A brain-to-text framework of decoding natural tonal sentences

Daohan Zhang[1,2,3#], Zhenjie Wang[4#], Youkun Qian[1,2,3#], Zehao Zhao[1,2,3], Yan Liu[1,2,3], Xiaotao Hao[1,2,3], Wanxin Li[1,2,3], Shuo Lu[5], Honglin Zhu[6], Luyao Chen[7], Kunyu Xu[8], Yuanning Li[4,9*], Jinsong Wu[1,2,3*], Junfeng Lu[1,2,3,8,10*]

Author affiliations:

[1]Department of Neurosurgery, Huashan Hospital, Shanghai Medical College, Fudan University, Shanghai 200040, China

[2]Shanghai Key Laboratory of Brain Function Restoration and Neural Regeneration, Shanghai 200040, China

[3]National Center for Neurological Disorders, Huashan Hospital, Shanghai Medical College, Fudan University, Shanghai 200040, China

[4]School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China

[5]Department of Chinese Language and Literature, Sun Yat-sen University, Guangzhou 510080, China

[6]Faculty of Life Sciences and Medicine, King's College London, London SE1 1UL, UK

[7]School of International Chinese Language Education, Beijing Normal University, Beijing 100875, China

[8]Institute of Modern Languages and Linguistics, Fudan University, Shanghai 200433, China

[9]State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai 201210, China

[10]MOE Frontiers Center for Brain Science, Huashan Hospital, Fudan University, Shanghai 200040, China

[#] These authors contributed equally to this work.

[*] Corresponding authors: Junfeng Lu (junfeng_lu@fudan.edu.cn), Jinsong Wu (wujinsong@huashan.org.cn), Yuanning Li (liyn2@shanghaitech.edu.cn)

Abstract

Speech brain-computer interfaces (BCIs) directly translate brain activity into speech sound and text, yet decoding tonal languages like Mandarin Chinese poses a significant, unexplored challenge. Despite successful cases in non-tonal languages, the complexities of Mandarin, with its distinct syllabic structures and pivotal lexical information conveyed through tonal nuances, present challenges in BCI decoding. Here we designed a brain-to-text framework to decode Mandarin tonal sentences from invasive neural recordings. Our modular approach dissects speech onset, base syllables, and lexical tones, integrating them with contextual information through Bayesian likelihood and the Viterbi decoder. The results demonstrate accurate tone and syllable decoding under variances in continuous naturalistic speech production, surpassing previous intracranial Mandarin tonal syllable decoders in decoding accuracy. We also verified the robustness of our decoding framework and showed that the model hyperparameters can be generalized across participants of varied gender, age, education backgrounds, pronunciation behaviors, and coverage of electrodes. Our pilot study shed lights on the feasibility of more generalizable brain-to-text decoding of natural tonal sentences from patients with high heterogeneities.

Key words

Electrocorticography (ECoG); Brain-Computer Interface (BCI); Tonal language; Natural speech; Neural Networks

