bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Verbal thinking in rhythm: motor-to-sensory transformation network mediates imagined singing

Yanzhu Li[1,2], Huan Luo[3] & Xing Tian[1,2]*

[1] New York University Shanghai, Shanghai, China

[2] NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai

[3] Department of Psychology, Peking University, China

* Corresponding author:

Xing Tian

New York University Shanghai

1555 Century Avenue, Room 1259

Shanghai, China 200122

xing.tian@nyu.edu

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Abstract

What enables us to think verbally? We hypothesized that the interaction between motor and sensory systems induced speech representation without external stimulation or overt articulation. This motor-to-sensory transformation formed the neural basis that enabled us to think verbally. Analogous to the frequency tracking of neural responses to auditory stimuli, we asked participants to imagine singing lyrics of famous songs rhythmically while their neural electro-magnetic signals were recorded using magnetoencephalography (MEG). We found that when participants imagined with less temporal variation, the neural oscillation at the delta band (same frequency band as the rhythm in the songs) showed more consistent phase coherence across trials. This neural phase tracking of imagined singing was observed in a frontal-parietal-temporal network – the proposed motor-to-sensory transformation pathway, including inferior frontal gyrus (IFG), insula, premotor, intra-parietal sulcus (IPS), temporal-parietal junction (TPJ), primary auditory cortex, and superior temporal gyrus and sulcus (STG & STS). These results suggest that neural oscillations can entrain the rhythm of our mental activity. The coherent activation in the motor-to-sensory transformation neural network mediates the internal construction of perceptual representation and forms the neural computation foundation for inner speech during verbal thinking.

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Introduction

'What is this paper about?' Probably you are asking this question now in your mind. We think in a verbal form all the time in everyday life. Verbal thinking is usually manifested in the form of inner speech – a type of mental imagery induced by covert speaking (Alderson-Day & Fernyhough, 2015; Sokolov, 2012; Tian & Poeppel, 2012). What enables the train of thought as inner speech?

Neural evidence suggest that modality-specific cortical processes mediate covert operations of mental functions. For example, previous studies have been demonstrated that mental imagery was mediated by neural activity in modality-specific cortices, such as motor system for motor imagery (Jeannerod, 1995; Porro et al., 1996) and sensory systems for visual imagery (Kosslyn et al., 1999; Wheeler, Petersen, & Buckner, 2000) and auditory imagery (Kraemer, Macrae, Green, & Kelley, 2005; Zatorre, Halpern, Perry, Meyer, & Evans, 1996).

Recently, the internal forward model has been proposed to internally link the motor and sensory systems (Wolpert & Ghahramani, 2000). The presupposition is the mechanism of *motor-to-sensory transformation* -- a copy of motor command, termed as efference copy, was internally sent to sensory regions to estimate the perceptual consequence of actions (Kawato, 1999; Schubotz, 2007). The motor-to-sensory transformation have been implicated in speech production, learning and control (Guenther, 1995; Hickok, 2012; Houde & Nagarajan, 2011; Liu & Tian, 2018; Zhen, Van Hedger, Heald, Goldin-Meadow, & Tian, 2019), and have been extended to

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

speech imagery (Jack et al., 2019; Tian, Ding, Teng, Bai, & Poeppel, 2018; Tian & Poeppel, 2010, 2012, 2013, 2014; Tian, Zarate, & Poeppel, 2016; Whitford et al., 2017). The operation of motor-to-sensory transformation has been suggested in a frontal-parietal-temporal network. Specifically, it was assumed that the motor system in the frontal lobe simulated the motor action, whereas the sensory systems in the parietal and temporal lobes estimated the possible perceptual changes caused by the action (Tian & Poeppel, 2010, 2012; Tian et al., 2016). Would the continuous simulation and estimation in the motor-to-sensory transformation network mediate inner speech during verbal thinking (Hesslow, 2002)?

Thinking verbally is similar to speech that is unfolding over time. The analysis of time-series information in speech perception has been investigated with a frequency-tagging paradigm. Using this paradigm, it has been demonstrated that neural oscillations can be temporally aligned to the frequency of acoustic features, such as speech envelop (Ding & Simon, 2012; Luo & Poeppel, 2007). The neural oscillations can also entrain to the perceptual and cognitive constructs, such as syllabic information (Buiatti, Peña, & Dehaene-Lambertz, 2009), music beats (Nozaradan, Peretz, Missal, & Mouraux, 2011; Nozaradan, Peretz, & Mouraux, 2012), and syntactic structures (Ding, Melloni, Zhang, Tian, & Poeppel, 2016). That is, the frequency of oscillation can mirror the rate of internal representation derived from external stimulation. Would neural oscillations track the representations that are constructed without external stimulation, such as inner speech during verbal thinking?

