Neural correlates of probabilistic information processing in gesture-speech integration

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

[1] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China 100101

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

[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China 200031

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

* Corresponding author:

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

16 Lincui Road, Chaoyang district, Beijing, China 100101

Abstract

How information of different modalities (e.g., speech and gesture) frame into a unified one remains unclear in the neurobiology of language. Leveraging the information-theoretic complexity metrics of entropy and mutual information (MI) to quantify gestures and speech, the present event-related potentials and chronometric double-pulse transcranial magnetic stimulation (TMS) study investigated the neural processes during probabilistic gesture-speech integration. In the early processing stage, we found a positive correlation of P1 amplitude with speech entropy and a negative correlation of N1-P2 amplitude with gesture entropy, while the TMS disruption effect of posterior middle temporal gyrus (pMTG) was negatively correlated with speech entropy. In the lexico-semantic processing stage, we observed a negative correlation of N400 amplitude with both gesture entropy and gesture-speech MI, a negative correlation of late positive component amplitude with speech entropy, together with correlations of the TMS interruption effect of inferior frontal gyrus (IFG) with gesture entropy, speech entropy and their MI. Our findings suggest a probabilistically updated lexical representation that was dynamically framed by top-down gesture constrain, bottom-up speech encoding and their interplay. A 5-step process in the pMTG-IFG-pMTG neural circuit is hypothesized for probabilistic gesture–speech processing, which promises a better understanding of the predictive multisensory language processing.

Keywords: gesture-speech integration; probabilistic representation; information theory; ERP; TMS

1. **Introduction**

In communication, gestures often co-exist with speech to convey a coherent sense. Gestures are regarded as 'part of language' (1) or functional equivalents of lexical units in spoken language that alternate and integrate with speech (2). It is suggested that information from gesture and speech is incorporated into a 'single unification space' during language comprehension (3-5). Empirical evidence has shown a bidirectional interaction between these two (6-8) as reflected by the N400 latency and amplitude (5), as well as common neural underpinnings (9-11). An unanswered question is how the spatial-motoric and synthetic gesture integrates with the linear-analytic speech into a unified representation of meaning during language comprehension.

Traditionally, language comprehension is understood as retrieving word meaning from memory and subsequently integrating them into a compositional representation of sentence meaning (12) in a manner that is governed by the apparatus of formal grammar. Recent research has conceptualized language processing as a sequence of mental states that are defined by a particular parser and function quantitively (13). According to this theory, there is inherent uncertainty in the information available to the senses, as well as in the encoding of that information by the sensory systems. Consequently, the neural representations that guide behavior are probabilistic (14). In fact, the probabilistic model has been applied to explain structural parsing principles (for a review, see (15)), word recognition (16), word processing (17) and syllable perception (18). Further studies have shown an updated probabilistic representation with the coming of each word in a sentence, the magnitude of which is correlated with the amplitude of the N400 event-related potential (ERP) component (19, 20) , the cortical response recorded from the high-density electrocorticography (ECoG) (21), and the brain oscillatory activities of theta (4-10 Hz) and delta (0.5-4 Hz) bands (22). However, empirical evidence of probabilistically updated representation during gesture-speech integration is still lacking.

The present study directly tested the neurocognitive processes underlying the probabilistic semantic integration of gesture and speech. To this end, we adopted the information-theoretic complexity metrics of *entropy* and *mutual information* (MI) to

mathematically quantify the information of gesture and speech during integration. Entropy captures the disorder or randomness of information and is used as a measurement of the uncertainty of probabilistic representation activated when an event occurs (23). MI illustrates the mutual constraint that the two variables impose on each other (18) and was used herein as an index of the magnitude of gesture–speech integration. In Experiment 1 (**Figure 1B**), ERPs with high temporal resolution were used to determine the time course of the gesture-speech integration process. Previous studies have reported various stages of gesture–speech integration, indexed by ERP components of the early N1–P1 and P2 sensory effects (24, 25), the N400 semantic anomaly effect (5, 25, 26), and the late positive component (LPC) restructuring effect (27, 28). Therefore, Experiment 1 hypothesized a proportional amplitude of different ERP components as a change of the gesture entropy, speech entropy or their MI. Using double-pulse transcranial magnetic stimulation (TMS) to locate the time windows (TWs) for the involvement of the posterior middle temporal gyrus (pMTG) and inferior frontal gyrus (IFG) in integration, Experiment 2 (**Figure 1C**) aimed to unpack the information processing within the pMTG–IFG–pMTG neural circuit during gesture–speech integration, as found in our prior study (10). We hypothesized different correlation patterns between the TW-selective disruption effect of the pMTG and IFG with the gesture entropy, speech entropy and MI. Together, the present study aimed to elucidate the probabilistic neural processing in gesture–speech integration by uncovering how ERP components and time-sensitive regional engagement were modulated by the information presented from gesture and speech.

2. Material and methods

2.1 Experiment 1

2.1.1 Participants

Twenty-two native Chinese participants signed written informed consent forms and took part in Experiment 1 (12 female; 19-29 years, mean = 21). All of the participants were right-handed (laterality quotient (LQ) = 88.49, SD = 12.65), had normal or corrected-to-normal vision and were paid ¥100 per hour for their participation. This study was approved by the Ethics Committee of the Institute of Psychology, Chinese Academy of Sciences.

2.1.2 Stimuli

Twenty gestures with 20 semantically congruent speech signals were used in the present study. The stimuli set were recorded from two native Chinese speakers (1 male, 1 female) and validated by replicating the semantic congruency effect with 30 participants. Results showed a significantly 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) (for details, see (10)). To create a semantic constrain of gesture upon the unfolding speech, the present study presented speech onset at the DP of gesture (**Figure 1A**), a magnitude of the minimal length of each gesture required for semantic identification that was determined by a gating pre-test with another 30 participants. On average, the DP of gestures was 183.78 ms (SD = 84.82) (for details, see (10)). Additionally, two separate pre-tests with 30 subjects in each (pre-test 1: 16 females, aged 19-35 years, SD = 11.03; pre-test 2: 15 females, aged 19-30 years, SD=10.26) were conducted to determine the nameability values of gesture and speech. Participants were presented with isolated gesture/speech and were asked to use one verb to describe the gesture they saw or the speech they heard. The nameability of gesture/speech was defined as the percentage of participants who provided the most frequently used label. The overall nameability of gesture (Mean =.574, SD =.271) and that of speech (Mean =.947, SD =.109) were calculated (**Table S1**).

To increase the statistical power, the gesture–speech combinations were sorted into 10 bins according to gesture entropy and MI (**S1**), from the highest (90%–100%) to the lowest (0–10%) (29). Given that there were only eight steps for speech entropy, pairs of gesture–speech items were sorted into 8 bins of speech entropy (**S1**). For each of the information model, each 10% bin was averaged first within and then across participants.

2.1.3 Experimental procedure

To test the automatic interaction of gesture and speech information, an irrelevant factor of gender congruency (e.g., a man making a gesture combined with a female voice) was created.

In total, 20 gender congruent and 20 gender incongruent trials were presented to each participant. Gesture–speech pairs were presented randomly using Presentation software (www.neurobs.com). Participants were asked to look at the screen but respond as quickly and accurately as possible only to the gender of the voice they heard. The RT and the button being pressed were recorded. Responses were made using the left and right index fingers, and key assignment was counterbalanced across participants. The experiment started with a fixation cross presented on the center of the screen, and the duration of the fixation cross was set to be random at 0.5-1.5 sec.

