A high-performance brain-to-sentence decoder for logosyllabic language

Chen Feng[1*], Lu Cao[2,*], Di Wu[3,*], En Zhang[4], Ting Wang[5,6], Xiaowei Jiang[7], Heng Ding[5,6], Chenhao Zhou[8], Jinbo Chen[3], Hui Wu[3], Siyu Lin[3], Qiming Hou[3], Junming Zhu[1,9,#], Jie Yang[3,#], Mohamad Sawan[3,#], Yue Zhang[2,#]

[1]Department of Neurosurgery, Zhejiang University School of Medicine Second Affiliated Hospital, Hangzhou, Zhejiang Province, China
[2]School of Engineering, Westlake University, Hangzhou, Zhejiang Province, China
[3]Center of Excellence in Biomedical Research on Advanced Integrated-on-chips Neurotechnologies (CenBRAIN), School of Engineering, Westlake University, Hangzhou, China
[4]State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China
[5]School of Foreign Languages, Tongji University, Shanghai, China
[6]Center for Speech and Language Processing, Tongji University, Shanghai,China
[7]Department of Bioengineering, University of Pennsylvania, Pennsylvania, USA
[8]Department of Psychology, New York University, State of New York, USA
[9]Clinical Research Center for Neurological Diseases of Zhejiang Province, Hangzhou, Zhejiang Province, China

[*] These authors contributed equally to the work.
[#]Corresponding Author.

Abstract

Many severe neurological diseases, such as stroke and amyotrophic lateral sclerosis, can lead to patients completely losing their ability to communicate[1,2]. Several language brain-computer interface systems have been shown to have the potential to help these patients regain their communication abilities by decoding speech or movement-related neural signals[3]. However, the aforementioned language brain-computer interfaces (BCIs) were all developed based on alphabetic linguistic systems without one specially designed for logosyllabic languages like Mandarin Chinese. Here, we established the first language BCI specifically designed for Chinese, decoding speech-related stereoelectroencephalography (sEEG) signals into sentences. Firstly, based on the acoustic features of full-spectrum Chinese syllable pronunciation, we constructed prediction models for three syllable elements (initials, tones, and finals). Subsequently, a language model transformed all the predicted syllable elements, integrating them with semantic information, to generate the most probable sentence. Ultimately, we achieved a high-performance decoder with a median character error rate of 29%, demonstrating preliminary potential for clinical application. Our research fills the gap in language BCI for logosyllabic languages and leverages a powerful language model to enhance the performance of the language BCI, offering new insights for future logosyllabic language neuroprosthesis studies.

Main

Previous language BCI efforts have demonstrated the potential to restore communication for speech-impaired patients, particularly within the scope of alphabetic linguistic systems, via establishing a mapping from brain activities to output formats such as hand-writing[4], spelling[5-7], typing[8,9] and speech[10-12]. In contrast to the alphabetic languages, where words are constructed from combinations of limited size alphabets (26 for English), logosyllabic languages employ logographic characters to represent either entire words or single morphemes. Logosyllabic languages such as Mandarin Chinese have an extensive inventory of over 50,000 logographic characters making spelling and typing unsuitable where patients are required to imagine choosing among different characters[13,14]. The intricate writing patterns of Chinese logographic characters render handwritten input impractical. Furthermore, the tonal nature of Mandarin Chinese introduces additional intricacy to speech decoding, as alterations in a word's tone can lead to a significant shift in its semantic meaning. To address this, we utilize *Pinyin*[15], a widely adopted phonetic text input system based on the Latin alphabet, to design a neuroprosthesis for Mandarin speech decoding. Pinyin has been developed to facilitate literacy and serve as a bridge between Mandarin pronunciation and characters[16]. It adopts a one-to-many mapping between a single *Pinyin* (combination of initials, finals, and tones) and multiple Chinese characters. Consequently, the surrounding context plays a pivotal role in selecting the correct character among homophones[17].