The aim of this study was using a frequency-tagging paradigm to investigate the

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

neural mechanisms that mediated inner speech and verbal thinking. We implemented a natural and rhythmic setting in which participants were asked to imagine singing lyrics of famous songs [Fig. 1a]. Unlike the frequency-tracking of passive listening to external stimuli that had a consistent rate across trials, the production rate in the active imagery inevitably had temporal variance across trials. We used two approaches to deal with temporal variation. First, the purpose of a musical context was to reduce the large temporal variability during imagery – participants would imagine singing in a more consistent rate compared with speaking the same lyrics. Second, we took advantage of the remaining temporal variation among trials in imagery [Fig. 1b-d]. The variation in reaction time correlated with the temporal consistency of neural responses across trials. If the neural oscillations tracked the rate of inner speech, the phase-coherence of neural responses would be different between two groups of trials that have different amount of temporal variation [Fig. 1e]. According to our hypothesis that the motor-to-sensory transformation neural network mediated the inner speech and verbal thinking, we further predicted that the different degrees of neural entrainment to the rate of inner speech between two groups of trials would be observed in specific areas in the frontal, parietal, and temporal regions [Fig. 1f], where the core computation of motor simulation and perceptual estimation in the motor-to-sensory transformation have been indicated (Tian & Poeppel, 2010, 2012; Tian et al., 2016).

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

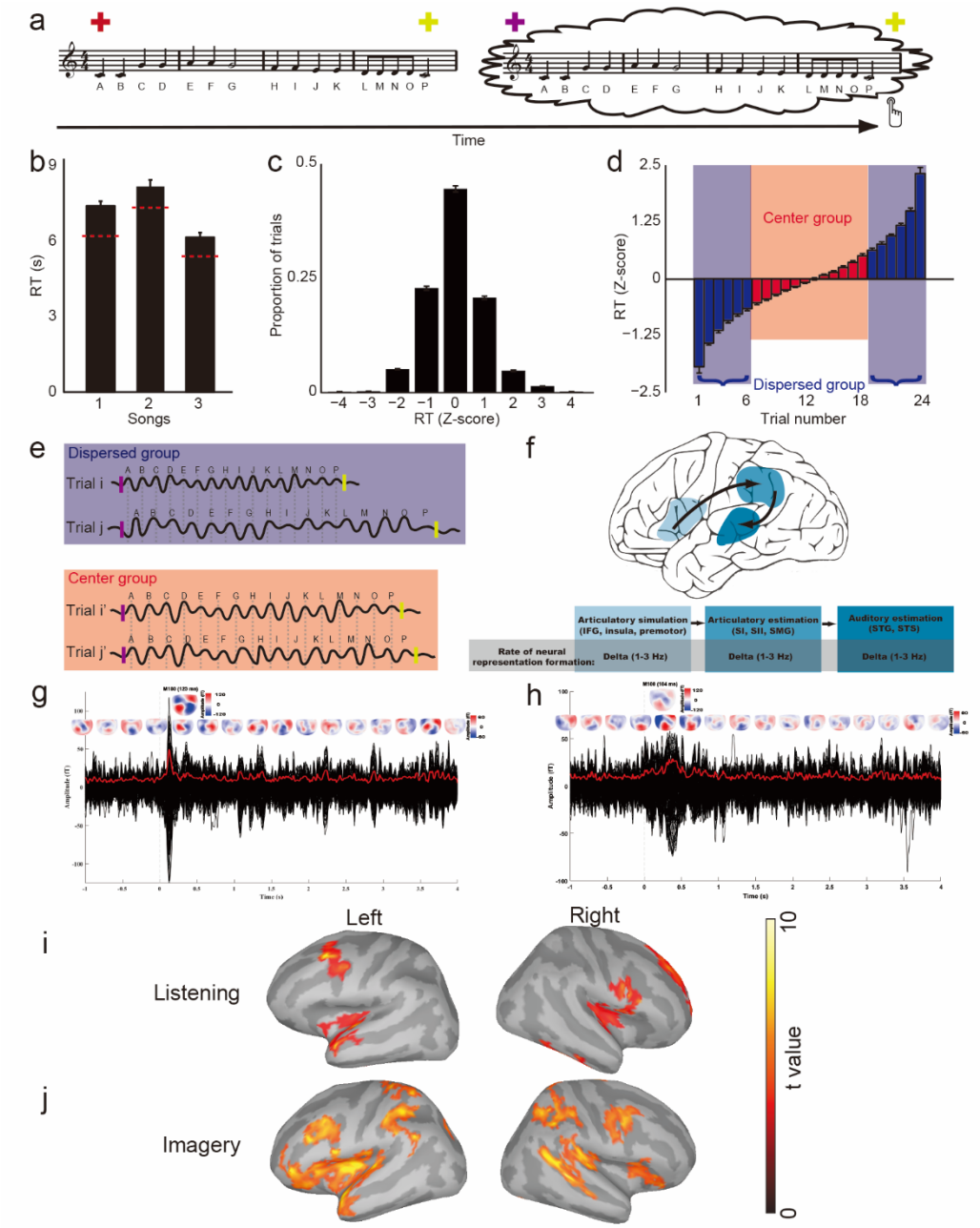

Figure 1. Neural oscillations entrain the rate of imagined singing. A) Experimental paradigm. According to the color of visual fixation, participants were asked to listen to the first sentence of three famous songs, followed by imagined singing the song they just heard. The Alphabet Song was used for illustration. Participants pressed a button to indicate the finish of imagery. B) Reaction time (RT) of imagined singing for three songs. The red dashed lines indicated the duration of the three songs. The duration of imagined singing was longer than the preceding auditory stimuli. C) Distribution of imagined singing RT. The z-scores of RT followed a normal distribution, with about half trials in the range of two standard deviations. D) Grouping of imagined singing trials. Twenty-four trials of each song were sorted ascendingly based on the z-scores and were separated into two groups. Twelve trials that were close to the mean RT were selected in the *center-group*, whereas the other twelve trials that were further away from the mean RT were included in the *dispersed-group*. E) Hypothesis about neural oscillation phase alignment across trials of imagined singing. Schematic display of two trials in each

