Progressive neural engagement within the IFG-pMTG circuit as gesture and speech entropy and MI advances

Wanying Zhao[1*]; Zhouyi Li[1,4]; Xiang Li[1,2]; Yi Du[1,2,3*]

[1] State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China 100101

[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing, China 100049

[3] Chinese Institute for Brain Research, Beijing, China 102206

[4] School of Psychology, Central China Normal University, Wuhan, China 430079

* Corresponding author:

Dr. Yi Du, Email: duyi@psych.ac.cn

16 Lincui Road, Chaoyang district, Beijing, China 100101

Dr. Wanying Zhao, Email: zhaowy@psych.ac.cn

16 Lincui Road, Chaoyang District, Beijing, China 100101

Abstract

Semantic representation emerges from distributed multisensory modalities, yet a comprehensive understanding of the functional changing pattern within convergence zones or hubs integrating multisensory semantic information remains elusive. In this study, employing information-theoretic metrics, we quantified gesture and speech information, alongside their interaction, utilizing entropy and mutual information (MI). Neural activities were assessed via interruption effects induced by High-Definition transcranial direct current stimulation (HD-tDCS). Additionally, chronometric double-pulse transcranial magnetic stimulation (TMS) and high-temporal event-related potentials were utilized to decipher dynamic neural changes resulting from various information contributors. Results showed gradual inhibition of both inferior frontal gyrus (IFG) and posterior middle temporal gyrus (pMTG) as degree of gesture-speech integration, indexed by MI, increased. Moreover, a time-sensitive and staged progression of neural engagement was observed, evidenced by distinct correlations between neural activity patterns and entropy measures of speech and gesture, as well as MI, across early sensory and lexico-semantic processing stages. These findings illuminate the gradual nature of neural activity during multisensory gesture-speech semantic processing, shaped by dynamic gesture constraints and speech encoding, thereby offering insights into the neural mechanisms underlying multisensory language processing.

Keywords: gesture-speech integration; pMTG-IFG circuit; information theory; multisensory; semantic; dual-stage modal

Introduction

Semantic representation, distinguished by its cohesive conceptual nature, emerges from distributed modality-specific regions. Consensus acknowledges the presence of 'convergence zones' within the temporal and inferior parietal areas [1], or the 'semantic hub' located in the anterior temporal lobe[2], pivotal for integrating, converging, or distilling multimodal inputs. Contemporary theories frame the semantic processing as a dynamic sequence of neural states[3], shaped by systems that are finely tuned to the statistical regularities inherent in sensory inputs[4]. These regularities enable the brain to evaluate, weight, and integrate multisensory information, optimizing the reliability of individual sensory signals[5]. However, sensory inputs available to the brain are often incomplete and uncertain, necessitating adaptive neural adjustments to resolve these ambiguities[6]. In this context, neuronal activity is thought to be linked to the probability density of sensory information, with higher levels of uncertainty resulting in the engagement of a broader population of neurons, thereby reflecting the brain's adaptive capacity to handle diverse possible interpretations[7,8]. Although the role of 'convergence zones' and 'semantic hubs' in integrating multimodal inputs is well established, the precise functional patterns of neural activity in response to the distribution of unified multisensory information—along with the influence of unisensory signals—remain poorly understood.

To this end, we developed an analytic approach to directly probe the cortical engagement during multisensory gesture-speech semantic integration. Even though gestures convey information in a global-synthetic way, while speech conveys information in a linear segmented way, there exists a bidirectional semantic influence between the two modalities[9,10]. Gesture is regarded as 'part of language'[11] or functional equivalents of lexical units that alternate and integrate with speech into a 'single unification space' to convey a coherent meaning[12-14]. Empirical studies have investigated the semantic integration between gesture and speech by manipulating their semantic relationship[15-18] and revealed a mutual interaction between them[19-21] as reflected by the N400 latency and amplitude[14] as well as common neural underpinnings in the left inferior frontal gyrus (IFG) and posterior middle temporal gyrus (pMTG)[15,22,23].

Building on these insights, the present study quantified the amount of information from both sources and their interaction adopting the information-theoretic complexity metrics of *entropy* and *mutual information* (MI). Unisensory Entropy measures the disorder or randomness of information and serves as an index of the uncertainty in modality-specific representations of gesture or speech activated by an event[24]. MI assesses share information between modalities[25], indicating multisensory convergence and acting as an index of gesture-speech integration.

To investigate the neural mechanisms underlying gesture-speech integration, we conducted three experiments to assess how neural activity correlates with distributed multisensory integration, quantified using information-theoretic measures of MI. Additionally, we examined the contributions of unisensory signals in this process, quantified through unisensory entropy. **Experiment 1** employed high-definition transcranial direct current stimulation (HD-tDCS) to administer Anodal, Cathodal and Sham stimulation to either the IFG or the pMTG. HD-tDCS induces membrane depolarization with anodal stimulation and membrane hyperpolarization with cathodal stimulation[26], thereby increasing or decreasing cortical excitability in the targeted brain area, respectively. This experiment aimed to determine whether the overall facilitation (Anodal-tDCS minus Sham-tDCS) and/or inhibitory (Cathodal-tDCS minus Sham-tDCS) of these integration hubs is modulated by the degree of gesture-speech integration, as measure by MI.

Given the differential involvement of the IFG and pMTG in gesture-speech integration, shaped by top-down gesture predictions and bottom-up speech processing [23], **Experiment 2** was designed to further assess whether the activity of these regions was associated with relevant informational matrices. Specifically, we applied inhibitory chronometric double-pulse transcranial magnetic stimulation (TMS) to specific temporal windows associated with integration processes in these regions[23], assessing whether the inhibitory effects of TMS were correlated with unisensory entropy or the multisensory convergence index (MI).

Experiment 3 complemented these investigations by focusing on the temporal dynamics of neural responses during semantic processing, leveraging high-temporal event-related potentials (ERPs). This experiment investigated how distinct information contributors modulated specific ERP components associated with semantic processing. These

components included the early sensory effects as P1 and N1–P2[27,28], the N400 semantic conflict effect[14,28,29], and the late positive component (LPC) reconstruction effect[30,31]. By integrating these ERP findings with results from Experiments 1 and 2, Experiment 3 aimed to provide a more comprehensive understanding of how gesture-speech integration is modulated by neural dynamics.

Material and methods

Participants

Ninety-eight young Chinese participants signed written informed consent forms and took part in the present study (Experiment 1: 29 females, 23 males, age = 20 ± 3.40 years; Experiment 2: 11 females, 13 males, age = 23 ± 4.88 years; Experiment 3: 12 females, 10 males, age = 21 ± 3.53 years). All of the participants were right-handed (Experiment 1: laterality quotient (LQ)[32] = 88.71 ± 13.14; Experiment 2: LQ = 89.02 ± 13.25; Experiment 3: LQ = 88.49 ± 12.65), had normal or corrected-to-normal vision and were paid ¥100 per hour for their participation. All experiments were approved by the Ethics Committee of the Institute of Psychology, Chinese Academy of Sciences.

Stimuli

Twenty gestures (**Appendix Table 1**) with 20 semantically congruent speech signals taken from previous study[23] were used. The stimuli set were recorded from two native Chinese speakers (1 male, 1 female). To validate the stimuli, 30 participants were recruited to replicate the multisensory index of semantic congruency effect, hypothesizing that reaction times for semantically incongruent gesture-speech pairs would be significantly longer than those for congruent pairs. The results confirmed this hypothesis, with a significantly ($t(29) = 7.16$, $p < 0.001$) larger reaction time when participants were asked to judge the gender of the speaker if gesture contained incongruent semantic information with speech (a ‘cut’ gesture paired with speech word ‘喷 pen1 (spray)’: mean = 554.51 ms, SE = 11.65) relative to when they were semantically congruent (a ‘cut’ gesture paired with ‘剪 jian3 (cut)’ word: mean = 533.90 ms, SE = 12.02)[23].

Additionally, two separate pre-tests with 30 subjects in each (pre-test 1: 16 females, 14

males, age = 24 ± 4.37 years; pre-test 2: 15 females, 15 males, age = 22 ± 3.26 years) were conducted to determine the comprehensive values of gesture and speech. Participants were presented with segments of increasing duration, beginning at 40 ms, and were prompted to provide a single verb to describe either the isolated gesture they observed (pre-test 1) or the isolated speech they heard (pre-test 2). For each gesture or speech, the action verb consistently provided by participants across four to six consecutive repetitions—with the number of repetitions varied to mitigate learning effects—was considered the comprehensive response for the gesture or speech. The initial instance duration was marked as the discrimination point (DP) for gesture (mean = 183.78 ± 84.82ms) or the identification point (IP) for speech (mean = 176.40 ± 66.21ms) (**Figure 1A top**).

To quantify information content, comprehensive responses for each item were converted into Shannon's entropy (H) as a measure of information richness (**Figure 1A bottom**). With no significant gender differences observed in both gesture ($t(20) = 0.21$, $p = 0.84$) and speech ($t(20) = 0.52$, $p = 0.61$), responses were aggregated across genders, resulting in 60 answers per item (**Appendix Table 2**). Here, p(xi) and p(yi) represent the distribution of 60 answers for a given gesture (**Appendix Table 2B**) and speech (**Appendix Table 2A**), respectively. High entropy indicates diverse answers, reflecting broad representation, while low entropy suggests focused lexical recognition for a specific item (**Figure 2B**). MI was used to measure the overlap between gesture and speech information, calculated by subtracting the entropy of the combined gesture-speech dataset (Entropy(gesture + speech)) from the sum of their individual entropies (Entropy(gesture) + Entropy(speech)) (see **Appendix Table 2C**). For specific gesture-speech combinations, equivalence between the combined entropy and the sum of individual entropies (gesture or speech) indicates absence of overlap in response sets. Conversely, significant overlap, denoted by a considerable number of shared responses between gesture and speech datasets, leads to a noticeable discrepancy between combined entropy and the sum of gesture and speech entropies. Elevated MI values thus signify substantial overlap, indicative of a robust mutual interaction between gesture and speech.

