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

Dr. Wanying Zhao, Email: zhaowy@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 sensory 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 constraint, 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

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.

Language processing has recently been conceptualized as a sequence of mental states that are defined by a particular parser and function quantitively[12]. 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[13]. In fact, the probabilistic model has been applied to explain perception and language processing[14,15], natural language comprehension[16], word recognition[17,18] and syllable perception[19]. 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[20,21] , the cortical response recorded from the high-density electrocorticography (ECoG)[22], and the brain oscillatory activities of theta (4-10 Hz) and delta (0.5-4 Hz) bands[23]. 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 [24]. MI illustrates the

mutual constraint that the two variables impose on each other[19] 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 the early N1–P1 and P2 sensory effects[25,26], the N400 semantic conflict effect[5,26,27], and the late positive component (LPC) reconstruction effect[28,29]. We thus hypothesized a proportional amplitude of different ERP components as a change of the gesture entropy, speech entropy or their MI. Experiment 2 (**Figure 1C**) used 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 gesture-speech integration, as found in our prior study[10]. We hypothesized that the TW-selective disruption effect of the pMTG and IFG would be differentially correlated with the gesture entropy, speech entropy and MI. Together, we 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.

Material and methods

Participants

Forty-six young Chinese participants signed written informed consent forms and took part in the present study (Experiment 1: N = 22, 12 females, 19-29 years, mean = 21; Experiment 2: N = 24, 11 females, 18-28 years, mean = 23). All of the participants were right-handed (Experiment 1: laterality quotient (LQ) = 88.49 ± 12.65; Experiment 2: LQ = 89.02 ± 13.25), 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.

Stimuli

Twenty gestures with 20 semantically congruent speech were used. 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)[10]. To mimic the natural settings where gestures always precede and provide semantic constraints upon the unfolding speech, the present study presented speech onset at the discrimination point (DP) of gesture (**Figure 1A**). The DP of gestures was determined by a gating pre-test with another 30 participants and defined as the minimal length of each gesture required for semantic identification. On average, the DP of gestures was 183.78 ms (SD = 84.82)[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) (**Figure 2A**) and that of speech (mean = .947, SD = .109) (**Figure 2B**) were calculated.

To capture the distribution of information in gestures and speech, nameabilities were transformed into Shannon’s entropy (H) using the relevant formula (**Figure 1D**). Here, p(xi) and p(yi) represent the nameability for a given gesture and speech, respectively. High entropy indicates that participants provided many answers for that gesture/speech with a low nameability score, implying a broad distribution of gesture/speech representation. In contrast, low entropy indicates that participants’ comprehension for that certain item was quite 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, the gesture–speech combinations were sorted into 10 bins according to gesture entropy (**Figure 2C**) and MI (**Figure 2D**), from the highest (90%–

100%) to the lowest (0–10%)[30]. Given that there were only eight steps for speech entropy, pairs of gesture–speech items were sorted into 8 bins of speech entropy (**Figure 2E**). For each of the information model, each 10% bin was averaged first within and then across participants.

Experimental procedure

To test the automatic interaction of gesture and speech, an irrelevant factor of gender congruency (e.g., a man making a gesture combined with a female voice) was created in both experiments. While only semantically congruent pairs were presented in Experiment 1, both semantically congruent and incongruent pairs were presented in Experiment 2 to examine the semantic congruency effect. 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. The experiment started with a fixation cross presented on the center of the screen, which lasted for 0.5-1.5 sec.

Experiment 1: Electroencephalogram (EEG) recording and data analysis

EEG were recorded from 48 Ag/AgCl electrodes mounted in a cap according to the 10-20 system[31], 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, 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. Artifacts were removed semiautomatically.

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 the onset of gesture stoke (**Figure 1B**). 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. 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. False discovery rate (FDR) correction was applied for multiple comparisons.