2.1.4 Electroencephalogram (EEG) recording

Data from Experiment 1 were recorded from 48 Ag/AgCl electrodes mounted in a cap according to the 10-20 system (48). EEG was amplified with a PORTI-32/MREFA amplifier (TMS International B.V., Enschede, NL) and digitized online at 500 Hz (bandpass, 0.01-70 Hz). Vertical and horizontal eye movements were measured with 4 electrodes placed above the left eyebrow and below the left orbital ridge and at the left and right external canthus. All electrodes were referenced online to the left mastoid. Electrode impedance was maintained below 5 KΩ.

2.1.5 Data 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. Next, the MATLAB-based toolbox EEGLAB was used (49). Epochs were time locked to the onset of speech and lasted for 1000 ms. To ensure that the baseline correction was clean with no stimulus presented, a 200 ms pre-stimulus baseline correction was applied before the onset of gesture stoke (**Figure 1B**). 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. Artifacts were removed semiautomatically. Overall, participants were excluded from future analysis if their rejected trials were more than 30% of the total number.

To select the activated component window more objectively, grand-average ERPs at electrode Cz for the higher half (60–100%) and the lower half (0–50%) were compared for

each model before being submitted to correlation analyses. Trials were averaged regardless of gender congruency as no difference was revealed by paired t-tests on each component. To analyze the scalp distribution of the ERP effect, 7 ROIs were defined: left anterior: F1, F3, F5, FC1, FC3, and FC5; left central: C1, C3, C5, CP1, CP3, and CP5; left posterior: P1, P3, P5, PO3, PO5, and O1; right anterior: F2, F4, F6, FC2, FC4, and FC6; right central: C2, C4, C6, CP2, CP4, and CP6; right posterior: P2, P4, P6, PO4, PO6, and O2; and midline electrodes: Fz, FCz, Cz, Pz, Oz, and CPz. Pearson correlations were calculated between each model (gesture entropy, speech entropy, and MI) and each ERP amplitude in each ROI, separately. FDR correction was applied for multiple comparisons.

2.2 Experiment 2

2.2.1 Participants

Another set of 24 participants was recruited for Experiment 2 (11 female; 18-28 years, mean = 23). All were right-handed (LQ = 89.02, SD = 13.25), had normal or corrected-to-normal vision, paid ¥100 per hour for their participation and were approved by the Ethics Committee of the Institute of Psychology, Chinese Academy of Sciences.

2.2.2 Experimental procedure

Similar as in Experiment 1, two factors were manipulated: semantic congruency and gender congruency. Value of MI only refers to trials when gesture contained semantic congruent information with speech. To eliminate the effect caused by stimuli in the semantically incongruent gesture-speech condition, the 20 gestures and 20 speech were combined into 10 pairs with a reverse combination of the semantic information (e.g., a ‘spray’ gesture while saying Mandarin speech word ‘剪 jian3 (cut)’ was paired with a ‘cut’ gesture while saying Mandarin speech word ‘喷 pen1 (spray)’). For each stimulation site, a total of 800 trials for 4 types of conditions (semantic congruency x gender congruency) and 5 TWs were further split into 3 blocks (320 trials in one block and 240 trials in the last two blocks). In each block, pairs of stimuli were kept intact. Participants completed all 3 blocks for each stimulation site, the order of which was counterbalanced using a Latin square design. During the experiment, participants were asked to look at the screen but respond only to the gender of the voice they heard.

185 **2.2.3 TMS protocol**

186 At an intensity of 50% of the maximum stimulator output, double-pulse TMS was delivered via
187 a 70 mm figure-eight coil using a Magstim Rapid² stimulator (Magstim, UK) over either the left
188 IFG (-62, 16, 22) in TW3 and TW6 or the left pMTG (-50, -56, 10) in TW1, TW2 and TW7
189 (**Figure 1C**). The coordinates of the left IFG and left pMTG corresponded to Montreal
190 Neurological Institute (MNI) coordinates. The vertex was used as a control site (for details,
191 see (10)).

192 **2.2.4 Data analysis**

193 All incorrect responses (922 out of the total number of 19200, 4.8% of trials) were excluded.
194 Since TMS upone the IFG and pMTG did not affect the gender congruency effect as shown in
195 our prior study (10), gender incongruent trials were also included for analysis. To eliminate the
196 influence of outliers, a 2 SD trimmed mean of RT for every participant in each session was
197 conducted. Correlations of the TMS effect (active TMS minus vertex TMS) over the semantic
198 congruency effect (RT difference between semantic congruent and incongruent trials) with the
199 three information models (gesture entropy, speech entropy and MI) were calculated. FDR
200 correction was applied where necessary.

201

202 **3. Results**

203 **3.1 Stimulus entropy and MI**

204 Twenty gestures and 20 corresponding speech signals were used in the present study. To
205 capture the distribution of information in gestures and speech, the nameability for each item,
206 which was defined as the percentage of participants who provided the most frequently used
207 label, was transformed into Shannon's entropy (H) using the relevant formula (**Figure 1D**).
208 Here, p(xi) and p(yi) represent the nameability for a given gesture and speech, respectively.
209 High entropy indicates that participants provided many answers for that gesture/speech with a
210 low nameability score, implying a broad distribution of gesture/speech representation. In
211 contrast, low entropy indicates that participants' comprehension for that certain item was quite
212 centralized, and the nameability score was high. Additionally, the interplay of gesture and

speech was captured by MI, which represents the amount of information obtained about gesture by observing speech, and vice versa. That is, the higher MI of gesture and speech, the stronger interplay and integration between the two.

To increase the statistical power, according to scores in each information model, stimuli were sorted into 10 bins for the gesture entropy and MI, and 8 bins for the speech entropy (see **Table S1**) (29). Correlation analyses were conducted between scores of each model and ERP amplitudes (Experiment 1), and between scores of each model and the TMS disruption effects of gesture-speech integration as indexed by reaction time (RT) difference (Experiment 2).

3.2 P1, N1-P2, N400 and LPC components were modulated by gesture and speech probabilities

In Experiment 1, electroencephalogram (EEG) data were recorded with epochs time locked to the onset of speech when subject were presented with semantically congruent gesture-speech pairs (**Figure 1B**). ERPs in relative to speech onset rather than gesture onset was focused as we aimed to assess the impacts of the gesture entropy and MI on speech processing when gestures always precede speech in natural settings. **Figure 2A** presents grand-average ERPs at the electrode Cz classified by either the gesture entropy, speech entropy or MI when each measurement was split into the higher half (60-100%) and the lower half (0-50%). Four ERP components were identified and put into the following analyses: the P1 component at 0-100 ms, the N1-P2 component at 150-250 ms, the N400 component at 250-450 ms and the LPC at 550-1000 ms. Notably, gender congruent and gender incongruent trials (see details in Methodology) were averaged, as the paired t-test failed to show a significant effect of gender congruency on the amplitude of P1 ($t(50) = 1.442$, $p = .156$), N1-P2 ($t(50) = -.539$, $p = 5.92$), N400 ($t(100) = .063$, $p = .95$), and LPC ($t(224) = .974$, $p = .331$). As shown, stimuli with the higher 50% speech entropy elicited a larger P1 amplitude ($t(50) = 24.18$, $p < .001$) and a smaller LPC amplitude ($t(224) = 46.83$, $p < .001$) than stimuli with the lower 50% speech entropy (**Figure 2A** middle). Meanwhile, a larger N400 amplitude ($t(100) = 12.43$, $p < .001$) and a larger N1-P2 amplitude ($t(50) = 14.58$, $p < .001$) were found by the lower 50% compared to the higher 50% gesture entropy (**Figure 2A** left). For the MI, a larger

N400 amplitude was found for stimuli of the lower 50% than the higher 50% ($t(100) = 6.45$, $p < .001$, **Figure 2A** right).

