

Tripartite organization of brain state dynamics underlying spoken narrative comprehension

Lanfang Liu, Jiahao Jiang, Hehui Li, Guosheng Ding 

Department of Psychology, School of Arts and Sciences, Beijing Normal University at Zhuhai, Zhuhai, China • Center for Cognition and Neuroergonomics, State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University at Zhuhai, Zhuhai, China • State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University & IDG/McGovern Institute for Brain Research, Beijing, China • Center for Brain Disorders and Cognitive Sciences, Shenzhen University, Shenzhen, China

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

 A newer version is available. [Read the latest version.](#)

Version of Record

January 21, 2025

Reviewed Preprint

v2 • January 6, 2025

Revised by authors

Reviewed Preprint

v1 • October 8, 2024

eLife Assessment

This study provides **important** insights into the brain activity and connectivity underlying speech comprehension, revealing three brain states. The authors present **compelling** evidence by leveraging hidden Markov modeling of fMRI data to link brain state dynamics to comprehension scores, though the functional role of these states remains under-explored. These findings advance our understanding of how brain state transitions in narrative comprehension relate to stimulus-specific features.

<https://doi.org/10.7554/eLife.99997.2.sa3>

Abstract

Speech comprehension involves the dynamic interplay of multiple cognitive processes, from basic sound perception, to linguistic encoding, and finally to complex semantic-conceptual interpretations. How the brain handles the diverse streams of information processing remains poorly understood. Applying Hidden Markov Modeling to fMRI data obtained during spoken narrative comprehension, we reveal that the whole brain networks predominantly oscillate within a tripartite latent state space. These states are respectively characterized by high activities in the sensory-motor (State #1), bilateral temporal (State #2), and DMN (State #3) regions, with State #2 acting as a transitional hub. The three states are selectively modulated by the acoustic, word-level semantic and clause-level semantic properties of the narrative. Moreover, the alignment with the best performer in brain state expression can predict participants' narrative comprehension scores. These results are reproducible with different brain network atlas and generalizable to two datasets consisting of young and older adults. Our study suggests that the brain underlies narrative comprehension by switching through a tripartite state space, with each state probably dedicated to a specific component of language faculty, and effective narrative comprehension relies on engaging those states in a timely manner.

Introduction

When listening to a speech, one adaptively samples information from external sound streams, converting them to linguistic expressions stored in the mental lexicon, and integrating those mental expressions with the internalized “mental world” to infer the semantic-pragmatic interpretations and intentions (Berwick, Friederici, Chomsky, & Bolhuis, 2013 [↗](#)). Crucially, those cognitive processes do not occur one after another in a fixed sequence, but are interwoven and occur in a fluid, dynamic manner. At one moment, you might detect the auditory cues such as volume and pitch in the speech. At another, you might recall memories or knowledge in relation to certain words just heard. To effectively understand the speech, you must flexibly and adaptively switch among those cognitive processes. The neural mechanism behind it is still elusive.

An emerging view suggests that flexible and adaptive cognitive functions arise from the dynamic brain, which transiently activates and coordinates distributed neural circuits in response to the changes in external environment and internal demands (Honey, Newman, & Schapiro, 2018 [↗](#); Kelso, 2012 [↗](#)). To capture the complex neural dynamics occurring across large-scale systems of the brain, researchers have conceptualized the brain’s activity as operating on a low-dimensional neural manifold. The dynamics of brain activity can then be modeled as a temporal trajectory within a latent state space, with each latent state characterized by a distinct pattern of brain activities and network connectivities (Langdon, Genkin, & Engel, 2023 [↗](#)). Employing statistical techniques for modeling dynamic systems such as Hidden Markov Modeling (HMM), recent studies have begun to explore the brain dynamics involved in narrative comprehension (Baldassano et al., 2017 [↗](#); Song, Park, Park, & Shim, 2021 [↗](#); Tang et al., 2023 [↗](#)) or movie viewing (Meer, Breakspear, Chang, Sonkusare, & Cocchi, 2020 [↗](#); Song, Shim, & Rosenberg, 2023 [↗](#)). It has been found that the whole brain systematically switches among a limited number of temporal clusters or latent states with distinct spatial features. Moreover, the switching of brain states was modulated by the time-varying stimuli features including event boundary (Baldassano et al., 2017 [↗](#)) and movie annotations (Meer et al., 2020 [↗](#)), and subjective experience including engagement (Meer et al., 2020 [↗](#)), attention fluctuations (Song et al., 2023 [↗](#)), emotional changes (Tan, Liu, & Zhang, 2022 [↗](#)) and narrative integration (Song et al., 2021 [↗](#)). Those findings demonstrate the functional relevance of brain state dynamics. Nevertheless, how neural state dynamics contribute to the different streams of cognitive processing that ebb and flow with the unfolding of speech is still elusive.

In this study, we explored how language comprehension arises from the dynamic interplay of large-scale brain networks. According to the psycholinguistic theory, the basic design of language faculty mainly comprises three modules (components): an external sensory-motor module, an internal conceptual-intentional module, and a basic linguistic module which represents mental expressions formed by syntactic rules and connects the other two modules (Berwick et al., 2013 [↗](#)). Built upon this theory, we hypothesize that brain dynamics underlying narrative comprehensions would predominantly oscillate within a tripartite latent state space, with each latent state primarily dedicated to a specific component of language faculty. Furthermore, we hypothesize that effective speech comprehension would rely on engaging these states in a timely manner.

To test the above tripartite-state-space hypothesis, we collected fMRI data from 64 young adults as they listened to 10-min real-life narratives. The HMM was applied to model the dynamics of whole-brain network activities. We expect the dynamics of whole-brain network activities would be optimally characterized by three latent states with distinct activity patterns. Specifically, one state would mainly activate the auditory and sensory-motor areas, contributing to the perceptual analyses of external sound streams. The second state would mainly activate the language network and frontal-parietal network, contributing to linguistic encoding and information integration. The third state would mainly activate the default mode networks (DMN), contributing to internalized

semantic-conceptual processing. Moreover, according to the proposed architecture of language components, we expect both the externally-oriented and internally-oriented states would be more likely to transit to the second state than directly transiting between each other. To further validate the functional nature of the three states, we investigated how the dynamic changes of state expression probabilities would be modulated by the temporal variation of speech properties. Three speech properties were analyzed, each reflecting progressively longer timescales of linguistic information: sound envelope, word-level semantic coherence, and clause-level semantic coherence. Based on previous findings that the timescales of information accumulation vary hierarchically from early sensory areas to higher-order areas (Hasson, Chen, & Honey, 2015 [↗](#); Lerner, Honey, Silbert, & Hasson, 2011 [↗](#)), we expect the occurrence probabilities of three brain states would be selectively modulated by these speech properties corresponding to distinct timescales. Finally, to probe the behavioral significance of the timing of brain states, we examined whether the alignment of a participant with the best performer in the time courses of brain state expression could predict his/her narrative comprehension score. To validate the robustness of results, we also conducted all the analyses using a replication dataset consisting of older adults.

Results

The brain reliably and robustly switches through three latent states

We applied the HMM to infer hidden brain states in 64 participants as they listened to one of three 10-min narratives. The observed variables were BOLD signal time series of nine networks obtained employing a state-of-the-art technique for cortical network communities detection (Ji et al., 2019 [↗](#)) (**Fig. S1** [↗](#), see supplementary material for details). Two criteria were comprehensively considered to determine the optimal number of latent states for the HMM. The first was the effectiveness of a model in capturing and separating patterns in the data, which was assessed by the clustering performance of the model. The second was the degree to which it aligns with prior knowledge about the data, which was evaluated by the model's ability to classify the three narratives. These dual criteria ensure that the selected model would be both statistically robust and cognitively sensible (Pohle, Langrock, Van Beest, & Schmidt, 2017 [↗](#)).

Across a range of candidate models with K from 2 to 10, the model's clustering performance tended to decrease with larger K, whereas the accuracy in classifying narrative contents tended to increase. The model with K=3 achieved the best overall performance, quantified by summed z scores (**Fig. 1** [↗](#)). When applying a different whole-brain parcellation scheme (Yeo-7 Networks atlas) to extract brain time series used for HMM inference, we also found the model with K=3 to be the optimal (Fig.S2). Moreover, when examining the replication dataset wherein participants' age, narrative contents as well as scanning duration differed substantially from those of the main dataset, we again found the model with K=3 to be optimal (**Fig. S3** [↗](#)). The robustness of findings suggests the tripartite state space likely captured some fundamental processes of the brain involved in narrative comprehension.

To further establish that the above tripartite-state organization was not trivial, we examined whether the three states would be reconstructed from smaller, more transient states. To this end, we applied a hierarchical clustering algorithm to the transition probability matrix derived from HMM models with 4, 10, and 12 states, and obtained three clusters for each. Those model orders were examined since they have been reported to be the optimal number in previous studies (Meer et al., 2020 [↗](#); Song et al., 2021 [↗](#); Song et al., 2023 [↗](#); Vidaurre, Smith, & Woolrich, 2017 [↗](#)). In this approach, states assigned to the same cluster were more likely to switch within themselves than switching to states belonging to other clusters, and such clusters have been called metastates in the literature (Vidaurre et al., 2017 [↗](#)). We then examined whether there was significant and exclusive correspondence between the clustered states and the three target states (from the HMM

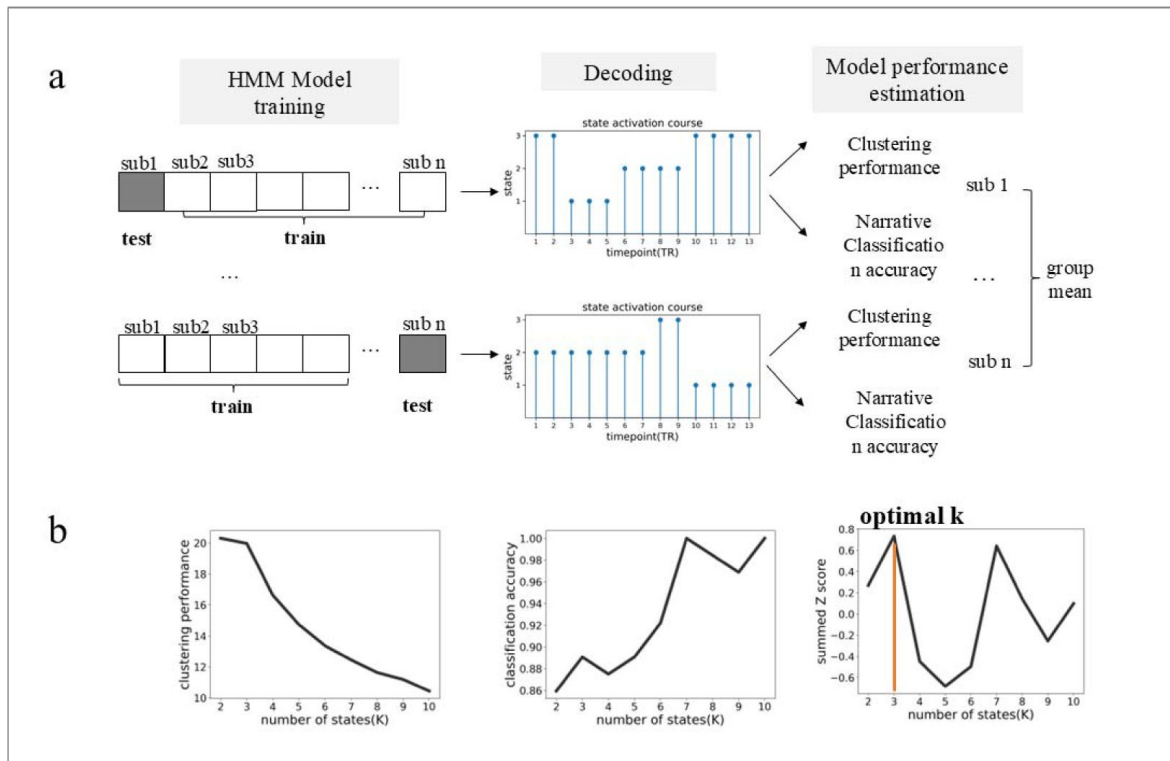


Figure 1.

Identifying the optimal number of brain latent states based on the criterion of statistical robustness and cognitive sensibility.

(a) For each candidate K ranging from 2 to 10, we trained an HMM model on $n-1$ subjects and applied it to decode the time course of state expression for the test subject. The decoded time course was then used to compute a Calinski-Harabasz score, with a larger value indicating better clustering performance, and to decipher which narrative (out of three) was heard by the subject. The two measurements were first assessed at the individual level and then averaged across participants. (b) Model performance as a function of K . With the increase of K , the model's clustering performance tended to decline while the ability to decipher narrative contents tended to improve. We combined the two indices by converting them independently to z scores and summed them up. Notably, at $K=3$, the summed z-score reached its highest point, therefore it was set as the optimal number of latent states.

with $K=3$). As anticipated, those clusters overlapped well with the target states in terms of both spatial activity patterns and the timing of state expression (**Fig. S4**), suggesting the tripartite-state organization is not trivial, but may reflect some fundamental processes of brain dynamics.

Three latent brain states have distinct spatial features

For each state, the HMM estimated its activity loadings on the nine networks and a functional connectivity matrix between these networks. We found the three latent states exhibited distinct activity patterns corresponding to the neural substrates for the three language components as suggested by the theory (Berwick et al., 2013). The first state (State #1) was characterized by relatively high activities in the auditory and somatomotor networks, along with low activities in the DMN and the cognitive control network (**Fig. 2**). This state seems to be associated with the external sensory-motor module of language faculty. The second state (State #2) was characterized by relatively high activities in the language and the frontal-parietal networks whereas low activities in the somatomotor and auditory networks, seemingly being associated with the basic linguistic component. The third state (State #3) was characterized by relatively high activities in the DMN and frontal-parietal networks whereas low activities in the auditory and language networks, seemingly being associated with the internal conceptual-intentional component. We observed similar activity patterns of latent states when using the Yeo-7 network atlas for brain parcellation (**Fig.S2**), and in the replication dataset consisting of older adults (**Fig.S3**).

State #2 acts as a “transitional hub” with high functional integration

According to the linguistic theory, the module for linguistic representation is located in the middle of the external and internal modules, having direct interactions with the other two modules (Berwick et al., 2013). If the hypothesis that the three brain states were associated with each module of the language faculty holds, we expect that State #1 and State#3 would be more likely to switch to State #2 than switching directly to each other. Moreover, the brain occupied by State #2 would exhibit the highest degree of information integration.

To test the first prediction, we examined the between-state switching matrix inferred by the HMM, which showed the probabilities of a state at each timepoint transitioning to another or staying in the same state at the next timepoint. Consistent with our prediction, both State #1 and #3 were more likely to switch to State #2 than switching directly to each other, i.e., State #2 acted as a transitional hub. To confirm that this state-switching tendency was driven by meaningful processes rather than occurring by chance, we made surrogate data by having the nine-network time series circular shifted independently 1000 times. In each iteration, we carried out an HMM analysis with $K=3$, and extracted the difference in transition probability if the inferred states exhibited a similar switching pattern as those from the experiment data. The results showed the differences in transition probabilities observed in our experiment, computed as $P_{(\text{State\#1} \rightarrow \text{State\#2})} - P_{(\text{State\#1} \rightarrow \text{State\#3})}$ and $P_{(\text{State\#3} \rightarrow \text{State\#2})} - P_{(\text{State\#3} \rightarrow \text{State\#1})}$, were respectively larger than 99.9 % and 94.4% of instances from the surrogate data (**Fig. 2**). In agreement with the between-state switching pattern, the brain spent most of the time on State #2 (mean FO = 46.7%), next on State #1 (mean FO =29%) and the least on State #3 (mean FO = 24.3 %). The same pattern was found in the dwelling time, with a group mean of 15.29s for State #2, 9.99s for State #1, and 9.68s for State #3. Both the FO and dwell time of State #2 were higher than 99.88% of instances in the surrogate data.