Experiment 2: TMS protocol and data analysis

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) over either the left IFG (-62, 16, 22) in TW3 (-40~0 ms in relative to speech identification point (IP)) and TW6 (80~120 ms,) or the left pMTG (-50, -56, 10) in TW1 (-120 ~ -80 ms), TW2 (-80 ~ -40 ms) and TW7 (120~160 ms). Those TWs were selected as they showed a TW-selective disruption of gesture-speech integration in our prior study (**Figure 1C**). The IFG/pMTG coordinates corresponded to Montreal Neurological Institute (MNI) coordinates. The vertex was used as a control site.

Incorrect responses (922 out of the total number of 19200, 4.8% of trials) were excluded. Since TMS did not affect the gender congruency effect as shown previously[10], gender incongruent trials were also included for analysis. We focused the current analysis on Pearson correlations of the TMS interruption effects (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. FDR correction was applied for multiple comparisons.

Results

Experiment 1: P1, N1-P2, N400 and LPC components were modulated by gesture and speech probabilities

Figure 3A 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%). As shown, stimuli with the higher 50% speech entropy elicited a larger P1 amplitude (0-100 ms) ($t(50) = 24.18$, $p < .001$) and a smaller LPC amplitude (550-1000 ms) ($t(224) = 46.83$, $p < .001$) than stimuli with the lower 50% speech entropy (**Figure 3A** middle). Meanwhile, a larger N400 amplitude (250-450 ms) ($t(100) = 12.43$, $p < .001$) and a larger N1-P2 amplitude (150-250 ms) ($t(50) = 14.58$, $p < .001$) were found by the lower 50% compared to the higher 50% gesture entropy (**Figure 3A** 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 3A** right). Notably, trials were averaged regardless of gender congruency as no difference was revealed by paired t-tests 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 ($t(224) = .974$, $p = .331$).

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 3B.**

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.

For 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 ($|r| < .68$).

For 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).

Experiment 2: Engagements of the left pMTG and IFG in integration were modulated by gesture and speech probabilities

As shown in **Figure 4**, a significant negative correlation was found between the TMS effect (a more negative TMS effect represents a stronger interruption of the integration effect) and the speech entropy when the pMTG was inhibited in TW2 (r = -.877, FDR-corrected *p* = .012). 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).

Discussion

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.

Empirical evidence has shown that 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]. 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[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 of integration with ERPs, when representations of gesture and speech were probabilistically presented. Early P1-N1 and P2 sensory effects linked to perception and attentional processes[26,32] was comprehended as a reflection of the early audiovisual gesture-speech integration in the sensory-perceptual processing chain[33]. 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. Nonetheless, activation of representation is modulated probabilistically. The input stimuli would 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) which indicate the change of lexical representation. Consistent with a recent ‘Sentence Gestalt’ account in multisensory information processing[21], our findings further confirm that the changed activation pattern can be induced from directions of both top-down and bottom-up gesture-speech processing.

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 reasonable to associate the ERP components with the cortical involvements that both were interrelated with the same information measure (e.g., gesture entropy). Consequently, referencing the recurrent neurocircuit connecting the left IFG and pMTG for semantic unification[34] , we extended the previously proposed two-stage gesture-speech integration circuit[10] into 5 steps (**Figure 5**). First, bottom-up speech processing mapping acoustic signal to its lexical representation was performed from the STG/S 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 5**①). Second, the gesture representation was activated in the pMTG and further exerted a top-down modulation over the phonological processing of speech in the STG/S[35]. The higher certainty of gesture, a larger modulation of gesture would be made upon speech, as indexed by a smaller gesture entropy with an enhanced N1-P2 amplitude (**Figure 5**②). 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 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 5**③). 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, a greater TMS disruption effect of the IFG and a reduced N400 amplitude indexing easier integration and less semantic conflict were observed (**Figure 5**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[3,36]. 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 5**⑤).

Limitations exist. The ERP components and cortical engagements were measured in two experiments and their link was built through mediating variables (e.g., entropy). We provided no direct evidence of the brain structures underpinning the corresponding ERPs, which needs to be clarified by future studies. Moreover, the causal effect of TMS on network reorganization and behavior is much more complex than previously thought[37], making the conclusion of the information flow within the pMTG–IFG-pMTG loop less certain.