Further correlation analyses found that the P1 amplitude significantly increased as the speech entropy increased in the left central (LC) region-of-interest (ROI) ($r = .825$, false discovery rate (FDR)-corrected $p = .036$), the right central (RC) ROI ($r = .850$, FDR-corrected $p = .024$), and the middleline (ML) ROI ($r = .833$, FDR-corrected $p = .03$). No correlation was found between the P1 amplitude and the gesture entropy ($|r| < 0.315$) or the MI ($|r| < 0.493$). Topographical voltage maps of the r values at various recording sites are shown in **Figure 2B.** Entropy has been considered to be a reflection of the strength of expectations about the upcoming word (30). The correlation of this early sensory response with the speech entropy is interpreted as a difficulty effect of the bottom-up phonological encoding and word access. That is, the larger speech entropy, the less speech intelligibility, therefore resulting in a larger P1 response.

For the N1-P2 amplitude, a negative correlation was found with the gesture entropy in the left anterior (LA) ROI ($r = -.84$, FDR-corrected $p = .014$) and the right anterior (RA) ROI ($r = -.765$, FDR-corrected $p = .035$). No correlation of the N1-P2 amplitude with the speech entropy ($|r| < .719$) or the MI ($|r| < .694$) has been spotted. Since N1-P2 represents the neural encoding of phonetic features, the correlation with the gesture entropy implies a top-down modulation of gesture over the phonological processing of speech. The smaller the gesture entropy, the higher intelligibility it presents, thus leading to a larger top-down modulation over speech phonological processing with an enlarged N1-P2 amplitude.

In terms of the N400 effect, a negative correlation between the N400 amplitude and the gesture entropy ($r = -.677$, FDR-corrected $p = .047$) and MI ($r = -.704$, FDR-corrected $p = .047$) was detected in the RA ROI. There was no correlation of the N400 effect with the speech entropy ($|r| < .68$). When gesture entropy is high, a broad distribution of gesture representations would be activated, resulting in weakened confidence of the predicted word and less cross-modal interplay with a reduced N400 amplitude. With the presentation of the coming word, the brain compares the semantic features being predicted by the top-down gesture and those being activated by the bottom-up speech, the larger the overlap, the smaller the amplitude of the scalp N400 (31), giving the inverse relation between the N400

effect and the MI of gesture and speech.

With regards to the LPC effect, a significant negative correlation with the speech entropy was found in ROIs of LA (, r = -.812, FDR-corrected *p* = .042); LC (r = -.858, FDR-corrected *p* = .018); left posterior (LP, r = -.822, FDR-corrected *p* = .036); RA (r = -.856, FDR-corrected *p* = .021) and ML (r = -.858, FDR-corrected *p* = .018). For the gesture entropy and MI, there were no significant correlation with the LPC (|r| < .55). The results were comprehended as that the representation activated by the bottom-up speech information would reweight the predicted candidates elicited by the top-down gesture information (32), which would give rise to a coherent meaning. With the centralization of the bottom-up speech representation, as quantified by a smaller entropy, confidence for the reanalysis of the predicted candidates becomes stronger, leading to a larger LPC amplitude.

3.3 Engagements of the left pMTG and IFG in integration were modulated by gesture and speech probabilities

Using the same stimuli as in Experiment 1, Experiment 2 manipulated the experimental factor of semantic congruency, with gesture containing either semantically congruent information or semantically incongruent information with the accompanying speech. Double-pulse TMS was delivered to either the left IFG or the left pMTG in one of the five 40-ms TWs in relative to speech identification point (IP) (TW1: -120 ~ -80 ms , TW2: -80 ~ -40 ms, TW3: -40~0 ms, TW6: 80~120 ms, and TW7: 120~160 ms), which showed a TW-selective disruption of gesture-speech integration in our prior study (**Figure 1C**) (for details, see (10)). We focused our current analysis on the correlation of the TMS interruption (active TMS minus vertex TMS) of the semantic congruency effect (RT of semantic incongruent minus RT of semantic congruent pairs) with the gesture entropy, speech entropy or MI. 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.

The results illustrated a significant negative correlation between the TMS effect and the speech entropy when the pMTG was inhibited in TW2 (r = -.877, FDR-corrected *p* = .012). Since a more negative TMS effect over the semantic congruency effect represents a stronger interruption of the integration effect, the negative correlation here indicates that the higher

uncertainty of speech, the stronger engagement of the pMTG in integrating gesture and speech segment before speech identification (TW2). Meanwhile, when the IFG activity was interrupted in TW6, a significant negative correlation was found between the TMS effect and the gesture entropy (r = -.876, FDR-corrected *p* = .003), speech entropy (r = -.736, FDR-corrected *p* = .037), and MI (r = -.761, FDR-corrected *p* = .017) (**Figure 3**). This suggests that the higher uncertainty of gesture or speech and higher MI of gesture and speech would result in stronger involvement of the IFG in multimodal integration after speech identification.

4. Discussion

In this study, we investigated how gesture-primes and speech were integrated for comprehension. By mathematically quantifying gesture and speech information with entropy and MI, we assessed directly the ERP correspondence (Experiment 1) and the regional involvement (Experiment 2) during probabilistic information processing in gesture–speech integration. From an integrated perspective, our results indicate dynamic neural stages within the pMTG-IFG-pMTG neurocircuit of gesture-speech integration that were differentially mediated by top-down gesture prediction (gesture entropy), bottom-up speech encoding (speech entropy) and their interplay (MI) in forming the probabilistically updated lexical representation. Our findings thus extend the current knowledge of how multimodal information is processed and integrated in the brain.

Semantic representation encodes knowledge of concepts that are experienced through multisensory channels abstract away from modality specific attributes (33) and are stored in transmodal regions (34, 35). Empirical evidence has shown that the process of gesture–speech integration involves the incorporation of information into a ‘single unification space’ (3-5) with common neural underpinnings in pMTG and IFG (9, 10). While pMTG is believed to be involved in perceptual matching and activation of a common stable representation, IFG has been reliably associated with selection, control and unification of semantic information (36). In fact, a recurrent neurocircuit connecting the left IFG and pMTG in semantic unification has been proposed extensively (37). In particular, by segmenting the possible integration period into eight TWs relative to the speech IP and applying inhibitory double-pulse TMS over each TW, our previous study hypothesized two stages within the pMTG-IFG-pMTG neurocircuit in

gesture–speech integration. In the first pre-lexical stage, semantic information extracted from gesture constrains the phonological encoding of speech, while in the post-lexical stage, the most feasible lexical candidate is selected, retrieved and unified with the gesture semantics to form a context-appropriate semantic representation (10). As an extension, the current study quantified the information of gesture and speech with entropy and MI, and discovered that TW-specific regional contribution of pMTG and IFG to gesture-speech semantic integration were distinctively correlated with the entropy of gestures and speech and their MI. In an early pre-lexical gesture-speech integration time window (TW2), speech entropy was negatively correlated with activity of pMTG, while in a later post-lexical time window of gesture-speech integration (TW6), a negative correlation of the IFG activity with both gesture entropy, speech entropy and their MI was found. Our results confirmed that semantic representation irrelevant of modalities builds up probabilistically within the neurocircuit connecting the transmodal regions of pMTG and IFG during gesture-speech integration.