In this study, we establish the language BCI by decoding full-spectrum pronunciation of the *Pinyin* syllables (This study was approved by the university hospital institutional ethics committee, and all participants signed informed consent). Specifically, we recorded stereoelectroencephalography (sEEG) data from four participants (2 males and 2 females, all young native Mandarin speakers) undergoing intracranial monitoring for epilepsy (Fig. 1a and 1c). They were asked to speak 407 unique *Pinyin* syllables (full-spectrum pronunciation covering almost all characters, the upper panel in Fig. 1b), and to read or imagine reading 100 Mandarin sentences (ranging from 2 to over 10 characters in length, the lower panel in Fig. 1b). We designed three prediction models (respectively for initial, tone and final) and a language model to form the first language BCI system specifically for logosyllabic languages (Fig. 1d). It is capable of decoding neural signals into any commonly used Mandarin sentences after a fixed and limited training period. Besides, benefiting from the capability to simultaneously record signals from multiple brain regions, we investigate the anatomical contributions for decoding, with a particular focus on deep brain structures and the non-dominant language hemisphere.

Result

Overview of decoder design

A three-phase (channel selection, syllable elements prediction and language model correction) decoding system is proposed for decoding the brain signal to the complete sentence (Fig. 1d). For each participant, we selected the channels associated with speech and excluded those implanted in visual-related brain regions and white matter regions (Fig. 1c). As for the syllable elements prediction phase, during the training process, three convolutional neural networks (CNNs) are trained to extract features representing the initials, tones and finals from the preprocessed sEEG signal segments of the syllables being spoken. For each syllable, we align the tone and final features with acoustic features extracted from the corresponding audio signal using neural-audio

regularization (NAR). Initial features are mapped into articulatory features, whose relations are further adjusted by a graph neural network (GNN). In the testing process of decoding complete sentences, we use the above three trained models to predict initials, vowels, and tones syllable by syllable.

There exists a gap between the *Pinyin* sequence and the Mandarin sentences[15]. Bridging this gap requires the capability to convert *Pinyin* into Mandarin characters (similar to the *Pinyin* input method) and the ability to consider the surrounding context (due to the homophones). Thus, in the third phase, a language model (LM) was established to output the most probable sentence based on the probability sequence of predicted elements of all syllables. The language model was trained on encyclopedia-style question-answer dialogues retrieved from Common Crawl. The probability sequence of top 3 initials, top 3 tones and all finals in the first final cluster form the output of the three syllable elements prediction models. Moreover, the language model can not only translate *Pinyin* sequences into Mandarin sentences, but also correct the syllable elements prediction errors and improve the performance of the system.