Additionally, the number of responses provided for each gesture and speech, as well as the total number of combined responses, were also recorded. The quantitative data for each

stimulus, including gesture entropy, speech entropy, joint entropy, MI, and the respective counts, are presented in **Appendix Table 3.**

To determine whether entropy or MI values corresponds to distinct neural changes, the current study first aggregated neural responses (including inhibition effects of tDCS and TMS or ERP amplitudes) that shared identical entropy or MI values, prior to conducting correlational analyses.

Experimental procedure

Given that gestures induce a semantic priming effect on concurrent speech[33], this study utilized a semantic priming paradigm in which speech onset was aligned with the DP of each gesture[23,33], the point at which the gesture transitions into a lexical form[34]. The gesture itself began at the stroke phase, a critical moment when the gesture conveys its primary semantic content[34].

An irrelevant factor of gender congruency (e.g., a man making a gesture combined with a female voice) was created[22,23,35]. This involved aligning the gender of the voice with the corresponding gender of the gesture in either a congruent (e.g., male voice paired with a male gesture) or incongruent (e.g., male voice paired with a female gesture) manner. This approach served as a direct control mechanism, facilitating the investigation of the automatic and implicit semantic interplay between gesture and speech[35]. In light of previous findings indicating a distinct TMS-disruption effect on the semantic congruency of gesture-speech interactions[23], both semantically congruent and incongruent pairs were included in Experiment 1 and Experiment 2. Experiment 3, conversely, exclusively utilized semantically congruent pairs to elucidate ERP metrics indicative of nuanced semantic progression.

Gesture–speech pairs were presented randomly using Presentation software (www.neurobs.com). Participants were asked to look at the screen but respond with both hands as quickly and accurately as possible merely to the gender of the voice they heard. The RT and the button being pressed were recorded. The experiment started with a fixation cross presented on the center of the screen, which lasted for 0.5-1.5 sec.

Experiment 1: HD-tDCS protocol and data analysis

Participants were divided into two groups, with each group undergoing HD-tDCS stimulation at different target sites (IFG or pMTG). Each participant completed three experimental sessions, spaced one week apart, during which 480 gesture-speech pairs were presented across various conditions. In each session, participants received one of three types of HD-tDCS stimulation: Anodal, Cathodal, or Sham. The order of stimulation site and type was counterbalanced using a Latin square design to control for potential order effects.

HD-tDCS protocol employed a constant current stimulator (The Starstim 8 system) delivering stimulation at an intensity of 2000mA. A 4 * 1 ring-based electrode montage was utilized, comprising a central electrode (stimulation) positioned directly over the target cortical area and four return electrodes encircling it to provide focused stimulation. Building on a meta-analysis of prior fMRI studies examining gesture-speech integration[22], we targeted Montreal Neurological Institute (MNI) coordinates for the left IFG at (-62, 16, 22) and the pMTG at (-50, -56, 10). In the stimulation protocol for HD-tDCS, the IFG was targeted using electrode F7 as the optimal cortical projection site[36], with four return electrodes placed at AF7, FC5, F9, and FT9. For the pMTG, TP7 was selected as the cortical projection site[36], with return electrodes positioned at C5, P5, T9, and P9. The stimulation parameters included a 20-minute duration with a 5-second fade-in and fade-out for both Anodal and Cathodal conditions. The Sham condition involved a 5-second fade-in followed by only 30 seconds of stimulation, then 19’20 minutes of no stimulation, and finally a 5-second fade-out (**Figure 1B**). Stimulation was controlled using NIC software, with participants blinded to the stimulation conditions.

All incorrect responses (702 out of the total number of 24960, 2.81% of trials) were excluded. To eliminate the influence of outliers, a 2SD trimmed mean for every participant in each session was also calculated. To examine the relationship between the degree of information and neural responses, we conducted Pearson correlation analyses using a sample of 20 sets. Neural responses were quantified based on the effects of HD-tDCS (active tDCS minus sham tDCS) on the semantic congruency effect, defined as the difference in reaction times between semantic incongruent and congruent conditions (Rt(incongruent) - Rt(congruent)). This effect served as an index of multisensory integration[35] within the left IFG and pMTG. The variation in information was assessed using three information-theoretic metrics. To account for potential confounds related to multiple candidate representations, we

conducted partial correlation analyses between the tDCS effects and gesture entropy, speech entropy, and MI, controlling for the number of responses provided for each gesture and speech, as well as the total number of combined responses. Given that HD-tDCS induces overall disruption at the targeted brain regions, we hypothesized that the neural activity within the left IFG and pMTG would be progressively affected by varying levels of multisensory convergence, as indexed by MI. Moreover, we hypothesized that the modulation of neural activity by MI would differ between the left IFG and pMTG, as reflected in the differential modulation of response numbers in the partial correlations, highlighting their distinct roles in semantic processing[37]. False discovery rate (FDR) correction was applied for multiple comparisons.

Experiment 2: TMS protocol and data analysis

Experiment 2 involved 800 gesture-speech pairs, presented across 15 blocks over three days, with one week between sessions. Stimulation was administered at three different sites (IFG, pMTG, or Vertex). Within the time windows (TWs) spanning the gesture-speech integration period, five TWs that exhibited selective disruption of integration were selected: TW1 (-120 to -80 ms relative to the speech identification point), TW2 (-80 to -40 ms), TW3 (-40 to 0 ms), TW6 (80 to 120 ms), and TW7 (120 to 160 ms)[23] (**Figure 1C**). The order of stimulation site and TW was counterbalanced using a Latin square design.

At an intensity of 50% of the maximum stimulator output, double-pulse TMS was delivered via a 70 mm figure-eight coil using a Magstim Rapid² stimulator (Magstim, UK). High-resolution (1 × 1 × 0.6 mm) T1-weighted MRI scans were obtained using a Siemens 3T Trio/Tim Scanner for image-guided TMS navigation. Frameless stereotaxic procedures (BrainSight 2; Rogue Research) allowed real-time stimulation monitoring. To ensure precision, individual anatomical images were manually registered by identifying the anterior and posterior commissures. Subject-specific target regions were defined using trajectory markers in the MNI coordinate system. Vertex was used as control.

All incorrect responses (922 out of the total number of 19200, 4.8% of trials) were excluded. We focused our analysis on Pearson correlations of the TMS interruption effects (active TMS minus vertex TMS) of the semantic congruency effect with the gesture entropy,

speech entropy or MI. To control for potential confounds, partial correlations were also performed between the TMS effects and gesture entropy, speech entropy, and MI, controlling for the number of responses for each gesture and speech, as well as the total number of combined responses. By doing this, we can determine how the time-sensitive contribution of the left IFG and pMTG to gesture–speech integration was affected by gesture and speech information distribution. FDR correction was applied for multiple comparisons.

Experiment 3: Electroencephalogram (EEG) recording and data analysis

Experiment 3, comprising a total of 1760 gesture-speech pairs, was completed in a single-day session. EEG were recorded from 48 Ag/AgCl electrodes mounted in a cap according to the 10-20 system[38], amplified with a PORTI-32/MREFA amplifier (TMS International B.V., Enschede, NL) and digitized online at 500 Hz (bandpass, 0.01-70 Hz). EEGLAB, a MATLAB toolbox, was used to analyze the EEG data[39]. Vertical and horizontal eye movements were measured with 4 electrodes placed above the left eyebrow, below the left orbital ridge and at bilateral external canthus. All electrodes were referenced online to the left mastoid. Electrode impedance was maintained below 5 KΩ. The average of the left and right mastoids was used for re-referencing. A high-pass filter with a cutoff of 0.05 Hz and a low-pass filter with a cutoff of 30 Hz were applied. Semi-automated artifact removal, including independent component analysis (ICA) for identifying components of eye blinks and muscle activity, was performed (**Figure 1D**). Participants with rejected trials exceeding 30% of their total were excluded from further analysis.

All incorrect responses were excluded (147 out of 1760, 8.35% of trials). To eliminate the influence of outliers, a 2 SD trimmed mean was calculated for every participant in each condition. Data were epoched from the onset of speech and lasted for 1000 ms. To ensure a clean baseline with no stimulus presented, a 200 ms pre-stimulus baseline correction was applied before gesture onset.

To consolidate the data, we conducted both a traditional region-of-interest (ROI) analysis, with ROIs defined based on a well-established work[40], and a cluster-based permutation approach, which utilizes data-driven permutations to enhance robustness and address multiple comparisons.

For the traditional ROI analysis, grand-average ERPs at electrode Cz were compared between the higher (≥50%) and lower (<50%) halves for gesture entropy (**Figure 5A1**), speech entropy (**Figure 5B1**), and MI (**Figure 5C1**). Consequently, four ERP components were determined: the P1 effect observed within the time window of 0-100 ms[27,28], the N1-P2 effect observed between 150-250ms[27,28], the N400 within the interval of 250-450ms[14,28,29], and the LPC spanning from 550-1000ms[30,31]. Additionally, seven regions-of-interest (ROIs) were defined in order to locate the modulation effect on each ERP component: left anterior (LA): F1, F3, F5, FC1, FC3, and FC5; left central (LC): C1, C3, C5, CP1, CP3, and CP5; left posterior (LP): P1, P3, P5, PO3, PO5, and O1; right anterior (RA): F2, F4, F6, FC2, FC4, and FC6; right central (RC): C2, C4, C6, CP2, CP4, and CP6; right posterior (RP): P2, P4, P6, PO4, PO6, and O2; and midline electrodes (ML): Fz, FCz, Cz, Pz, Oz, and CPz[40].

Subsequently, cluster-based permutation tests[41] in Fieldtrip was further used to determine the significant clusters of adjacent time points and electrodes of ERP amplitude between the higher and lower halves of gesture entropy, speech entropy and MI, respectively. The electrode-level type I error threshold was set to 0.025. Cluster-level statistic was estimated through 5000 Monte Carlo simulations, where the cluster-level statistic is the sum of T-values for each stimulus within a cluster. The cluster-level type I error threshold was set to 0.05. Clusters with a p-value less than the critical alpha-level are considered to be conditionally different.