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

group. The short bar indicated the beginning and end of a trial. The wave lines represented neural oscillations. The trials in the *dispersed-group* had different durations so that temporal variance was large. The phase of neural oscillation that corresponded to the construction of syllabic representation during imagined singing did not align across trials. Whereas in the *center-group*, the temporal variance was small across trials, so that the phase of neural oscillation was more coherent across trials. F) Hypothesis about phase coherence in the motor-to-sensory transformation network during imagined singing. The motor-to-sensory transformation network was assumed to be a frontal-parietal-temporal network, including the inferior frontal gyrus (IFG), insular cortex (INS) and premotor/supplementary motor area (SMA) in the frontal lobe for simulating articulation; somatosensory areas (SI, SII), supramarginal gyrus (SMG) and its adjacent parietal operculum (OP), angular gyrus (AG) and temporal-parietal junction (TPJ) in the parietal lobe for estimating somatosensory consequence; as well as the superior temporal gyrus and sulcus (STG & STS) with a possibility of extension to the Heschl's gyrus (HG) in the temporal lobe for estimating auditory consequence. The more consistent phase coherence at the delta band (1-3 Hz) – the rate of imagined singing of the three songs – were predicted to be observed in the motor-to-sensory network. G) Waveforms and topographies in the listening condition (Alphabet Song). The vertical dotted line at time 0 indicated the onset of the auditory stimuli. Each black line represented waveform responses from a sensor and the red bold line represented the root-mean-square (RMS) waveform across all sensors. Topographies were plotted every 333ms from -1000ms to 4000ms. A clear auditory onset event-related response (M100, the single topography in the upper row) was observed. H) Waveforms and topographies in the imagery condition. Similar depicting form as in G). The vertical dotted line at time 0 indicated the onset of imagined singing. No repetitive patterns in topographies across the period. A similar event-related response in the range of M100 latency as in the listening condition was observed (the single topography in the upper row). I) Phase-coherence results in listening conditions. Neural entrainment was observed in the HG and its adjacent aSTG and pINS, as well as premotor areas. J) Phase coherence results in the imagery conditions. More consistent neural entrainment at the delta band was observed in the proposed motor-to-sensory network, including frontal areas (IFG, aINS and premotor), parietal areas (intraparietal sulcus (IPS) and TPJ), and temporal areas (HG, aSTG, and m&pSTS).

Materials and methods

Participants: Sixteen volunteers (7 males, mean age = 25 years, range from 19 to 32 years) participated in this experiment with monetary compensation. All participants were right-handed native English speakers and without history of neurological disorders. This experiment was approved by the Institutional Review Board (IRB) at New York University.

Materials: Female vocals were recorded for the first sentences of four familiar songs

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

(Alphabet Song: 6.24 s, 16 syllables, 2.56 syllables/sec; Itsy Bitsy Spider: 7.38 s, 23 syllables, 3.12 syllables/sec; Take Me out to the Ball Game: 5.42 s, 13 syllables, 2.40 syllables/sec; and Twinkle Twinkle Little Star: 6.10s, 14 syllables, 2.30 syllables/sec). All songs were recorded with a sampling rate of 44.1 kHz. During the experiment, stimuli were normalized and delivered at about 70 dB SPL via plastic air tubes connected to foam earpieces (E-A-R Tone Gold 3A Insert earphones, Aearo Technologies Auditory Systems).

Procedure: A fixation cross was presented in the center of the screen throughout the experiment. Participants were asked to listen to one of the four songs when the color of fixation was red. The fixation changed to yellow after the auditory stimulus offset. After 1.5s, the fixation changed to purple, and participants were required to imagine singing the song they just heard. They were asked to covertly reproduce the song using the same rhythm and speed as the preceding auditory stimuli. Participants pressed a button to indicate the finish of imagery. Reaction time was recorded. After the button-press, the fixation turned into yellow and stayed on screen for 1.5s-2.5s (with an increment of 0.333s) till the next trial began. During the imagined singing, participants were required to refrain from any overt movement, and not to produce any sounds. A video camera and a microphone were used to monitor any overt movement and vocalization throughout the experiment.

Four blocks were included in this experiment, with 24 trials in each block (6 trials per song in each block, 24 trials per song in total). The presentation order was

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

randomized. Participants familiarized with the experimental procedure before the experiment.