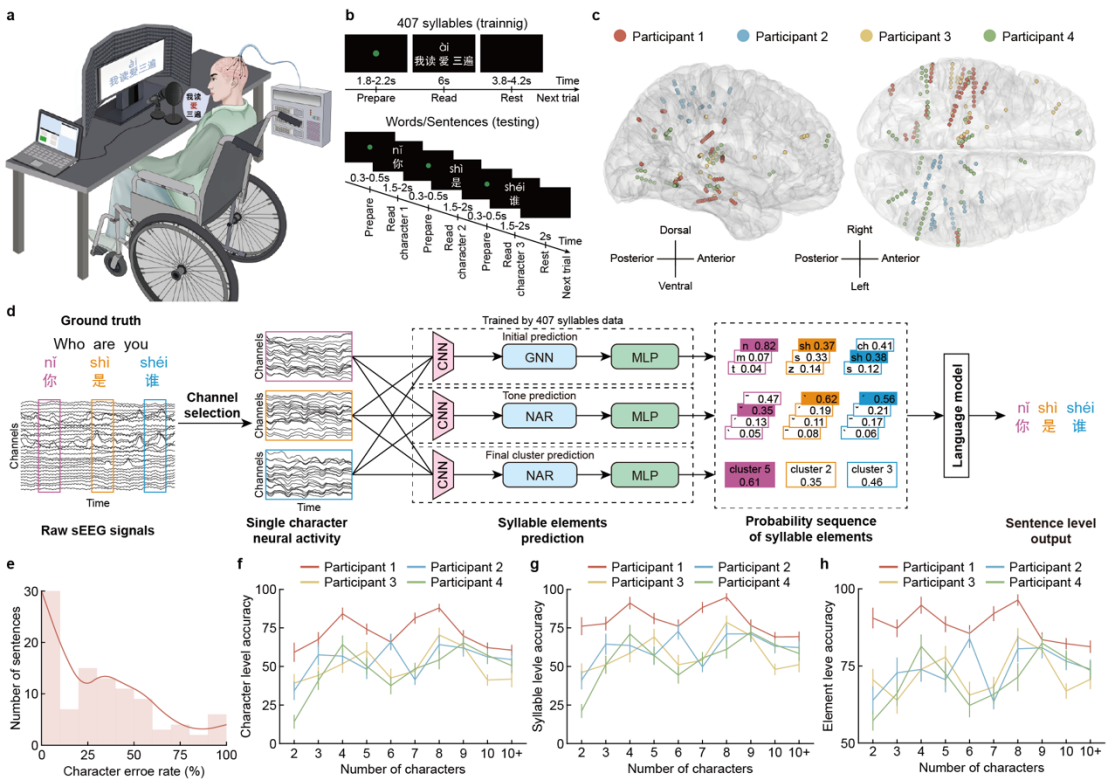

Fig. 1 | Design and performance of brain-to-sentence decoder for logosyllabic language.

Performance of brain-to-sentence decoder

To fully replicate the decoding capability of the system in the natural Mandarin environment, we randomly selected words or sentences for daily use ranging in length from 2 characters to more than 10 characters, taking 10 of each length for a total of 100 test samples. First, we analyzed the distribution of the average character error rate for each sentence (Fig. 1e). For participant 1, 30.00% of the decoded sentences were entirely accurate, with a median character error rate of 29.00%. Native Chinese speakers can comprehend the majority of the decoded sentences. We recruited 25

native Mandarin speakers to compare the semantics of all decoded sentences with the original sentences, and 58.45% of the decoded sentences were considered to be exactly the same in meaning as the original sentences, with the median character error rate reduced to 21.60% (For participant 1). The low error rate and intelligibility level of sentences decoded from neural signals is already sufficiently meaningful and practical for applications.

Then, we calculate the accuracy of the decoded sentences at three levels, which are character, syllable, and element levels, to comprehensively evaluate the decoding performance. We observed that the average character level accuracy reached 72.09%, 54.87%, 50.94%, and 49.08% for four participants respectively (Fig. 1f), with higher accuracy occurring for moderate length sentences (approximately 4-9 characters). The average accuracy of syllables showed no significant (p=0.17) deviation from the accuracy of character (Fig. 1g), registering at 72.87%, 54.99%, 51.26%, and 49.15%, respectively. This indicates that almost all the correct syllables were translated into the correct characters with a powerful language model. Only a few instances of homophone conversion errors (such as '他, tā, [tʰa], he' and '她, tā, [tʰa], she'), which does not affect the comprehension of the sentences for native Mandarin speakers. For the accuracy of syllable element, they are 84.56%, 70.82%, 69.03%and 67.88%, respectively, significantly higher than the accuracy without the language model (p<0.001), which confirms that the language model plays a significant role in correcting errors from the prediction models of initial, tone, and final.