Paired t-tests were conducted to compare the lower and upper halves of each information model for the averaged amplitude within each ROI or cluster across the four ERP time windows, separately. Pearson correlations were computed between each model value and the averaged ERP amplitudes in each ROI or cluster. Additionally, partial correlations were conducted, accounting for the number of responses for each respective metric. FDR correction was applied for multiple comparisons.

Results

Experiment 1: Modulation of left pMTG and IFG engagement by gradual changes in gesture-speech semantic information

In the IFG, one-way ANOVA examining the effects of three tDCS conditions (Anodal, Cathodal, or Sham) on semantic congruency (RT (semantic incongruent) – RT (semantic congruent)) demonstrated a significant main effect of stimulation condition ($F(2, 75) = 3.673$, $p = 0.030$, ηp2 = 0.089). Post hoc paired t-tests indicated a significantly reduced semantic congruency effect between the Cathodal condition and the Sham condition ($t(26) = -3.296$, $p = 0.003$, 95% CI = [-11.488, 4.896]) (**Figure 3A left**). Subsequent Pearson correlation analysis revealed that the reduced semantic congruency effect was progressively associated with the MI, evidenced by a significant correlation between the Cathodal-tDCS effect (Cathodal-tDCS minus Sham-tDCS) and MI ($r = -0.595$, $p = 0.007$, 95% CI = [-0.995, -0.195]) (**Figure 3B**). Additionally, a similar correlation was observed between the Cathodal-tDCS effect and the total response number ($r = -0.543$, $p = 0.016$, 95% CI = [-0.961, -0.125]).

However, partial correlation analysis, controlling for the total response number, revealed that the initially significant correlation between the Cathodal-tDCS effect and MI was no longer significant ($r = -0.303$, $p = 0.222$, 95% CI = [-0.770, 0.164]). This suggests that the observed relationship between Cathodal-tDCS and MI may be confounded by semantic control difficulty, as reflected by the total number of responses. Specifically, the reduced activity in the IFG under Cathodal-tDCS may be driven by variations in the difficulty of semantic control rather than a direct modulation of MI.

In the pMTG, a one-way ANOVA assessing the effects of three tDCS conditions on semantic congruency also revealed a significant main effect of stimulation condition ($F(2, 75) = 3.250$, $p = 0.044$, ηp2 = 0.080). Subsequent paired t-tests identified a significantly reduced semantic congruency effect between the Cathodal condition and the Sham condition ($t(25) = -2.740$, $p = 0.011$, 95% CI = [-11.915, 6.435]) (**Figure 3A right**). Moreover, a significant correlation was observed between the Cathodal-tDCS effect and MI ($r = -0.457$, $p = 0.049$, 95% CI = [-0.900, -0.014]) (**Figure 3B**).

Importantly, the reduced activity in the pMTG under Cathodal-tDCS was not influenced by the total response number, as indicated by the non-significant correlation ($r = -0.253$, $p = 0.295$, 95% CI = [-0.735, 0.229]). This finding was further corroborated by the unchanged significance in the partial correlation between Cathodal-tDCS and MI, when controlling for the total response number ($r = -0.472$, $p = 0.048$, 95% CI = [-0.903, -0.041]).

RTs of congruent and incongruent trials of IFG and pMTG in each of the stimulation conditions were shown in **Appendix Table 4A.**

Experiment 2: Time-sensitive modulation of left pMTG and IFG engagements by gradual changes in gesture-speech semantic information

A 2 (TMS effect: active - Vertex) × 5 (TW) ANOVA on semantic congruency revealed a significant interaction between TMS effect and TW ($F(3.589, 82.538) = 3.273$, $p = 0.019$, ηp2 = 0.125). Further t-tests identified a significant TMS effect over the pMTG in TW1 ($t(23) = -3.068$, $p = 0.005$, 95% CI = [-6.838, 0.702]), TW2 ($t(23) = -2.923$, $p = 0.008$, 95% CI = [-6.490, 0.644]), and TW7 ($t(23) = -2.005$, $p = 0.047$, 95% CI = [-5.628, 1.618]). In contrast, a significant TMS effect over the IFG was found in TW3 ($t(23) = -2.335$, $p = 0.029$, 95% CI = [-5.928, 1.258]), and TW6 ($t(23) = -4.839$, $p < 0.001$, 95% CI = [-7.617, -2.061]) (**Figure 4A**). Raw RTs of congruent and incongruent trials were shown in **Appendix Table 4B.**

Additionally, a significant negative correlation was found between the TMS effect (a larger negative TMS effect signifies a greater disruption of the integration process) and speech entropy when the pMTG was inhibited in TW2 ($r = -0.792$, $p = 0.004$, 95% CI = [-1.252, -0.331]). Meanwhile, when the IFG activity was interrupted in TW6, a significant negative correlation was found between the TMS effect and gesture entropy ($r = -0.539$, $p = 0.014$, 95% CI = [-0.956, -0.122]), speech entropy ($r = -0.664$, $p = 0.026$, 95% CI = [-1.255, -0.073]), and MI ($r = -0.677$, $p = 0.001$, 95% CI = [-1.054, -0.300]) (**Figure 4B**).

Notably, inhibition of pMTG activity in TW2 was not influenced by the number of speech responses ($r = -0.539$, $p = 0.087$, 95% CI = [-1.145, 0.067]). However, the number of speech responses did affect the modulation of speech entropy on the pMTG inhibition effect in TW2. This was evidenced by the non-significant partial correlation between pMTG inhibition and speech entropy when controlling for speech response number ($r = -0.218$, $p = 0.545$, 95% CI = [-0.563, 0.127]).

In contrast, the interrupted IFG activity in TW6 appeared to be consistently influenced by the confound of semantic control difficulty. This was reflected in the significant correlation with both gesture response number ($r = -0.480$, $p = 0.032$, 95% CI = [-904, -0.056]), speech response number ($r = -0.729$, $p = 0.011$, 95% CI = [-1.221, -0.237]), and total response

number (r = -0.591, p = 0.008, 95% CI = [-0.993, -0.189]). Additionally, partial correlation analyses revealed non-significant relationship between interrupted IFG activity in TW6 and gesture entropy (r = -0.369, p = 0.120, 95% CI = [-0.810, -0.072]), speech entropy (r = -0.455, p = 0.187, 95% CI = [-1.072, 0.162]), and MI (r = -0.410, p = 0.091, 95% CI = [-0.856, -0.036]) when controlling for response numbers.

Experiment 3: Temporal modulation of P1, N1-P2, N400 and LPC components by gradual changes in gesture-speech semantic information

Topographical maps illustrating amplitude differences between the lower and higher halves of speech entropy demonstrate a central-posterior P1 amplitude (0-100 ms, **Figure 5B**). Aligning with prior findings[27], the paired t-tests demonstrated a significantly larger P1 amplitude within the ML ROI (t(22) = 2.510, p = 0.020, 95% confidence interval (CI) = [1.66, 3.36]) when contrasting stimuli with higher 50% speech entropy against those with lower 50% speech entropy (**Figure 5D1 left**). Subsequent correlation analyses unveiled a significant increase in the P1 amplitude with the rise in speech entropy within the ML ROI (r = 0.609, p = 0.047, 95% CI = [0.039, 1.179], **Figure 5D1 right**). Furthermore, a cluster of neighboring time-electrode samples exhibited a significant contrast between the lower 50% and higher 50% of speech entropy, revealing a P1 effect spanning 16 to 78 ms at specific electrodes (FC2, FCz, C1, C2, Cz, and CPz, **Figure 5D2 middle**) (t(22) = 2.754, p = 0.004, 95% confidence interval (CI) = [1.65, 3.86], **Figure 5D2 left**), with a significant correlation with speech entropy (r = 0.636, p = 0.035, 95% CI = [0.081, 1.191], **Figure 5D2 right**).

Additionally, topographical maps comparing the lower 50% and higher 50% gesture entropy revealed a frontal N1-P2 amplitude (150-250 ms, **Figure 5A**). In accordance with previous findings on bilateral frontal N1-P2 amplitude[27], paired t-tests displayed a significantly larger amplitude for stimuli with lower 50% gesture entropy than with higher 50% entropy in both ROIs of LA (t(22) = 2.820, p = 0.011, 95% CI = [2.21, 3.43]) and RA (t(22) = 2.223, p = 0.038, 95% CI = [1.56, 2.89]) (**Figure 5E1 left**). Moreover, a negative correlation was found between N1-P2 amplitude and gesture entropy in both ROIs of LA (r = -0.465, p = 0.039, 95% CI = [-0.87, -0.06]) and RA (r = -0.465, p = 0.039, 95% CI = [-0.88, -0.05]) (**Figure 5E1 right**). Additionally, through a cluster-permutation test, the N1-P2 effect was identified between 184

to 202 ms at electrodes FC4, FC6, C2, C4, C6, and CP4 (**Figure 5E2 middle**) ($t(22) = 2.638$, $p = 0.015$, 95% CI = [1.79, 3.48], (**Figure 5E2 left**)), exhibiting a significant correlation with gesture entropy ($r = -0.485$, $p = 0.030$, 95% CI = [-0.91, -0.06], **Figure 5E2 right**).

Furthermore, in line with prior research[42], a left-frontal N400 amplitude (250-450 ms) was discerned from topographical maps of gesture entropy (**Figure 5A**). Specifically, stimuli with lower 50% values of gesture entropy elicited a larger N400 amplitude in the LA ROI compared to those with higher 50% values ($t(22) = 2.455$, $p = 0.023$, 95% CI = [1.95, 2.96], **Figure 5F1 left**). Concurrently, a negative correlation was noted between the N400 amplitude and gesture entropy ($r = -0.480$, $p = 0.032$, 95% CI = [-0.94, -0.03], **Figure 5F1 right**) within the LA ROI. The identified clusters showing the N400 effect for gesture entropy (282 – 318 ms at electrodes FC1, FCz, C1, and Cz, **Figure 5F2 middle**) ($t(22) = 2.828$, $p = 0.010$, 95% CI = [2.02, 3.64], **Figure 5F2 left**) also exhibited significant correlation between the N400 amplitude and gesture entropy ($r = -0.445$, $p = 0.049$, 95% CI = [-0.88, -0.01], **Figure 5F2 right**).