Conclusion

In summary, with 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.

Acknowledgments

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).

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. 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).
13. K. Armeni, R. M. Willems, S. L. Frank, Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neurosci Biobehav R* **83**, 579-588 (2017).
14. N. Chater, C. D. Manning, Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences* **10**, 335-344 (2006).
15. F. Lange, M. Heilbron, P. Kok, How Do Expectations Shape Perception? *Trends in cognitive sciences* 10.1016/j.tics.2018.06.002, 22 (2018).
16. M. Heilbron, K. Armeni, J.-M. Schoffelen, P. Hagoort, F. P. de Lange, A hierarchy of linguistic predictions during natural language comprehension. *P Natl Acad Sci USA* **119**, e2201968119-e2201968119 (2022).
17. D. Norris, The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review* **113**, 327-357 (2006).
18. B. R. Payne, C. L. Lee, K. D. Federmeier, Revisiting the incremental effects of context on

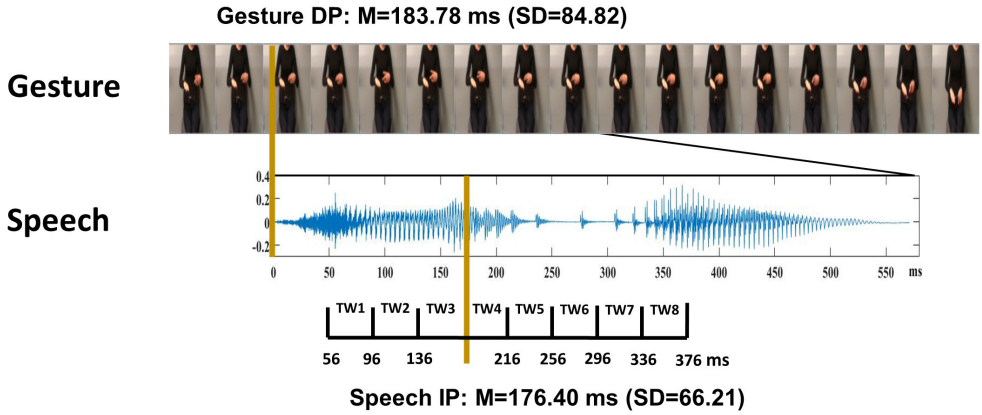
word processing: Evidence from single-word event-related brain potentials. *Psychophysiology* **52**, 1456-1469 (2015).

19. 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).
20. 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).
21. 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).
22. 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).
23. P. W. Donhauser, S. Baillet, Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron* **105**, 385-+ (2020).
24. C. E. Shannon, A Mathematical Theory of Communication. *Bell Syst Tech J* **27**, 379-423 (1948).
25. K. D. Federmeier, H. Mai, M. Kutas, Both sides get the point: hemispheric sensitivities to sentential constraint. *Mem Cognit* **33**, 871-886 (2005).
26. S. D. Kelly, C. Kravitz, M. Hopkins, Neural correlates of bimodal speech and gesture comprehension. *Brain and Language* **89**, 253-260 (2004).
27. Y. C. Wu, S. Coulson, Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology* **42**, 654-667 (2005).
28. 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).
29. 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).
30. 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).
31. 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).
32. L. Fadiga, L. Craighero, E. Olivier, Human motor cortex excitability during the perception of others' action. *Curr Opin Neurobiol* **15**, 213-218 (2005).
33. 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).
34. P. Hagoort, MUC (Memory, Unification, Control) and beyond. *Frontiers in Psychology* **4**, 416 (2013).
35. 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).
36. C. M. J. Y. Tesink *et al.*, Unification of Speaker and Meaning in Language Comprehension:

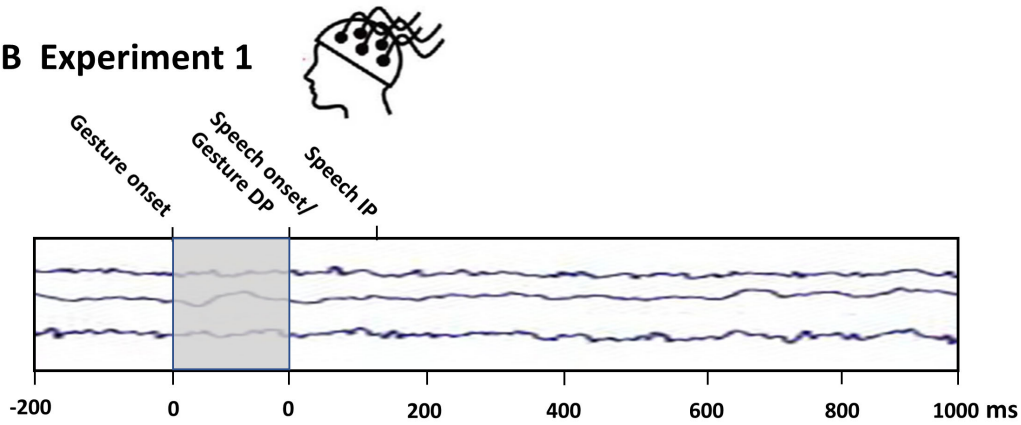
An fMRI Study. *J Cognitive Neurosci* **21**, 2085-2099 (2009).

37. T. O. Bergmann, G. Hartwigsen, Inferring causality from noninvasive brain stimulation in cognitive neuroscience. *Journal of Cognitive Neuroscience* **33**, 195-225 (2021).

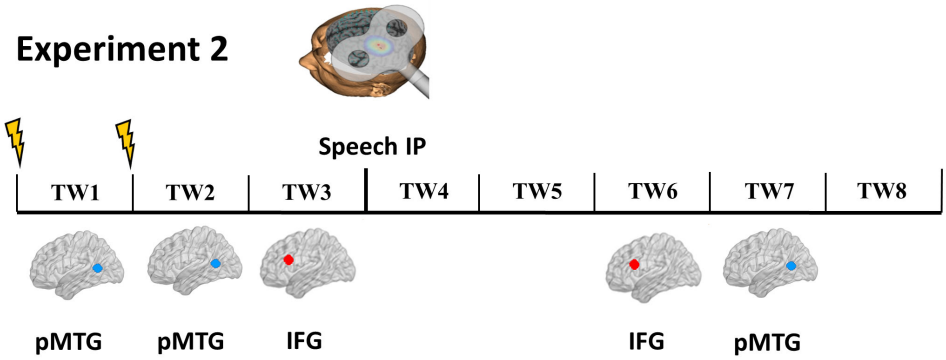
A Experimental stimuli



B Experiment 1

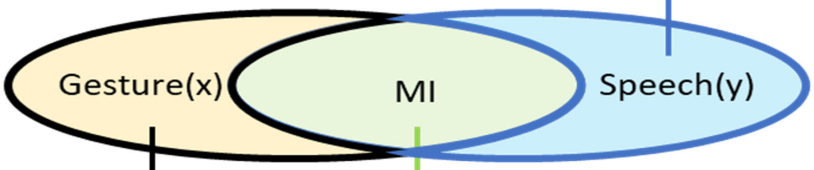


C Experiment 2



D Quantification formulas

$$H(speech) = -\sum_{i=1}^{n} p(yi)\log(p(yi));$$



$$H(gesture) = -\sum_{i=1}^{n} p(xi)\log(p(xi));$$

$$I(gesture; speech) = -\sum_{i=1}^{n} p(xi)\log(p(xi)) - \{-\sum_{i=1}^{n} P(xi, yi)\log(p(xi|yi))\};$$

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.