Introduction

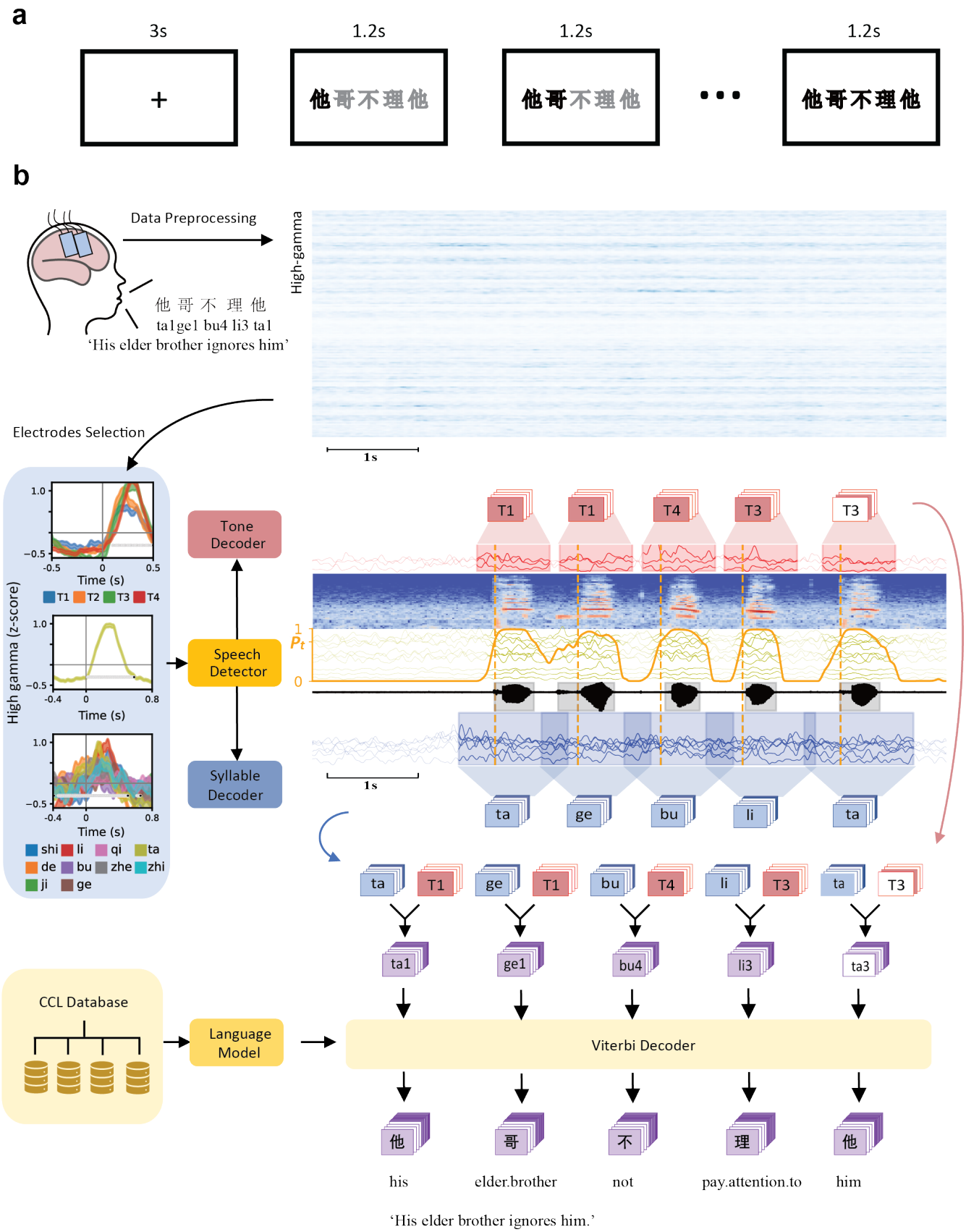
Sentence is the basic language unit embodies our construal of representational meaning and interpersonal meaning, which constitutes the basis for daily communication[1]. Several recent studies have shown the feasibility of decoding or synthesis sentences of non-tonal languages, such as English[2-10] and Japanese[11], using intracranial neural recordings, such as electrocorticography (ECoG). These studies have primarily relied on decoding the spatiotemporal neural patterns associated with articulatory movements—such as those of the lips, tongue, and larynx—in the ventral sensorimotor cortex during intended speech production. These advances not only provide new methods in anarthria treatment[12] but also increase the communication efficiency of speech brain-computer-interface (BCI).