Similarly, a left-frontal N400 amplitude (250-450 ms)[42] was discerned from topographical maps for MI (**Figure 5C)**. A larger N400 amplitude in the LA ROI was observed for stimuli with lower 50% values of MI compared to those with higher 50% values ($t(22) = 3.00$, $p = 0.007$, 95% CI = [2.54, 3.46], **Figure 5G1 left**). This was accompanied by a significant negative correlation between N400 amplitude and MI ($r = -0.504$, $p = 0.028$, 95% CI = [-0.97, -0.04], **Figure 5G1 right**) within the LA ROI. The N400 effect for MI, observed in the 294–306 ms window at electrodes F1, F3, Fz, FC1, FC3, FCz, and C1 (**Figure 5G2 middle**) ($t(22) = 2.461$, $p = 0.023$, 95% CI = [1.62, 3.30], **Figure 5G2 left**), also showed a significant negative correlation with MI ($r = -0.569$, $p = 0.011$, 95% CI = [-0.98, -0.16], **Figure 5G2 right**).

Finally, consistent with previous findings[30], an anterior LPC effect (550-1000 ms) was observed in topographical maps comparing stimuli with lower and higher 50% speech entropy (**Figure 5B**). The reduced LPC amplitude was evident in the paired t-tests conducted in ROIs of LA ($t(22) = 2.614$, $p = 0.016$, 95% CI = [1.88, 3.35]); LC ($t(22) = 2.592$, $p = 0.017$, 95% CI = [1.83, 3.35]); RA ($t(22) = 2.520$, $p = 0.020$, 95% CI = [1.84, 3.24]); and ML ($t(22) = 2.267$, $p = 0.034$, 95% CI = [1.44, 3.10]) (**Figure 5H1 left**). Simultaneously, a marked negative

correlation with speech entropy was evidenced in ROIs of LA (r = -0.836, p = 0.001, 95% CI = [-1.26, -0.42]); LC (r = -0.762, p = 0.006, 95% CI = [-1.23, -0.30]); RA (r = -0.774, p = 0.005, 95% CI = [-1.23, -0.32]) and ML (r = -0.730, p = 0.011, 95% CI = [-1.22, -0.24]) (**Figure 5H1 right**). Additionally, a cluster with the LPC effect (644 - 688 ms at electrodes Cz, CPz, P1, and Pz, **Figure 5H2 middle**) (t(22) = 2.754, p = 0.012, 95% CI = [1.50, 4.01], **Figure 5H2 left**) displayed a significant correlation with speech entropy (r = -0.699, p = 0.017, 95% CI = [-1.24, -0.16], **Figure 5H2 right**).

To clarify potential confounds of semantic control difficulty, partial correlation analyses were conducted to examine the relationship between the elicited ERP components and the relevant information matrices, controlling for response numbers. Results consistently indicated modulation by response numbers in the relationship of ERP components with the information matrix, as evidenced by the non-significant partial correlations between the P1 amplitude (P1 component over ML: r = -0.574, p = 0.082, 95% CI = [-1.141, -0.007]) and the P1 cluster (r = -0.503, p = 0.138, 95% CI = [-1.102, 0.096]) with speech entropy; the N1-P2 amplitude (N1-P2 component over LA: r = -0.080, p = 0.746, 95% CI = [-0.554, 0.394]) and N1-P2 cluster (r = -0.179, p = 0.464, 95% CI = [-0.647, 0.289]) with gesture entropy; the N400 amplitude (N400 component over LA: r = 0.264, p = 0.247, 95% CI = [-0.195,0.723]) and N400 cluster (r = 0.394, p = 0.095, 95% CI = [-0.043, 0.831]) with gesture entropy; the N400 amplitude (N400 component over LA: r = -0.134, p = 0.595, 95% CI = [-0.620, 0.352]) and N400 cluster (r = -0.034, p = 0.894, 95% CI = [-0.524,0.456]) with MI; and the LPC amplitude (LPC component over LA: r = -0.428, p = 0.217, 95% CI = [-1.054, 0.198]) and LPC cluster (r = -0.202, p = 0.575, 95% CI = [-0.881, 0.477]) with speech entropy.

Discussion

Through mathematical quantification of gesture and speech information using entropy and mutual information (MI), we examined the functional pattern and dynamic neural structure underlying multisensory semantic integration. Our results, for the first time, revealed that the inhibition effect of cathodal-tDCS on the pMTG and IFG correlated with the degree of gesture-speech multisensory convergence, as indexed by MI (**Experiment 1**). Moreover, the gradual

neural engagement was found to be time-sensitive and staged, as evidenced by the selectively interrupted time windows (**Experiment 2**) and the distinct correlated ERP components (**Experiment 3**), which were modulated by different information contributors, including unisensory entropy or multisensory MI. These findings significantly expand our understanding of the cortical foundations of statistically regularized multisensory semantic information.

It is widely acknowledged that a single, amodal system mediates the interactions among perceptual representations of different modalities[5,43,44]. Moreover, observations have suggested that semantic dementia patients experience increasing overregularization of their conceptual knowledge due to the progressive deterioration of this amodal system[45]. Consistent with this, the present study provides robust evidence, through the application of HD-tDCS and TMS, that the integration hubs for gesture and speech—the pMTG and IFG—operate in an incremental manner. This is supported by the progressive inhibition effect observed in these brain areas as the entropy and mutual information of gesture and speech advances.

Moreover, by dividing the potential integration period into eight time windows (TW) relative to the speech identification point (IP) and administering inhibitory double-pulse TMS across each TW, the current study attributed the gradual TMS-selective regional inhibition to distinct information sources. In TW2 of gesture-speech integration, which precedes the speech identification point[23] and represents a pre-lexical stage, the suppression effect observed in the pMTG was correlated with speech entropy. Conversely, during TW6, which follows the speech identification point[23] and represents a post-lexical stage, the IFG interruption effect was influenced by both gesture entropy, speech entropy, and their MI. A dual-stage pMTG-IFG-pMTG neurocircuit loop during gesture-speech integration has been proposed previous[23]. As an extension, the present study unveils a staged accumulation of engagement within the neurocircuit linking the transmodal regions of pMTG and IFG, arising from distinct contributors of information.

Furthermore, we disentangled the sub-processes of integration with high-temporal ERPs, when representations of gesture and speech were variously presented. Early P1-N1 and P2

sensory effects linked to perception and attentional processes[28,46] was comprehended as a reflection of the early audiovisual gesture-speech integration in the sensory-perceptual processing chain[47]. Note that a semantic priming paradigm was adopted here to create a top-down prediction of gesture over speech. The observed positive correlation of the P1 effect with speech entropy and the negative correlation of the N1-P2 effect with gesture entropy suggest that the early interaction of gesture-speech information was modulated by both top-down gesture prediction and bottom-up speech processing. Additionally, the lexico-semantic effect of the N400 and the LPC were differentially mediated by top-down gesture prediction, bottom-up speech encoding and their interaction: the N400 was negatively correlated with both the gesture entropy and MI, but the LPC was negatively correlated only with the speech entropy.

The varying contributions of unisensory gesture-speech information and the convergence of multisensory inputs, as reflected in the correlation between distinct ERP components and TMS time windows (TMS TWs), are consistent with recent models suggesting that multisensory processing involves parallel detection of modality-specific information and hierarchical integration across multiple neural levels[4,48]. These processes are further characterized by coordination across multiple temporal scales[49]. Building on this, the present study offers additional evidence that the multi-level nature of gesture-speech processing is statistically structured, as measured by information matrix of unisensory entropy and multisensory convergence index of MI, the input of either source would activate a distributed representation, resulting in progressively functioning neural responses.

Given that control processes are intrinsically integrated with semantic processing[50], a distributed semantic representation enables dynamic modulation of access to and manipulation of meaningful information, thereby facilitating flexible control over the diverse possibilities inherent in a concept. Accordingly, an increased number of candidate responses amplifies the control demands necessary to resolve competing semantic representations. This effect was observed in the present study, where the association of the information matrix with the tDCS effect in IFG, the inhibition of pMTG activity in TW2, disruption of IFG activity in TW6, and modulation of four distinct ERP components collectively demonstrated that response quantity modulates neural activity. These results underscore the intricate interplay

between the difficulty of semantic representation and the control pressures that shape the resulting neural responses.

The IFG and pMTG, central components of the semantic control network, have been extensively implicated in previous research [50-52]. While the role of the IFG in managing both unisensory information and multisensory convergence remains consistent, as evidenced by the confounding difficulty results across Experiments 1 and 2, the current study highlights a more context-dependent function for the pMTG. Specifically, although the pMTG is well-established in the processing of distributed speech information, the multisensory convergence, indexed by MI, did not evoke the same control-related modulation in pMTG activity. These findings suggest that, while the pMTG is critical to semantic processing, its engagement in control processes is likely modulated by the specific nature of the sensory inputs involved.

Considering the close alignment of the ERP components with the TWs of TMS effect, it is reasonable to speculate the ERP components with the cortical involvements (**Figure 6**). Consequently, referencing the recurrent neurocircuit connecting the left IFG and pMTG for semantic unification[53], we extended the previously proposed two-stage gesture-speech integration circuit[23] into sequential steps. First, bottom-up speech processing mapping acoustic signal to its lexical representation was performed to the pMTG. The larger speech entropy was, the greater effort was made during the matching of the acoustic input with its stored lexical representation, thus leading to a larger involvement of the pMTG at pre-lexical stage (TW2) and a larger P1 effect (**Figure 6①**). Second, the gesture representation was activated in the pMTG and further exerted a top-down modulation over the phonological processing of speech [54]. The higher the certainty of gesture is, a larger modulation of gesture would be made upon speech, as indexed by a smaller gesture entropy with an enhanced N1-P2 amplitude (**Figure 6②**). Third, information was relayed from the pMTG to the IFG for sustained activation, during which a semantic constraint from gesture has been made on the semantic retrieval of speech. Greater TMS inhibitory effect over the IFG at post-lexical stage (TW6) accompanying with a reduced N400 amplitude were found with the increase of gesture entropy, when the representation of gesture was wildly distributed and the constrain over the following speech was weak **(Figure 6③**). Fourth, the activated speech representation was

compared with that of the gesture in the IFG. At this stage, the larger, overlapped neural populations activated by gesture and speech as indexed by a larger MI is, a greater TMS disruption effect of the IFG and a reduced N400 amplitude indexing easier integration and less semantic conflict were observed (**Figure 6④**). Last, the activated speech representation would disambiguate and reanalyze the semantic information and further unify into a coherent comprehension in the pMTG[12,37]. As speech entropy increases, indicating greater uncertainty in the information provided by speech, more cognitive effort is directed towards selecting the targeted semantic representation. This leads to enhanced involvement of the IFG and a corresponding reduction in LPC amplitude (**Figure 6⑤**).