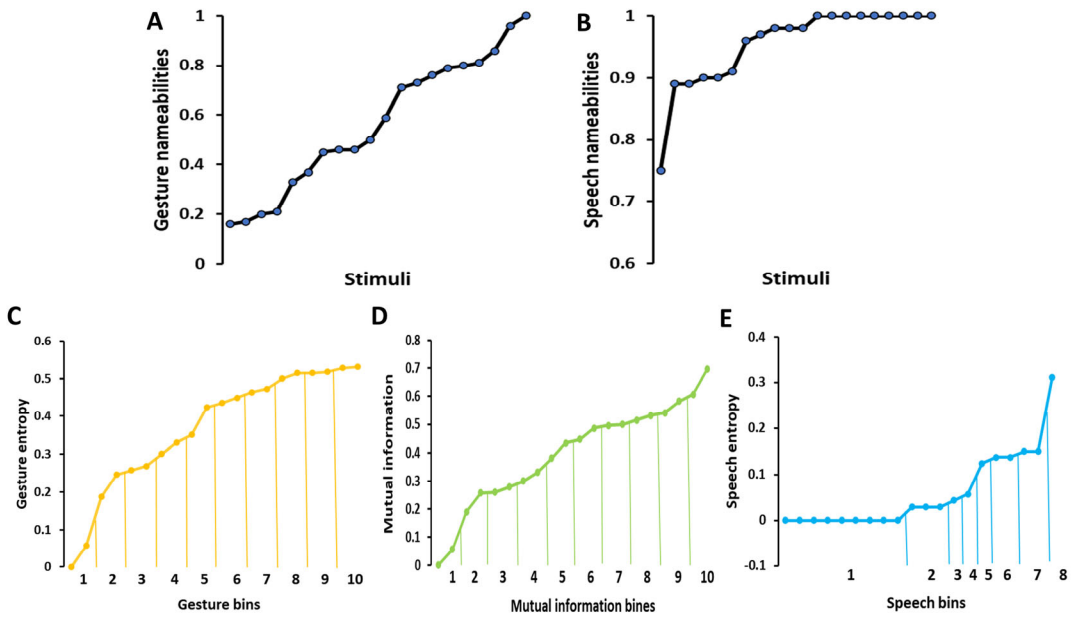


Figure 2. Distributions of gesture and speech nameabilities and transformed Shannon's entropy and MI. (**A** and **B**) Nameabilities of 20 gestures (**A**) and 20 speech words (**B**) were determined as the percentage of participants (N = 30) that provided the most frequently used label when they were asked to use one verb to describe the gesture they saw or the speech they heard. Overall, a mean of .574 (SD =.271) for gesture and .947 (SD =.109) for speech were calculated. (**C** and **E**) Entropy transformed from nameabilities using relevant formula. Gesture (**C**) and speech (**E**) entropy captures the randomness of gestural or speech information and represents the uncertainty of probabilistic representation activated when a certain stimulus occurs. Ranked from low to high, entropy was sorted into 10 bins for gesture entropy and 8 bins for speech entropy with corresponding stimuli. (**D**) MI captures the interplay of gesture and speech and represents distribution of the amount of information obtained about gesture by observing speech. 10 bins of MI were sorted according to the MI score.