Moreover, we disentangled the sub-processes taken place during gesture-speech comprehension with the use of high temporal resolution of ERPs, when representations of gesture and speech were probabilistically presented. Four ERP components were segmented: a P1-N1-P2 complex that was considered to be responsible for speech phonological encoding, and a N400 component followed by the LPC component that reflect the lexico-semantic processing stage. Early P1-N1 and P2 sensory effects in gesture-speech integration has been reported in previous studies and were linked to perception and attentional processes (25, 38). Here, the P1-N1-P2 complex was comprehended as a reflection of the early audiovisual gesture-speech integration taken place during the sensory and perceptual processing chain (39). Note that a semantic priming paradigm was adopted here by presenting the onset of speech to be at the gesture DP (9, 10), during which a semantic top-down prediction of gesture over speech has been created. The early interaction of gesture-speech information happened with both top-down gesture prediction and bottom-up speech processing, as shown by a positive correlation of the P1 effect with speech entropy and a negative correlation of the N1-P2 effect with gesture entropy. Additionally, the lexico-semantic effect of the N400 and the LPC that have consistently been found in previous studies investigating gesture-speech integration (5, 25-28) were differentially mediated by top-down gesture prediction, bottom-up

speech encoding and their interaction. As reflected by a negatively correlation between the N400 effect with both the gesture entropy and their MI, as well as a correlation of the LPC with the speech entropy. In the present study, activation of representation of meaning is modulated probabilistically. The input stimuli activate a dynamically distributed neural landscape, the state of which constructs probabilistically as measured by entropy and MI and correlates with the electrophysiological signals (N400 and LPC) indicating a change of lexical representation. Consistent with a recent 'Sentence Gestalt' account proposed by Rabovsky (2018) in the multisensory gesture-speech information processing (20), our findings further indicate that the changed activation pattern was induced from directions of both top-down and bottom-up gesture-speech processing. This process was associated with the N400 effect, followed by the LPC which might reflect an update weighing of the semantic cues that provided the changes of activation pattern.

Considering the close alignment of the ERP components with the TWs of TMS effect, i.e. the P1 component (0-100 ms of speech onset) was aligned with TW2 (96-136 ms of speech processing), and the N400 effect (250-450 ms of speech onset) coincided with TW6 (256-296 ms of speech processing) (**Figure 1A**), it is quite reasonable to associate the ERP components with the cortical involvements that both were interrelated with the same information measure (e.g., gesture entropy). Hence, hypothesis was made that the early sensory effect of the P1 component was associated mainly with the pMTG due to their temporal co-occurrence and correlation with the speech entropy, while the N400 component was a reflection of the IFG activity because of their temporal overlap and correlation with both the gesture entropy and MI (**Figure 4**). Specifically, double-pulse TMS over the supramodal pMTG could both feedforward and feedback affect the phonological processing of speech in the superior temporal gyrus/sulcus (STG/S) as indexed by the P1 component and the N1-P2 component respectively (40). The hypothesis that the IFG is the major generator of the N400 component is also consistent with previous findings using the semantic priming paradigm (41), especially when the stimulus-onset asynchrony of two different modalities is long enough and the prime word is fully processed. During that process, IFG activity is considered to ease the effort of word selection or facilitate the controlled semantic retrieval that is made available by the top-down prediction from the context or primes over the target word (42). Besides,

393 although the LPC component lagged behind TW6 in which the IFG was involved, the LPC
394 amplitude and the TMS effect of the IFG were co-related with the speech entropy. A link of the
395 LPC to the IFG has been provided by several studies (for a review, see (43)), in which the IFG
396 is considered to integrate the current stimulus with prior context into an updated utterance
397 representation. Similarly, the present study reported a significant correlation of IFG with the
398 speech entropy, which we comprehended as a reanalysis of semantic information, and further
399 led to a coherent understanding.

400 In emphasizing the probabilistic notion in gesture-speech integration, we recognized a
401 distributional property of the language input that is inherently graded and/or stochastic. Here,
402 we estimated the probabilities of gesture and speech by computing the amount of information
403 conveyed by each modality with entropy and MI and revealed distinct contributions of gesture
404 and speech information within both the early speech processing stage and the late lexico-
405 semantic processing stage. Referencing the recurrent neurocircuit connecting the left IFG and
406 pMTG for semantic unification (44) , we extended the two-stage gesture-speech integration
407 circuit proposed in our previous study (10) into 5 steps (**Figure 4**). First, bottom-up speech
408 processing mapping acoustic signal to its lexical representation was performed from the
409 STG/S to the pMTG. The larger speech entropy was, the greater effort was made during the
410 matching of the acoustic input with its stored lexical representation, thus leading to a larger
411 involvement of the pMTG at pre-lexical stage (TW2) and a larger P1 effect (**Figure 4**①).
412 Second, the gesture representation was activated in the pMTG and further exerted a top-
413 down modulation over the phonological processing of speech in the STG/S (45). The higher
414 certainty of gesture, a larger modulation of gesture would be made upon speech, as indexed
415 by a smaller gesture entropy with an enhanced N1-P2 amplitude (**Figure 4**②). Third,
416 information was relayed from the pMTG to the IFG for sustained activation, during which a
417 semantic constraint from gesture has been made on the semantic retrieval of speech. Greater
418 TMS effect over the IFG at post-lexical stage (TW6) accompanying with a reduced N400
419 amplitude were found with the increase of gesture entropy, when the representation of
420 gesture was wildly distributed and the constrain over the following speech was weak **(Figure**
421 **4**③). Fourth, the activated speech representation was compared with that of the gesture in
422 the IFG. At this stage, the larger overlapped neural populations activated by gesture and

speech as indexed by a larger MI, a greater TMS disruption effect of the IFG and a reduced N400 amplitude indexing easier integration and less semantic conflict were observed (**Figure 4**④). Last, the activated speech representation would disambiguate and reanalyze the semantic information that was stored in the IFG and further unify into a coherent comprehension in the pMTG (3, 46). The more uncertain information being provided by speech, as indicated by an increased speech entropy, a stronger reweighting effect was made over the activated semantic information, resulting in a strengthened involvement of the IFG as well as a reduced LPC amplitude (**Figure 4**⑤).

Certain limitations should be noted. Firstly, the ERP components and the cortical engagements were measured in two separate experiments and their link was reflected through mediating variables of the entropy and MI. We provided no direct evidence of the brain structures underpinning the corresponding ERP components, and we even failed to find the TMS evidence responsible for the top-down gesture effect on the N1-P2 component. Hence, the cortical implementation of the ERP components in gesture–speech integration needs to be clarified by future studies. Secondly, the causal effect of TMS on behavioral performance is much more complex than previously thought (47), making the conclusion of the information flow within the pMTG–IFG-pMTG loop less convincing due to a close functional connection between the two locations. On the one hand, we cannot tell whether the significantly reduced activity in the left IFG was due to a solo contribution of the IFG or a joint involvement of the left pMTG and the left IFG in gesture–speech integration (9). On the other hand, neither could we exclude the involvement of other brain areas, such as the anterior temporal lobe (34), which may modulate activities in the IFG and pMTG.