Note that the sequential cortical involvement and ERP components discussed above are derived from a deliberate alignment of speech onset with gesture DP, creating an artificial priming effect with gesture semantically preceding speech. Caution is advised when generalizing these findings to the spontaneous gesture-speech relationships, although gestures naturally precede speech[34]. Furthermore, MI quantifies overlap in gesture-speech integration, primarily when gestures convey redundant meaning. Consequently, the conclusions drawn in this study are constrained to contexts in which gestures serve to reinforce the meaning of the speech. Future research should aim to explore the neural responses in cases where gestures convey supplementary, rather than redundant, semantic information.

Limitations exist. ERP components and cortical engagements were linked through intermediary variables of entropy and MI. Dissociations were observed between ERP components and cortical engagement. Importantly, there is no direct evidence of the brain structures underpinning the corresponding ERPs, necessitating clarification in future studies. Additionally, not all influenced TWs exhibited significant associations with entropy and MI. While HD-tDCS and TMS may impact functionally and anatomically connected brain regions[55,56], whether the absence of influence in certain TWs can be attributed to compensation by other connected brain areas, such as angular gyrus[57] or anterior temporal lobe[58], warrants further investigation. Therefore, caution is needed when interpreting the causal relationship between inhibition effects of brain stimulation and information-theoretic

metrics (entropy and MI). Finally, the current study incorporated a restricted set of entropy and MI measures. The generalizability of the findings should be assessed in future studies using a more extensive range of matrices.

In summary, utilizing information-theoretic complexity metrics such as entropy and mutual information (MI), our study demonstrates that multisensory semantic processing, involving gesture and speech, gives rise to dynamically evolving representations through the interplay between gesture-primed prediction and speech presentation. This process correlates with the progressive engagement of the pMTG-IFG-pMTG circuit and various ERP components. These findings significantly advancing our understanding of the neural mechanisms underlying multisensory semantic integration.

Acknowledgments

This research was supported by grants from the STI 2030—Major Projects 2021ZD0201500, the National Natural Science Foundation of China (31822024, 31800964), the Scientific Foundation of Institute of Psychology, Chinese Academy of Sciences (E2CX3625CX), and the Strategic Priority Research Program of Chinese Academy of Sciences (XDB32010300).

Author contributions

Conceptualization, W.Y.Z. and Y.D.; Investigation, W.Y.Z. and Z.Y.L.; Formal Analysis, W.Y.Z. and Z.Y.L.; Methodology, W.Y.Z. and Z.Y.L.; Validation, Z.Y.L. and X.L.; Visualization, W.Y.Z. and Z.Y.L. and X.L.; Funding Acquisition, W.Y.Z. and Y.D.; Supervision, Y.D.; Project administration, Y.D.; Writing – Original Draft, W.Y.Z.; Writing – Review & Editing, W.Y.Z., Z.Y.L., X.L., and Y.D.

Competing interests

The authors declare no competing interests.

References

1. Damasio, H., Grabowski, T.J., Tranel, D., Hichwa, R.D., and Damasio, A.R. (1996). A neural basis for lexical retrieval. Nature *380*, 499-505. DOI 10.1038/380499a0.
2. Patterson, K., Nestor, P.J., and Rogers, T.T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. Nature Reviews Neuroscience *8*, 976-987. 10.1038/nrn2277.
3. Brennan, J.R., Stabler, E.P., Van Wagenen, S.E., Luh, W.M., and Hale, J.T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. Brain and Language *157*, 81-94. 10.1016/j.bandl.2016.04.008.
4. Benetti, S., Ferrari, A., and Pavani, F. (2023). Multimodal processing in face-to-face interactions: A bridging link between psycholinguistics and sensory neuroscience. Front Hum Neurosci *17*, 1108354. 10.3389/fnhum.2023.1108354.
5. Noppeney, U. (2021). Perceptual Inference, Learning, and Attention in a Multisensory World. Annual Review of Neuroscience, Vol 44, 2021 *44*, 449-473. 10.1146/annurev-neuro-100120-085519.
6. Ma, W.J., and Jazayeri, M. (2014). Neural coding of uncertainty and probability. Annu Rev Neurosci *37*, 205-220. 10.1146/annurev-neuro-071013-014017.
7. Fischer, B.J., and Pena, J.L. (2011). Owl's behavior and neural representation predicted by Bayesian inference. Nat Neurosci *14*, 1061-1066. 10.1038/nn.2872.
8. Ganguli, D., and Simoncelli, E.P. (2014). Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. Neural Comput *26*, 2103-2134. 10.1162/NECO_a_00638.

9. Hostetter, A., and Mainela-Arnold, E. (2015). Gestures occur with spatial and Motoric knowledge: It's more than just coincidence. Perspectives on Language Learning and Education *22*, 42-49. doi:10.1044/lle22.2.42.

10. McNeill, D. (2005). Gesture and though (University of Chicago Press). 10.7208/chicago/9780226514642.001.0001.

11. Kendon, A. (1997). Gesture. Annu Rev Anthropol *26*, 109-128. 10.1146/annurev.anthro.26.1.109.

12. Hagoort, P. (2005). On broca, brain, and binding: a new framework. Trends in Cognitive Sciences *9*, 416-423. 10.1016/j.tics.2005.07.004.

13. Hagoort, P., Hald, L., Bastiaansen, M., and Petersson, K.M. (2004). Integration of word meaning and world knowledge in language comprehension. Science *304*, 438-441. 10.1126/science.1095455.

14. Ozyurek, A., Willems, R.M., Kita, S., and Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. J Cognitive Neurosci *19*, 605-616. 10.1162/jocn.2007.19.4.605.

15. Willems, R.M., Ozyurek, A., and Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. Neuroimage *47*, 1992-2004. 10.1016/j.neuroimage.2009.05.066.

16. Drijvers, L., Jensen, O., and Spaak, E. (2021). Rapid invisible frequency tagging reveals nonlinear integration of auditory and visual information. Human Brain Mapping *42*, 1138-1152. 10.1002/hbm.25282.

17. Drijvers, L., and Ozyurek, A. (2018). Native language status of the listener modulates

637 the neural integration of speech and iconic gestures in clear and adverse listening
638 conditions. Brain and Language *177*, 7-17. 10.1016/j.bandl.2018.01.003.
639 18. Drijvers, L., van der Plas, M., Ozyurek, A., and Jensen, O. (2019). Native and non-
640 native listeners show similar yet distinct oscillatory dynamics when using gestures to
641 access speech in noise. Neuroimage *194*, 55-67. 10.1016/j.neuroimage.2019.03.032.
642 19. Holle, H., and Gunter, T.C. (2007). The role of iconic gestures in speech
643 disambiguation: ERP evidence. J Cognitive Neurosci *19*, 1175-1192.
644 10.1162/jocn.2007.19.7.1175.
645 20. Kita, S., and Ozyurek, A. (2003). What does cross-linguistic variation in semantic
646 coordination of speech and gesture reveal?: Evidence for an interface representation
647 of spatial thinking and speaking. J Mem Lang *48*, 16-32. 10.1016/S0749-
648 596x(02)00505-3.
649 21. Bernardis, P., and Gentilucci, M. (2006). Speech and gesture share the same
650 communication system. Neuropsychologia *44*, 178-190.
651 10.1016/j.neuropsychologia.2005.05.007.
652 22. Zhao, W.Y., Riggs, K., Schindler, I., and Holle, H. (2018). Transcranial magnetic
653 stimulation over left inferior frontal and posterior temporal cortex disrupts gesture-
654 speech integration. Journal of Neuroscience *38*, 1891-1900. 10.1523/Jneurosci.1748-
655 17.2017.
656 23. Zhao, W., Li, Y., and Du, Y. (2021). TMS reveals dynamic interaction between inferior
657 frontal gyrus and posterior middle temporal gyrus in gesture-speech semantic
658 integration. The Journal of Neuroscience, 10356-10364. 10.1523/jneurosci.1355-

659 21.2021.

660 24. Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst Tech J *27*,
661 379-423. 10.1002/j.1538-7305.1948.tb01338.x.

662 25. Tremblay, P., Deschamps, I., Baroni, M., and Hasson, U. (2016). Neural sensitivity to
663 syllable frequency and mutual information in speech perception and production.
664 Neuroimage *136*, 106-121. 10.1016/j.neuroimage.2016.05.018.

665 26. Bikson, M., Inoue, M., Akiyama, H., Deans, J.K., Fox, J.E., Miyakawa, H., and
666 Jefferys, J.G.R. (2004). Effects of uniform extracellular DC electric fields on
667 excitability in rat hippocampal slices. J Physiol-London *557*, 175-190.
668 10.1113/jphysiol.2003.055772.

669 27. Federmeier, K.D., Mai, H., and Kutas, M. (2005). Both sides get the point:
670 hemispheric sensitivities to sentential constraint. Memory & Cognition *33*, 871-886.
671 10.3758/bf03193082.

672 28. Kelly, S.D., Kravitz, C., and Hopkins, M. (2004). Neural correlates of bimodal speech
673 and gesture comprehension. Brain and Language *89*, 253-260. 10.1016/s0093-
674 934x(03)00335-3.

675 29. Wu, Y.C., and Coulson, S. (2005). Meaningful gestures: Electrophysiological indices
676 of iconic gesture comprehension. Psychophysiology *42*, 654-667. 10.1111/j.1469-
677 8986.2005.00356.x.