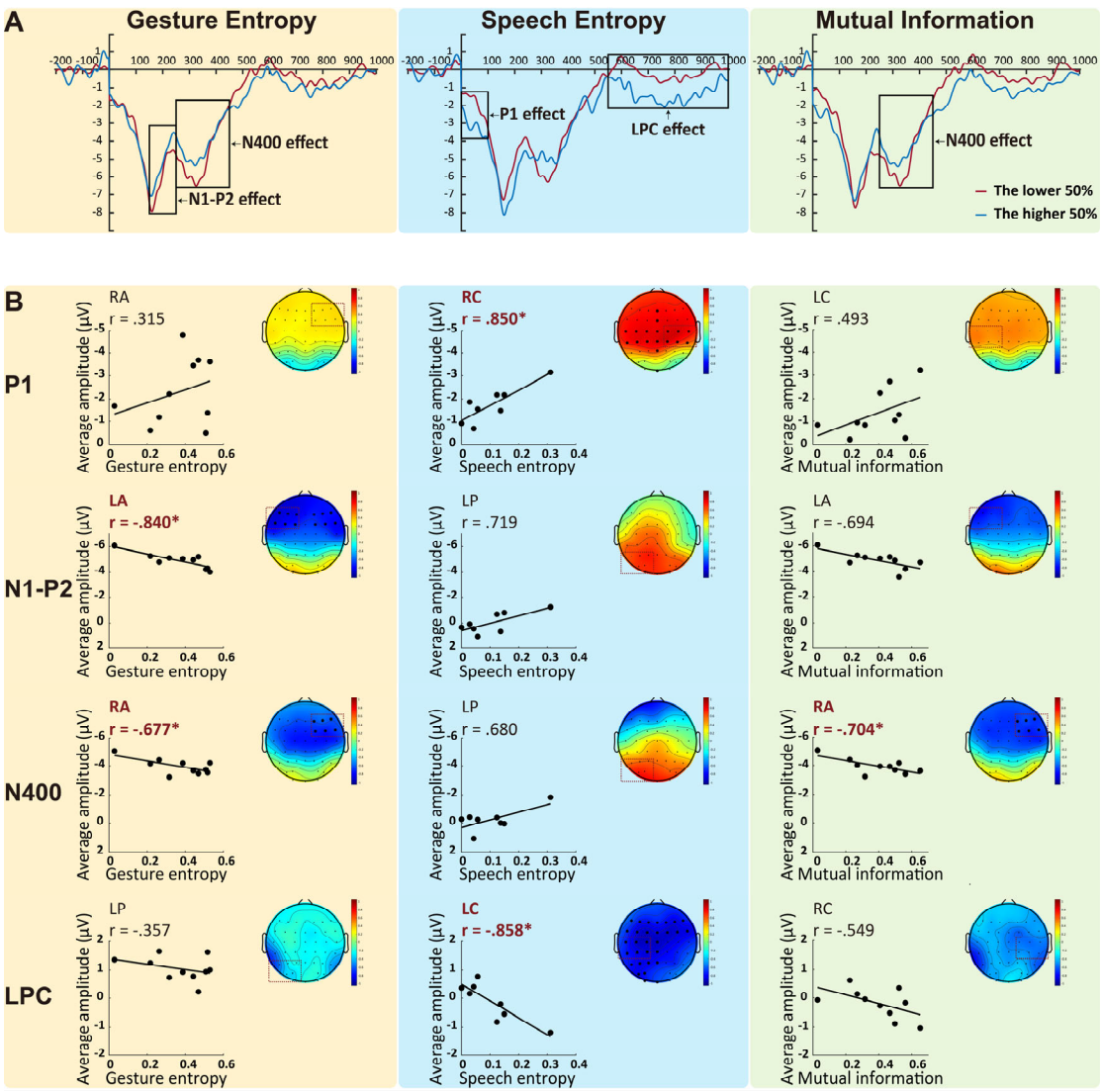


Figure 3. ERP results. (**A**) Grand-average ERPs at the Cz electrode comparing trials with the lower 50% (red lines) and the higher 50% (blue lines) of gesture entropy, speech entropy or MI. Four ERP components in which there was a significant difference between the bottom and top halves were selected (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.