To test the second prediction, we applied the graph theoretical analyses to assess the global efficiency and modularity of the whole-brain networks when occupied by each of the three states. Consistent with our prediction, when occupied by State#2, the brain exhibited significantly higher global efficiency than when occupied by the other two states (t values > 4.67, $ps < 10^{-4}$). An opposite pattern was found in network modularity (t values < -5.82, $ps < 10^{-6}$). These results indicate that, when occupied by State #2, the whole-brain networks were well connected to enable

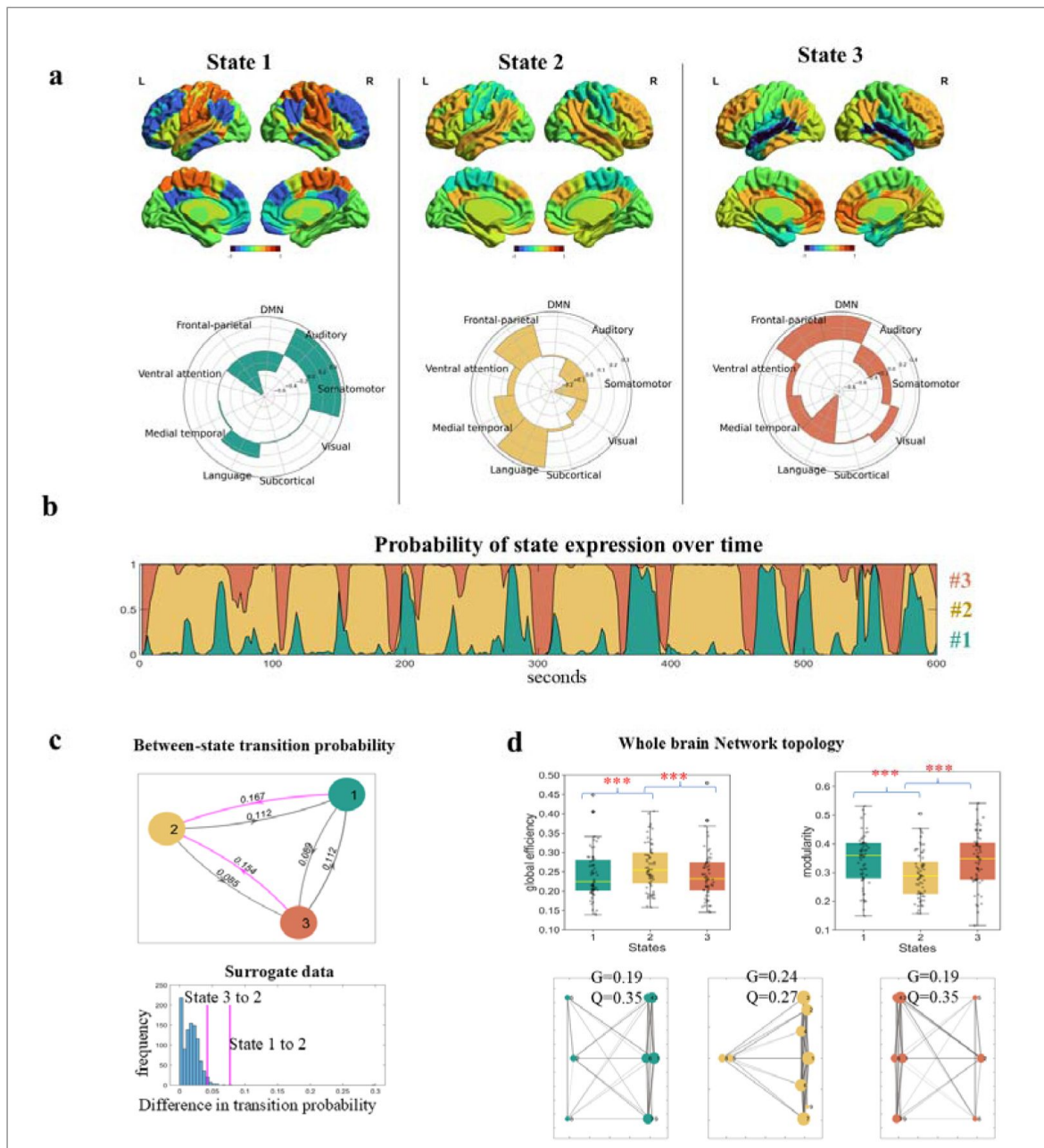


Figure 2.

Spatial and temporal features of latent states revealed by HMM.

(a) The activity loadings of each state on the nine networks. These activity values preserved the scale of the original BOLD time courses which were z-score normalized. (b) The ebb and flow of state expression over the time course of narrative comprehension, plotted using data from a representative participant. The curves of the three states are stacked showing the relative strength of activation probability at each time interval. (c) Between-state transition probabilities. Both State #1 and State #3 were more likely to switch to State #2 (denoted by the pink lines) than switching directly to each other. The differences in transition probabilities were larger than most of the instances from surrogate data. (d). Topological properties of whole-brain networks when occupied by each of the three states. Brain occupied by State #2 demonstrated the highest global efficiency (G) and the lowest modularity (Q). The upper panel shows the results of graph constructed using state-specific time series extracted from individual participants. The lower panel shows results of graph constructed using FC matrix derived from the HMM. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.0005$ for t -tests.

efficient information integration across distinct functional systems. In contrast, when occupied by State #1 and State #3, the whole-brain networks were well separated which enabled functional specialization. The findings were consistent whether using state-specific time series from individual participants to construct the FC matrix or taking the FC matrix derived from the HMM (Fig. 2 [↗](#)).

The between-state switching and topological properties are replicable using the different brain network parcellation scheme (Fig.S2) and generalizable to the replication dataset consisting of the older adults (Fig.S3).

Expression of brain states is selectively modulated by narrative properties

To more directly establish the association of the three brain states to the theoretical language modules, we investigated how the expression of brain states would be modulated by changes in the stimulus properties as the narrative progressed. Three distinct stimuli properties presumably reflecting an increasingly deeper level of information conveyed by the narrative were extracted, including speech envelope, word-level semantic coherence and clause-level semantic coherence.

Speech envelope captures the smoothed, slow amplitude fluctuations of the speech signal over time, which is the perceptual property of the stimuli. We observed a consistent positive correlation between the speech envelope and the expression probability of State#1 across participants, with a group mean of 0.035. This correlation was significantly greater than zero ($t_{(63)} = 2.67$, $p = 0.009$, FDR corrected) and surpassed 99.8% of the 5,000 instances from permutation tests where the state time courses were randomly circularly shifted, yielding a mean of -0.0005 (Fig. 3 [↗](#)). A similar effect was observed for State #2, with a group mean $r = 0.034$ ($t_{(63)} = 2.61$, $p = 0.011$, FDR corrected).

Word-level semantic coherence was evaluated using cosine similarity between embedding vectors for each word and the preceding word. Among the three states, only the expression probability of State #2 showed a consistent correlation with word-level semantic coherence across participants, with a group mean of 0.030 (one-sample t-test against zero: $t_{(63)} = 2.48$, $p = 0.015$, FDR corrected). This group mean value was significantly greater than 99.28% of instances from permutation tests, which produced a mean $r = 0.001$.

Clause-level semantic coherence was evaluated using cosine similarity between embedding vectors for each clause and the preceding clause. Among the three states, only the expression probability of State #3 exhibited a consistent correlation with clause-level semantic coherence, with a group mean of 0.031 (one-sample t-test against zero: $t_{(63)} = 2.89$, $p = 0.005$, FDR corrected). This group mean value was significantly greater than 99.8% of instances from the permuted data, which yielded a mean $r = 0.0001$.

We noted that the correlation value between individual brain state dynamics and speech features was relatively low (group averaged $r \approx 0.03$), which is also the case in previous studies (e.g., [Fernandino, Tong, Conant, Humphries, & Binder, 2022 \[↗\]\(#\)](#); [Oota et al., 2022 \[↗\]\(#\)](#)). This low correlation is likely due to the high level of noise inherent in fMRI data. Brain activity fluctuations are shaped by a variety of factors, including task-related cognitive processing, internal thoughts, physiological states, as well as arousal and vigilance. Additionally, the speech features we measured may account for only a small portion of the cognitive processes underlying the task. As a result, the variance in speech features may only explain a limited portion of the overall variance in brain state fluctuations. Nonetheless, the brain-stimuli correlation achieves statistical significance, and is reproducible across different populations and narratives, providing robust support for the reliability of the results.

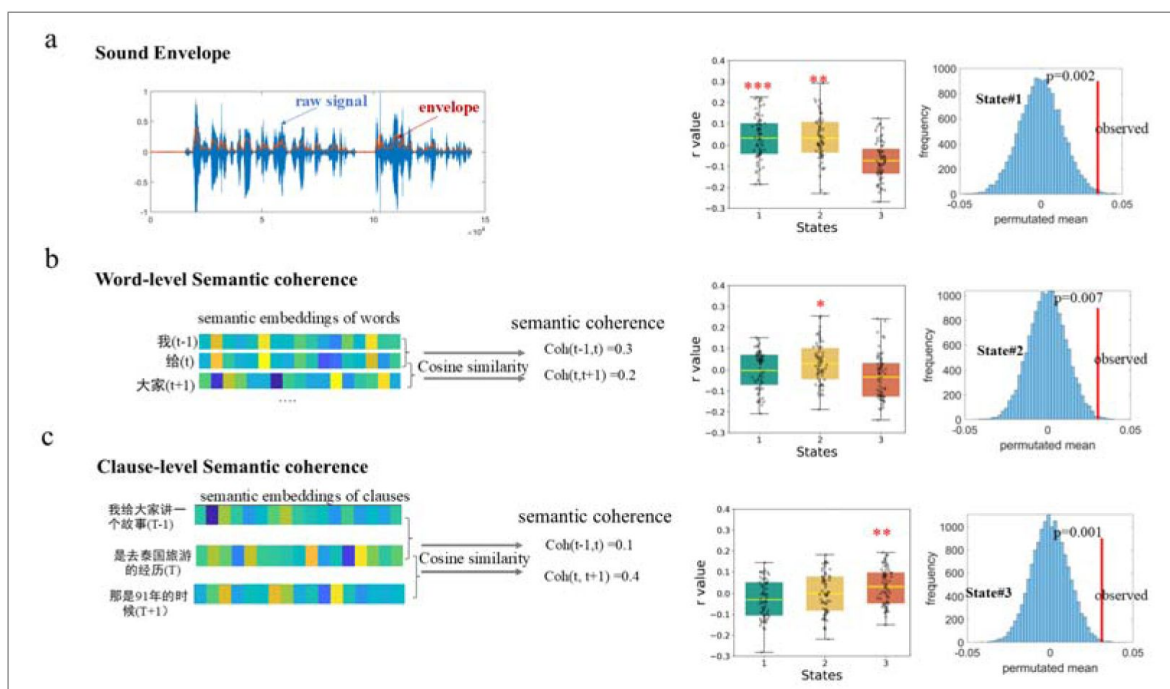


Figure 3.

Selective modulation of state expression by different narrative features.

(a) The expression probability of State #1, as well as of State #2, was positively modulated by the temporal envelope of speech. The blue shaded area represents the original amplitude of a speech signal clip, while the orange line shows the smoothed, low-frequency contour of amplitude changes over time (i.e., the sound envelope) **(b)** Only the expression probability of State #2 was modulated by word-level semantic coherence. **(c)** Only the expression probability of State #3 was modulated by clause-level semantic coherence. Semantic coherence was measured by cosine similarity between the embeddings (obtained by BERT) of each word (or clause) and the word (clause) immediately before it. Those effects were greater than most instances from permutation where the time courses of state expression were randomly shuffled 5000 times. $*p < 0.05$, $**p < 0.01$, $***p < 0.005$ for t -tests.

In sum, the selective modulation by the different aspects of narrative properties provides further evidence supporting the functional relevance of three latent brain states to different language components. These results are replicated with the different brain network atlas. On the replication dataset, we also observed selective modulation effects of speech envelope on State#1 and word-level semantic coherence on State#2; however, no modulation effect of clause-level semantic coherence was found (**Fig. S3** [↗](#)).

Inter-subject correlation in brain state dynamics predict task performance

The above results have demonstrated the functional relevance of the tripartite state space to narrative comprehension. Next, we tested the hypothesis that effective narrative comprehension would rely on engaging these states in a timely manner. To tackle this question, we measured the alignment of brain state fluctuation between each participant (except for the best performers) with that of the best performer(s), then we used the inter-brain alignment index to predict participants' comprehension scores. The best performer was the one (or those) who achieved the highest comprehension score within the subgroup of participants exposed to the same narrative. The rationale is that, if effective comprehension relies on the brain to turn into specific patterns at the right times, the best performer would demonstrate the most "accurate" pattern. Consequently, participants whose brain state fluctuations deviated more (or less alignment) from the "accurate" pattern were anticipated to perform less effectively in the task.

As anticipated, alignments with the best performer(s) in both the State#1 and State#2 were significantly correlated with participants' comprehension scores (Pearson's $r_{(54)} = 0.31$ and 0.36 , respectively). A marginally significant correlation was also found in the alignment of State#3 ($r_{(54)} = 0.22$, $p = 0.10$). As an alternative, we also took the group-mean time courses of brain states expression as the most "accurate" pattern and recalculated the inter-brain alignment value. Even stronger correlations were found between individual-to-group alignments and comprehension scores in all three states ($r_{(62)} = 0.425$, 0.507 , 0.269 for State#1, #2, and #3 respectively) (**Fig. 4** [↗](#)).

In previous studies, the similarity of brain activities across subjects has usually been interpreted as reflecting the inter-subject similarity in the fluctuation of task engagement or attention (Nanni-Zepeda et al., 2024 [↗](#); Ohad & Yeshurun, 2023 [↗](#)), which in turn may be associated with the individual similarity in task performance. We examined whether the above association of inter-subject alignment in brain states with behavior was merely an epiphenomenon of overall task engagement. It is well known that continuous self-reports on task engagement may severely disrupt the ongoing processing of prolonged naturalistic stimuli. As an alternative, studies have demonstrated that head movement serves as a reliable time-resolved indicator for task engagement, with greater task engagement accompanied by decreased movement (Ballenghein, Megalakaki, & Baccino, 2019 [↗](#); Greipl, Bernecker, & Ninaus, 2021 [↗](#); Kaakinen, Ballenghein, Tissier, & Baccino, 2018 [↗](#)). Leveraging this, we computed inter-subject correlations (ISC) in the trajectory of head movement (quantified by framewise displacement) during the fMRI scanning as a proxy for inter-subject similarity in task engagement. Congruent with our assumption, similarities with the best performer in terms of head movement trajectory were indeed positively correlated with participants' comprehension scores ($r_{(54)} = 0.33$, $p = 0.01$) (**Fig. 4** [↗](#)). After adjusting the effect of head movement by applying partial correlation, the positive correlation between the inter-subject alignment in brain states and comprehension scores remained robust (partial r values > 0.29 , $ps < 0.04$). These findings suggest that the inter-subject alignments in brain states were unlikely merely the byproduct of shared levels of task engagement, but instead reflected the commonality in neural processes that directly influence narrative comprehension.