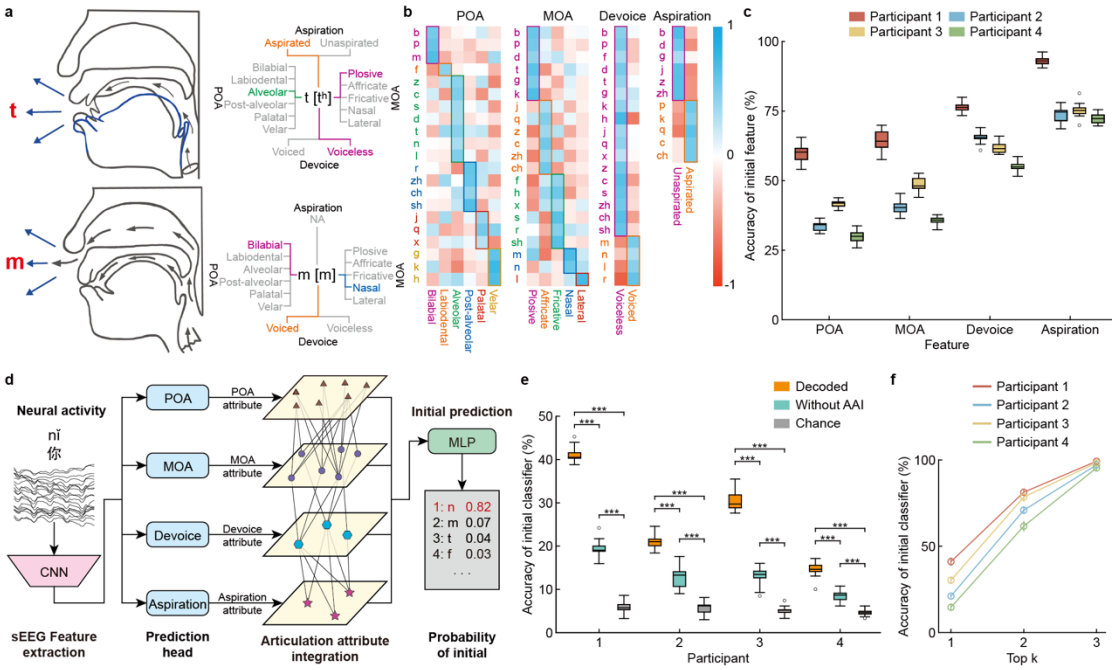

Fig 2. | Design motivation and performance of initial prediction model.

Initial prediction model design

In the Chinese *Pinyin* system, the initial consonant is termed initials, and there are a total of 21 distinct initials. The articulation of initials is produced by overcoming the obstruction of airflow as it passes through the oral or nasal cavity. Four features, place of articulation (POA, including bilabial, labiodental, alveolar, retroflex, alveolo-palatal and velar), manner of articulation (MOA, including plosive, affricate, fricative, nasal and lateral), devoice (including voiced and voiceless) and

aspiration (including unaspirated and aspirated) uniquely describe the articulation process of each Mandarin initial consonants (Table 1). Figure 2a shows the articulation process and corresponding articulation features of two initials, which is ‘t’ ([t^h]) and ‘m’ ([m]).

Table 1. Mandarin Chinese initials

Manner of articulation		Plosive		Affricate		Fricative		Nasal	Lateral
Devoice		Voiceless		Voiceless		Voiceless	Voiced	Voiced	Voiced
Aspiration		Unaspirated	Aspirated	Unaspirated	Aspirated	NA	NA	NA	NA
Place of articulation	Bilabial	b [p]	p [p^h]					m [m]	
	Labiodental					f [f]			
	Alveolar	d [t]	t [t^h]	z [ts]	c [ts^h]	s [s]		n [n]	l [l]
	Retroflex			zh [ʈʂ]	ch [$ʈʂ^h$]	sh [ʂ]	r [ʐ]		
	Alveolo-palatal			j [tɕ]	q [$tɕ^h$]	x [ɕ]			
	Velar	g [k]	k [k^h]			h [x]			

To assess the similarity in neural activity during the articulation of initials with the same features, we quantified the correlation between specific initials and the average of their corresponding category (Fig. 2b). The results indicate a positive correlation (r=0.52, r=0.54, r=0.49 and r=0.59 for POA, MOA, devoice and aspiration, respectively) between the articulation of initial consonants and the average articulation of consonants of the same type. This can be observed from four diagonal lines in the heatmap of four features. Moreover, the correlation is significantly higher than that with mismatched articulation features (p<0.001). This suggests that the recorded sEEG signals encompass relevant information characterizing categorical articulation features. To quantify the carried information about the articulation features, we trained classification models to decode the four features from each trial of all initials using sEEG signals from selected channels (Fig. 2c). These classifiers showed significant information about four features for all participants with all decoding accuracy significantly higher than corresponding chance level.

Consequently, we build the initial prediction model by identifying the acoustic-phonetic features (Fig. 2d). We first map the brain signals to representations in the acoustic-phonetic feature spaces of POA, MOA, aspiration and devoice. Given that the voicing of initials results from the interactions among diverse voicing features, we address this inherent relationship with a GNN as an articulation attribute integration (AAI). As a result, the accuracy rates from the model were significantly above the chance level (p<0.001 for all participants), with scores of 41.07%, 21.10%, 30.32%, and 14.65% for the four participants, respectively, and the chance level is 5%. However, the sole prediction result from the model is evidently insufficient in accuracy for input into the language model for sentence transformation. Therefore, we calculated the accuracy considering multiple candidate outputs from

the model. Notably, there was a significant increase in prediction accuracy as the number of candidate results grew (all $p<0.001$). In the top 3 most probable predictions, the probability of containing the correct initial for the four participants was 99.24%, 97.12%, 98.39%, and 95.03%, respectively. In summary, the initial prediction model can identify accurate initial results in most trials based on sEEG signals from selected channels, and its near-perfect top-3 accuracy narrows the possible initial candidates from 21 down to 3. This reduces the computational burden on the third phase of the decoding system and enhances the efficiency of the selection process of language model.

