, P. (2017). Functional and spatial segregation within the inferior frontal and superior temporal cortices during listening, articulation imagery, and production of vowels. *Scientific Reports*, 7(1), 17029. <https://doi.org/10.1038/s41598-017-17314-0>
- Rogalsky, C., Poppa, T., Chen, K. H., Anderson, S. W., Damasio, H., Love, T., & Hickok, G. (2015). Speech repetition

- as a window on the neurobiology of auditory-motor integration for speech: A voxel-based lesion symptom mapping study. *Neuropsychologia*, 71(1), 18–27. <https://doi.org/10.1016/j.neuropsychologia.2015.03.012>
- Stokes, R. C., Venezia, J. H., & Hickok, G. J. (2019). The motor system's [modest] contribution to speech perception. *Psychonomic Bulletin Review*, 26(4), 1354–1366. <https://doi.org/10.3758/s13423-019-01580-2>
- Tian, X., Ding, N., Teng, X., Bai, F., & Poeppel, D. (2018). Imagined speech influences perceived loudness of sound. *Nature Human Behaviour*, 2(3), 225–234. <https://doi.org/10.1038/s41562-018-0305-8>
- Tian, X., & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology*, 1(3), 255–262. <https://doi.org/10.3389/fpsyg.2010.00166>
- Tian, X., & Poeppel, D. (2012). Mental imagery of speech: Linking motor and perceptual systems through internal simulation and estimation. *Frontiers in Human Neuroscience*, 6(6), 314. <https://doi.org/10.3389/fnhum.2012.00314>
- Tian, X., & Poeppel, D. (2013). The effect of imagination on stimulation: The functional specificity of efference copies in speech processing. *Journal of Cognitive Neuroscience*, 25(7), 1020–1036. https://doi.org/10.1162/jocn_a_00381
- Tian, X., & Poeppel, D. (2015). Dynamics of self-monitoring and error detection in speech production: Evidence from mental imagery and MEG. *Journal of Cognitive Neuroscience*, 27(2), 352–364. https://doi.org/10.1162/jocn_a_00692
- Tian, X., Zarate, J. M., & Poeppel, D. (2016). Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex*, 77(7), 1–12. <https://doi.org/10.1016/j.cortex.2016.01.002>
- Tremblay, S., Shiller, D. M., & Ostry, D. J. J. N. (2003). Somatosensory basis of speech production. *Nature*, 423(6942), 866–869. <https://doi.org/10.1038/nature01710>
- Whitford, T. J., Jack, B. N., Pearson, D., Griffiths, O., Luque, D., Harris, A. W., Spencer, K. M., & Le Pelley, M. E. (2017). Neurophysiological evidence of efference copies to inner speech. *eLife*, 6(2017), e28197. <https://doi.org/10.7554/eLife.28197>
- Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, 3(Suppl), 1212–1217. <https://doi.org/10.1038/81497>
- Yee, E., Drucker, D. M., & Thompson-Schill, S. L. (2010). fMRI-adaptation evidence of overlapping neural representations for objects related in function or manipulation. *NeuroImage*, 50(2), 753–763. <https://doi.org/10.1016/j.neuroimage.2009.12.036>
- Zhao, L., Chen, C., Shao, L., Wang, Y., Xiao, X., Chen, C., ... Xue, G. (2016). Orthographic and phonological representations in the fusiform cortex. *Cerebral Cortex*, 27(11), 5197–5210. <http://doi.org/10.1093/cercor/bhw300>
- Zhen, A., Van Hedger, S., Heald, S., Goldin-Meadow, S., & Tian, X. (2019). Manual directional gestures facilitate cross-modal perceptual learning. *Cognition*, 187(2019), 178–187. <https://doi.org/10.1016/j.cognition.2019.03.004>