As a comparison, we investigated whether individual differences in the FO and dwell time of latent states were associated with individual difference in narrative comprehension. No significant result was found on any of the three states (r values < 0.15 , $ps > 0.23$). Taken together,

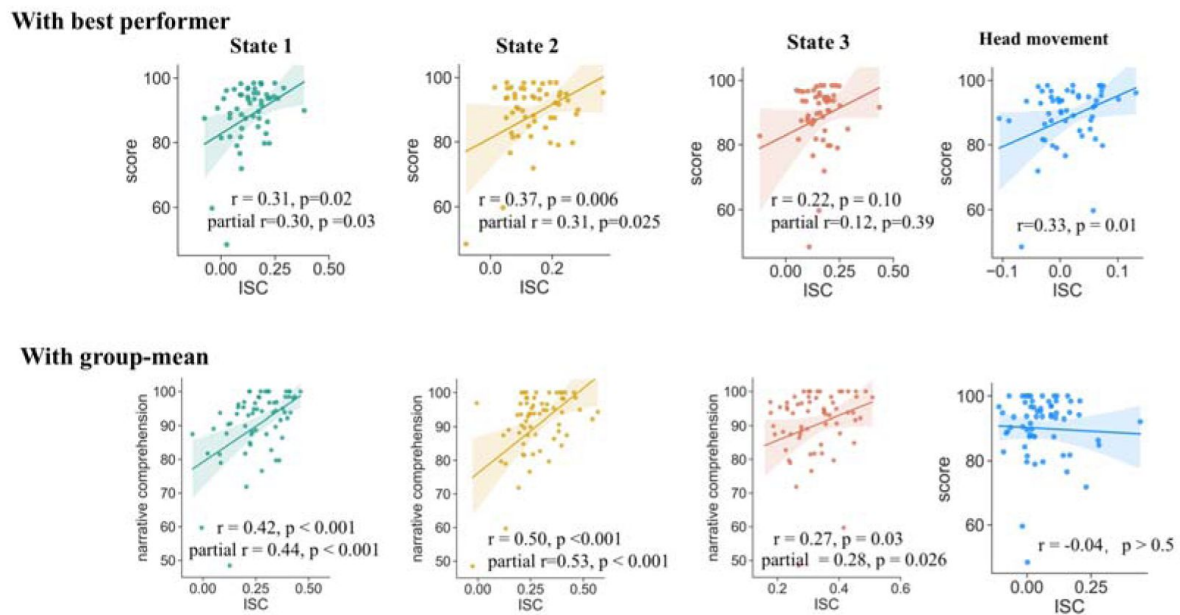


Figure 4.

Correlation of state expression with behavior.

Participants' alignment with both the best performer(s) and the group mean in terms of brain state expression predicted their narrative comprehension scores. The alignment with the best performer in head movement trajectory, which probably reflected inter-subject similarity in the fluctuation of task engagement or attention, also correlated with narrative comprehension. After adjusting this effect using partial correlation, the significant correlations between inter-subject alignment in states expression and narrative comprehension still existed.

these findings suggest that timely engagement with specific brain states, rather than the overall magnitude of engagement in those states, is crucial for narrative comprehension.

These findings were replicable with a different scheme for brain network parcellation (Fig.S3). However, in the replication dataset, we did not find significant positive correlations between inter-subject alignment in brain states and narrative comprehension. This may be due to substantial heterogeneity among older adults. Despite engaging in the same task, older adults exhibit considerable inter-subject variation in cortical morphology, function, and brain metabolism (Yu, 2024 [↗](#)). These factors can diminish the inter-subject correlation of brain state dynamics—indeed, ISCs among older adults were significantly lower than those among younger adults (Fig. S5 [↗](#)) — and potentially reduce ISC's sensitivity to individual differences in task performance.

Comparison between conditions

The above results have revealed a tripartite latent space of whole-brain dynamics, with each state probably subserving a different cognitive component underlying narrative comprehension. Is this temporospatial organization a task-free, intrinsic organization of the dynamic brain, or mainly driven by language processing? To address this question, we compared the brain states involved in narrative comprehension task with those of the same participants when they listened to an unintelligible narrative (told in Mongolian, MG) and during rest. Note, the involvement in linguistic computations decreased monotonically across the three conditions.

HMMs with $K=3$ were conducted separately for the resting and MG conditions. To establish correspondence between states in the resting and MG conditions with those in the task condition, we examined the correlation in network activities for each state identified in the rest (treated as “candidate state”) with each of the three states identified in the task (treated as “predefined state”). The predefined state that a candidate state most closely resembled was designated as its matching state.

For the resting condition, two candidate states showed the highest similarity to State #2 in the task condition. The most similar candidate ($r_{(7)} = 0.778$) was labeled as State #2. The other candidate was assigned to the predefined state (State #3) with the second-highest correlation, denoted with a prime symbol (State #3') to indicate this confusion. The last candidate state matched best with State #1 without any confusion. Analysis of the between-state transition matrix revealed that, unlike in the task condition, State #3 acted as the transitional hub in the resting condition (Fig. 5a [↗](#)).

For the MG condition, two candidate states also showed the highest similarity to State #2. The most similar candidate ($r_{(7)} = 0.687$) was labeled as State #2, while the other was labeled as State #1'. The final candidate state best matched with State #3 without any ambiguity. Consistent with the task condition, State #2 served as the transitional hub under the MG condition (Fig. 5b [↗](#)).

Notably, the FOs of State#2 monotonically increased across the three conditions: resting < MG (incomprehensible speech) < task (narrative comprehension). In contrast, the FOs of the State#3 monotonically decreased: resting > MG > task (Fig. 5c [↗](#)). A similar pattern was found on the dwelling time. These findings provide additional evidence supporting that State#2 was associated with linguistic computations, while State#3 was associated with internalized mental activities. Together, these results suggest that the tripartite latent space of whole-brain dynamics is mainly driven by language processing.

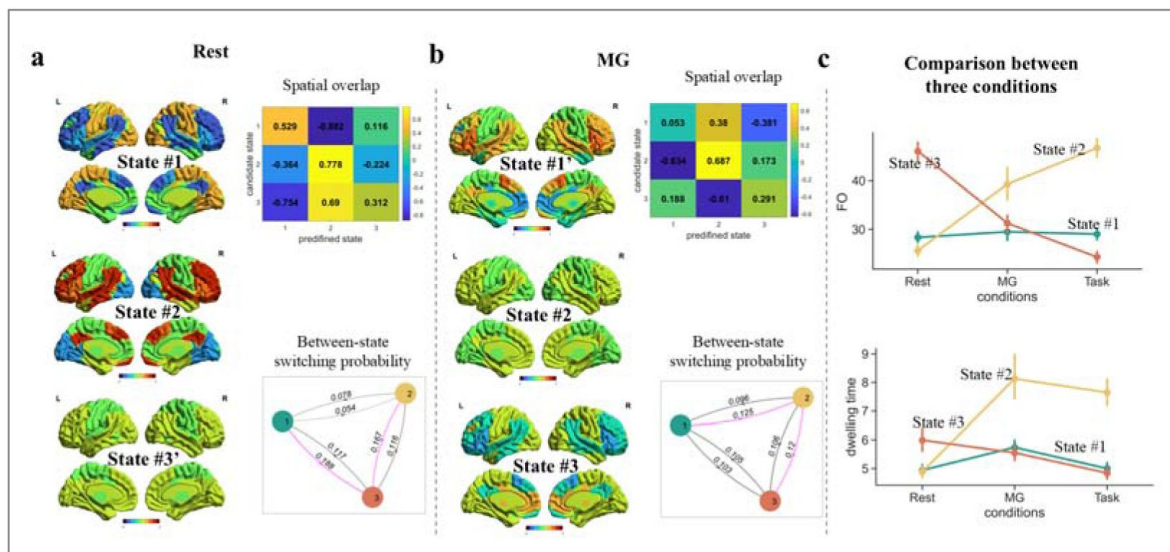


Figure 5.

Comparisons of brain states across conditions.

(a) During rest, the activity patterns of latent states were similar to those during narrative comprehension, but State #3 became the transitional hub. (b) When listening to the unintelligible narrative (in Mongolian, MG), the activity patterns of latent states varied substantially from that during narrative comprehension, but State#2 was still the transitional hub. (c) The fractional occupation of State#2 increased with greater involvement in linguistic computations, while that of State#3 decreased. A similar pattern was found on the dwelling time of states.

Discussion

Speech comprehension is a sophisticated cognitive task that requires the dynamic interplay of various processes, from basic sound perception to complex semantic-pragmatic interpretations, all of which are fluidly coordinated in real time as the speech unfolds. Here, we explored how the brain transiently activates and coordinates distributed neural networks to support the diverse cognitive streams underlying spoken narrative comprehension. Applying HMM, we found that the brain reliably and robustly switches through three latent states, which were characterized respectively by high activities in the sensory-motor (State #1), bilateral temporal-frontal (State #2), and DMN (State #3) regions. Among them, State #2 occurred most frequently, acted as a “transitional hub”, and was characterized by the highest level of functional integration. Furthermore, the three states were selectively modulated by the perceptual, word-level semantic and clause-level semantic properties of the speech. Importantly, participants’ alignments with the best performers on the time courses of brain states expression predicted their narrative comprehension scores, indicating effective speech comprehension relies on engaging the specific brain states in a timely manner. Finally, by comparing the comprehension task with the resting and the unintelligible speech conditions, we demonstrated that the tripartite latent state space was mainly driven by language processing.

A set of results convergently suggest that the tripartite state space is not incidental, but likely reflects a fundamental principle governing the brain dynamics underlying narrative comprehension. First, among a range of candidate models with the number of states ranging from 2 to 10, the model with $K=3$ performed the best in terms of separating patterns in the data and decoding narrative contents, being both statistically robust and cognitively sensible. Second, the tripartite latent state space was replicable with different network atlas and generalizable to different datasets. Especially, despite that the two datasets vary substantially in terms of participants (young versus older adults), the contents of narratives and data length, the two groups still exhibited highly similar brain temporal organization that was best captured by the three latent states. Moreover, the spatial and temporal patterns of the tripartite state space can be hierarchically reconstructed from more nuanced state patterns, being “metastates” of the brain.

Intriguingly, the characteristics of the three latent states align well with the theoretical framework concerning the basic design of language faculty (Berwick et al., 2013). The State#1, which was featured by relatively high activities in the auditory and somatomotor networks and more likely to occur when speech sounds were louder, probably corresponds to the external sensory-motor component of language faculty. State#2, which was featured by relatively high activities in the bilateral temporal and the frontal-parietal networks and more likely to occur when the inputting word was semantically related to the word immediately before it, probably corresponds to the basic linguistic component. State#3, which was featured by relatively high activities in the DMN and frontal-parietal networks and more likely to occur when the inputting clause was semantically related to the clause immediately before it, probably corresponds to the higher-level semantic-conceptual component. Moreover, State #2 acted as a transitional hub, with both State #1 and State #3 were more likely to switch to it than switching directly within them. This directed switching pattern was not incidental, as the observed discrepancy in switching probability was greater than most instances from the surrogate data. Additionally, when occupied by State #2, whole brain networks exhibited a higher level of information integration (quantified by global efficiency) than when occupied by the other two states. These patterns are also consistent with the theoretical prediction that the basic linguistic module is located in the middle of the external and internal modules, having direct interactions with the other two. Collectively, these findings demonstrate specific relationship between the tripartite brain latent states and three critical components of language cognition, going beyond the account of arousal or attentional fluctuation for brain state dynamics (Meer et al., 2020; Song et al., 2023; Taghia et al., 2018).

The activity patterns of the brain latent states associated with each of the theoretical components of language faculty are consistent with the findings from earlier studies that have mainly focused on the averaged brain activities over time (Ferstl, Neumann, Bogler, & von Cramon, 2008 [↗](#); Price, 2012 [↗](#)). Extending prior work, our study may provide novel insights about how the different streams of cognitive processing are temporally organized with the unfolding of speech. Specifically, the time-varying probabilities of latent states indicate that the associated cognitive processes underlying speech comprehension may not operate in parallel with equal priority or occur one after another. Instead, while all processes are simultaneously engaged, one process dominates over others and this dominance changes over time, taking the form of mode-switching. This is consistent with the emerging view that internal/external switching processes of neural circuits drive learning (Honey et al., 2018 [↗](#)). Furthermore, we found it was the alignment with the best performer in the time courses of state expression, rather than the overall occupancy of latent states, that positively correlated with participants' task performance, suggesting recruiting these states in a timely manner is the key to effective speech comprehension. These findings may provide a useful guide to understand the development of language ability as well as language disorders.

Our study provides a unifying perspective for two prevailing approaches aiming to understand how the brain produces cognition. The modular approach postulates the brain areas to act as independent processors for a specific aspect of complex cognitive functions, contributing to much of our current knowledge of the relationship between brain and behavior. However, this approach has been criticized for ignoring the multifunctionality of brain structures (Fuster, 2000 [↗](#)). Alternatively, the network approach, which has been growing rapidly in recent years, posits that cognitive functions arise from dynamic interactions within and between distributed brain systems (Bressler & Menon, 2010 [↗](#)). While revealing valuable insights into the operation rules of the brain, the network approach seems to only provide a general descriptive model. It lacks mechanism accounts for how the interactions of large-scale brain networks give rise to the different streams of information processing involved in a cognitive task. Here, by demonstrating the multistability of large-scale brain networks and establishing the close relationship between specific latent states to specific language components, our study raises a hypothesis that could reconcile the modular and the dynamic network approaches to understand the brain function. Specifically, for a given task, the brain follows modular organization where different regions specialize in specific functions. However, the importance of these regions dynamically changes in response to external environment and internal demands. Accordingly, goal-directed behaviors arise from the precise temporal coordination of different functional modules (Vyas, Golub, Sussillo, & Shenoy, 2020 [↗](#)). To test this possibility, future studies could combine fMRI and neuroregulation techniques and assess the change in state dynamics and behavioral performance as a result of intervention.

Conclusion

In sum, our study reveals that the brain involved in narrative comprehension predominantly oscillates within a tripartite latent state space. The spatial and topological characteristics of these states correspond well to the three core components of language faculty as specified in the theory (Berwick et al., 2013 [↗](#)). Moreover, we demonstrate that effective speech comprehension relies on engaging these brain states in a timely manner. These results are largely reproducible with different brain network parcellation schemes, and generalizable to two replication datasets consisting of young and older adults. The findings establish the link of brain dynamics with both ongoing cognitive processing and behavioral outcomes, providing a mechanistic account of how language comprehension arises from the dynamic interplay of large-scale brain networks.

Materials and methods

Participants and experiment procedure

The main dataset consisted of 64 Chinese college students (33 males, aged 19-27 years) scanned with fMRI while listening to a 10-minute narrative in Chinese. The speech played to each participant was randomly chosen from three real-life stories told by a female college student. After the scanning, participants were asked to recall the narrative as detailed as possible, and then answered several questions regarding narrative contents that were not recalled. Two experimenters then independently rated the degree of narrative comprehension for each participant based on the interview. A more detailed description about the experiment is provided in the supplementary material.

The replication dataset came from 30 healthy older adults (12 males, aged 53-75 years) recruited from the residential community near the college. During fMRI scanning, each participant listened to two real-life stories told by a 62-year-old woman, presented with and without background noise. After omitting those with large head movements, a total of 50 runs of fMRI scans were included for subsequence analyses.

This research was approved by the Reviewer Board of Southwest University in China. Written informed consent was obtained from all participants before the study. The dataset has been used in our previous work addressing different questions (Liu et al., 2022 [DOI](#); Liu et al., 2020 [DOI](#)).

MRI acquisition and preprocessing

We used a 3T Siemens Trio scanner in the MRI Center of the Southwest University of China to collect imaging data. Functional images were acquired employing a gradient echo-planar imaging sequence with the specified parameters: repetition time = 2000 ms, echo time = 30 ms, flip angle = 90°, field of view = 220 mm², matrix size = 64 × 64, 32 interleaved slice, voxel size = 3.44 × 3.44 × 3.99 mm³. Structural images were acquired using a MPRAGE sequence with the following parameters: repetition time = 2530 ms, echo time = 3.39 ms, flip angle = 7°, FOV = 256 mm², scan order = interleaved, matrix size = 256 × 256, and voxel size = 1.0 × 1.0 × 1.33 mm³. The preprocessing pipeline includes slice-timing correction, spatial realignment, co-registration to the individual participants' anatomical maps, normalization to the Montreal Neurological Institute (MNI) space, resampling into a 3 × 3 × 3 mm³ voxel size, and smoothing (FWHM = 7mm). The resulting images underwent additional processing, including detrending, nuisance variable regression and high-pass filtering (1/128 Hz). Data with head movement exceeding 3 degrees or 3 mm were omitted from further analyses.