Behavioral analysis: The song of 'Twinkle Twinkle Little Star' was for a research question independent from this study. Therefore, only three songs were used in further analysis. The mean reaction time (RT) was obtained for the imagery of each song. The RT data were further transferred into z-scores. The distribution of z-scores was obtained and averaged across three imagery songs. The RT z-scores of 24 trials were ranked from shortest to longest for each song and averaged across three songs. Two groups were formed based on RT ranking: the *centered group* consisted of 12 trials close to the mean RT, whereas the *dispersed group* comprised the other 12 trials that were farther away from the mean RT.

MEG recording: Neuromagnetic signals were measured using a 157-channel whole head axial gradiometer system (KIT, Kanazawa, Japan). Five electromagnetic coils were attached to each participant's head to monitor head position during MEG recording. The locations of the coils were determined with respect to three anatomical landmarks (nasion, left and right preauricular points) on the scalp using 3D digitizer software (Source Signal Imaging, Inc.) and digitizing hardware (Polhemus, Inc.). The coils were localized to the MEG sensors, at both the beginning and the end of the experiment. The MEG data were acquired with a sampling frequency of 1000 Hz, filtered online between 1 Hz and 200 Hz, with a notch at 60 Hz.

MEG analysis:

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Raw data were noise-reduced offline using the continuously adjusted least square method (Adachi, Shimogawara, Higuchi, Haruta, & Ochiai, 2001) in MEG160 software (MEG Laboratory 2.001 M, Yokogawa Corporation, Eagle Technology Corporation, Kanazawa Institute of Technology). We rejected the artifacts caused by eye movement and cardiac activity with the independent component analysis (ICA). Epochs were extracted for both trials in the listening and imagery conditions, with each epoch of 6000 ms in duration (including 2000 ms pre-stimulus and 4000ms post-stimulus period). For the listening conditions, 24 trials of each song were grouped and formed three *within-groups*. Furthermore, eight trails were randomly sampled from 24 trials of each song and yielded a new group of 24 trials (*between-group*). This sampling procedure was conducted three times to form three *between-groups*. The sampling was without replacement so that each trial was used only once in the three *between-groups*. For imagery conditions, MEG trials were separated into *center-groups* and *dispersed-groups* for each song according to the RT z-scores (refer to the above behavioral analysis).

Fast-Fourier-transform (FFT) was applied on each trial with a 500 ms time window in steps of 200 ms. The phase values were extracted at each time point and frequency. The inter-trial phase coherence (ITC) was calculated as Eq. 1 (Luo & Poeppel, 2007), for each of the *within-groups* and *between-groups* in the listening conditions, and for each of the *center-groups* and *dispersed-groups* in the imagery conditions.

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

$$ITC(t,f) = \left(\frac{\sum_{j=1}^{N} \cos\theta_j(t,f)}{N} \right)^2 + \left(\frac{\sum_{j=1}^{N} \sin\theta_j(t,f)}{N} \right)^2 \quad \text{Eq. 1}$$

The ITC characterizes the consistency of the temporal (phase) neural responses across trials. If the phase responses were identical across trials, the ITC value would be 1. The ITC values were averaged in the Delta band (1-3 Hz), according to our hypothesis that neural responses would track the rhythm in acoustic signals and in imagery at the syllabic rate of 2-3 Hz. Further, the ITC values were averaged over time (0-4 s) and across three songs to yield a single value in every MEG channel for the *within-group* and *between-group* in the listening conditions, and for each of the *center-group* and *dispersed-group* in the imagery conditions. For the *between-group* in the listening condition, the random grouping procedure was repeated 100 times and yield 100 ITC values. The 95th percentile of ITC was chosen for further analysis.

Distributed source localization of ITC was obtained by using the Brainstorm software (Tadel, Baillet, Mosher, Pantazis, & Leahy, 2011). The cortical surface was reconstructed from individual structural MRI using Freesurfer (Martinos Center for Biomedical Imaging, Massachusetts General Hospital, MA). Current sources were represented by 15,002 vertices. Overlapping spheres method was used to compute the individual forward model (Tadel et al., 2011). The inverse solution was calculated by approximating the spatiotemporal activity distribution that best explains the ITC value. Dynamic statistical parametric mapping (dSPM) (Dale et al., 2000) were calculated using the noise covariance matrix estimated with the 1000ms pre-stimulus

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

period. To compute and visualize the group results, each participant's cortical surface was inflated and flattened (Dale, Fischl, & Sereno, 1999) and morphed to a grand average surface (Fischl, Sereno, & Dale, 1999). Source data was spatially smoothed using a Gaussian smoothing function from implemental SurfStat package in Brainstorm.

The non-parametric cluster-based permutation test (Maris & Oostenveld, 2007) was used to assess the significant differences between groups in the source space (Oostenveld, Fries, Maris, & Schoffelen, 2011). For the listening conditions, the ITC values of the *within-group* were compared with the 95th percentile of ITC values of the *between-group*. For the imagery conditions, the ITC values of the *center-group* were compared with *dispersed-group*. The empirical statistics was first obtained by a two-tailed paired t-test with two or more adjacent significant vertices. Next, a null distribution was formed by randomly shuffling the group labels 1000 times. Cluster-level FDR corrected results were obtained by comparing the empirical statistics with the null distribution (alpha=0.05 for the listening conditions, and alpha=0.001 for the imagery conditions).