5. Conclusion

In summary, with the help of the information-theoretic complexity metrics of entropy and MI, we show that in multisensory semantic processing of gesture and speech, a probabilistically updated representation is constructed in a dynamic interplay between gesture-primed prediction and speech presentation. The probabilistic updating process is connected to different ERP components and sequential cortical involvement of the pMTG-IFG-pMTG circuit.

A 5-step neural stage model is proposed for processing probabilistic gesture–speech information, which advances our knowledge about the dynamic multisensory integration and probabilistic language processing.

Funding: This research was supported by grants from the National Key Research and Development Program of China (2021ZD0201501), 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.; Formal Analysis, W.Y.Z.; Methodology, W.Y.Z.; Validation, X.L.; Visualization, W.Y.Z. 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., X.L. and Y.D.

Declaration of interests: The authors declare no competing interests.

Figure legends

Figure 1. Experimental design, stimulus characteristics, and quantification formulas. (**A**) Experimental stimuli. Twenty gestures were paired with 20 relevant speech stimuli. Two gating studies 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. Eight time windows (TWs, duration = 40 ms) were segmented in relative to the speech IP. **(B)** Procedure of Experiment 1. 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. (**C**) Procedure of Experiment 2. Among the eight TWs, five (TW1, TW2, TW3, TW6, and TW7) were chosen based on the significant results in our prior study (10). Double-pulse TMS was delivered over each of the TW of either the pMTG or the IFG. (**C**) Quantification formulas. Another two separate pre-tests were carried out to determine the nameability of isolated gestures and speech, which was defined as the percentage of participants that provided the mostly used label. Nameability for each item (*p(xi)* represents nameability of gesture and *p(yi)* for nameability of speech) was transformed into Shannon's entropy (H) and the MI (I (gesture; speech)) to capture the contribution of gesture and speech in multisensory information processing.

Figure 2. ERP results. (**A**) Grand-average ERPs at the Cz electrode comparing trials with the lower 50% (red lines) and the higher 50% (blue lines) of gesture entropy, speech entropy or MI. Four ERP components in which there was a significant difference between the bottom and top halves were selected (paired t test, p values corrected by FDR). (**B**) Correlations of ERP amplitudes with three information models with best-fitting regression lines and topographical voltage maps. Significant correlations are marked in red with spotted subset of electrodes (Spearman correlation analysis, p values corrected by FDR). The solid rectangles demarcate ROIs that showed a maximal value of correlation. LC: left central; LP: left posterior;

RA: right anterior; RC: right central.

Figure 3. Cortical engagement correlated with gesture entropy, speech entropy and MI. Correlations of the TMS effect over the semantic congruency effect with three information models are displayed with best-fitting regression lines. Significant correlations are marked in red**.** * p <.05, ** p <.01 after FDR correction.

Figure 4. Proposed stages for probabilistic processing of gesture–speech information. Correlations of cortical involvement over semantic congruency effect with three information models were indicated by the Blue line and the Red lines, which reflect the TMS-disruption of the pMTG and the IFG activities respectively. Black lines show the strongest correlations of the ERP amplitudes with three information models. * p <.05, ** p <.01 after FDR correction.

References

1. A. Kendon, Gesture. *Annu Rev Anthropol* **26**, 109-128 (1997).
2. T. Slamacazacu, Nonverbal Components of Oral Communication and the Concept of So-Called Mixed-Syntax - a Synthesis. *Degres*, E1-E24 (1993).
3. P. Hagoort, On Broca, brain, and binding: a new framework. *Trends Cogn Sci* **9**, 416-423 (2005).
4. P. Hagoort, L. Hald, M. Bastiaansen, K. M. Petersson, Integration of word meaning and world knowledge in language comprehension. *Science* **304**, 438-441 (2004).
5. A. Ozyurek, R. M. Willems, S. Kita, P. Hagoort, On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *J Cognitive Neurosci* **19**, 605-616 (2007).
6. H. Holle, T. C. Gunter, The role of iconic gestures in speech disambiguation: ERP evidence. *J Cognitive Neurosci* **19**, 1175-1192 (2007).
7. S. Kita, A. Ozyurek, What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *J Mem Lang* **48**, 16-32 (2003).
8. P. Bernardis, M. Gentilucci, Speech and gesture share the same communication system. *Neuropsychologia* **44**, 178-190 (2006).
9. W. Y. Zhao, K. Riggs, I. Schindler, H. Holle, Transcranial magnetic stimulation over left inferior frontal and posterior temporal cortex disrupts gesture-speech integration. *Journal of Neuroscience* **38**, 1891-1900 (2018).
10. W. Zhao, Y. Li, Y. Du, TMS reveals dynamic interaction between inferior frontal gyrus and posterior middle temporal gyrus in gesture-speech semantic integration. *The Journal of Neuroscience* 10.1523/jneurosci.1355-21.2021, 10356-10364 (2021).
11. R. M. Willems, A. Ozyurek, P. Hagoort, Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *Neuroimage* **47**, 1992-2004 (2009).
12. R. Jackendoff, Foundations of Language: Brain, Meaning, Grammar, Evolution. (2002).
13. J. R. Brennan, E. P. Stabler, S. E. Van Wagenen, W. M. Luh, J. T. Hale, Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language* **157**, 81-94 (2016).
14. K. Armeni, R. M. Willems, S. L. Frank, Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience Biobehavioral Review* **83**, 579-588 (2017).
15. N. Chater, C. D. Manning, Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences* **10**, 335-344 (2006).
16. D. Norris, The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review* **113**, 327-357 (2006).
17. B. R. Payne, C. L. Lee, K. D. Federmeier, Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology* **52**, 1456-1469 (2015).
18. P. Tremblay, I. Deschamps, M. Baroni, U. Hasson, Neural sensitivity to syllable frequency

and mutual information in speech perception and production. *Neuroimage* **136**, 106-121 (2016).