Data analyses

Whole-brain parcellation

The inference for brain dynamic states was conducted at the whole-brain network level. Currently, most brain functional networks reported in the literature are made based on resting-state fMRI data. To better capture the brain network organization during the task, we conducted brain network parcellation applying a state-of-the-art method proposed by (Ji et al., 2019 [DOI](#)), using data from the 64 participants engaged in narrative comprehension. This method employs multiple quality control metrics to ensure the stability and reliability of the network partition, and most importantly, uses parameter optimization guided by well-established neurobiological principles (e.g., the separation of sensory and motor systems). A detailed description of network partition is presented in the supplementary material. The network detection approach identified a total of 11

networks (Fig. S1). Two networks were discarded due to comprising too few nodes (less than three) and nine networks were included for further analyses. By reference to the functional decoding results using NeuroSynth (<https://www.neurosynth.org>), we tentatively labeled the nine networks as the auditory, visual, somatomotor, bilateral language, medial temporal, frontal-parietal, ventral attention, subcortical, and default mode networks. To test the robustness of findings, we also adopted the seven-network atlas (Yeo et al., 2011) for whole-brain parcellation and reconduted the main analyses.

Brain state inference using Hidden Markov Model

We applied Hidden Markov model (HMM) to infer latent brain states during narrative comprehension using the HMM-MAR toolbox (<https://github.com/OHBA-analysis/HMM-MAR>). The HMM model assumes that the observed data are generated through a finite number of latent states, and each state can be characterized respectively by a multivariate Gaussian distribution with mean and covariance. The BOLD time series were first standardized within each participant and each network. Then HMM was fitted using concatenated data from all participants, such that unified brain states could be obtained.

The number of latent states (represented by “K”) is a crucial aspect of the HMM, and it needs to be predetermined before fitting the model. Two criteria were considered to determine the optimal K. The first was a model’s clustering performance, which reflects how well the model can capture and separate different patterns in the data. The second criterion was how well the model aligned with existing knowledge about the data. This criterion was evaluated by the ability of a trained HMM model to decode the narrative content heard by unseen participants. This dual criterion ensures that the selected number of brain states (K) for the HMM is both statistically robust and cognitively meaningful (Pohle et al., 2017). The clustering performance and prediction accuracy were assessed through a leave-one-out cross-validation strategy. In this approach, we trained the HMM using data from all participants except one. For each candidate K, we repeated the training process 10 times, and the instance with the smallest free energy was selected for decoding the latent state sequence of the left-out participant. Utilizing the decoded latent state sequence, along with the participant’s network time series, we calculated the Calinski-Harabasz score as an indicator of the model’s clustering performance, with a higher score indicating better clustering performance. Further, to assess the model’s decoding capability, we applied a K-nearest neighbor algorithm utilizing the decoded latent state sequences to classify which of the three narratives the left-out participant was listening to. A higher accuracy indicates the model has well captured the task information in the data. Both the Calinski-Harabasz score and narrative classification accuracy were acquired from each participant and then averaged across the group. To combine the two criteria, we first converted the Calinski-Harabasz score and narrative decoding accuracy independently to Z scores and then summed them up to create a single composite score.

We repeated the above cross-validation procedure across a range of K from 2 to 10. The K with the largest composite score was set to be the optimal number of HMM states representing the brain dynamics during the narrative comprehension task. Upon determining the optimal number of states, we reconduted the HMM on the data from all participants and chose the instance with the largest model evidence (lowest free energy) from 10 iterations as the final result.

To demonstrate that those hidden states identified by the above analyses was not trivial but potentially reflected several fundamental processes of the dynamic brain, we explored whether they can be reconstructed from smaller, more nuanced patterns using hierarchically clustering (see supplementary material for details).

Analyses of brain state properties

The HMM model generated, for each state, a group-level activation map and a functional connectivity matrix, as well as a between-state transition probability matrix. With these parameters, the probability of each state being active (or expressed) at each time point and the most likely sequence of states (referred to as the Viterbi path) were estimated for each participant. Based on the Viterbi path, the total time spent on each state over the entire duration (referred to as fractional occupancy, FO) and the duration for which a state continuously persisted before switching to another one (referred to as dwell time) were computed for each participant.

Next, we conducted a graph theoretical analysis to assess the degree of functional integration and segregation of the whole brain when occupied by a specific state. For each participant and each state, a weighted and undirected graph was constructed in which the nine networks were represented as nodes, and FCs estimated using network time series corresponding to the specific state were represented as edges. Employing Brain Connectivity toolbox (Whitfield-Gabrieli & Nieto-Castanon, 2012 [\[1\]](#)), we computed network global efficiency as the measurement for functional integration. Functional segregation was measured by a network modularity score using Louvain algorithm with a resolution parameter $\gamma=1$. Then t-tests were employed to examine the differences in these indices across the three states. For validation purpose, we additionally computed the two graph theoretical indices using the state-specific FC matrices estimated by the HMM.

Surrogate data generation and permutation test

To ascertain that the trend in between-state transition was not by chance, we generated surrogate data by having the 9-network time series circular shifted independently. In this approach, the meaningful covariance between networks was disrupted while the temporal characteristics of the time series were retained (Song et al., 2023 [\[2\]](#)). On each permuted data, we conducted an HMM analysis with the optimal K (i.e., $K=3$). If there were two states where both showed a higher probability of transitioning to a third state compared to directly transitioning between them, this instance would be taken as exhibiting a similar switching pattern as to the experiment result. Then the associated differences in the transition probabilities were extracted and averaged between two pairs. Otherwise, the difference for this instance was set to zero. This step ensured that only meaningful differences in transition probabilities were considered. By repeating this procedure 1000 times, we obtained a null distribution for the discrepancy in state transition probabilities.

Modulation of brain state activation by time-varying stimuli features

To gain more insights into the functional nature of brain dynamic states, we investigated how narrative properties would modulate the probability of a neural state being expressed in individuals. Specifically, we focused on three different stimuli properties which were assumed to reflect an increasingly “deeper” level of information conveyed by the narrative, including the temporal changes in acoustic property, and semantic coherence at the word level and at the clause level.

To characterize the temporal variation of acoustic property, we derived the temporal envelope of each story using Hilbert transform, which reflects the overall fluctuation of voice amplitude. The speech envelope was then convolved with the canonical hemodynamic response function (HRF) and down-sampled to 0.5 Hz (the same resolution as the fMRI acquisition). To characterize the temporal variation of semantic coherence, we first transcribed the speech to texts and retrieved the semantic representations for each word applying a large language model BERT that was pretrained on a large-scale Chinese corpus (Cui, Che, Liu, & Yang, 2021 [\[3\]](#); Devlin, Chang, Lee, &

Toutanova, 2018 [\[4\]](#)). The output from the last layer of the model was used as word embedding. To avoid overfitting, we further decomposed the high-dimensional embedding vectors ($N=768$) with principal component analysis (PCA) and retained the first 50 PCs (Goldstein et al., 2024 [\[4\]](#); Goldstein et al., 2022 [\[4\]](#)). Next, a vector of word-level semantic coherence was generated for each narrative by computing the cosine similarity between the embeddings of every word and the word immediately before it. After aligning the onset time of words using Praat (<https://www.fon.hum.uva.nl/praat/> [\[4\]](#)), the semantic coherence vector was convolved with HRF and down-sampled to 0.5Hz. Clauses were encoded by two researchers, each including 8-9 characters on average. The semantic representations for clauses were obtained by averaging the embedding vectors of words within a clause. Using the same method, a vector for clause-level semantic coherence was generated for each narrative.

We first computed Pearson's correlation between the vector of narrative properties and the vector of state expression probability at the individual level. To infer significance, the group mean of correlation values across participants was compared to a null distribution generated by 5000 permutations. For each iteration, the time courses of brain state expression were randomly circular-shuffled, and the correlation between narrative property vector and the shuffled time course of state expression was re-calculated and averaged over participants to create a random value. An empirical p -value was determined by the proportion of values from the 5000 iterations that were larger than the original group-mean value. FDR correction was used to account for multiple comparisons.

Correlation of latent state dynamics with behavior

To assess the importance of the timing of brain latent states to behavior, we examined whether the alignment of participant's brain state fluctuations with that of the best performer could predict their narrative comprehension scores. The best performer was the one (or those) who scored the highest in the narrative recall task within the subgroup of participants exposed to the same narrative. For each narrative, if there were more than one best performer, we first assessed the alignment between a participant with each of them, and then got the average. For the non-best-performers ($N=56$), their brain alignment with the best performer(s) was measured by Pearson's correlation using the time course of state expression probability. After that, we computed a Pearson's correlation between inter-brain alignments and participants' comprehension scores. Considering that the best performer may lack of representativeness, we also measured the alignment of each participant's brain state fluctuations with the group mean. The inter-brain alignment was assessed by iteratively leaving out one participant and calculating the Pearson correlation between the time course of state expression probabilities for the left-out participant and the average time course for the remaining participants. All participants were engaged in the same narrative during this process.

As a comparison, we also investigated whether the overall engagement of a brain state was associated with task performance. For this purpose, we examined the correlation between participants' FO and dwell time in each state and their comprehension scores.

Compare brain states across conditions

Finally, we investigated whether the temporal organization of brain dynamics observed during narrative comprehension was mainly driven by language processing, or instead an intrinsic organization of the dynamic brain. For this purpose, we analyzed the fMRI data from the same group of participants at rest and when listening to an unintelligible narrative told in a foreign language (Mongolian). The scanning parameters as well as the scanning length were identical to the main experiment. The HMM with $K=3$ was conducted separately for the two conditions, then the resulting three brain states were mapped to the corresponding states from the narrative comprehension condition by maximizing the similarity in state activity patterns.

Establishing state correspondence between analyses

State correspondence between the two datasets and across different conditions was established based on spatial overlap. To assess this overlap, we first extracted each state's activity values in the nine networks derived from the HMM. Then, for each state identified in other analyses (treated as a candidate state), we calculated its correlation with the three states from the main analyses (treated as the predefined states) in terms of network activity patterns. The predefined state that a candidate state most closely resembled was designated to be its matching state. For instance, if a candidate state showed the strongest correlation with State #1, it was labeled as State #1 accordingly. In cases where two candidate states showed the highest similarity to the same predefined state (e.g., State #1), the candidate state with the strongest similarity was labeled as State #1, and the other was assigned to the predefined state with the next highest correlation. A prime symbol (e.g., State #2') was used to indicate cases where such confusion occurred.

Given that the spatial layout of the 9-network atlas used in the main analysis does not align well with the Yeo-7 network atlas, we established state correspondence between the two parcellation schemes using evidence of temporal overlap. Specifically, we calculated Pearson's correlation between each state derived from the Yeo-7 network scheme (treated as a "candidate state") and the three predefined states from the main analysis, based on time courses of state expression probabilities. For each state, the time courses from all 64 participants were concatenated, resulting in 19,200 (300×64) time points. The predefined state most similar to each candidate state was designated as its corresponding state.

To assess the statistical significance of the spatial (and temporal) overlap, we utilized the HMM results from 1,000 permutations, where participants' BOLD time courses were circularly shifted and HMMs were conducted for each permutation. Applying the same state-correspondence strategy described above, we generated a null distribution representing the matches between candidate states and states in the surrogate data.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (NSFC: 32471100), Guangdong Basic and Applied Basic Research Foundation (2024A1515030046). No conflict of interest is declared. We greatly appreciate the two anonymous reviewers for their insightful comments and constructive suggestions.

Data availability

The raw and processed fMRI data are available on OpenNeuro: <https://openneuro.org/datasets/ds005623>¹. The experimental stimuli and behavioral data are provided on [Github](#).

Supplementary Material

1. Experimental procedure

Stimuli

The young adults in the main dataset listened to one of three real-life stories told in Mandarin by a female college student. These stories described her experiences with a graduate admission interview, a hiking trip, and her first time taking a flight. The older adults serving as the replication dataset listened to one of two real-life stories told by a 62-year-old woman in a Chinese dialect, describing her experience in Thailand and riding a warship, respectively. The recordings were collected when the speakers were scanned with fMRI. To obtain audio recordings with good quality, we used a noise-canceling microphone (FOMRI-III, Optoacoustics Ltd., Or-Yehuda, Israel) positioned above the mouth of the speaker and further de-noised the audio recordings offline using Adobe Audition 3.0 (Adobe Systems Inc., USA).

Behavioral assessment for the level of speech understanding

After the scanning session, participants were asked to retell the story in detail. Following the free recall, experimenters asked additional questions from a prepared list about details not previously mentioned. The list contained 9-10 questions for each story, which addressed key events from different parts of the narrative (e.g., “Can you describe the preparations she made for the hike?” or “What happened on her way to the hotel?”). Correct answers needed to cover critical details, including actions, characters, locations, time, and motivations. Two independent raters listened to the recordings of participants’ free recall and their responses to the experimenters, assigning a score for each question based on the completeness of the key points provided. The comprehension score for each participant was calculated as a percentage of the total possible points. Given the high inter-rater reliability (Spearman’s $r(62) = 0.74$), the average score from both raters was used to quantify each participant’s level of story comprehension.

The audio files of those stories, along with their corresponding transcriptions and sample questions used in the comprehension test, are available at: <https://github.com/liulanfang11/Tripartite-brain-states/upload/main> .

2 Functional network detections

Network detection was performed following the method proposed by (Ji et al., 2019 ) using fMRI data from the 64 participants engaged in narrative comprehension. First, the mean time series of 246 regions defined by the Brainnetome atlas (Fan et al., 2016 ) was extracted, and a functional connectivity (FC) matrix was computed for each participant and then averaged across them. This atlas covers both cortical and subcortical regions and is made based on both anatomical and functional connectivity (FC) patterns. Next, community detection was performed on the group-averaged FC matrix applying the Louvain clustering algorithm in the Brain connectivity toolbox (<https://sites.google.com/site/bctnet/> ) . Three criteria were taken into account when determining the Gamma parameter in the algorithm, including (1) separation of primary sensory-motor network (visual, auditory and somatomotor) from all other networks (i.e., neurobiologically sensible); (2) high similarity of network partitions across nearby parameters (i.e., statistically stable); and (3) high within-network connectivity relative to between-network connectivity (i.e., high modularity).

A set of gamma values ranging from 1.2 to 2.5, with a step size of 0.01, were tested. For every tested gamma, we ran the algorithm 1,000 times and measured how consistent a given partition was to every other partition using a z-rand score. Each z-rand score averaged across the iterations was then multiplied by its corresponding modularity score to find a modularity-weighted z-rand score. Finally, the gamma value of 2.5 was selected, as it corresponded to the peak of the modularity-weighted z-rand score and satisfied the three criteria, resulting in a plausible number of networks, including the primary sensory/motor networks. We implemented network detection using code published by a prior study (Barnett et al., 2021 ) .

A total of 11 networks were obtained by the above method. Two networks with fewer than three nodes were discarded, resulting in nine networks for subsequent analyses.