Results

The reaction time of imagery suggested that the duration of imagined singing was longer than the duration of auditory stimuli [Fig.1b] [one sample t-test; for song 1, $t(15) = 6.04$, $p < 0.001$; for song 2, $t(15) = 2.64$, $p = 0.02$; for song 3, $t(15) = 2.82$, $p = 0.01$]. Repeat measures one-way ANOVA did not reveal differences in the increase of

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

duration among imagined singing of the three songs [$F(2) = 2.17$, $p = 0.126$], suggesting the slow-down of speed in imagery plus motor responses of button press were consistent during imagery of all songs. The distribution of RT [Fig. 1c] followed a normal distribution [chi-square goodness-of-fit test, $\chi^2(4) = 2.30$, $p = 0.68$] and revealed that about half of trials falling within two (plus and minus) standard deviations. Two groups of trials, on the basis of the variation from the mean of RT, were formed for further analysis of MEG responses of imagined singing: for imagery of each song, 12 trials that were close to the mean were included in the *center-group*, whereas the other 12 trials that were farther away from the RT mean were included in the *dispersed-group* [Fig. 1d].

We first examed the MEG in the temporal domain. The responses time-locked to the onset of auditory stimuli (event-related responses) revealed a clear peak and topography of M100 auditory response (Fig. 1g). Moreover, in the imagined singing condition, a topographic pattern that was similar to the M100 auditory response was observed around the similar latency (Fig. 1h), even though no external auditory stimulus was presented in the imagined conditions. This similar event-related auditory responses in the listening and imagery conditions were consistent with our previous findings (Tian & Poeppel, 2010) and suggested that auditory cortices were activated during imagery conditions. Moreover, no repetitive patterns were observed in the time course of listening or imagery (Fig. 1g&h), suggesting that the tracking of the acoustic stream or the rate of imagery was not by the response magnitude.

We further investigated the neural tracking in the spectral domain. The MEG

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

responses in the listening conditions showed neural tracking of rhythm in the songs. The phase-coherence analysis revealed that the significant differences ($p_{corr(FDR)} < 0.05$) between the ITC of the *within-group* and *between-group* were localized mostly in the primary auditory cortex (HG) and its adjacent areas including left anterior superior temporal gyrus (laSTG) and sulcus (laSTS), and bilateral posterior insula (pINS) [Fig. 1i]. These results suggested that auditory systems can reliably follow the rhythm in the acoustic signals. These results demonstrated the validity and accuracy of source localization based on phase coherence.

For the imagery condition, comparison between the ITC in the *center-group* and *dispersed-group* revealed three significant clusters ($p_{corr(FDR)}<0.01$) in the frontal, parietal and temporal regions [Fig. 1j]. Specifically, in the frontal region, more consistent phase coherence was observed in the left inferior frontal gyrus (lIFG), insular cortex, and left middle frontal gyrus (lMFG) and sulcus (lMFS), as well as the right premotor cortex (rPreM). In the parietal region, the differences were in the bilateral intra-parietal sulcus (IPS) and left temporal-parietal junction (lTPJ). In the temporal region, more consistent phase coherence was located in the bilateral HG, laSTG and right middle and posterior superior temporal sulcus (rmSTS, rpSTS). These activation patterns during imagined singing were consistent with the proposed core computational regions in motor-to-sensory transformation [Fig. 1f], and confirmed the motor-based prediction pathway (Tian & Poeppel, 2010, 2012; Tian et al., 2016).

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Discussion

We investigated the function and dynamics of neural network that mediated the inner speech in verbal thinking. With an imagery singing paradigm, we found that frontal-parietal-temporal regions in the proposed motor-to-sensory network collaboratively synchronized at the rate of inner speech. These results suggest that neural oscillations can entrain the rhythm of mental activity. The synchronized neural activity in the motor-to-sensory transformation network mediates the inner speech in verbal thinking.

Three advances are in this study. First, we used sentences in the lyrics of famous songs as experimental stimuli. Our experimental setting was naturalistic, and the results were more generalizable to everyday situations. Second, we implemented a frequency-tagging procedure. Participants were asked to imagined singing according to the given rhythm of songs. We used sophisticated phase-coherence analysis to probe the dynamics of mental activity. In this way, the neural dynamics during inner speech in verbal thinking was obtained. Third, the MEG recording was at the system level, so that the source localization can provide a whole-brain analysis, similar as studies using neuroimaging methods (Hurlburt, Alderson-Day, Kühn, & Fernyhough, 2016; McGuire et al., 1995; McGuire et al., 1996). Moreover, combined with the analyses on neural oscillations, both functional and anatomical aspects of motor-to-sensory transformation network were investigated.