678 30. Fritz, I., Kita, S., Littlemore, J., and Krott, A. (2021). Multimodal language processing:
679 How preceding discourse constrains gesture interpretation and affects gesture
680 integration when gestures do not synchronise with semantic affiliates. J Mem Lang

681 *117*, 104191. 10.1016/j.jml.2020.104191.

682 31. Gunter, T.C., and Weinbrenner, J.E.D. (2017). When to take a gesture seriously: On
683 how we use and prioritize communicative cues. J Cognitive Neurosci *29*, 1355-1367.
684 10.1162/jocn_a_01125.

685 32. Oldfield, R.C. (1971). The assessment and analysis of handedness: the Edinburgh
686 inventory. Neuropsychologia *9*, 97-113. 10.1016/0028-3932(71)90067-4.

687 33. Zhao, W. (2023). TMS reveals a two-stage priming circuit of gesture-speech
688 integration. Front Psychol *14*, 1156087. 10.3389/fpsyg.2023.1156087.

689 34. McNeill, D. (1992). Hand and mind : what gestures reveal about thought (University of
690 Chicago Press). 10.2307/1576015.

691 35. Kelly, S.D., Creigh, P., and Bartolotti, J. (2010). Integrating speech and iconic
692 gestures in a Stroop-like task: Evidence for automatic processing. Journal of
693 Cognitive Neuroscience
694 *22*, 683-694. 10.1162/jocn.2009.21254.

695 36. Koessler, L., Maillard, L., Benhadid, A., Vignal, J.P., Felblinger, J., Vespignani, H.,
696 and Braun, M. (2009). Automated cortical projection of EEG sensors: Anatomical
697 correlation via the international 10-10 system. Neuroimage *46*, 64-72.
698 10.1016/j.neuroimage.2009.02.006.

699 37. Tesink, C.M.J.Y., Petersson, K.M., van Berkum, J.J.A., van den Brink, D., Buitelaar,
700 J.K., and Hagoort, P. (2009). Unification of speaker and meaning in language
701 comprehension: An fMRI study. J Cognitive Neurosci *21*, 2085-2099.
702 10.1162/jocn.2008.21161.

703 38. Nuwer, M.R., Comi, G., Emerson, R., Fuglsang-Frederiksen, A., Guerit, J.M., Hinrichs,
704 H., Ikeda, A., Luccas, F.J., and Rappelsberger, P. (1999). IFCN standards for digital
705 recording of clinical EEG. The International Federation of Clinical Neurophysiology.
706 Electroencephalogr Clin Neurophysiol Suppl *52*, 11-14. 10.1016/S0013-
707 4694(97)00106-5.
708 39. Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of
709 single-trial EEG dynamics including independent component analysis. J Neurosci
710 Methods *134*, 9-21. 10.1016/j.jneumeth.2003.10.009.
711 40. Habets, B., Kita, S., Shao, Z.S., Ozyurek, A., and Hagoort, P. (2011). The Role of
712 Synchrony and Ambiguity in Speech-Gesture Integration during Comprehension. J
713 Cognitive Neurosci *23*, 1845-1854. 10.1162/jocn.2010.21462.
714 41. Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). FieldTrip: Open
715 Source Software for Advanced Analysis of MEG, EEG, and Invasive
716 Electrophysiological Data. Computational Intelligence and Neuroscience *2011*,
717 156869. 10.1155/2011/156869.
718 42. Kutas, M., and Federmeier, K.D. (2011). Thirty Years and Counting: Finding Meaning
719 in the N400 Component of the Event-Related Brain Potential (ERP). Annual Review
720 of Psychology, Vol 62 *62*, 621-647. 10.1146/annurev.psych.093008.131123.
721 43. Rogers, T.T., Ralph, M.A.L., Garrard, P., Bozeat, S., McClelland, J.L., Hodges, J.R.,
722 and Patterson, K. (2004). Structure and deterioration of semantic memory: A
723 neuropsychological and computational investigation. Psychological Review *111*, 205-
724 235. 10.1037/0033-295x.111.1.205.

44. Ralph, M.A.L., Jefferies, E., Patterson, K., and Rogers, T.T. (2017). The neural and computational bases of semantic cognition. Nature Reviews Neuroscience *18*, 42-55. 10.1038/nrn.2016.150.

45. Rogers, T.T., Hodges, J.R., Ralph, M.A.L., and Patterson, K. (2003). Object recognition under semantic impairment: The effects of conceptual regularities on perceptual decisions. Lang Cognitive Proc *18*, 625-662. 10.1080/01690960344000053.

46. Fadiga, L., Craighero, L., and Olivier, E. (2005). Human motor cortex excitability during the perception of others' action. Current Opinion in Neurobiology *15*, 213-218. 10.1016/j.conb.2005.03.013.

47. Giard, M.H., and Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. J Cognitive Neurosci *11*, 473-490. 10.1162/089892999563544.

48. Meijer, G.T., Mertens, P.E.C., Pennartz, C.M.A., Olcese, U., and Lansink, C.S. (2019). The circuit architecture of cortical multisensory processing: Distinct functions jointly operating within a common anatomical network. Prog Neurobiol *174*, 1-15. 10.1016/j.pneurobio.2019.01.004.

49. Senkowski, D., and Engel, A.K. (2024). Multi-timescale neural dynamics for multisensory integration. Nat Rev Neurosci *25*, 625-642. 10.1038/s41583-024-00845-7.

50. Jackson, R.L. (2021). The neural correlates of semantic control revisited. Neuroimage *224*, 117444. 10.1016/j.neuroimage.2020.117444.

51. Jefferies, E. (2013). The neural basis of semantic cognition: converging evidence from neuropsychology, neuroimaging and TMS. Cortex *49*, 611-625. 10.1016/j.cortex.2012.10.008.

52. Noonan, K.A., Jefferies, E., Visser, M., and Lambon Ralph, M.A. (2013). Going beyond inferior prefrontal involvement in semantic control: evidence for the additional contribution of dorsal angular gyrus and posterior middle temporal cortex. J Cogn Neurosci *25*, 1824-1850. 10.1162/jocn_a_00442.

53. Hagoort, P. (2013). MUC (Memory, Unification, Control) and beyond. Frontiers in Psychology *4*, 416. 10.3389/fpsyg.2013.00416.

54. Bizley, J.K., Maddox, R.K., and Lee, A.K.C. (2016). Defining auditory-visual objects: Behavioral tests and physiological mechanisms. Trends in Neurosciences *39*, 74-85. 10.1016/j.tins.2015.12.007.

55. Hartwigsen, G., Bzdok, D., Klein, M., Wawrzyniak, M., Stockert, A., Wrede, K., Classen, J., and Saur, D. (2017). Rapid short-term reorganization in the language network. Elife *6*. 10.7554/eLife.25964.

56. Jackson, R.L., Hoffman, P., Pobric, G., and Ralph, M.A.L. (2016). The semantic network at work and rest: Differential connectivity of anterior temporal lobe subregions. Journal of Neuroscience *36*, 1490-1501. 10.1523/JNEUROSCI.2999-15.2016.

57. Humphreys, G. F., Lambon Ralph, M. A., & Simons, J. S. (2021). A Unifying Account of Angular Gyrus Contributions to Episodic and Semantic Cognition. Trends in neurosciences, 44(6), 452–463. https://doi.org/10.1016/j.tins.2021.01.006

58. Bonner, M. F., & Price, A. R. (2013). Where is the anterior temporal lobe and what does it do?. The Journal of neuroscience : the official journal of the Society for

Neuroscience, 33(10), 4213–4215. https://doi.org/10.1523/JNEUROSCI.0041-13.2013

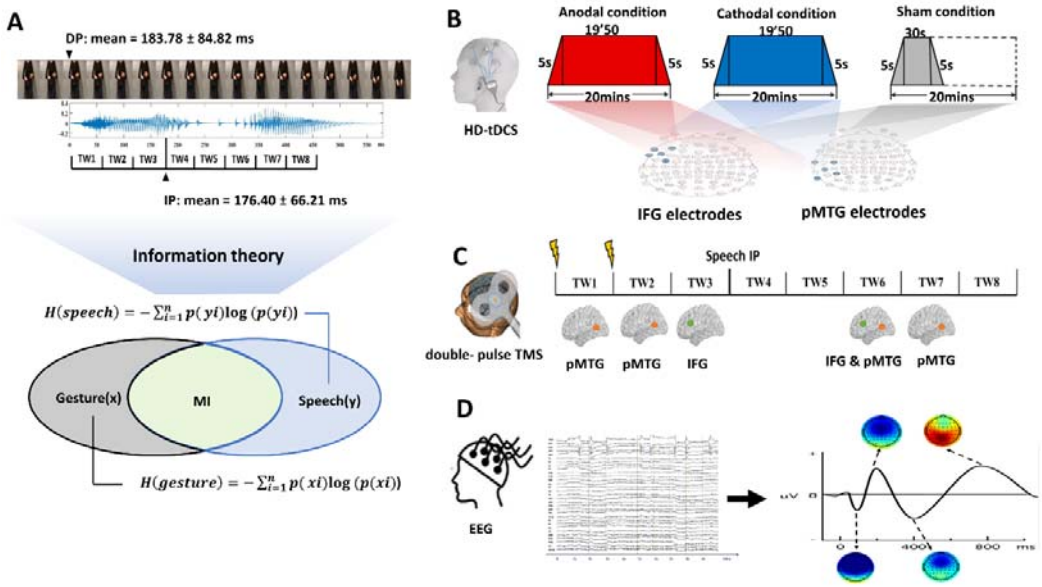

Figure 1. Experimental design, and stimulus characteristics.

(**A**) **Experimental stimuli.** Twenty gestures were paired with 20 relevant speech stimuli. Two separate Pre-tests were executed to define the minimal length of each gesture and speech required for semantic identification, namely, the discrimination point (DP) of gesture and the identification point (IP) of speech. Overall, a mean of 183.78 ms (SD = 84.82) was found for the DP of gestures and the IP of speech was 176.40 ms (SD = 66.21). The onset of speech was set at the gesture DP. Responses for each item were assessed utilizing information-theoretic complexity metrics to quantify the information content of both gesture and speech during integration, employing entropy and MI.