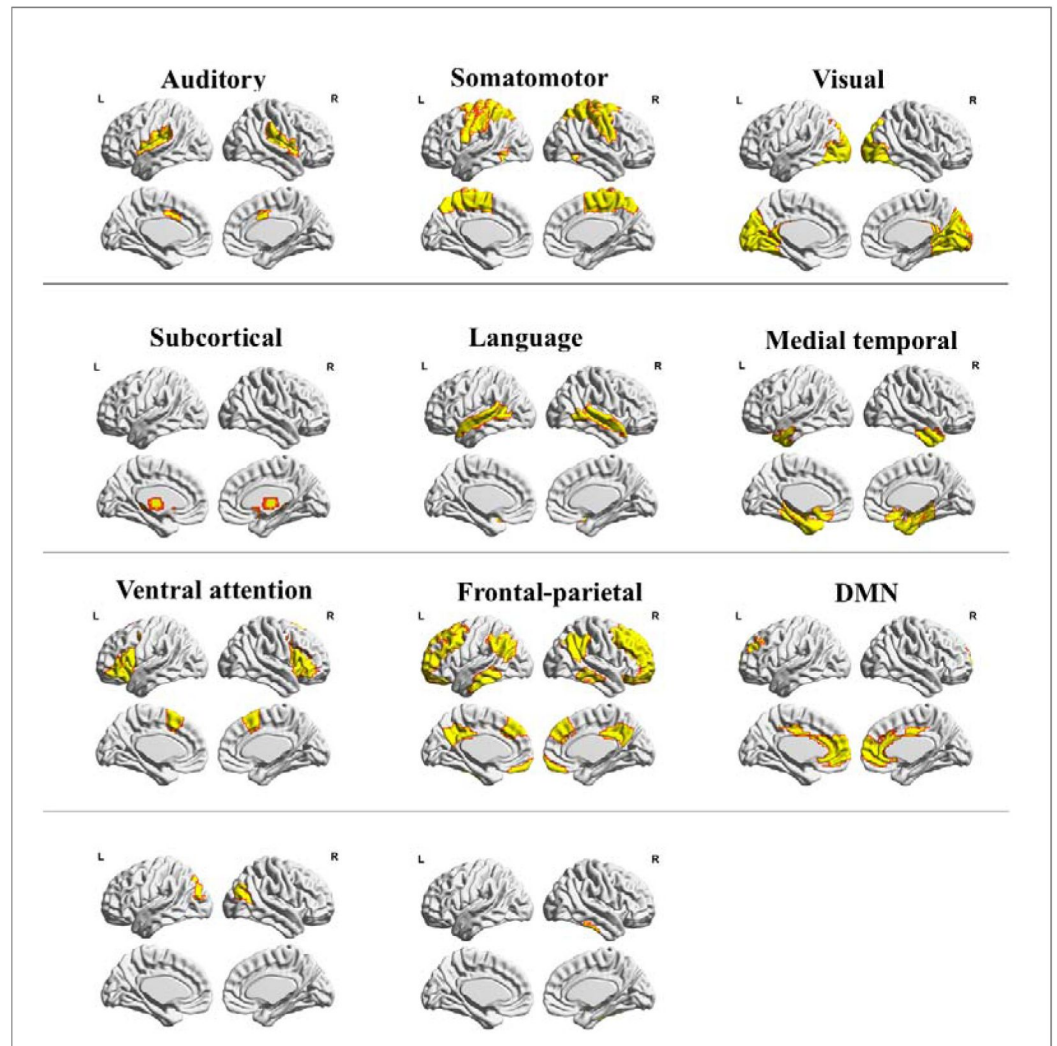


Figure S1.

Spatial maps of the 11 networks derived from whole-brain parcellation using data from 64 participants engaged in narrative comprehension. The two unlabeled networks consisted of only one or two nodes each and were therefore excluded from further analyses.

3. Replicating main results with the Yeo-7 network atlas

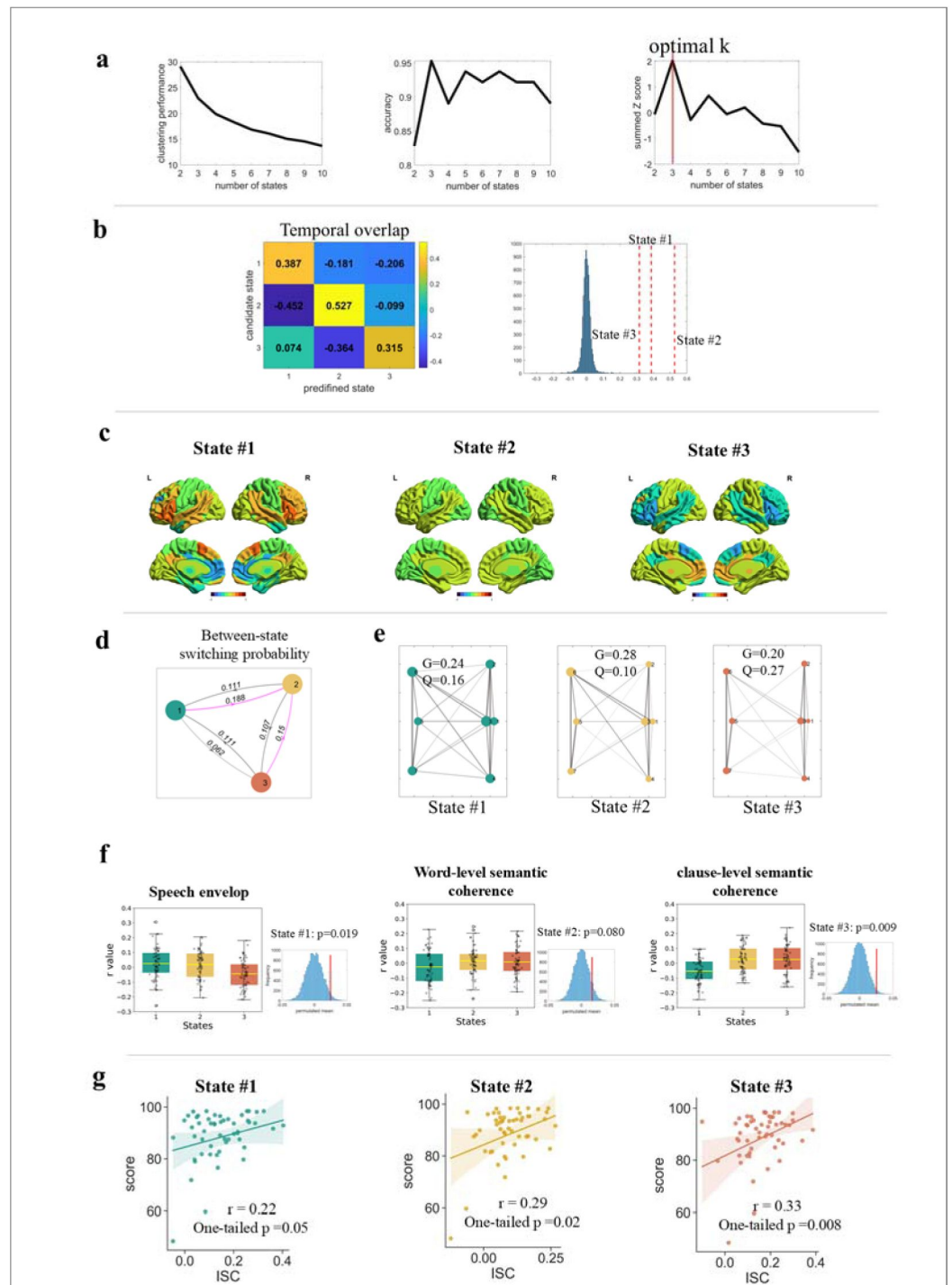


Fig. S2.

Replicating the findings with the 7-network atlas to parcellate brain networks. **(a)** Model selection. The model with $K=3$ achieved the overall best performance in terms of clustering and accuracy in classifying three narratives. **(b)** Establishing State correspondence between parcellation scheme. Each of the three states from the 7-network scheme (treated as candidate states) was best matched to State #1, State #2, and State #3 from the 9-network

scheme (treated as predefined states), respectively. The degree of actual overlap for the three states exceeded 100%, 100% and 99.97% of the permutation instances. **(c)** Activity patterns of latent states. The network activity values preserved the scale of the original BOLD time courses which were z-scored normalized. **(d)** Between-state switching probabilities. State#2 was the transitional hub. **(e)** Topological properties of whole brain networks when occupied by each state. At State#2, the brain exhibited the highest global efficiency and lowest modularity. **(f)** Modulation of state expression probability by narrative properties. **(g)** Alignment with the best performer predicted participants' narrative comprehension performance.

4. Replicating the major results using a dataset of older adults

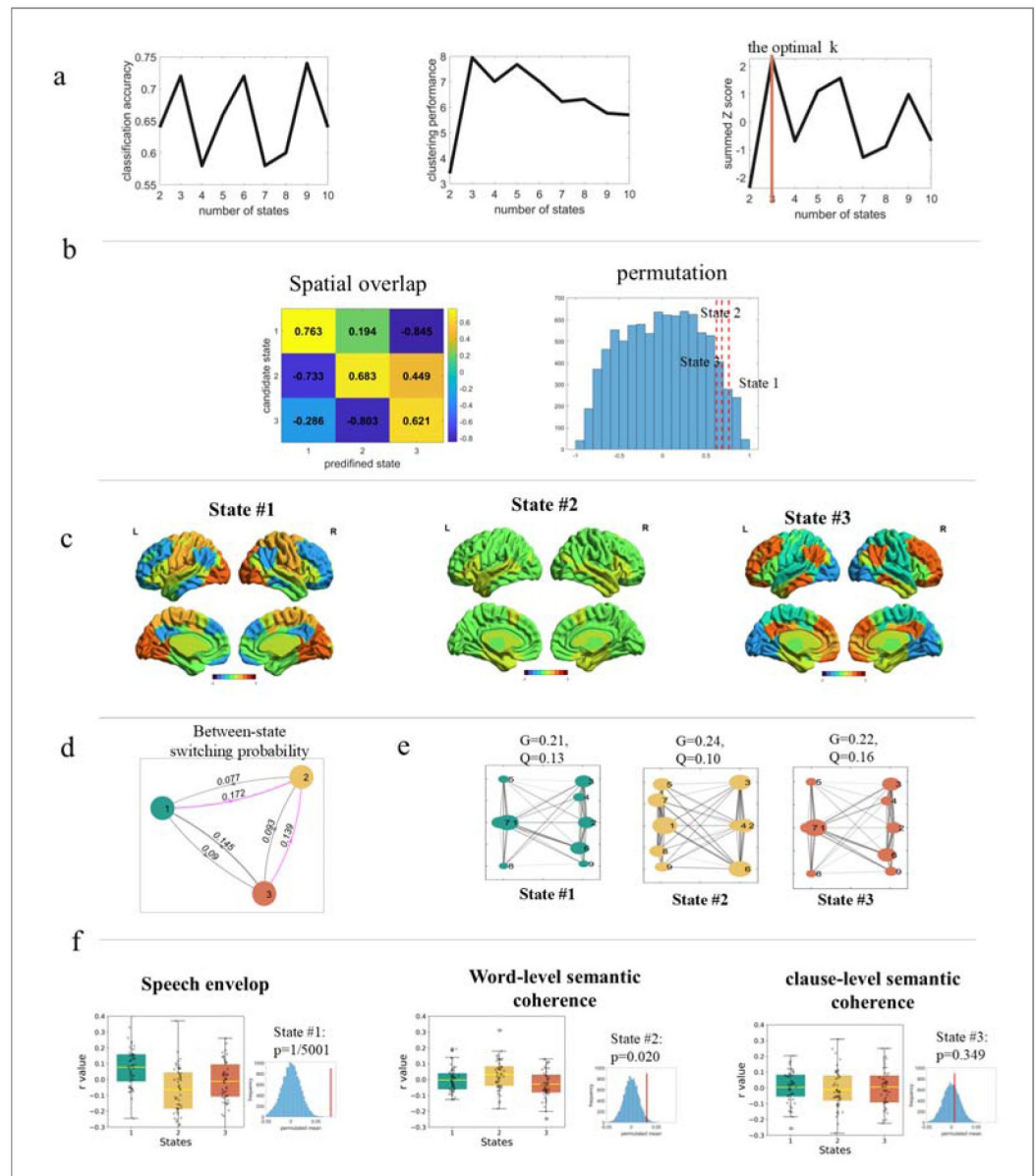


Fig. S3.

Replicating the major results on a different dataset consisting of older adults. **(a)** Model selection. The model with $K=3$ achieved the best overall performance in terms of clustering and accuracy in classifying the two narratives. **(b)** Establishing State correspondence between datasets. Each of the three states from the older-adult dataset (treated as candidate states) was best matched to State #1, State #2, and State #3 from the young-adult dataset (treated as predefined states), respectively. The degree of actual overlap for the three states exceeded 95.7%, 93.04% and 90.30% of the permutation instances. **(c)** Activity patterns of latent states. The network activity values preserved the scale of the original BOLD time courses which were z-scored normalized **(d)** Between-state switching probabilities. State#2 was the transitional hub. **(e)** Topological properties of whole brain networks when occupied by each state. At State#2, the brain exhibited the highest global efficiency and lowest modularity. **(f)** Modulation of state expression probability by narrative properties.

5. Reconstruction of latent states using Hierarchical clustering

We explored whether the latent states derived from HMM with $K=3$ can be reconstructed from smaller, more nuanced patterns. To this end, we applied an agglomerative hierarchical clustering algorithm to the transition probability matrix derived from HMM with $K = 4, 10$, and 12 , respectively, and obtained three clusters from each. Next, we compared the clusters with the target states (i.e., those resulting from the HMM with $K=3$) in terms of similarities in activity patterns (spatial overlap) and the time course of state expression (temporal overlap).

To evaluate the spatial overlap, we first averaged the activation values across those states belonging to the same cluster, merging them into a single new state. Then we assessed the similarity in the activity patterns between the merged states and the target states using Pearson's correlation. To evaluate temporal overlap, we first substituted states belonging to the same cluster with the newly formed one, then computed Jaccard Similarity between the sequences of the new states and the sequences of the targeted states.

There was a clear and exclusive correspondence between clusters reconstructed from both the 4-state and 10-state models and the predefined target states in terms of activity patterns ($r_{(6)}$ ranges from 0.72 to 0.98). The timing of state expression was also well aligned between the reconstructed model and the target model (more than 77% overlap across a total of 19, 200 time points for 64 participants). For the 12-states model, we also found two reconstructed clusters resembling State#1 and State#3 (Fig.S2). Probe on the cluster that deviated most from the target states showed that it consisted of nine states, possibly capturing too many nuanced patterns of neural dynamics. Indeed, when splitting this “big” cluster into two smaller ones, one of them demonstrated significant similarity to State#2 ($r_{(7)}=0.75$).

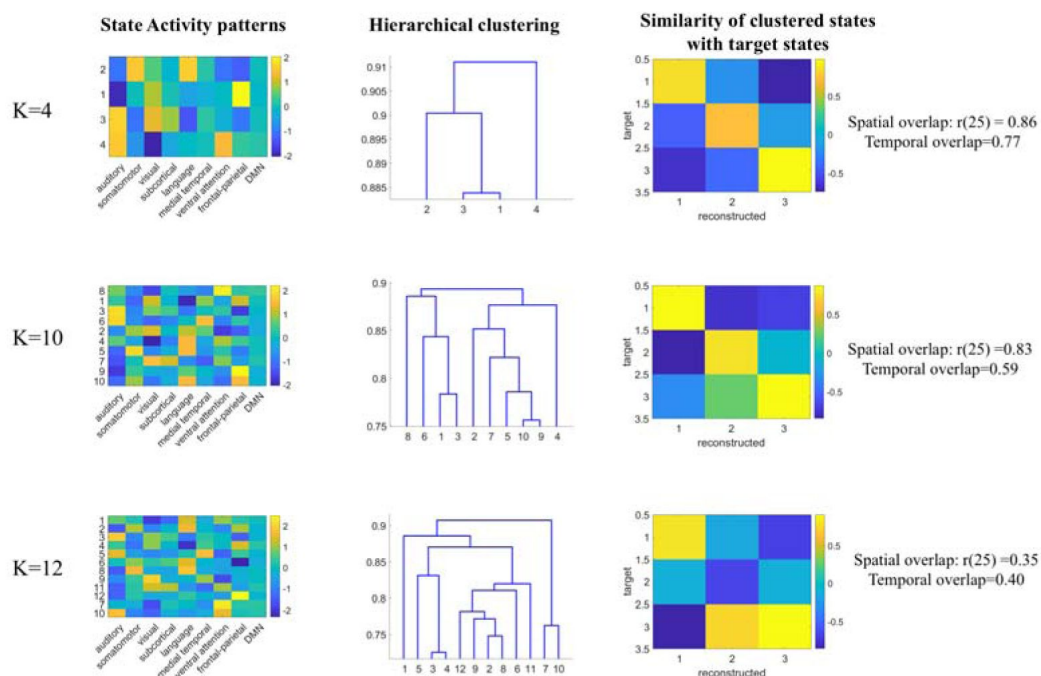


Figure S4.

Reconstructing the tripartite-state space from smaller states.

Left and middle panels: States inferred by HMM with $K=4, 10$, and 12 were hierarchically grouped into three clusters based on between-states transition probability matrix. Right panels: The reconstructed states (clusters) from the 4- and 10-state models show high and exclusive similarity to the three states (targets) inferred by HMM with $K=3$.