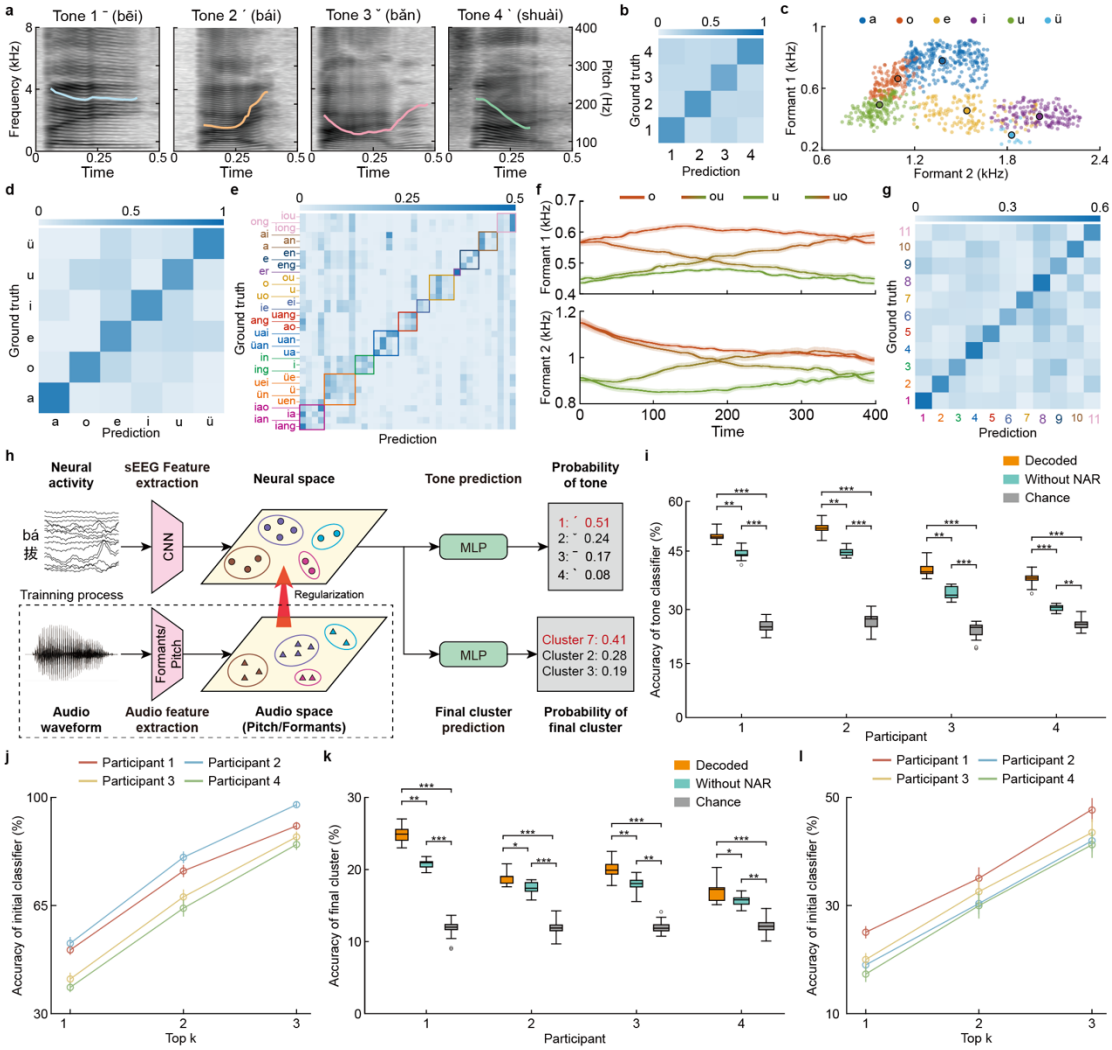

Fig. 3 | Design motivation and performance of tone and final cluster prediction models.