Using the functional constraints of phase-coherence in neural oscillations, the

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

observed frontal-parietal-temporal network during imagined singing was consistent with the proposed motor-to-sensory transformation network. The observed IFG, premotor and insular cortices in the frontal region in this study were involved in articulatory preparation during overt (Bohland, Bullock, & Guenther, 2010; Bohland & Guenther, 2006; Buiatti et al., 2009) and covert speech (McCarthy, Blamire, Rothman, Gruetter, & Shulman, 1993; Papathanassiou et al., 2000). The responses in these frontal cortices were also consistent with findings in speech imagery, suggesting the function of motor simulation (Tian et al., 2016). In the parietal region, the observation of TPJ in this study, an area closed to SMG, PO and Angular Gyrus (AG), has been suggested for sensorimotor integration and goal-directed prediction-based speech feedback control (Alexandrou, Saarinen, Mäkelä, Kujala, & Salmelin, 2017; Behroozmand et al., 2018; Cogan et al., 2014; Rong, Isenberg, Sun, & Hickok, 2018). Similar parietal areas of TPJ, SMG, PO, and adjacent IPS were also observed during speech imagery, suggesting possible functions for estimating somatosensory consequences of actions (Tian & Poeppel, 2010; Tian et al., 2016).

The observations in the temporal region further support the motor-to-sensory transformation during inner speech in verbal thinking. Comparing the phase-coherence results in imagined singing with those in listening conditions, the observations were overlapped in the temporal regions of primary and secondary auditory cortices. The STG was commonly observed during musical imagery (Zatorre & Halpern, 2005). The activation of auditory imagery can extend to the HG (Kraemer et al., 2005). The observation of HG during inner speech in this study was consistent

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

with the hypothesis that high task demand drove auditory estimation down to primary sensory area (Tian et al., 2018). The rSTS was only observed in imagined singing but not in listening conditions, which was consistent with previous findings (Tian et al., 2016), suggesting a possible specific functional role of STS in auditory imagery. The additional frontal (IFG, INS) and parietal (TPJ, IPS) activations in imagined singing suggest that auditory representation, similar as the representation established in perception, can be constructed via the motor-to-sensory transformation pathway (Tian & Poeppel, 2010, 2012; Tian et al., 2016).

This study provides hints about the possible functions of neural oscillations on perception. Many studies demonstrated that neural oscillations could entrain speech signals (Ding & Simon, 2012; Luo & Poeppel, 2007). However, it is still in debate whether the entrainment is driven by the stimuli features or is modulated by a top-down factor on intrinsic oscillations (Ding & Simon, 2014). Previous frequency-tagging experiments used external stimuli and investigated how neural oscillations track physical features. The perceptual constructs are derived from external stimuli features. It is hard, if not impossible, to separate perception from stimuli, and hence cannot solve the debate. Our results of imagined singing without external stimulation suggest that the phase of neural oscillations can be aligned to internal thought. That is, the phase of neural oscillations can be modulated by internally constructed representation during rhythmic mental imagery. These results support the view that a top-down factor modulates intrinsic neural oscillations.

These results may have impacts on practical and clinical domains. The motor-to-

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

sensory transformation network for imagined speech may implicate novel strategies for building brain-computer interface (BCI). Previous direct BCI mostly focuses on the motor system (Pfurtscheller & Neuper, 2001). Our findings of synchronized neural activity across motor and sensory domains during mental imagery suggest possible updates of decoding algorithms from a system-level and multi-modal perspective, which is demonstrated in a recent advance (Anumanchipalli, Chartier, & Chang, 2019). Moreover, these results may offer insights into the functional and anatomical foundations of auditory hallucination. We have hypothesized that from a cognitive perspective, auditory hallucination may be caused by incorrect source monitoring of internally self-induced auditory representation (Tian & Poeppel, 2012). These results of synchronized neural activity in the frontal-parietal-temporal network suggest the possible neural pathways for internal generation of auditory representation. These results are consistent with the neural modulation treatment for auditory hallucination that targets the electric stimulation at the motor-to-sensory transformation network (Yang et al., 2019).

Using a frequency-tagging imagined singing paradigm, we observed that thought modulated the phase of neural oscillations at an internal rate of thinking. The synchronized activity spanned across dedicated frontal-parietal-temporal regions that indicated the motor-to-sensory transformation network. The coherent activation in the motor-to-sensory transformation network mediates the internal construction of perceptual representation and forms the neural computation foundation for inner speech during verbal thinking.

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Acknowledgment

We thank Ling Liu for her help on the source localization analysis. This study was supported by the National Natural Science Foundation of China (NSFC) 31871131, the Major Program of Science and Technology Commission of Shanghai Municipality (STCSM) 17JC1404104, and the Program of Introducing Talents of Discipline to Universities, Base B16018 to Xing Tian, and NIH 5R01DC005660 to David Poeppel.

References

Adachi, Y., Shimogawara, M., Higuchi, M., Haruta, Y., & Ochiai, M. (2001). Reduction of non-periodic environmental magnetic noise in MEG measurement by continuously adjusted least squares method. *IEEE Transactions on Applied Superconductivity, 11*(1), 669-672.

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin, 141*(5), 931.

Alexandrou, A. M., Saarinen, T., Mäkelä, S., Kujala, J., & Salmelin, R. (2017). The right hemisphere is highlighted in connected natural speech production and perception. *NeuroImage, 152*, 628-638.

Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature, 568*(7753), 493.

Behroozmand, R., Phillip, L., Johari, K., Bonilha, L., Rorden, C., Hickok, G., & Fridriksson, J. (2018). Sensorimotor impairment of speech auditory feedback processing in aphasia. *NeuroImage, 165*, 102-111.

Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *Journal of cognitive neuroscience, 22*(7), 1504-1529.

Bohland, J. W., & Guenther, F. H. (2006). An fMRI investigation of syllable sequence production. *NeuroImage, 32*(2), 821-841.

Buiatti, M., Peña, M., & Dehaene-Lambertz, G. (2009). Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *NeuroImage, 44*(2), 509-519.

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.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage, 9*(2), 179-194.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron, 26*(1), 55-67.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience, 19*(1), 158.

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences, 109*(29), 11854-11859.

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in human neuroscience, 8*, 311.

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. *NeuroImage, 9*(2), 195-207.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological review, 102*(3), 594.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in cognitive sciences, 6*(6), 242-247.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature reviews neuroscience, 13*(2), 135.

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in human neuroscience, 5*, 82.

Hurlburt, R. T., Alderson-Day, B., Kühn, S., & Fernyhough, C. (2016). Exploring the ecological validity of thinking on demand: neural correlates of elicited vs. spontaneously occurring inner speech. *PloS one, 11*(2), e0147932.

Jack, B. N., Le Pelley, M. E., Han, N., Harris, A. W., Spencer, K. M., & Whitford, T. J. (2019). Inner speech is accompanied by a temporally-precise and content-specific corollary discharge. *NeuroImage, 198*, 170-180.

Jeannerod, M. (1995). Mental imagery in the motor context. *Neuropsychologia, 33*(11), 1419-1432.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current opinion in neurobiology, 9*(6), 718-727.

Kosslyn, S. M., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J., Ganis, G., . . . Alpert, N. (1999). The role of area 17 in visual imagery: convergent evidence from PET and rTMS. *Science, 284*(5411), 167-170.

Kraemer, D. J., Macrae, C. N., Green, A. E., & Kelley, W. M. (2005). Musical imagery: sound of silence activates auditory cortex. *Nature, 434*(7030), 158.

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), 11910.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron, 54*(6), 1001-1010.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods, 164*(1), 177-190.

McCarthy, G., Blamire, A. M., Rothman, D. L., Gruetter, R., & Shulman, R. G. (1993). Echo-planar magnetic resonance imaging studies of frontal cortex activation during word generation in humans. *Proceedings of the National Academy of Sciences, 90*(11), 4952-4956.

McGuire, P., David, A., Murray, R., Frackowiak, R., Frith, C., Wright, I., & Silbersweig, D. (1995). Abnormal monitoring of inner speech: a physiological basis for auditory hallucinations. *The Lancet, 346*(8975), 596-600.

McGuire, P., Silbersweig, D., Murray, R., David, A., Frackowiak, R., & Frith, C. (1996). Functional anatomy of inner speech and auditory verbal imagery. *Psychological medicine, 26*(1), 29-38.

Nozaradan, S., Peretz, I., Missal, M., & Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *Journal of Neuroscience, 31*(28), 10234-10240.

Nozaradan, S., Peretz, I., & Mouraux, A. (2012). Selective neuronal entrainment to the beat and meter

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

embedded in a musical rhythm. *Journal of Neuroscience, 32*(49), 17572-17581.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience, 2011*, 1.

Papathanassiou, D., Etard, O., Mellet, E., Zago, L., Mazoyer, B., & Tzourio-Mazoyer, N. (2000). A common language network for comprehension and production: a contribution to the definition of language epicenters with PET. *NeuroImage, 11*(4), 347-357.

Pfurtscheller, G., & Neuper, C. (2001). Motor imagery and direct brain-computer communication. *Proceedings of the IEEE, 89*(7), 1123-1134.

Porro, C. A., Francescato, M. P., Cettolo, V., Diamond, M. E., Baraldi, P., Zuiani, C., . . . Di Prampero, P. E. (1996). Primary motor and sensory cortex activation during motor performance and motor imagery: a functional magnetic resonance imaging study. *Journal of Neuroscience, 16*(23), 7688-7698.

Rong, F., Isenberg, A. L., Sun, E., & Hickok, G. (2018). The neuroanatomy of speech sequencing at the syllable level. *PloS one, 13*(10), e0196381.

Schubotz, R. I. (2007). Prediction of external events with our motor system: towards a new framework. *Trends in cognitive sciences, 11*(5), 211-218.

Sokolov, A. (2012). *Inner speech and thought*: Springer Science & Business Media.

Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D., & Leahy, R. M. (2011). Brainstorm: a user-friendly application for MEG/EEG analysis. *Computational intelligence and neuroscience, 2011*, 8.

Tian, X., Ding, N., Teng, X., Bai, F., & Poeppel, D. (2018). Imagined speech influences perceived loudness of sound. *Nature Human Behaviour, 2*(3), 225.

Tian, X., & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology, 1*, 166.