6. Between-group comparison in inter-subject correlation (ISC) in brain state dynamics

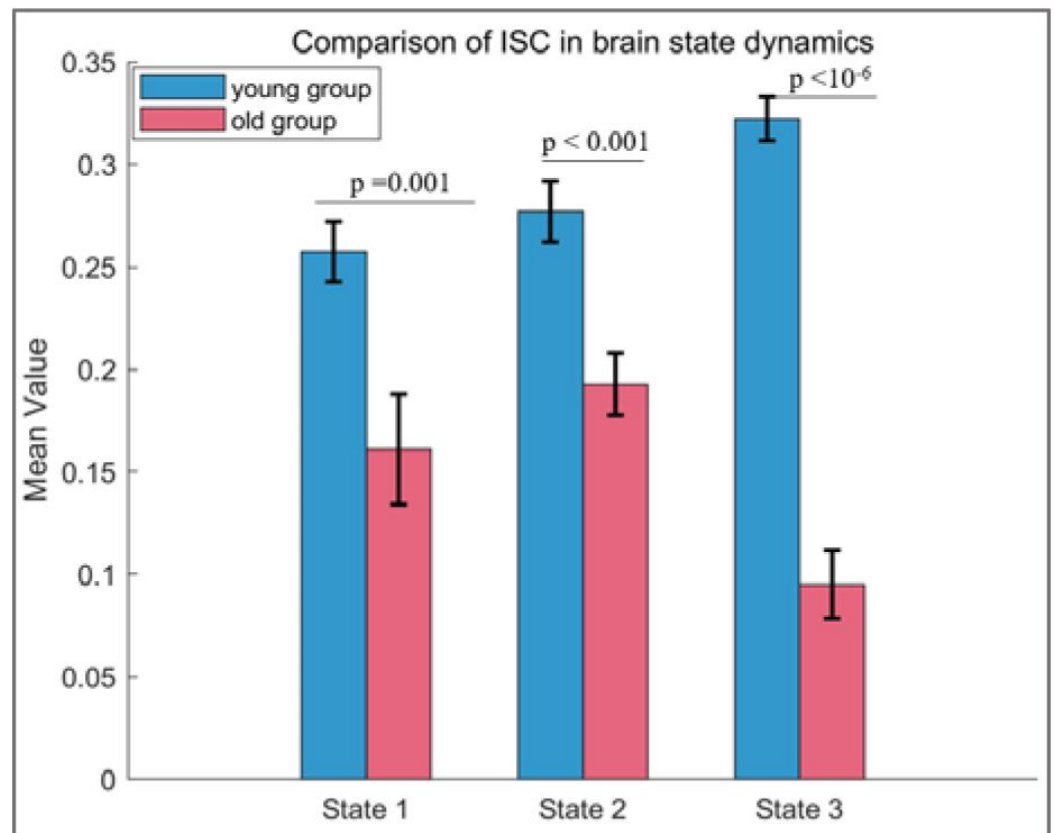


Figure S5.

Comparison in inter-subject correlation (ISC) in brain state dynamics between the young and older age groups. ISC was assessed by iteratively leaving out one participant at a time. For each left-out participant, we calculated the Pearson correlation between their time course of state expression probabilities and the average time course of the remaining participants, all of whom were all engaged in the same narrative. For all three states, the ISCs were significantly higher among the young adults than that among the older adults. For State #1: $t_{(59, 49)} = 3.32$, $p = 0.001$; for State #2: $t_{(59, 49)} = 3.94$, $p < 0.001$; and for State #3: $t_{(59, 49)} = 11.96$, $p < 10^{-6}$.

References

- Baldassano C., Chen J., Zadbood A., Pillow J. W., Hasson U., Norman K. A (2017) **Discovering event structure in continuous narrative perception and memory** *Neuron* **95**:709–721
- Ballenghein U., Megalakaki O., Baccino T (2019) **Cognitive engagement in emotional text reading: concurrent recordings of eye movements and head motion** *Cognition and Emotion*
- Barnett A. J., Reilly W., Dimsdale-Zucker H. R., Mizrak E., Reagh Z., Ranganath C (2021) **Intrinsic connectivity reveals functionally distinct cortico-hippocampal networks in the human brain** *PLoS Biology* **19** <https://doi.org/10.1371/journal.pbio.3001275>
- Berwick R. C., Friederici A. D., Chomsky N., Bolhuis J. J (2013) **Evolution, brain, and the nature of language** *Trends in Cognitive Sciences* **17**:89–98 <https://doi.org/10.1016/j.tics.2012.12.002>
- Bressler S. L., Menon V (2010) **Large-scale brain networks in cognition: emerging methods and principles** *Trends in Cognitive Sciences* **14**:277–290
- Cui Y., Che W., Liu T., Qin B., Yang Z (2021) **Pre-training with whole word masking for chinese bert.** *IEEE/ACM Transactions on Audio Speech, and Language Processing* **29**:3504–3514
- Devlin J., Chang M.-W., Lee K., Toutanova K (2018) **Bert: Pre-training of deep bidirectional transformers for language understanding** *arXiv*
- Fan L., Li H., Zhuo J., Zhang Y., Wang J., Chen L., Jiang T (2016) **The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture** *Cerebral Cortex* **26**:3508–3526 <https://doi.org/10.1093/cercor/bhw157>
- Fernandino L., Tong J.-Q., Conant L. L., Humphries C. J., Binder J. R (2022) **Decoding the information structure underlying the neural representation of concepts** *Proceedings of the national academy of sciences* **119** <https://doi.org/10.1073/pnas.2108091119>
- Ferstl E. C., Neumann J., Bogler C., von Cramon D. Y. (2008) **The extended language network: A meta-analysis of neuroimaging studies on text comprehension** *Human Brain Mapping* **29**:581–593 <https://doi.org/10.1002/hbm.20422>
- Fuster J. M (2000) **The module: crisis of a paradigm** *Neuron* **26**:51–53
- Goldstein A., Grinstein-Dabush A., Schain M., Wang H., Hong Z., Aubrey B., Hasson U (2024) **Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns** *Nature Communications* **15** <https://doi.org/10.1038/s41467-024-46631-y>
- Goldstein A., Zada Z., Buchnik E., Schain M., Price A., Aubrey B., Hasson U (2022) **Shared computational principles for language processing in humans and deep language models** *Nature Neuroscience* **25**:369–380 <https://doi.org/10.1038/s41593-022-01026-4>

- Greipl S., Bernecker K., Ninaus M (2021) **Facial and bodily expressions of emotional engagement: How dynamic measures reflect the use of game elements and subjective experience of emotions and effort** *Proceedings of the ACM on Human-Computer Interaction* **5**:1–25
- Hasson U., Chen J., Honey C. J (2015) **Hierarchical process memory: memory as an integral component of information processing** *Trends in Cognitive Sciences* **19**:304–313
- Honey C. J., Newman E. L., Schapiro A. C (2018) **Switching between internal and external modes: A multiscale learning principle** *Netw Neurosci* **1**:339–356 https://doi.org/10.1162/NETN_a_00024
- Ji J. L., Spronk M., Kulkarni K., Repovš G., Anticevic A., Cole M. W (2019) **Mapping the human brain's cortical-subcortical functional network organization** *NeuroImage* **185**:35–57 <https://doi.org/10.1016/j.neuroimage.2018.10.006>
- Kaakinen J. K., Ballenghein U., Tissier G., Baccino T (2018) **Fluctuation in cognitive engagement during reading: Evidence from concurrent recordings of postural and eye movements.** *Journal of Experimental Psychology: Learning Memory, and Cognition* **44**
- Kelso J. A. S (2012) **Multistability and metastability: understanding dynamic coordination in the brain** *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**:906–918 <https://doi.org/10.1098/rstb.2011.0351>
- Langdon C., Genkin M., Engel T. A (2023) **A unifying perspective on neural manifolds and circuits for cognition** *Nat Rev Neurosci* **24**:363–377 <https://doi.org/10.1038/s41583-023-00693-x>
- Lerner Y., Honey C. J., Silbert L. J., Hasson U (2011) **Topographic mapping of a hierarchy of temporal receptive windows using a narrated story** *Journal of Neuroscience* **31**:2906–2915 <https://doi.org/10.1523/JNEUROSCI.3684-10.2011>
- Liu L., Li H., Ren Z., Zhou Q., Zhang Y., Lu C., ..., Ding G. (2022) **The “two-brain” approach reveals the active role of task-deactivated default mode network in speech comprehension** *Cerebral Cortex* **32**:4869–4884
- Liu L., Zhang Y., Zhou Q., Garrett D. D., Lu C., Chen A., ..., Ding G. (2020) **Auditory– Articulatory Neural Alignment between Listener and Speaker during Verbal Communication** *Cerebral Cortex* **30**:942–951 <https://doi.org/10.1093/cercor/bhz138>
- Meer J. N. V., Breakspear M., Chang L. J., Sonkusare S., Cocchi L (2020) **Movie viewing elicits rich and reliable brain state dynamics** *Nat Commun* **11** <https://doi.org/10.1038/s41467-020-18717-w>
- Nanni-Zepeda M., DeGutis J., Wu C., Rothlein D., Fan Y., Grimm S., ..., Zuberer A. (2024) **Neural signatures of shared subjective affective engagement and disengagement during movie viewing** *Human Brain Mapping* **45** <https://doi.org/10.1002/hbm.26622>
- Ohad T., Yeshurun Y (2023) **Neural synchronization as a function of engagement with the narrative** *Neuroimage* **276** <https://doi.org/10.1016/j.neuroimage.2023.120215>
- Oota S. R., Arora J., Agarwal V., Marreddy M., Gupta M., Surampudi B. R (2022) **Neural language taskonomy: Which NLP tasks are the most predictive of fMRI brain activity?** *arXiv*

- Pohle J., Langrock R., Van Beest F. M., Schmidt N. M. (2017) **Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement.** *Journal of Agricultural Biological and Environmental Statistics* **22**:270–293
- Price C. J. (2012) **A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading** *Neuroimage* **62**:816–847
- Song H., Park B. Y., Park H., Shim W. M (2021) **Cognitive and Neural State Dynamics of Narrative Comprehension** *J Neurosci* **41**:8972–8990 <https://doi.org/10.1523/JNEUROSCI.0037-21.2021>
- Song H., Shim W. M., Rosenberg M. D (2023) **Large-scale neural dynamics in a shared low-dimensional state space reflect cognitive and attentional dynamics** *Elife* **12** <https://doi.org/10.7554/eLife.85487>
- Taghia J., Cai W., Ryali S., Kochalka J., Nicholas J., Chen T., Menon V (2018) **Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition** *Nature Communications* **9** <https://doi.org/10.1038/s41467-018-04723-6>
- Tan C., Liu X., Zhang G (2022) **Inferring Brain State Dynamics Underlying Naturalistic Stimuli Evoked Emotion Changes With dHA-HMM** *Neuroinformatics* **20**:737–753 <https://doi.org/10.1007/s12021-022-09568-5>
- Tang X., Zhang J., Liu L., Yang M., Li S., Chen J., ..., Ding G. (2023) **Distinct brain state dynamics of native and second language processing during narrative listening in late bilinguals** *Neuroimage* **280** <https://doi.org/10.1016/j.neuroimage.2023.120359>
- Vidaurre D., Smith S. M., Woolrich M. W (2017) **Brain network dynamics are hierarchically organized in time** *Proc Natl Acad Sci U S A* **114**:12827–12832 <https://doi.org/10.1073/pnas.1705120114>
- Vyas S., Golub M. D., Sussillo D., Shenoy K. V (2020) **Computation Through Neural Population Dynamics** *Annu Rev Neurosci* **43**:249–275 <https://doi.org/10.1146/annurev-neuro-092619-094115>
- Whitfield-Gabrieli S., Nieto-Castanon A (2012) **Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks** *Brain Connectivity* **2**:125–141
- Yeo B. T., Krienen F. M., Sepulcre J., Sabuncu M. R., Lashkari D., Hollinshead M., ..., Polimeni J. R. (2011) **The organization of the human cerebral cortex estimated by intrinsic functional connectivity** *Journal of neurophysiology*
- Yu J (2024) **Age-related decrease in inter-subject similarity of cortical morphology and task and resting-state functional connectivity** *GeroScience* **46**:697–711

Author information

Lanfang Liu

Department of Psychology, School of Arts and Sciences, Beijing Normal University at Zhuhai, Zhuhai, China, Center for Cognition and Neuroergonomics, State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University at Zhuhai, Zhuhai, China

ORCID iD: [0000-0002-1448-9009](https://orcid.org/0000-0002-1448-9009)

Jiahao Jiang

State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University & IDG/McGovern Institute for Brain Research, Beijing, China

ORCID iD: [0000-0001-6335-391X](https://orcid.org/0000-0001-6335-391X)

Hehui Li

Center for Brain Disorders and Cognitive Sciences, Shenzhen University, Shenzhen, China

Guosheng Ding

State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University & IDG/McGovern Institute for Brain Research, Beijing, China

For correspondence: dinggsh@bnu.edu.cn

Editors

Reviewing Editor

Andrea Martin

Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

Senior Editor

Barbara Shinn-Cunningham

Carnegie Mellon University, Pittsburgh, United States of America

Reviewer #1 (Public review):

Summary:

Liu and colleagues applied the hidden Markov model on fMRI to show three brain states underlying speech comprehension. Many interesting findings were presented: brain state dynamics were related to various speech and semantic properties, timely expression of brain states (rather than their occurrence probabilities) was correlated with better comprehension, and the estimated brain states were specific to speech comprehension but not at rest or when listening to non-comprehensible speech.

Strengths:

Recently, the HMM has been applied to many fMRI studies, including movie watching and rest. The authors cleverly used the HMM to test the external/linguistic/internal processing theory that was suggested in comprehension literature. I appreciated the way the authors theoretically grounded their hypotheses and reviewed relevant papers that used the HMM on other naturalistic datasets. The manuscript was well written, the analyses were sound, and the results had clear implications.

<https://doi.org/10.7554/eLife.99997.2.sa2>

Reviewer #2 (Public review):

Liu et al. applied hidden Markov models (HMM) to fMRI data from 64 participants listening to audio stories. The authors identified three brain states, characterized by specific patterns of activity and connectivity, that the brain transitions between during story listening. Drawing on a theoretical framework proposed by Berwick et al. (TICS 2023), the authors interpret

these states as corresponding to external sensory-motor processing (State 1), lexical processing (State 2), and internal mental representations (State 3). States 1 and 3 were more likely to transition to State 2 than between one another, suggesting that State 2 acts as a transition hub between states. Participants whose brain state trajectories closely matched those of an individual with high comprehension scores tended to have higher comprehension scores themselves, suggesting that optimal transitions between brain states facilitated narrative comprehension.

Overall, the conclusions of the paper are well-supported by the data. Several recent studies (e.g., Song, Shim, and Rosenberg, eLife, 2023) have found that the brain transitions between a small number of states; however, the functional role of these states remains under-explored. An important contribution of this paper is that it relates the expression of brain states to specific features of the stimulus in a manner that is consistent with theoretical predictions.

The correlation between narrative features and brain state expression was reliable, but relatively low (~ 0.03). As discussed in the manuscript, this could be due to measurement noise, as well as narrative features accounting for a small proportion of cognitive processes underlying the brain states.

A strength of the paper is that the authors repeated the HMM analyses across different tasks (Figure 5) and an independent dataset (Figure S3) and found that the data was consistently best fit by 3 brain states. Across tasks, however, the spatial regions associated with each state varied. For example, state 2 during narrative comprehension was similar to both states 2 and 3 during rest (Fig. 5A), suggesting that the organization of the three states was task dependent.

The three states identified in the manuscript correspond rather well to areas with short, medium, and long temporal timescales (see Hasson, Chen & Honey, TiCs, 2015). Given the relationship with behavior, where State 1 responds to acoustic properties, State 2 responds to word-level properties, and State 3 responds to clause-level properties, a "single-process" account where the states differ in terms of the temporal window for which one needs to integrate information over may offer a more parsimonious account than a multi-process account where the states correspond to distinct processes. This possibility is mentioned briefly in the introduction, but not developed further.

<https://doi.org/10.7554/eLife.99997.2.sa1>

Author response:

The following is the authors' response to the original reviews.

Public Reviews:

Reviewer #1 (Public review):

Summary:

Liu and colleagues applied the hidden Markov model on fMRI to show three brain states underlying speech comprehension. Many interesting findings were presented: brain state dynamics were related to various speech and semantic properties, timely expression of brain states (rather than their occurrence probabilities) was correlated with better comprehension, and the estimated brain states were specific to speech comprehension but not at rest or when listening to non-comprehensible speech.