19. A. Hodapp, M. Rabovsky, The N400 ERP component reflects an error-based implicit learning signal during language comprehension. *European Journal of Neuroscience* **54**, 7125-7140 (2021).
20. M. Rabovsky, S. S. Hansen, J. L. McClelland, Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nat Hum Behav* **2**, 693-705 (2018).
21. M. K. Leonard, K. E. Bouchard, C. Tang, E. F. Chang, Dynamic Encoding of Speech Sequence Probability in Human Temporal Cortex. *Journal of Neuroscience* **35**, 7203-7214 (2015).
22. P. W. Donhauser, S. Baillet, Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron* **105**, 385-+ (2020).
23. C. E. Shannon, A Mathematical Theory of Communication. *Bell Syst Tech J* **27**, 379-423 (1948).
24. K. D. Federmeier, H. Mai, M. Kutas, Both sides get the point: hemispheric sensitivities to sentential constraint. *Mem Cognit* **33**, 871-886 (2005).
25. S. D. Kelly, C. Kravitz, M. Hopkins, Neural correlates of bimodal speech and gesture comprehension. *Brain and Language* **89**, 253-260 (2004).
26. Y. C. Wu, S. Coulson, Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology* **42**, 654-667 (2005).
27. I. Fritz, S. Kita, J. Littlemore, A. Krott, Multimodal language processing: How preceding discourse constrains gesture interpretation and affects gesture integration when gestures do not synchronise with semantic affiliates. *J Mem Lang* **117**, 104191 (2021).
28. T. C. Gunter, J. E. D. Weinbrenner, When to Take a Gesture Seriously: On How We Use and Prioritize Communicative Cues. *J Cognitive Neurosci* **29**, 1355-1367 (2017).
29. K. A. DeLong, T. P. Urbach, M. Kutas, Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience* **8**, 1117-1121 (2005).
30. R. M. Willems, S. L. Frank, A. D. Nijhof, P. Hagoort, A. van den Bosch, Prediction During Natural Language Comprehension. *Cerebral Cortex* **26**, 2506-2516 (2016).
31. G. Baggio, P. Hagoort, The balance between memory and unification in semantics: A dynamic account of the N400. *Lang Cognitive Proc* **26**, 1338-1367 (2011).
32. R. Levy, Expectation-based syntactic comprehension. *Cognition* **106**, 1126-1177 (2008).
33. A. Caramazza, A. E. Hillis, B. C. Rapp, C. Romani, The multiple semantics hypothesis: Multiple confusions? *Cognitive Neuropsychology* **7**, 161-189 (1990).
34. K. Patterson, P. J. Nestor, T. T. Rogers, Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience* **8**, 976-987 (2007).
35. M. A. L. Ralph, E. Jefferies, K. Patterson, T. T. Rogers, The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience* **18**, 42-55 (2017).
36. X. Xu, I. L. Hanganu-Opatz, M. Bieler, Cross-talk of low-level sensory and high-level cognitive processing: Development, mechanisms, and relevance for cross-modal abilities of the brain. *Frontiers in Neurorobotics* **14**, 7 (2020).
37. H. Damasio, T J. Grabowski, D. Tranel, R D. Hichwa, A R. Damasio, A neural basis for

lexical retrieval. *Nature* **380**, 499-505 (1996).
38. L. Fadiga, L. Craighero, E. Olivier, Human motor cortex excitability during the perception of others' action. *Curr Opin Neurobiol* **15**, 213-218 (2005).
39. M. H. Giard, F. Peronnet, Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *J Cognitive Neurosci* **11**, 473-490 (1999).
40. L. H. Arnal, B. Morillon, C. A. Kell, A. L. Giraud, Dual Neural Routing of Visual Facilitation in Speech Processing. *Journal of Neuroscience* **29**, 13445-13453 (2009).
41. B. T. Gold *et al.*, Dissociation of automatic and strategic lexical-semantics: Functional magnetic resonance imaging evidence for differing roles of multiple frontotemporal regions. *Journal of Neuroscience* **26**, 6523-6532 (2006).
42. G. R. Kuperberg, T. Sitnikova, B. M. Lakshmanan, Neuroanatomical distinctions within the semantic system during sentence comprehension: Evidence from functional magnetic resonance imaging. *Neuroimage* **40**, 367-388 (2008).
43. H. Brouwer, M. W. Crocker, N. J. Venhuizen, J. C. J. Hoeks, A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cogn Sci* **41 Suppl 6**, 1318-1352 (2017).
44. P. Hagoort, MUC (Memory, Unification, Control) and beyond. *Frontiers in Psychology* **4**, 416 (2013).
45. J. K. Bizley, R. K. Maddox, A. K. C. Lee, Defining auditory-visual objects: Behavioral tests and physiological mechanisms. *Trends in Neurosciences* **39**, 74-85 (2016).
46. C. M. J. Y. Tesink *et al.*, Unification of Speaker and Meaning in Language Comprehension: An fMRI Study. *J Cognitive Neurosci* **21**, 2085-2099 (2009).
47. T. O. Bergmann, G. Hartwigsen, Inferring causality from noninvasive brain stimulation in cognitive neuroscience. *Journal of Cognitive Neuroscience* . **33**, 195-225 (2021).
48. M. R. Nuwer *et al.*, IFCN standards for digital recording of clinical EEG. The International Federation of Clinical Neurophysiology. *Electroencephalogr Clin Neurophysiol Suppl* **52**, 11-14 (1999).
49. A. Delorme, S. Makeig, EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* **134**, 9-21 (2004).

Figure 1

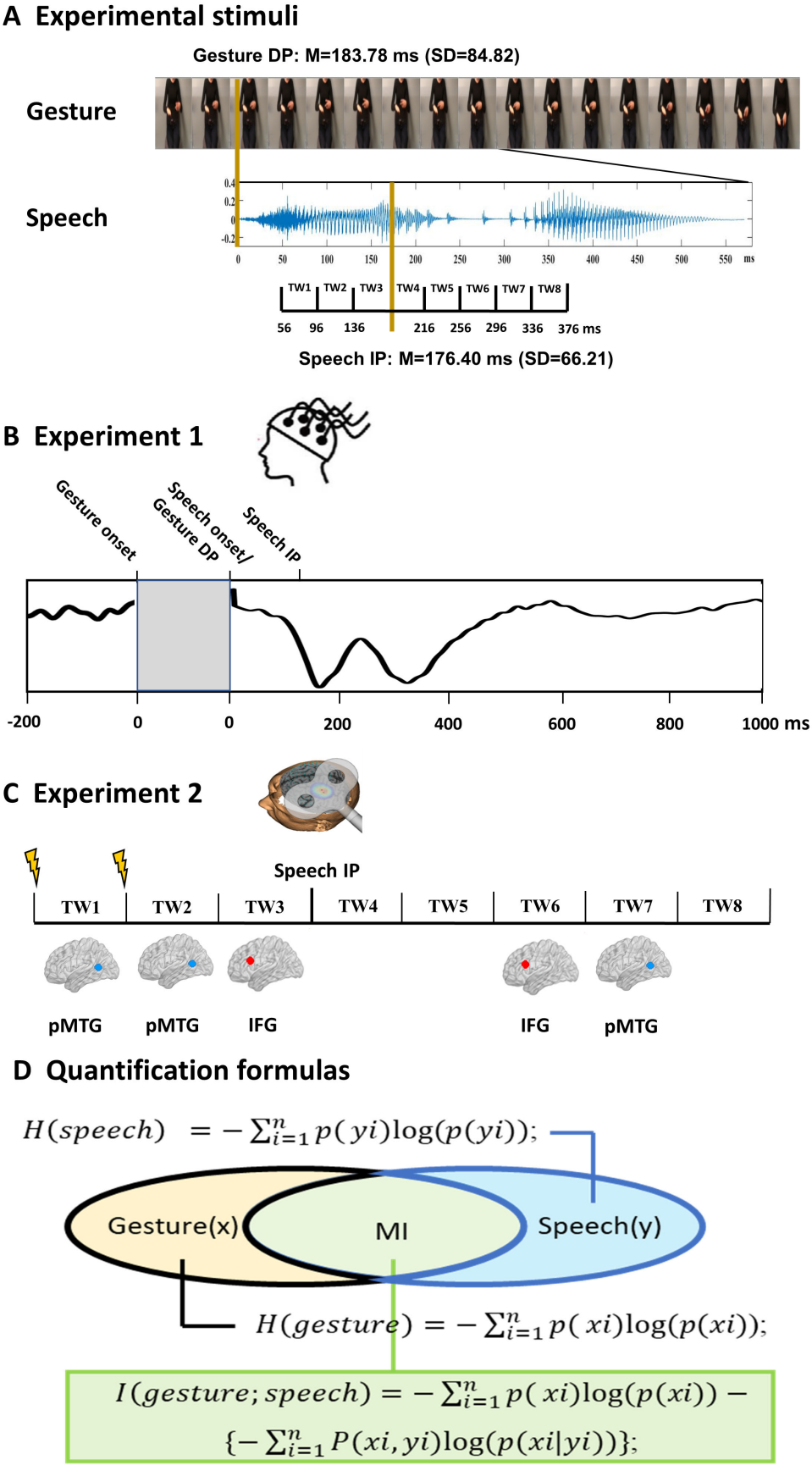
A Experimental stimuli
Gesture DP: M=183.78 ms (SD=84.82)
Gesture
Speech
TW1 TW2 TW3 TW4 TW5 TW6 TW7 TW8
56 96 136 216 256 296 336 376 ms
Speech IP: M=176.40 ms (SD=66.21)
B Experiment 1
Gesture onset
Speech onset/ Gesture DP
Speech IP
-200 0 0 200 400 600 800 1000 ms
C Experiment 2
Speech IP
TW1 TW2 TW3 TW4 TW5 TW6 TW7 TW8
pMTG pMTG IFG IFG pMTG
D Quantification formulas
$H(speech) = -\sum_{i=1}^{n} p(yi)\log(p(yi));$
Gesture(x)
MI
Speech(y)
$H(gesture) = -\sum_{i=1}^{n} p(xi)\log(p(xi));$
$I(gesture; speech) = -\sum_{i=1}^{n} p(xi)\log(p(xi)) - \{-\sum_{i=1}^{n} P(xi, yi)\log(p(xi|yi))\};$