Tone prediction model design

Tone is a unique and vital component within a complete Mandarin syllable[18]. It is borne on the finals and can distinguish lexical and grammatical meaning. We visualized the changes in pitch over time for these four tones from Mandarin (Fig. 3a). The pitch variations for the four tones are markedly different, and they remain relatively stable across different trials of the same tone. To quantitatively distinguish between the four tones based on pitch, we trained an SVM classification model using all trial pitch data (combining the same tones borne on different rhymes). The four different tones

can be distinguished from the perspective of pitch variations, and the values on the diagonal are significantly higher than the off-diagonal values in Fig. 3b ($p=0.017$). The final classification accuracy was 57.51%, significantly above the chance level ($p<0.001$). Consequently, we incorporated pitch into the training process of the tone prediction model (Fig. 3h). Aligning the distribution of sEEG data in the neural space with the distribution of audio data, particularly concerning pitch features, enhances the extraction of neural features for tone categorization. We incorporate a neural-audio regularization (NAR) term based on similarity measurement in addition to the categorical cross-entropy loss during model training. Audio signals will solely be utilized in the decoder training process and not during testing.

The tone prediction model established using the NAR achieved prediction accuracies for the four participants of 50.60%, 52.75%, 41.19%, and 38.48%, respectively, all of which were significantly higher than the chance level ($p<0.001$ for all participants, Fig. 3i). Moreover, these results were also significantly higher than the accuracies from models without the NAR, underscoring the critical role of NAR in the model. Similarly, a singular prediction cannot encompass all the correct tones. Thus, we opted to select the top 3 most probable tones as the element sequence for input into the third phase of the decoding system to enhance the accuracy of final sentence prediction. The top 3 tones cover the correct tones with probabilities of 90.77%, 97.66%, 87.21% and 84.76% for four participants, respectively (Fig. 3j). It is worth noting that the probability of tones in the sequence will also be delivered to the language model for more reference information.

Identification and prediction of final cluster

The finals within Mandarin syllables are extremely complex, consisting of 6 simple finals and 29 compound finals. The production of these finals involves not only the vibration of the vocal cords but also adjustments to vertical and anteroposterior positions of the tongue to articulate different finals. Correspondingly, the formant features F1 and F2 in the finals' frequency spectrum directly relate to the vertical and anteroposterior positions of the tongue, respectively. We visualized the formant features F1 and F2 for all trials of simple finals, revealing a clear clustering effect. To quantitatively distinguish among the simple finals using F1 and F2, we trained an SVM classification model to identify the simple final. This model effectively differentiated the audio of the 6 simple finals (Fig. 3d), and the classification accuracy stood at 66.35%, significantly above the chance level ($p<0.001$).

However, once the compound finals were incorporated, the classification results deteriorated (15.46%, compared to the chance level, $p= 0.14$), eliminating the clear diagonal previously observed (Fig. 3e). This decline in performance can be attributed to the dynamic nature of compound finals articulation, which involves transitioning from one articulatory shape to another. For instance, we visualized the changes in F1 and F2 during the articulation of the simple finals 'o' and 'u' alongside the compound finals 'ou' and 'uo' (Fig. 3f). The compound final 'ou' is articulated by transitioning from the sound of 'o' to 'u', while 'uo' undergoes the opposite transition, making their articulatory configurations exceedingly similar and hence challenging to distinguish. Therefore, based on the resonance peaks of the finals in the audio, we employed the k-means algorithm to cluster finals with similar articulatory configurations, ultimately forming 11 distinct finals clusters (Fig. 3e). The SVM classification models were used to identify the final clusters, achieving an accuracy of 40.86%,

which is significantly higher than the chance level (p=0.013).

Extracting the formant features from audio has proven beneficial for classifying final clusters. Therefore, we also employed the model used to predict tones with NAR to predict the final clusters based on sEEG signals (Fig. 3h). Given the complexity of finals, the model achieved accuracy of 25.10%, 19.02%, 20.04%, and 17.33% for four patients, respectively, all of which were significantly higher than the chance level. As the number of candidate clusters increased, the probability of including the correct final also rose (Fig. 3i). However, considering that a single cluster already contains multiple finals, to alleviate the selection pressure on the third phase of the decoding system, we still opted to input all finals from the top-predicted cluster into the element sequence. Although the final cluster prediction model couldn't achieve very high accuracy, in Mandarin Chinese, initials and tones can encapsulate most of the information. This reduces, to a certain extent, the impact of incorrect vowels on the decoding system's sentence prediction.