Strengths:

Recently, the HMM has been applied to many fMRI studies, including movie watching and rest. The authors cleverly used the HMM to test the external/linguistic/internal processing theory that was suggested in comprehension literature. I appreciated the way the authors theoretically grounded their hypotheses and reviewed relevant papers that used the HMM on other naturalistic datasets. The manuscript was well written, the analyses were sound, and the results had clear implications.

Weaknesses:

Further details are needed for the experimental procedure, adjustments needed for statistics/analyses, and the interpretation/rationale is needed for the results.

For the Experimental Procedure, we will provide a more detailed description about stimuli, and the comprehension test, and upload the audio files and corresponding transcriptions as the supplementary dataset.

For statistics/analyses, we have reproduced the states' spatial maps using unnormalized activity pattern. For the resting state, we observed a state resembling the baseline state described in Song, Shim, & Rosenberg (2023). However, for the speech comprehension task, all three states were characterized by network activities varying largely from zero. In addition, we have re-generated the null distribution for behavior/brain state correlations using circular shift. The results are largely consistent with the previous findings. We have also made some other adjustment to the analyses or add some new analyses as recommended by the reviewer. We will revise the manuscript to incorporate these changes.

For the interpretation/rationale: We will add a more detailed interpretation for the association between state occurrence and semantic coherence. Briefly speaking, higher semantic coherence may allow for the brain to better accumulate information over time.

State #2 seems to be involved in the integration of information at shorter timescales (hundreds of milliseconds) while State #3 seems to be involved in the longer timescales (seconds).

We greatly appreciate the reviewer for the insightful comments and constructive suggestions.

Reviewer #2 (Public review):

Liu et al. applied hidden Markov models (HMM) to fMRI data from 64 participants listening to audio stories. The authors identified three brain states, characterized by specific patterns of activity and connectivity, that the brain transitions between during story listening. Drawing on a theoretical framework proposed by Berwick et al. (TICS 2023), the authors interpret these states as corresponding to external sensory-motor processing (State 1), lexical processing (State 2), and internal mental representations (State 3). States 1 and 3 were more likely to transition to State 2 than between one another, suggesting that State 2 acts as a transition hub between states. Participants whose brain state trajectories closely matched those of an individual with high comprehension scores tended to have higher comprehension scores themselves, suggesting that optimal transitions between brain states facilitated narrative comprehension.

Overall, the conclusions of the paper are well-supported by the data. Several recent studies (e.g., Song, Shim, and Rosenberg, eLife, 2023) have found that the brain transitions between a small number of states; however, the functional role of these states remains under-explored. An important contribution of this paper is that it relates the expression of brain states to specific features of the stimulus in a manner that is consistent with theoretical predictions.

(1) It is worth noting, however, that the correlation between narrative features and brain state expression (as shown in Figure 3) is relatively low (~ 0.03). Additionally, it was unclear if the temporal correlation of the brain state expression was considered when generating the null distribution. It would be helpful to clarify whether the brain state expression time courses were circularly shifted when generating the null.

In the revision, we generated the null distribution by circularly shifting the state time courses. The results remain consistent with our previous findings: $p = 0.002$ for the speech envelope, $p = 0.007$ for word-level coherence, and $p = 0.001$ for clause-level coherence.

We note that in other studies which examined the relationship between brain activity and word embedding features, the group-mean correlation values are similarly low but statistically significant and theoretically meaningful (e.g., Fernandino et al., 2022; Oota et al., 2022). We think these relatively low correlations are primarily due to the high level of noise inherent in neural data. Brain activity fluctuations are shaped by a variety of factors, including task-related cognitive processing, internal thoughts, physiological states, as well as arousal and vigilance. Additionally, the narrative features we measured may account for only a small portion of the cognitive processes occurring during the task. As a result, the variance in narrative features can only explain a limited portion of the overall variance in brain activity fluctuations.

We will replace Figure 3 and the related supplementary figures with new ones, in which the null distribution is generated via circular shift. Furthermore, we will expand our discussion to address why the observed brain-stimuli correlations are relatively small, despite their statistical significance.

(2) A strength of the paper is that the authors repeated the HMM analyses across different tasks (Figure 5) and an independent dataset (Figure S3) and found that the data was consistently best fit by 3 brain states. However, it was not entirely clear to me how well the 3 states identified in these other analyses matched the brain states reported in the main analyses. In particular, the confusion matrices shown in Figure 5 and Figure S3 suggests that that states were confusable across studies (State 2 vs. State 3 in Fig. 5A and S3A, State 1 vs. State 2 in Figure 5B). I don't think this takes away from the main results, but it does call into question the generalizability of the brain states across tasks and populations.

We identified matching states across analyses based on similarity in the activity patterns of the nine networks. For each candidate state identified in other analyses, we calculate the correlation between its network activity pattern and the three predefined states from the main analysis, and set the one it most closely resembled to be its matching state. For instance, if a candidate state showed the highest correlation with State #1, it was labelled State #1 accordingly.

Each column in the confusion matrix depicts the similarity of each candidate state with the three predefined states. In Figure S3 (analysis for the replication dataset), the highest similarity occurred along the diagonal of the confusion matrix. This means that each of the three candidate states was best matched to State #1, State #2, and State #3, respectively, maintaining a one-to-one correspondence between the states from two analyses.

For the comparison of speech comprehension task with the resting and the incomprehensible speech condition, there was some degree of overlap or "confusion."

In Figure 5A, there were two candidate states showing the highest similarity to State #2. In this case, we labelled the candidate state with the strongest similarity as State #2, while the other candidate state is assigned as State #3 based on the ranking of similarity. This strategy

was also applied to naming of states for the incomprehensible condition. The observed confusion supports the idea that the tripartite-state space is not an intrinsic, task-free property. To make the labeling clearer in the presentation of results, we will use a prime symbol (e.g., State #3') to indicate cases where such confusion occurred, helping to distinguish these ambiguous matches.

(3) The three states identified in the manuscript correspond rather well to areas with short, medium, and long temporal timescales (see Hasson, Chen & Honey, TiCs, 2015).

Given the relationship with behavior, where State 1 responds to acoustic properties, State 2 responds to word-level properties, and State 3 responds to clause-level properties, the authors may want to consider a "single-process" account where the states differ in terms of the temporal window for which one needs to integrate information over, rather than a multi-process account where the states correspond to distinct processes.

The temporal window hypothesis provides a more fitting explanation for our results. Based on the spatial maps and their modulation by speech features, States #1, #2, and #3 seem to correspond to short, medium, and long processing timescales, respectively. We will update the discussion to reflect this interpretation.

We sincerely appreciate the constructive suggestions from the two anonymous reviewers, which have been highly valuable in improving the quality of the manuscript.

Recommendations for the authors:

Reviewer #1 (Recommendations for the authors):

(1) The "Participants and experimental procedure" section deserves more details. I've checked Liu et al. (2020), and the dataset contained 43 participants aged 20-75 years, whereas this study contained data from 64 young adults and 30 old adult samples. The previous dataset seems to have two stories, whereas this study seems to have three. Please be specific, given that the dataset does not seem the same. Could the authors also include more descriptions of what the auditory stories were? For example, what were the contents, and how were they recorded?

The citation is partially incorrect. The dataset of young adults is shared with our work published in (2022). The 64 participants listened to one of three stories told by a female college student in Mandarin, recounting her real-life experience of hiking, a graduate admission interview, and her first time taking a flight, respectively. The sample of older adults is from our work published in (2020), which includes 30 older adults and additionally 13 young adults. The stimuli in this case were two stories told by an older woman in a Chinese dialect, describing her experience in Thailand and riding a warship, respectively. Since we aim to explore whether the main results can be replicated on a different age group, we excluded the 13 young adults from the analysis.

All the stories were recorded during fMRI scanning using a noise-canceling microphone (FOMRI-III; Optoacoustics Ltd, Or-Yehuda, Israel) positioned above the speaker's mouth. The audio recordings were subsequently processed offline with Adobe Audition 3.0 (Adobe Systems Inc., USA) to further eliminate MRI scanner noise.

In the revised manuscript, we have updated the citation, and provided a more detailed description of the stimuli in the supplementary material. We have also uploaded the audio files along with their corresponding transcriptions to GitHub.

(2) I am curious about individual differences in comprehension scores. Did participants have less comprehension of the audio-narrated story because the story was a hard-

to comprehend narrative or because the audio quality was low? Could the authors share examples of comprehension tests?

We believe two factors contribute to the individual differences in comprehension scores. First, the audio quality is indeed moderately lower than in daily life story-listening conditions. This is because those stories were recorded and played during fMRI scanning. Although a noise-canceling equipment was used, there were still some noises accompanying the speech, which may have made speech perception and comprehension more difficult than usual.

Second, the comprehension test measured how much information about the story (including both main themes and details) participants could recall. Specifically, participants were asked to retell the stories in detail immediately after the scanning session. Following this free recall, the experimenters posed a few additional questions drawn from a pre-prepared list, targeting information not mentioned in their recall. If participants experienced lapses of attention or did not store the incoming information into memory promptly, they might fail to recall the relevant content. In several studies, such a task has been called a narrative recall test. However, memory plays a crucial role in real-time speech comprehension, while comprehension affects the depth of processing during memory encoding, thereby influencing subsequent recall performance. To align with prior work (e.g., Stephens et al., 2010) and our previous publications, we chose to refer to this task as narrative comprehension.

In the revised manuscript, we have provided a detailed description about the comprehension test (Line 907-933) and share the examples on GitHub.

(3) Regarding Figure 3, what does it mean for a state occurrence to follow semantic coherence? Is there a theoretical reason why semantic coherence was measured and related to brain state dynamics? A related empirical question is: is it more likely for the brain states to transition from one state to another when nearby time points share low semantic similarity compared to chance?

We analyzed semantic coherence and sound envelope as they capture different layers of linguistic and acoustic structure that unfold over varying temporal scales. Changes in the sound envelope typically occur on the order of milliseconds to a few hundred milliseconds, changes in word-level semantic coherence span approximately 0.24 ± 0.15 seconds, and changes in clause-level semantic coherence extend to 3.2 ± 1.7 seconds. Previous theory and empirical studies suggest that the timescales of information accumulation vary hierarchically, progressing from early sensory areas to higher-order areas (Hasson et al., 2015; Lerner et al., 2011). Based on this work, we anticipate that the three brain states, which are respectively associated with the auditory and sensory motor network, the language network and the DMN, would be selectively modulated by these speech properties corresponding to distinct timescales.

Accordingly, when a state occurrence aligns with (clause-level) semantic coherence, it suggests that this state is engaged in processing information accumulated at the clause level (i.e., its semantic relationship). Higher coherence facilitates better accumulation, making it more likely for the associated brain state to be activated.

We analyzed the relationship between state transition probability and semantic coherence, but did not find significant results. Here, the transition probability was calculated as $\text{Gamma}(t) - \text{Gamma}(t-1)$, where Gamma refers to the state occurrence probability. The lack of significant findings may be because brain state transitions are driven primarily by more slowly changing factors. Indeed, we found the average dwell time of the three states ranges from 9.66 to 15.29s, which is a much slower temporal dynamics compared to the relatively rapid shifts in acoustic/semantic properties.

In the revised version, we have updated the *Introduction* to clarify the rational for selecting the three speech properties and to explore their relationship with brain dynamics (Line 111-118)

(4) When running the HMM, the authors iterated K of 2 to 10 and K = 4, 10, and 12. However, the input features of the model consist of only 9 functional networks. Given that the HMM is designed to find low-dimensional latent state sequences, the choice of the number of latent states being higher than the number of input features sounds odd to me - to my speculation, it is bound to generate almost the exact same states as 9 networks and/or duplicates of the same state. I suggest limiting the K iterations from 2 to 8. For replication with Yeo et al.'s 7 networks, K iteration should also be limited to K of less than 7, or optionally, Yeo's 7 network scheme could be replaced with a 17network scheme.

We understand your concern. However, the determination of the number (K) of hidden states is not directly related to the number of features (in this case, the number of networks), but rather depends on the complexity of the time series and the number of underlying patterns. Given that each state corresponds to a distinct combination of the features, even a small number of features can be used to model a system with complex temporal behaviors and multiple states. For instance, for a system with n features, assuming each is a binary variable (0 or 1), there are maximally 2^n possible underlying states.

In our study, we recorded brain activity over 300 time points and used the 9 networks as features. At different time points, the brain can exhibit distinct spatial configurations, reflected in the relative activity levels of the nine networks and their interactions. To accurately capture the temporal dynamics of brain activity, it is essential to explore models that allow for more states than the number of features. We note that in other HMM studies, researchers have also explored states more than the number of networks to find the best number of hidden states (e.g., Ahrends et al., 2022; Stevner et al., 2019).

Furthermore, Ahrends et al. (2022) suggested that “Based on the HCP-dataset, we estimate as a rule of thumb that the ratio of observations to free parameters per state should not be inferior to 200”, where free parameters per state is $[K * (K - 1) + (K - 1) * K * N * (N + 1) / 2] / K$. According to this, there should be above 10, 980 observations when the number of states (K) is 10 (the maximal number in our study) and the number of networks (N) is 9. In our group-level HMM model, there were 64 (valid runs) * 300 (TR) = 19200 observations for young adults, and 50 (valid runs) * 210 (TR) = 10500 observations for older adults. Aside from the older adults' data being slightly insufficient (4.37% less than the suggestion), all other hyperparameter combinations in this study meet the recommended number of observations.

(5) In Figure 2, the authors write that the states' spatial maps were normalized for visualization purposes. Could the authors also show visualization of brain states that are not normalized? The reason why I ask is, for example, in Song, Shim, & Rosenberg (2023), the base state was observed which had activity levels all close to the mean (which is 0 because the BOLD activity was normalized). If the activity patterns of this brain state were to be normalized after state estimation, the base state would have looked drastically different than what is reported.

We derived the spatial maps of the states using unnormalized activity patterns, with the BOLD signals Z-score normalized to a mean of zero. Under the speech comprehension task, the three states exhibited relatively large fluctuations in network activity levels. The activity ranges were as follows: [-0.71 to 0.51] for State #1, [-0.26 to 0.30] for State #2, and [-0.82 to 0.40] for State #3. For the resting state, we observed a state resembling the baseline state as described in Song, Shim, & Rosenberg (2023), with activity values ranging from -0.133 to 0.09.

In the revision, we have replaced the states' spatial maps with versions showing unnormalized activity patterns.

(6) In line 297, the authors speculate that "This may be because there is too much heterogeneity among the older adults". To support this speculation, the authors can calculate the overall ISC of brain state dynamics among older adults and compare it to the ISC estimated from younger adults.

We analyzed the overall ISC of brain state dynamics, and found the ISC was indeed significantly lower among the older adults than that among the younger adults. We have revised this statement as follows:

These factors can diminish the inter-subject correlation of brain state dynamics— indeed, ISCs among older adults were significantly lower than those among younger adults (Figure S5)—and reduce ISC's sensitivity to individual differences in task performance (Line 321-326).

Other comments:

(7) In Figure 4, the authors showed a significant positive correlation between head movement ISC with the best performer and comprehension scores. Does the average head movement of all individuals negatively correlate with comprehension scores, given that the authors argue that "greater task engagement is accompanied by decreased movement"?

We examined the relationship between participants' average head movement across the comprehension task and their comprehension scores. There was no significant correlation ($r = 0.041$, $p = 0.74$). In the literature (e.g., Ballenghein et al., 2019), the relationship between task engagement and head movement was also assessed at the moment-by-moment level, rather than by using time-averaged data.

Real-time head movements reflect fluctuations in task engagement and cognitive state. In contrast, mean head movement, as a static measure, fails to capture these changes, and thus is not effective in predicting task performance.

(8) The authors write the older adults sample, the "independent dataset". Technically, however, this dataset cannot be independent because they were collected at the same time by the same research group. I would advise replacing the word independent to something like second dataset or replication dataset.

We have replaced the phrase “independent dataset” with “replication dataset”.