Figure 1. Experimental design, stimulus characteristics, and quantification formulas. (**A**) Experimental stimuli. Twenty gestures were paired with 20 relevant speech stimuli. Two gating studies 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. Eight time windows (TWs, duration = 40 ms) were segmented in relative to the speech IP. **(B)** Procedure of Experiment 1. 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. (**C**) Procedure of Experiment 2. Among the eight TWs, five (TW1, TW2, TW3, TW6, and TW7) were chosen based on the significant results in our prior study (10). Double-pulse TMS was delivered over each of the TW of either the pMTG or the IFG. (**C**) Quantification formulas. Another two separate pre-tests were carried out to determine the nameability of isolated gestures and speech, which was defined as the percentage of participants that provided the mostly used label. Nameability for each item (*p(xi)* represents nameability of gesture and *p(yi)* for nameability of speech) was transformed into Shannon's entropy (H) and the MI (I (gesture; speech)) to capture the contribution of gesture and speech in multisensory information processing.

Figure 2

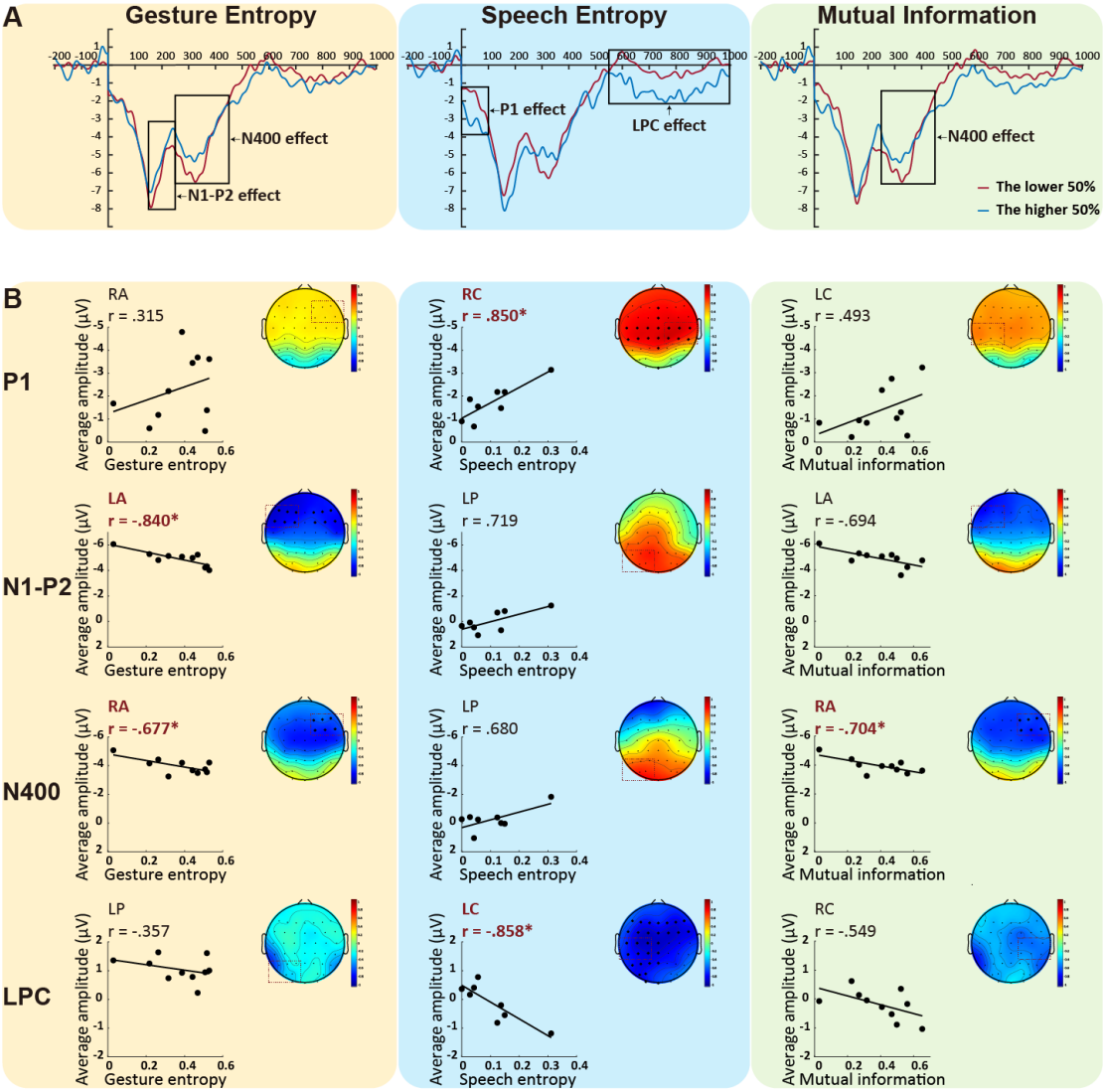

Figure 2. ERP results. (**A**) Grand-average ERPs at the Cz electrode comparing trials with the lower 50% (red lines) and the higher 50% (blue lines) of gesture entropy, speech entropy or MI. Four ERP components in which there was a significant difference between the bottom and top halves were selected (paired t test, p values corrected by FDR). (**B**) Correlations of ERP amplitudes with three information models with best-fitting regression lines and topographical voltage maps. Significant correlations are marked in red with spotted subset of electrodes (Spearman correlation analysis, p values corrected by FDR). The solid rectangles demarcate ROIs that showed a maximal value of correlation. LC: left central; LP: left posterior; RA: right anterior; RC: right central.