Discussion

In our research, we have introduced a high-performance language BCI approach specifically designed for decoding logosyllabic languages, such as Mandarin Chinese. Logosyllabic languages, with their logographic characters, present unique challenges in language processing. In contrast to the alphabetic system, where words are constructed from combinations of letters, individual characters in logosyllabic languages often represent entire words or morphemes. Additionally, tonal languages like Mandarin Chinese add another layer of complexity, as the tone of a word can completely alter its meaning. Based on the above features of Mandarin, we chose the 407 monosyllabic words as the training dataset since it includes all possible valid syllables in Mandarin, accounts for tonal variations, and covers combinations of all initials and finals, which may encompass all phonemic variations. Additionally, we used monosyllable with tone as the modeling unit because previous unit selection study has proved that the use of syllable with tone effectively increases the speech recognition accuracy of Mandarin Chinese compared to other units such as characters or phonemes[19]. Since around 60–70% of the world's languages are tonal[20], which is not reflected in the mainstream research on high-freedom language brain-computer interfaces, our study of Mandarin Chinese contributed fundamental insights to the broader landscape of tonal language interface design and brain-computer system development.

In this study, we did not attempt to decode the neutral tones, which may raise questions among some researchers. This decision was made considering that except for a small number of particles, most of the neutral tones are contextually derived from their full tone counterparts[21], changing their surface phonetic properties as a function of the preceding tones in vocal speech[22] but being not affected in the writing system as used in our study. Nonetheless, future investigations might delve further into the nuances of neutral tones for a more comprehensive understanding of their role in Chinese brain-computer interface research. We observed notable variations in the decoding performance when distinguishing between single finals, compound finals, and nasal finals, with the latter two showing less favorable results. This may be attributed to the finer timing and position properties of articulatory gestures during a continuous shift of vowels and nasals in compound finals and nasal finals compared to single finals[23]. Thus, we implemented a k-means clustering approach, utilizing F1 and F2 frequency parameters, to classify these non-single finals into 11 clusters. These

clusters share similar acoustic properties and articulatory patterns, such as 'in' and 'ing,' grouped based on the nucleus in the structure of finals, as the nucleus bears highest sonority scale[24]. Such groupings could potentially simplify the decoding process. Furthermore, when using Pinyin input, the accuracy requirements for finals are relatively lenient. These considerations provide a foundation for further investigations into the complexities of decoding compound finals and nasal finals in the context of Chinese brain-computer interface systems.

The automatic correction capability of the language model, coupled with its understanding of Chinese text, allows for a relatively flexible Pinyin sequence input. This enhances the error tolerance of the three element predictors in the second phase of decoder, which in turn significantly improves the accuracy of the final sentence prediction. Our designed language model is intended for open and unrestricted Chinese language communication. While the language model corrects the output, there may still be some disparity from the user's intended meaning. However, when using speech neuroprostheses in specific environments and scenarios, such as in the domestic settings of individuals with disabilities, the language model can be trained on a restricted, predefined corpus to yield more accurate output results.

Addressing the challenges of mitigating the need for extensive data acquisition to facilitate cross-day model transfer was beyond the immediate scope of this study. Nonetheless, this endeavor stands as a pivotal requirement in the development of a genuinely pragmatic and impactful system. Furthermore, it is essential to verify the online decoding performance, despite the promising outcomes of our offline analyses. Recent large language models benefit from extensive training data, but combining data across subjects and time points for neurophysiological signals does not yield commensurate enhancements due to inherent cross-patient and cross-day variations. Considering the model adaptation in a lightweight manner to ensure cross-day accuracy and enhancing decoder performance through the utilization of multi-subject data, the critical endeavor of optimizing the efficient utilization of data from diverse days and individuals remains a key focus for forthcoming research.

Reference

1 Beukelman, D. R., Fager, S., Ball, L. & Dietz, A. AAC for adults with acquired neurological conditions: a review. *Augment Altern Commun* **23**, 230-242 (2007).