(B) Procedure of Experiment 1. HD-tDCS, including Anodal, Cathodal, or Sham conditions, was administered to the IFG or pMTG) using a 4 * 1 ring-based electrode montage. Electrode F7 targeted the IFG, with return electrodes placed on AF7, FC5, F9, and FT9. For pMTG stimulation, TP7 was targeted, with return electrodes positioned on C5, P5, T9, and P9. Sessions lasted 20 minutes, with a 5-second fade-in and fade-out, while the Sham condition involved only 30 seconds of stimulation.

(**C**) **Procedure of Experiment 2.** Eight time windows (TWs, duration = 40 ms) were segmented in relative to the speech IP. Among the eight TWs, five (TW1, TW2, TW3, TW6, and TW7) were chosen based on the significant results in our prior study[23]. Double-pulse TMS was delivered over each of the TW of either the pMTG or the IFG.

(D) Procedure of Experiment 3. Semantically congruent gesture-speech pairs were presented randomly with Electroencephalogram (EEG) recorded simultaneously. Epochs were time locked to the onset of speech and lasted for 1000 ms. A 200 ms pre-stimulus baseline correction was applied before the onset of gesture stoke. Various elicited components were hypothesized.

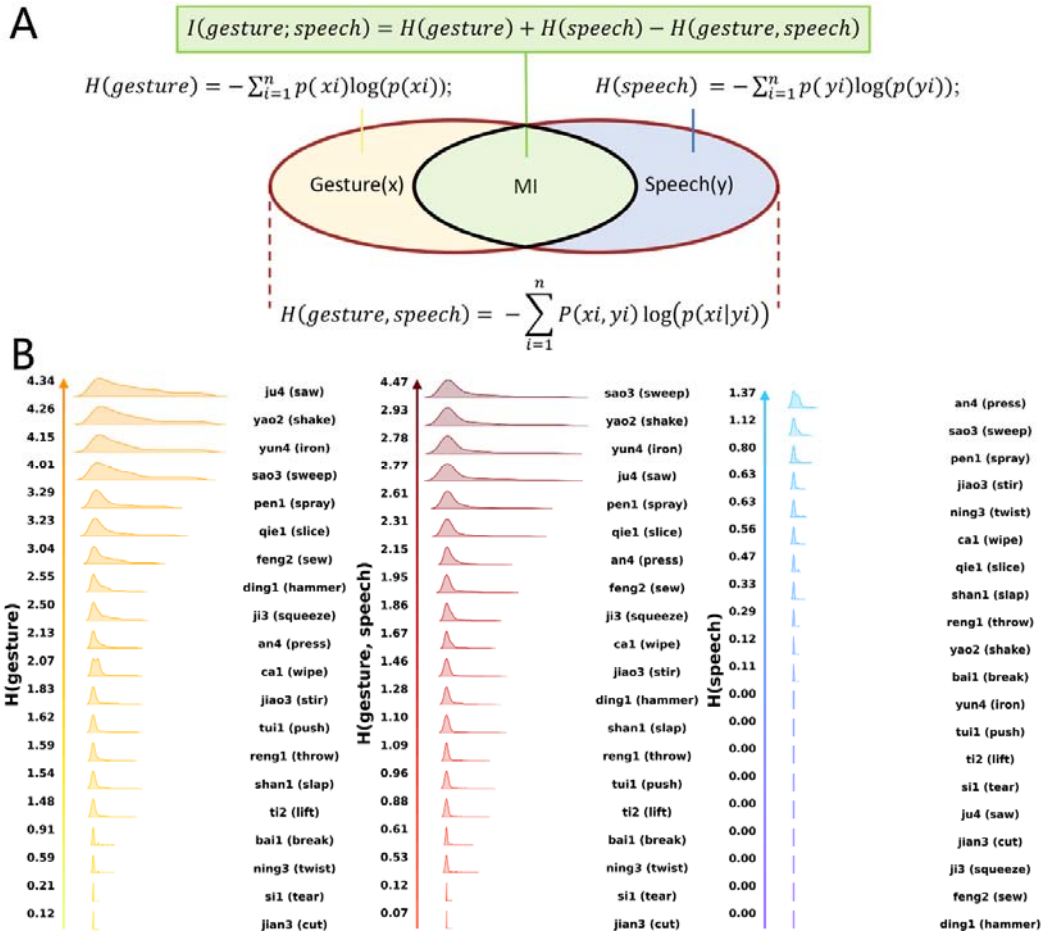

Figure 2. Quantification formulas (A) and distributions of each stimulus in Shannon's entropy (B).

Two separate pre-tests (N = 30) were conducted to assign a single verb for describing each of the isolated 20 gestures and 20 speech items. Responses provided for each item were transformed into Shannon's entropy using a relative quantification formula. Gesture (**B left**) and speech (**B right**) entropy quantify the randomness of gestural or speech information, representing the uncertainty of probabilistic representation activated when a specific stimulus occurs. Joint entropy (**B middle**) captures the widespread nature of the two sources of information combined. Mutual information (MI) was calculated as the difference between joint entropy with gesture entropy and speech entropy combined (**A**), thereby capturing the overlap of gesture and speech and representing semantic integration.

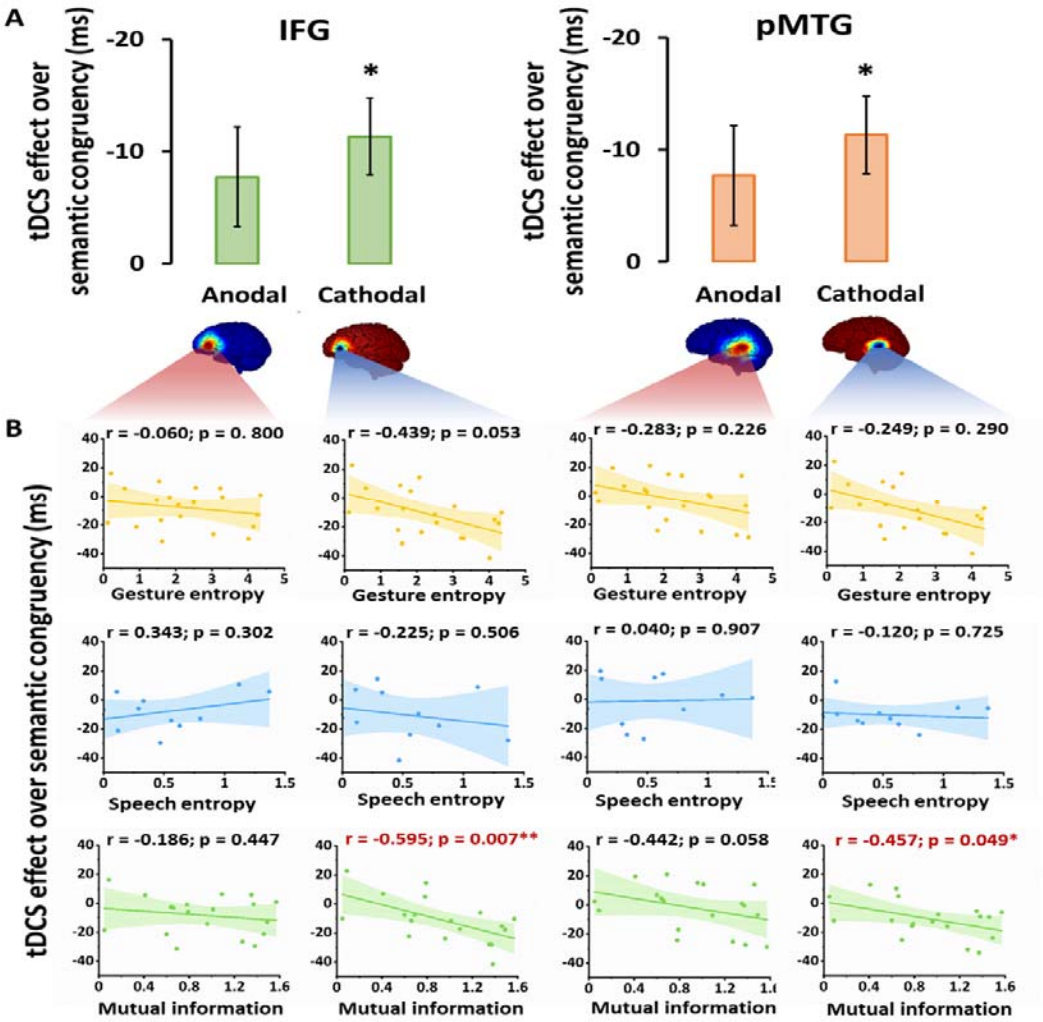

Figure 3. tDCS effect over semantic congruency.

(A) tDCS effect was defined as active-tDCS minus sham-tDCS. The semantic congruency effect was calculated as the reaction time (RT) difference between semantically incongruent and semantically congruent pairs (Rt(incongruent) - Rt(congruent)).

(B) Correlations of the tDCS effect over the semantic congruency effect with three information models (gesture entropy, speech entropy and MI) are displayed with best-fitting regression lines. Significant correlations are marked in red**.** * $p < 0.05$, ** $p < 0.01$ after FDR correction.