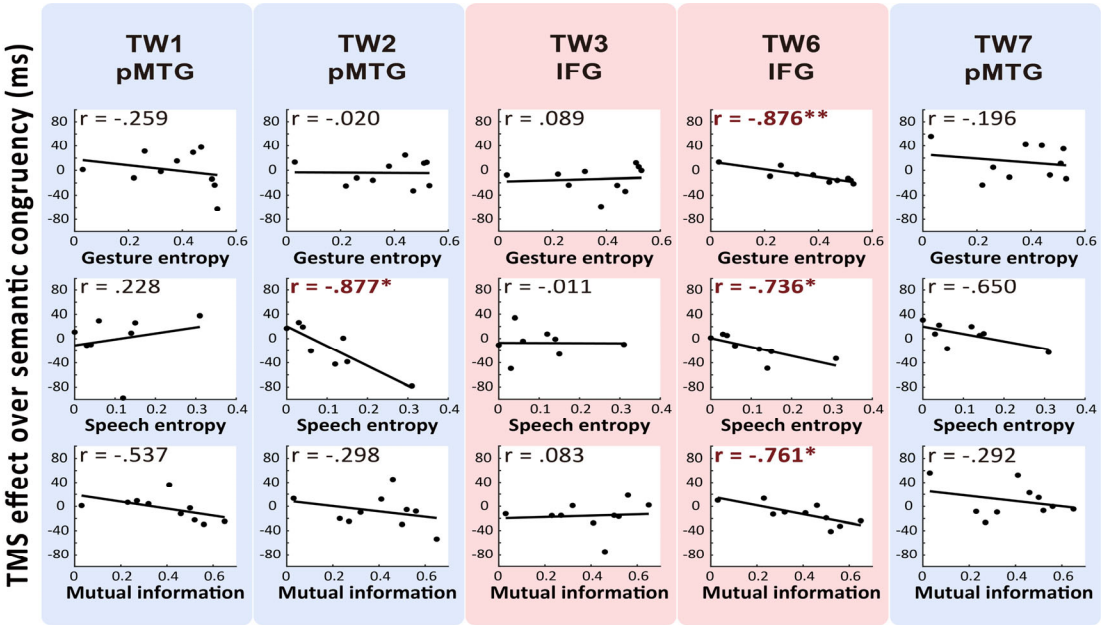


Figure 4. Cortical engagement correlated with gesture entropy, speech entropy and MI. Correlations of the TMS effect over the semantic congruency effect with three information models are displayed with best-fitting regression lines. Significant correlations are marked in red. * p <.05, ** p <.01 after FDR correction.

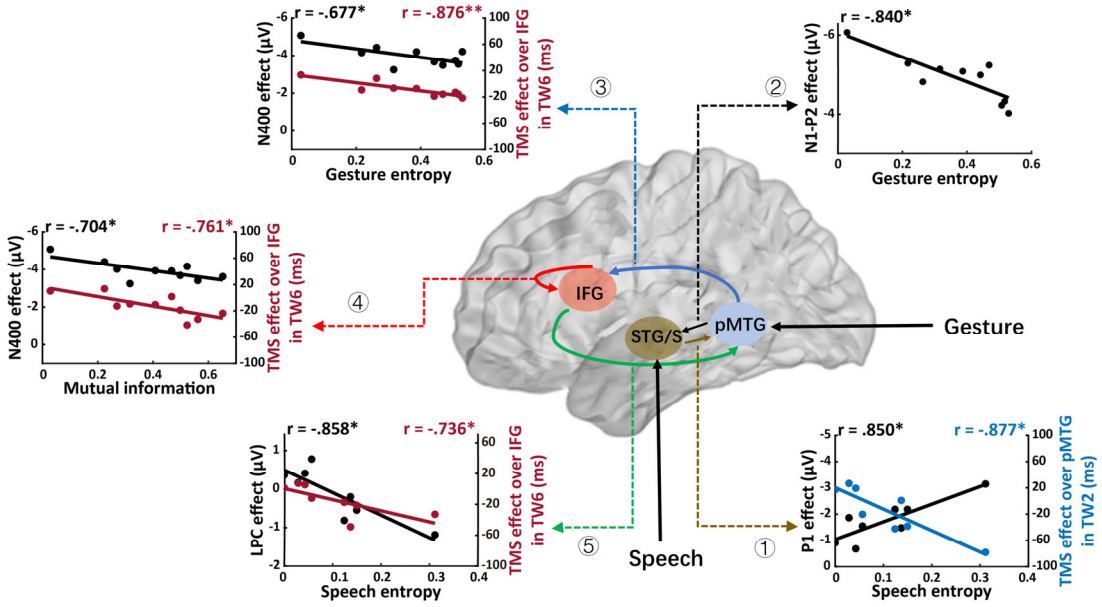


Figure 5. Proposed stages for probabilistic processing of gesture–speech information. Correlations of cortical involvement over semantic congruency effect with three information models were indicated by the Blue line and the Red lines, which reflect the TMS-disruption of the pMTG and the IFG activities respectively. Black lines show the strongest correlations of the ERP amplitudes with three information models. * p <.05, ** p <.01 after FDR correction.