2 Felgoise, S. H., Zaccheo, V., Duff, J. & Simmons, Z. Verbal communication impacts quality of life in patients with amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Frontotemporal Degener* **17**, 179-183, doi:10.3109/21678421.2015.1125499 (2016).

3 Cooney, C., Folli, R. & Coyle, D. Neurolinguistics Research Advancing Development of a Direct-Speech Brain-Computer Interface. *iScience* **8**, 103-125, doi:10.1016/j.isci.2018.09.016 (2018).

4 Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M. & Shenoy, K. V. High-performance brain-to-text communication via handwriting. *Nature* **593**, 249-254, doi:10.1038/s41586-021-03506-2 (2021).

5 Chaudhary, U. *et al.* Spelling interface using intracortical signals in a completely locked-in patient enabled via auditory neurofeedback training. *Nat Commun* **13**, 1236, doi:10.1038/s41467-022-28859-8 (2022).

6 Metzger, S. L. *et al.* Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. *Nat Commun* **13**, 6510, doi:10.1038/s41467-022-33611-3 (2022).

7 Oxley, T. J. *et al.* Motor neuroprosthesis implanted with neurointerventional surgery improves capacity for activities of daily living tasks in severe paralysis: first in-human experience. *J Neurointerv Surg* **13**, 102-108, doi:10.1136/neurintsurg-2020-016862 (2021).

8 Vansteensel, M. J. *et al.* Fully Implanted Brain-Computer Interface in a Locked-In Patient with ALS. *N Engl J Med* **375**, 2060-2066, doi:10.1056/NEJMoa1608085 (2016).

9 Pandarinath, C. *et al.* High performance communication by people with paralysis using an intracortical brain-computer interface. *Elife* **6**, doi:10.7554/eLife.18554 (2017).

10 Willett, F. R. *et al.* A high-performance speech neuroprosthesis. *Nature* **620**, 1031-1036, doi:10.1038/s41586-023-06377-x (2023).

11 Metzger, S. L. *et al.* A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* **620**, 1037-1046, doi:10.1038/s41586-023-06443-4 (2023).

12 Herff, C. *et al.* Brain-to-text: decoding spoken phrases from phone representations in the brain. *Front Neurosci* **9**, 217, doi:10.3389/fnins.2015.00217 (2015).

13 Duanmu, S. *The phonology of standard Chinese*. (OUP Oxford, 2007).

14 Shen, H. H. & Bear, D. R. Development of orthographic skills in Chinese children. *Reading and Writing* **13**, 197-236, doi:10.1023/A:1026484207650 (2000).

15 Wang, W. The Chinese Language. *Scientific American* **228**, 50-60, doi:10.1038/scientificamerican0273-50 (1973).

16 Tan, L. H., Xu, M., Chang, C. Q. & Siok, W. T. China' s language input system in the digital age affects children' s reading development. *Proceedings of the National Academy of Sciences* **110**, 1119-1123, doi:doi:10.1073/pnas.1213586110 (2013).

17 Li, P. & Yip, M. C. Context effects and the processing of spoken homophones. *Reading and Writing: An Interdisciplinary Journal* **10**, 223-243, doi:10.1023/A:1008091816322 (1998).

18 Liu, Y. *et al.* Decoding and synthesizing tonal language speech from brain activity. *Sci Adv* **9**, eadh0478, doi:10.1126/sciadv.adh0478 (2023).

19 Fu, L., Li, X. & Zi, L. Research on Modeling Units of Transformer Transducer for Mandarin Speech Recognition. *ArXiv* **abs/2004.13522** (2020).

20 Yip, M. J. W. *Tone*. (Cambridge University Press, 2002).

21 Shen, X. S. Mandarin neutral tone revisited. *Acta linguistica hafniensia* **24**, 131-152 (1992).

22 Chao, Y. R. *A grammar of spoken Chinese*. (Univ of California Press, 1968).

23 Cai, S., Ghosh, S. S., Guenther, F. H. & Perkell, J. S. Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong/iau/and its pattern of generalization. *The Journal of the Acoustical Society of America* **128**, 2033-2048 (2010).

24 Yavas, M. S. & Gogate, L. J. Phoneme awareness in children: A function of sonority. *Journal of psycholinguistic research* **28**, 245-260 (1999).