Figure 3

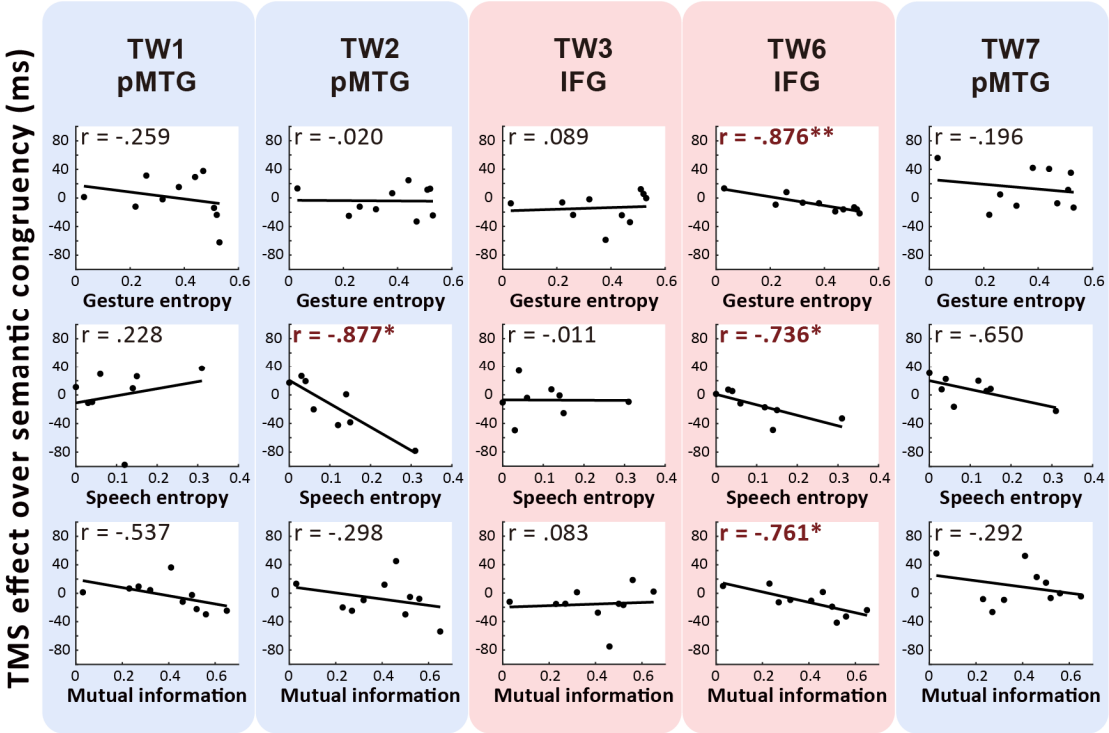

Figure 3. Cortical engagement correlated with gesture entropy, speech entropy and MI. Correlations of the TMS effect over the semantic congruency effect with three information models are displayed with best-fitting regression lines. Significant correlations are marked in red. * p <.05, ** p <.01 after FDR correction.

Figure 4

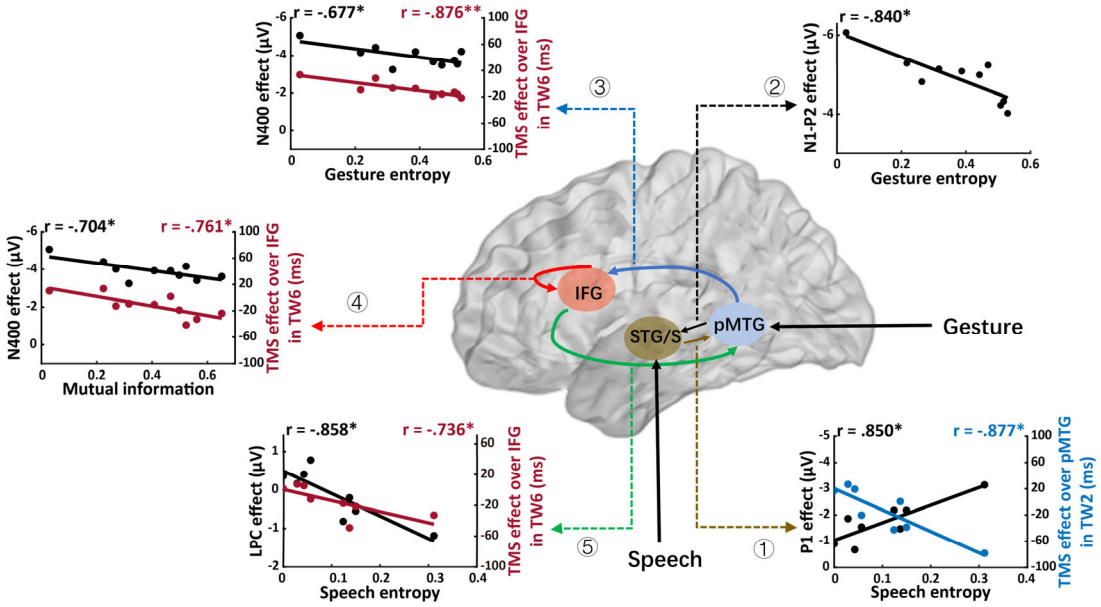

Figure 4. Proposed stages for probabilistic processing of gesture–speech information. Correlations of cortical involvement over semantic congruency effect with three information models were indicated by the Blue line and the Red lines, which reflect the TMS-disruption of the pMTG and the IFG activities respectively. Black lines show the strongest correlations of the ERP amplitudes with three information models. * p <.05, ** p <.01 after FDR correction.

Supplementary Table 1. Stimuli characteristics.

Item	pair	Gesture nameability	Speech nameability	Gesture entropy	Speech entropy	MI	Gesture bins	Speech bins	MI bins
an4 (press)	1	0.59	0.9	0.449	0.137	0.496	6	6	7
yun4 (iron)	1	0.16	0.98	0.423	0.029	0.447	5	2	6
bai1 (break)	2	0.86	0.98	0.187	0.029	0.191	2	2	2
yao2 (shake)	2	0.2	0.98	0.464	0.029	0.487	7	2	6
ca1 (wipe)	3	0.46	0.9	0.515	0.137	0.582	8	6	9
reng1 (throw)	3	0.73	1	0.331	0.000	0.331	4	1	4
ding1(hammer)	4	0.5	1	0.500	0.000	0.500	8	1	7
tui1 (push)	4	0.76	1	0.301	0.000	0.301	4	1	4
ti2 (lift)	5	0.37	1	0.531	0.000	0.531	10	1	8
feng2 (sew)	5	0.8	1	0.258	0.000	0.258	3	1	2
ji3 (squeeze)	6	0.46	1	0.515	0.000	0.515	9	1	8
si1 (tear)	6	0.96	1	0.057	0.000	0.057	1	1	1
jiao3 (stir)	7	1	1	0.000	0.000	0.000	1	1	1
shan1 (slap)	7	0.21	0.75	0.473	0.311	0.697	7	8	10
ju4 (saw)	8	0.71	0.89	0.351	0.150	0.381	5	7	5
ning2 (twist)	8	0.79	0.96	0.269	0.057	0.279	3	4	3
sao3 (sweep)	9	0.17	1	0.435	0.000	0.435	6	1	5
jian3 (cut)	9	0.81	0.89	0.246	0.150	0.260	2	7	3
pen1 (spray)	10	0.45	0.97	0.518	0.043	0.541	9	3	9
qie1 (cut)	10	0.33	0.91	0.528	0.124	0.607	10	5	10