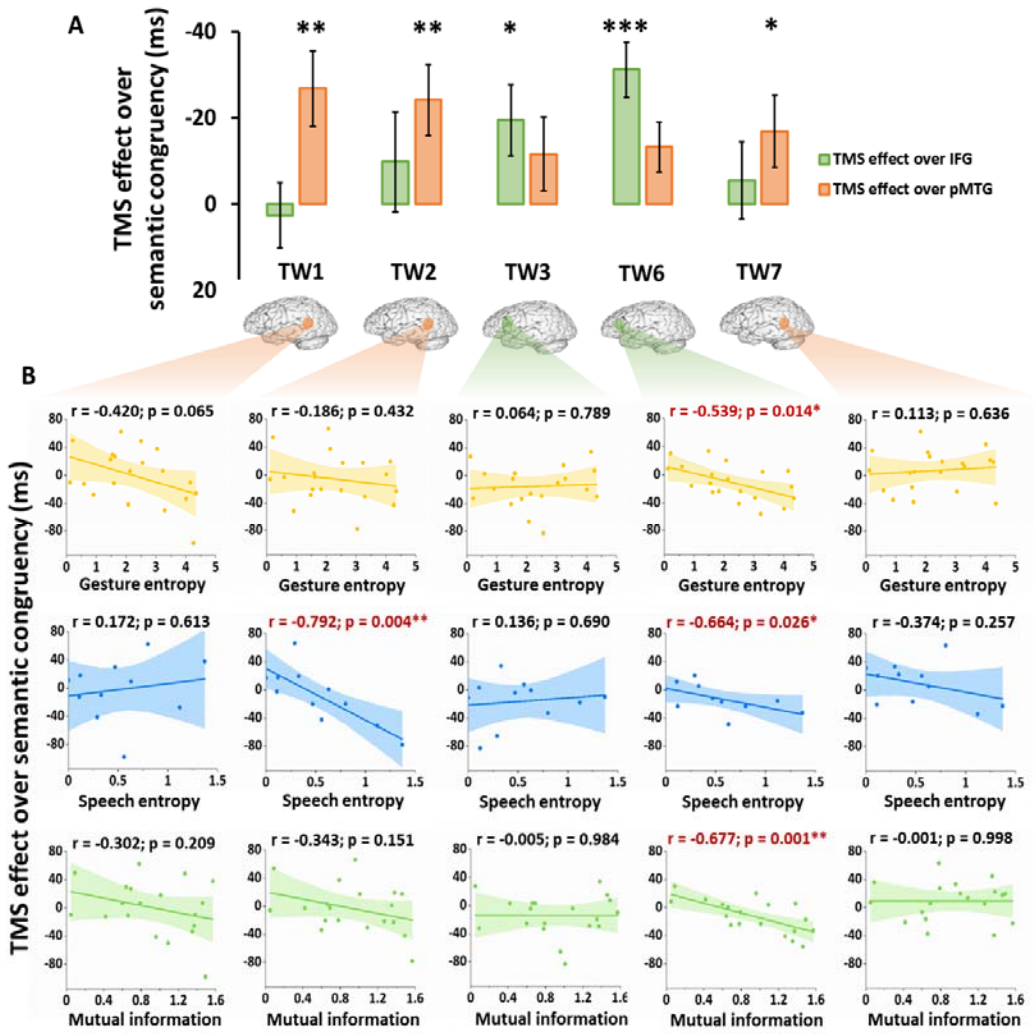

Figure 4. TMS impacts on semantic congruency effect across various time windows (TW).

(A) Five time windows (TWs) showing selective disruption of gesture-speech integration were chosen: TW1 (-120 to -80 ms relative to speech identification point), TW2 (-80 to -40 ms), TW3 (-40 to 0 ms), TW6 (80 to 120 ms), and TW7 (120 to 160 ms). TMS effect was defined as active-TMS minus vertex-TMS. The semantic congruency effect was calculated as the reaction time (RT) difference between semantically incongruent and semantically congruent pairs.

(B) Correlations of the TMS effect over the semantic congruency effect with three information models (gesture entropy, speech entropy and MI) are displayed with best-fitting regression lines. Significant correlations are marked in red**.** * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ after FDR correction.

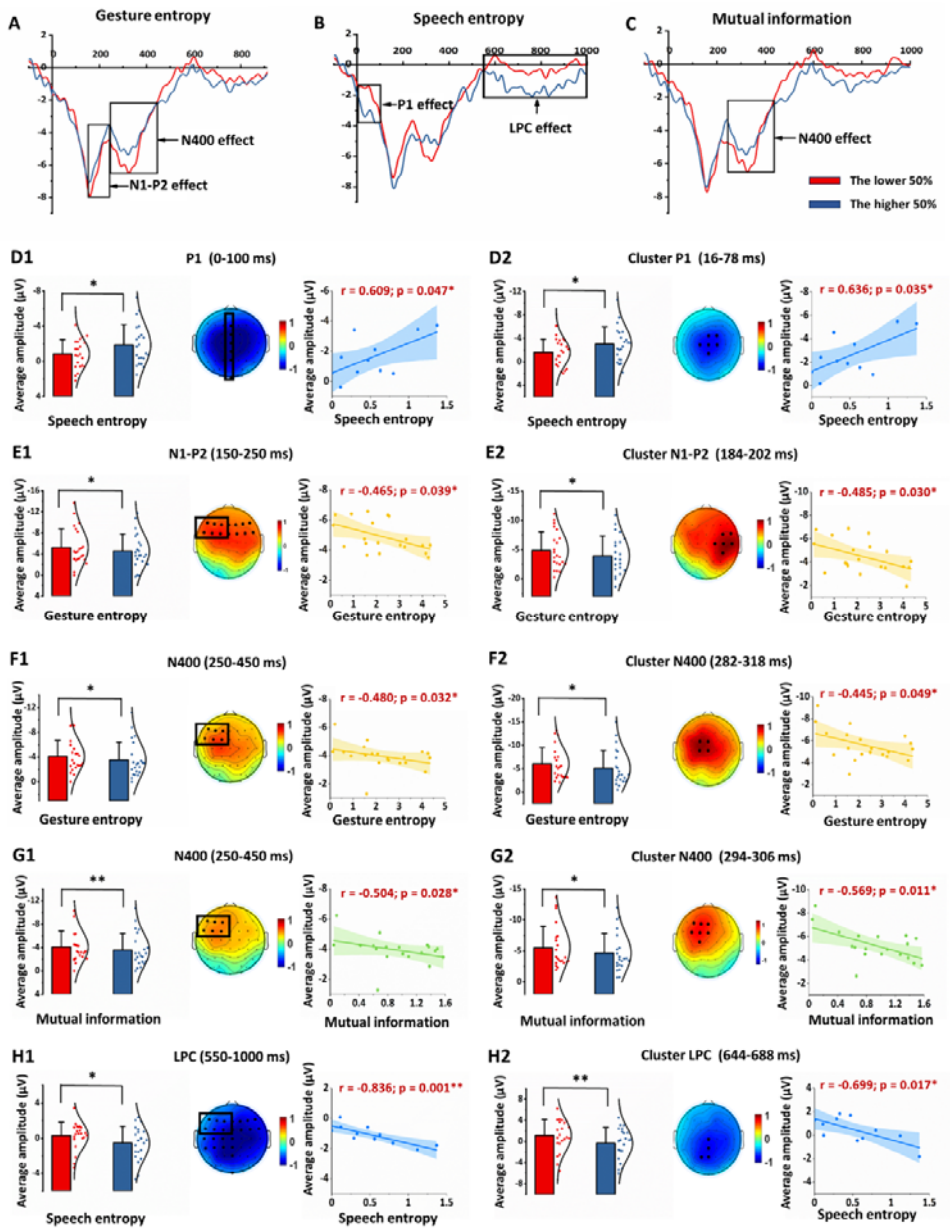

Figure 5. ERP results of gesture entropy (A), speech entropy (B) or MI (C). Four ERP components were identified from grand-average ERPs at the Cz electrode, contrasting trials with the lower 50% (red lines) and the higher 50% (blue lines) of gesture entropy, speech entropy or MI. Clusters of adjacent time points and electrodes were subsequently identified within each component using a cluster-based permutation test. Topographical maps depict amplitude differences between the lower and higher halves of each information model, with significant ROIs (**D1-H1 middle**) or electrode clusters (**D2-H2 middle**) highlighted in black. Solid rectangles delineating the ROIs that exhibited the maximal correlation and paired t-values (**D1-H1 middle**). T-test comparisons with normal distribution lines (**left**) and correlations with best-fitting regression lines (**right**) are calculated and illustrated between the average ERP amplitude within the rectangular ROI (**D1-H1**) or the

elicited clusters (**D2-H2**) and the three information models individually. * $p < 0.05$, ** $p < 0.01$ after FDR correction.

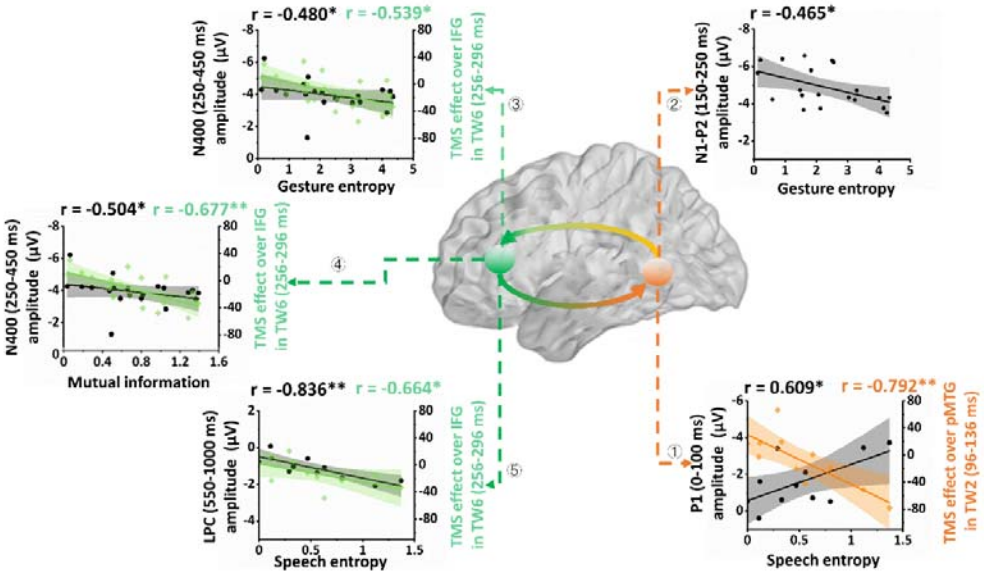

Figure 6. Progressive processing stages of gesture–speech information within the pMTG-IFG loop.

Correlations between the TMS disruption effect of pMTG and IFG with three information models are represented by the orange line and the green lines, respectively. Black lines denote the strongest correlations of ROI averaged ERP components with three information models. * $p < 0.05$, ** $p < 0.01$ after FDR correction.