Tian, X., & Poeppel, D. (2012). Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Frontiers in human neuroscience, 6*, 314.

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.

Tian, X., & Poeppel, D. (2014). Dynamics of self-monitoring and error detection in speech production: evidence from mental imagery and MEG. *Journal of cognitive neuroscience, 27*(2), 352-364.

Tian, X., Zarate, J. M., & Poeppel, D. (2016). Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex, 77*, 1-12.

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences, 97*(20), 11125-11129.

Whitford, T. J., Jack, B. N., Pearson, D., Griffiths, O., Luque, D., Harris, A. W., . . . Le Pelley, M. E. (2017). Neurophysiological evidence of efference copies to inner speech. *Elife, 6*, e28197.

Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature neuroscience, 3*(11s), 1212.

Yang, F., Fang, X., Tang, W., Hui, L., Chen, Y., Zhang, C., & Tian, X. (2019). Effects and potential mechanisms of transcranial direct current stimulation (tDCS) on auditory hallucinations: A meta-analysis. *Psychiatry research*.

Zatorre, R. J., & Halpern, A. R. (2005). Mental concerts: musical imagery and auditory cortex. *Neuron, 47*(1), 9-12.

Zatorre, R. J., Halpern, A. R., Perry, D. W., Meyer, E., & Evans, A. C. (1996). Hearing in the mind's ear: a

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

PET investigation of musical imagery and perception. *Journal of cognitive neuroscience, 8*(1), 29-46.

Zhen, A., Van Hedger, S., Heald, S., Goldin-Meadow, S., & Tian, X. (2019). Manual directional gestures facilitate cross-modal perceptual learning. *Cognition, 187*, 178-187.

Figure captions

Figure 1. Neural oscillations entrain the rate of imagined singing. A) Experimental paradigm. According to the color of visual fixation, participants were asked to listen to the first sentence of three famous songs, followed by imagined singing the song they just heard. The Alphabet Song was used for illustration. Participants pressed a button to indicate the finish of imagery. B) Reaction time (RT) of imagined singing for three songs. The red dashed lines indicated the duration of the three songs. The duration of imagined singing was longer than the preceding auditory stimuli. C) Distribution of imagined singing RT. The z-scores of RT followed a normal distribution, with about half trials in the range of two standard deviations. D) Grouping of imagined singing trials. Twenty-four trials of each song were sorted ascendingly based on the z-scores and were separated into two groups. Twelve trials that were close to the mean RT were selected in the *center-group*, whereas the other twelve trials that were further away from the mean RT were included in the *dispersed-group*. E) Hypothesis about neural oscillation phase alignment across trials of imagined singing. Schematic display of two trials in each group. The short bar indicated the beginning and end of a trial. The wave lines represented neural oscillations. The trials in the *dispersed-group* had different durations so that temporal variance was large. The phase of neural oscillation that corresponded to the construction of syllabic representation during imagined singing did not align across trials. Whereas in the *center-group*, the temporal variance was small across trials, so that the phase of neural oscillation was more coherent across trials. F) Hypothesis about phase coherence in the motor-to-sensory transformation network during imagined singing. The motor-to-sensory transformation network was assumed to be a frontal-parietal-temporal network, including the inferior frontal gyrus (IFG), insular cortex (INS) and premotor/supplementary motor area (SMA) in the frontal lobe for simulating articulation; somatosensory areas (SI, SII), supramarginal gyrus (SMG) and its adjacent parietal operculum (OP), angular gyrus (AG) and temporal-parietal junction (TPJ) in the parietal lobe for estimating somatosensory consequence; as well as the superior temporal gyrus and sulcus (STG & STS) with a possibility of extension to the Heschl's gyrus (HG) in the temporal lobe for estimating auditory consequence. The more consistent phase coherence at the delta band (1-3 Hz) – the rate of imagined singing of the three songs – were predicted to be observed in the motor-to-sensory network. G) Waveforms and topographies in the listening condition (Alphabet Song). The vertical dotted line at time 0 indicated the onset of the auditory stimuli. Each black line represented waveform responses from a sensor and the red bold line represented the root-mean-square (RMS) waveform across all sensors. Topographies were plotted every 333ms from -1000ms to 4000ms. A clear auditory onset event-related response (M100, the single topography in the upper row) was observed. H) Waveforms and topographies in the imagery condition. Similar depicting form as in G). The vertical dotted line at time 0 indicated the onset of imagined singing. No repetitive patterns in topographies across the period. A similar event-related response in the range of M100 latency as in the listening condition was observed (the single topography in the upper row). I) Phase-

bioRxiv preprint first posted online Sep. 16, 2019; doi: http://dx.doi.org/10.1101/771196. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

coherence results in listening conditions. Neural entrainment was observed in the HG and its adjacent aSTG and pINS, as well as premotor areas. J) Phase coherence results in the imagery conditions. More consistent neural entrainment at the delta band was observed in the proposed motor-to-sensory network, including frontal areas (IFG, aINS and premotor), parietal areas (intraparietal sulcus (IPS) and TPJ), and temporal areas (HG, aSTG, and m&pSTS).

Figures