(9) Pertaining to a paragraph starting in line 586: For non-parametric permutation tests, the authors note that the time courses of brain state expression were "randomly shuffled". How was this random shuffling done: was this circular-shifted randomly, or were the values within the time course literally shuffled? The latter approach, literal shuffling of the values, does not make a fair null distribution because it does not retain temporal regularities (autocorrelation) that are intrinsic to the fMRI signals. Thus, I suggest replacing all non-parametric permutation tests with random circular shifting of the time series (np. roll in python).

In the original manuscript, the time course was literally shuffled. In the revised version, we circular-shifted the time course randomly (circshift.m in Matlab) to generate the null distribution. The results remain consistent with our previous findings: $p = 0.002$ for the

speech envelope, $p = 0.007$ for word-level coherence, and $p = 0.001$ for clause-level coherence (Line 230-235).

(10) The p value calculation should be $p = (1 + \#(\text{chance} \geq \text{observed})) / (1 + \#(\text{iterations}))$ for one-tailed test and $p = (1 + \#(\text{abs}(\text{chance}) \geq \text{abs}(\text{observed}))) / (1 + \#(\text{iterations}))$ for two-tailed test. Thus, if 5,000 iterations were run and none of the chances were higher than the actual observation, the p -value is $p = 1/5001$, which is the minimal value it can achieve.

Have corrected.

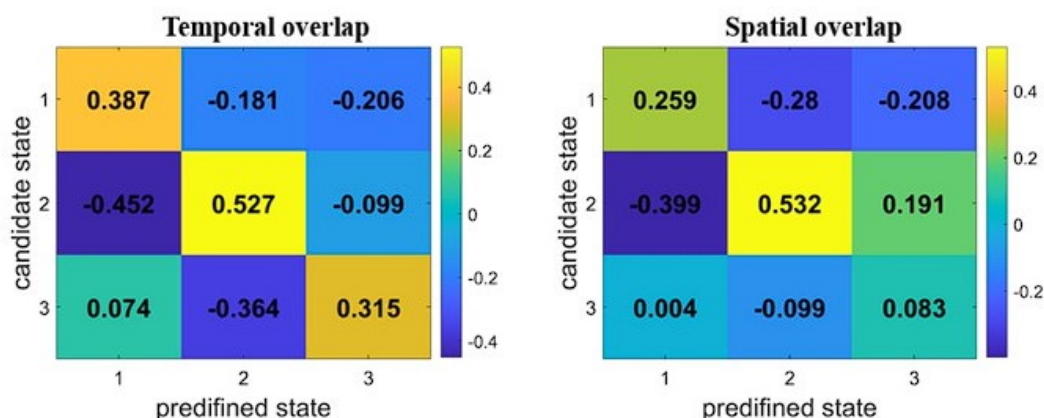
(11) State 3 in Figure S2 does not resemble State 3 of the main result. Could the authors explain why they corresponded State 3 of the Yeo-7 scheme to State 3 of the nine-parcellation scheme, perhaps using evidence of spatial overlap?

The correspondence of states between the two schemes was established using evidence of state expression time course.

To assess temporal overlap, we calculated Pearson's correlation between each candidate state obtained by the Yeo-7 scheme and the three predefined states obtained by the nine-network parcellation scheme in terms of state expression probabilities. The time courses of the 64 participants were concatenated, resulting in 19200 (300×64) time points for each state. The one that the candidate state most closely resembled was set to be its corresponding state. For instance, if a candidate state showed the highest correlation with State #1, it was labelled State #1 accordingly. As demonstrated in the confusion matrix, each of the three candidate states was best matched to State #1, State #2, and State #3, respectively, maintaining a one-to-one correspondence between the states from the two schemes.

We also assessed the spatial overlap between the two schemes. First, a state activity value was assigned to each voxel across the whole brain (including a total of 34,892 voxels covered by both parcellation schemes). This is done for each brain state. Next, we calculated Spearman's correlation between each candidate state obtained by the Yeo-7 scheme and the three predefined states obtained by the nine-network scheme in terms of whole-brain activities. The pattern of spatial overlap is consistent with the pattern of temporal overlap, such that each of the three candidate states was best matched to State #1, State #2, and State #3, respectively.

Author response image 1.

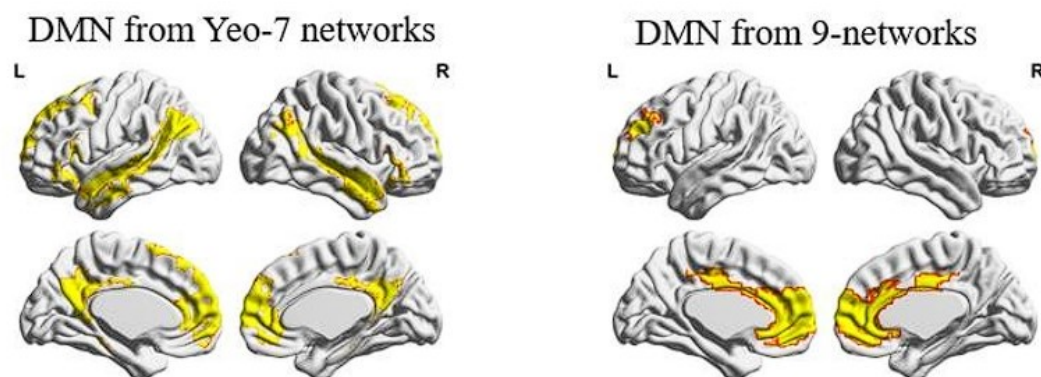


We noted that the networks between the two schemes are not well aligned in their spatial location, especially for the DMN (as shown below). This may lead to the low spatial overlap of

State #3, which is dominated by DMN activity. Consequently, establishing state correspondence based on temporal information is more appropriate in this context. We therefore only reported the results of temporal overlap in the manuscript.

We have added a paragraph in the main text for “Establishing state correspondence between analyses” (Line 672-699). We have also updated the associated figures (Fig.S2, Fig.S3 and Fig.5)

Author response image 2.



(12) Line 839: *gamma parameter, on a step size of?*

(16) Figure 3. Please add a legend in the "Sound envelope" graph what green and blue lines indicate. The authors write $Coh(t)$ and $Coh(t, t+1)$ at the top and $Coh(t)$ and $Coh(t+1)$ at the bottom. Please be consistent with the labeling. Shouldn't they be $Coh(t-1, t)$ and $Coh(t, t+1)$ to be exact for both?

Have corrected.

(17) In line 226, is this one-sample *t*-test compared to zero? If so, please write it inside the parentheses. In line 227, the authors write "slightly weaker"; however, since this is not statistically warranted, I suggest removing the word "slightly weaker" and just noting significance in both States 1 and 2.

Have corrected.

(18) In line 288, please fix "we also whether".

Have corrected.

(19) In Figure 2C, what do pink lines in the transition matrix indicate? Are they colored just to show authors' interests, or do they indicate statistical significance? Please write it in the figure legend.

Yes, the pink lines indicate a meaningful trend, showing that the between-state transition probabilities are significantly higher than those in permutation.

We have added this information to the figure legend.

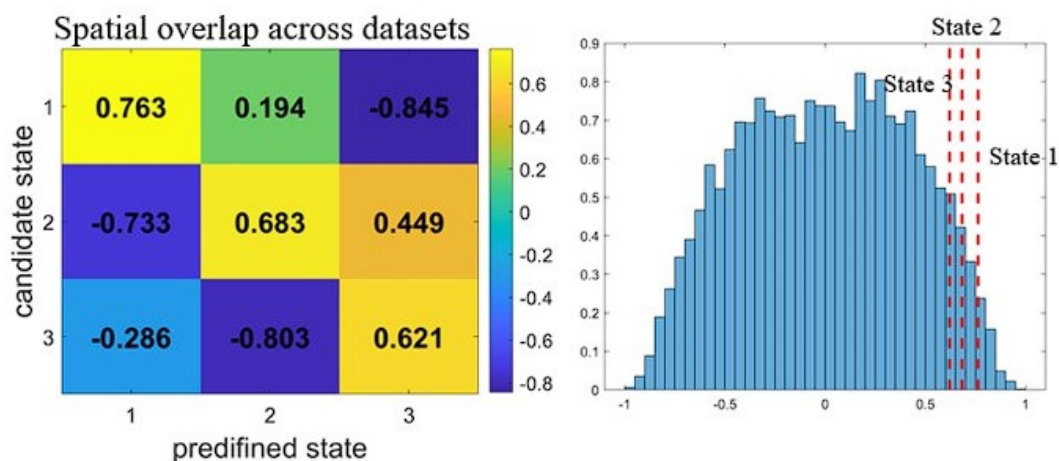
Reviewer #2 (Recommendations for the authors):

(1) It is unclear how the correspondence between states across different conditions and datasets was computed. Given the spatial autocorrelation of brain maps, I recommend reporting the Dice coefficient along with a spin-test permutation to test for statistical significance.

The state correspondence between different conditions and between the two datasets are established using evidence of spatial overlap. The spatial overlap between states was quantified by Pearson's correlation using the activity values (derived from HMM) of the nine networks. For each candidate state identified in other analyses (for the Rest, MG and older-adult datasets), we calculate the correlation between its network activity pattern and the three predefined states from the main analysis (for the young-adults dataset), and set the one it most closely resembled to be its matching state. For instance, if a candidate state showed the highest correlation with State #1, it was labelled State #1 accordingly.

For the comparison between the young and older adults' datasets (as shown below), the largest spatial overlap occurred along the diagonal of the confusion matrix, with high correlation values. This means that each of the three candidate states was best matched to State #1, State #2, and State #3, respectively, maintaining a one-to-one correspondence between the states from the two datasets. As the HMM is modelled at the level of networks which lack accurate coordinates, we did not apply the spin-test to assess the statistical significance of overlap. Instead, we extracted the state activity patterns from the 1000 permutations (wherein the original BOLD time courses were circularly shifted and an HMM was conducted) for the older-adults dataset. Applying the similar state-correspondence strategy, we generated a null distribution of spatial overlap. The real overlap of the three states was greater than and 97.97%, 95.34% and 92.39% instances from the permutation (as shown below).

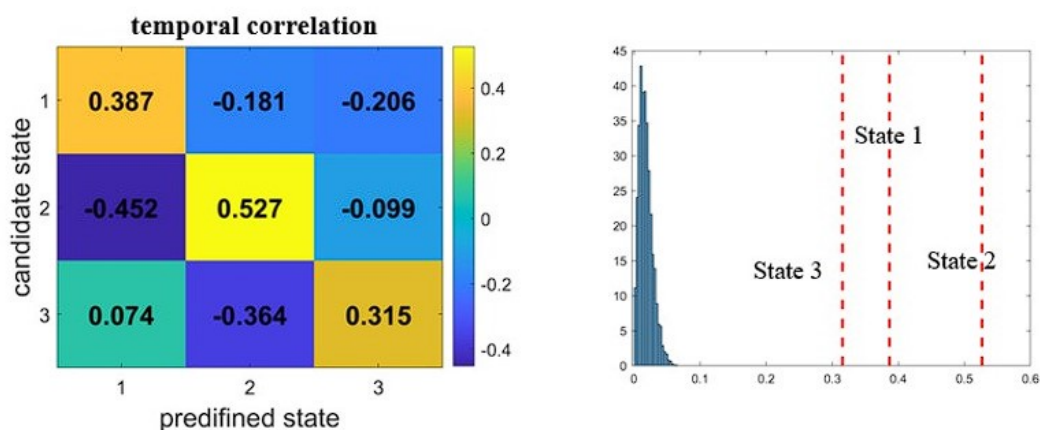
Author response image 3.



For the comparison of main task with the resting and the incomprehensible speech condition, there was some degree of confusion: there were two candidate states showing the highest similarity to State #2. In this case, we labeled the most similar candidate as State #2. The other candidate was then assigned to the predefined state with which it had the second-highest correlation. We used a prime symbol (e.g., State #3') to denote cases where such confusion occurred. These findings support our conclusion that the tripartite-organization of brain states is not a task-free, intrinsic property.

When establishing the correspondence between the Yeo-7 network and the nine network parcellation schemes, we primarily relied on evidence from temporal overlap measures, as a clear network-level alignment between the two parcellation schemes is lacking. Temporal overlap was quantified by calculating the correlation of state occurrence probabilities between the two schemes. To achieve this, we concatenated the time courses of 64 participants, resulting in a time series consisting of 19,200 time points (300 time points per participant) for each state. Each of the three candidate states from the Yeo-7 network scheme was best matched to State #1, State #2, and State #3 from the main analyses, respectively. To determine the statistical significance of the temporal overlap, we circularly shifted each participant's time course of state expression obtained from the Yeo-7 network scheme for 1000 times. Applying the same strategy to find the matching states, we generated a null distribution of overlap. The real overlap was much higher than the instances from permutation.

Author response image 4.



In the revision, we have provided detailed description for how the state correspondence is established and reported the statistical significance of those correspondence (Line 671-699). The associated figures have also been updated (Fig.5, Fig. S2 and Fig.S3).

- (2) Please clarify if circle-shifting was applied to the state expression time course when generating the null distribution for behavior-brain state correlations reported in Figure
- (3). This seems important to control for the temporal autocorrelation in the time courses.

We have updated the results by using circle-shifting to generated the null distribution. The results are largely consistent with the previous on without circular shifting (Line 230-242).

- (3) Figure 3: What does the green shaded area around the sound envelope represent? In the caption, specify whether the red line in the null distributions indicates the mean or median R between brain state expression and narrative features. It would also be beneficial to report this value in the main text.

The green shaded area indicated the original amplitude of speech signal, while blue line indicates the smoothed, low-frequency contour of amplitude changes over time (i.e., speech envelope). We have updated the figure and explained this in the figure caption.

The red line in the null distributions indicates the R between brain state expression and narrative features for the real data. and reported the mean R of the permutation in the main text.

(4) *The manuscript is missing a data availability statement (<https://elifesciences.org/inside-elife/51839f0a/for-authors-updates-to-elife-s-datasharing-policies>).*

We have added a statement of data availability in the revision, as follows:

“The raw and processed fMRI data are available on OpenNeuro: <https://openneuro.org/datasets/ds005623>. The experimental stimuli, behavioral data and main scripts used in the analyses are provided on Github.”

(5) *There is a typo in line 102 (“perceptual alalyses”).*

Have corrected.

We sincerely thank the two reviewers for their constructive feedback, thorough review, and the time they dedicated to improving our work.

Reference:

Ahrends, C., Stevner, A., Pervaiz, U., Kringelbach, M. L., Vuust, P., Woolrich, M. W., & Vidaurre, D. (2022). Data and model considerations for estimating timevarying functional connectivity in fMRI. *Neuroimage*, 252, 119026.

Ballenghein, U., Megalakaki, O., & Baccino, T. (2019). Cognitive engagement in emotional text reading: concurrent recordings of eye movements and head motion. *Cognition and Emotion*.

Fernandino, L., Tong, J.-Q., Conant, L. L., Humphries, C. J., & Binder, J. R. (2022). Decoding the information structure underlying the neural representation of concepts. *Proceedings of the national academy of sciences*, 119(6), e2108091119. <https://doi.org/10.1073/pnas.2108091119>

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: memory as an integral component of information processing. *Trends in Cognitive Sciences*, 19(6), 304-313.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story [Article]. *Journal of Neuroscience*, 31(8), 2906-2915. <https://doi.org/10.1523/JNEUROSCI.3684-10.2011>

Liu, L., Li, H., Ren, Z., Zhou, Q., Zhang, Y., Lu, C., Qiu, J., Chen, H., & Ding, G. (2022). The “two-brain” approach reveals the active role of task-deactivated default mode network in speech comprehension. *Cerebral Cortex*, 32(21), 4869-4884.

Liu, L., Zhang, Y., Zhou, Q., Garrett, D. D., Lu, C., Chen, A., Qiu, J., & Ding, G. (2020). Auditory–Articulatory Neural Alignment between Listener and Speaker during Verbal Communication. *Cerebral Cortex*, 30(3), 942-951. <https://doi.org/10.1093/cercor/bhz138>

<https://doi.org/10.7554/eLife.99997.2.sa0>