However, decoding tonal sentences is still a largely-unexplored work. More than 60% of the languages in the world are tonal[13], with approximately 2 billion people speaking tonal languages, including most Sino-Tibetan languages and the entire Tai-Kadai family[14]. The distinct feature of these languages is the use of pitch to distinguish lexical and grammatical meaning. While prior research has investigated decoding stereotypical instances of lexical tones from neural activity for monosyllabic speech[15], decoding continuous tonal sentences is still a challenging issue. Unlike the relatively stable acoustic cues in canonical forms, natural speech introduces substantial variability in tone components. According to Fujisaki model, these components, typically represented by base frequency (F0) contours, encompass base frequency, phrase variations, and accent components[16,17]. Furthermore, the influence of tone sandhi—alterations in a morpheme's tone due to syntactic context—adds an additional layer of complexity[18]. Altogether, these variances make decoding tonal sentences more complicated than both non-tonal sentences and isolated tonal syllables.

Besides, existing invasive language BCI usually reported success in individual cases, usually with intense hyperparameter optimizing[5,9,10]. Few studies have tested the replication or generalization of the decoding framework. It remains uncertain whether the same set of model design hyperparameters—such as the number of convolution and recurrent blocks, hidden variables per layer, and onset detection thresholds—will generalize across different subjects. Consequently, verifying the generalizability of these published frameworks, especially when applied to patients exhibiting high heterogeneity, remains elusive. In clinical settings, different patients need to restore speech function via speech BCI have varied pronunciation behaviors, which is especially obvious among tonal languages such as Mandarin Chinese. Mandarin speakers, influenced by regional dialects such as Northern, Jianghuai, and subgroups like Wu-Tai, Shanghainese-influenced, Northeastern Mandarin, demonstrate distinct tonal and syllabic variations[19]. Though sharing the same written form of Chinese word, speakers of these mandarin branches have distinct pronunciation behaviors on both tones and syllables. While these distinct pronunciation behaviors or preferences do not affect speakers' daily communication, they add significant challenges to the decoding task. In addition to the variance in pronunciation behaviors, the inherent variations in language-functional cortical areas also contribute to the inter-subject variances. It is impossible for invasive devices to collect identical neural signals from different patients even if the placement of invasive electrodes is stringently controlled. Whether and to what extent these bias from both language behaviors and varied neural signals will affect the robustness of language decoding framework remain unknown. This uncertainty challenges the eventual generalization of speech BCI across patients with high heterogeneity.

In this study, we aim to decode Mandarin tonal sentences from invasive neural recordings using high-density ECoG. We used different neural network modules targeting functionally distinct

neural populations in speech-related brain areas. These modules detected the onsets of the utterance of each individual Chinese character and then decode tone labels and syllable labels in parallel, mirroring the inherit parallel coding of the articulation process in tonal languages[20-22]. Subsequently, a language model was used to calculate the Bayesian likelihood of the entire sentence from the probability distribution of tonal syllable sequences, integrating the contextual and prior information (**Fig.1**). Moreover, we established a versatile framework using a set of standardized hyperparameters, eliminating additional needs for hyperparameter optimization, and assessed its potential for generalization by testing across multiple patients. This approach decoded Mandarin tonal sentences across diverse patient profiles without specific hyperparameter adjustments.



76 **Fig 1 Schematic overview of the brain-to-text decoder for natural speech of a tonal language.**

a. Schematic overview of the tonal speech production task. Each participant was guided by a visual cue to produce one of the ten sentences. Each trial began with a fixation cross at the center of the screen for 3 seconds, and the sentence was shown in the middle of the screen in grey text. With each word turning black for 1.2 seconds consecutively from the beginning to the end of the sentence, the participant was instructed to pronounce the sentence following these go cues, the pace of the speech was not strictly aligned with the visual cues.

b. During the experiment, neural activity from speech-related cortex was recorded using implanted high-density electrocorticography array when the participant was instructed to read sentences consisted of words from a predefined vocabulary set of 40 words. The preprocessed neural signals from responsive electrodes were sampled by sliding window and was sent to the speech detection module (speech detector) to detect the onsets of words. The peri-onset neural activity within a fixed time window was used to compute the base syllable probability (across 10 possible syllables) and lexical tone probability (across 4 tones) of the word via a tone decoder and a syllable decoder respectively. A Viterbi decoding algorithm used these probabilities in conjunction with word-sequence probabilities from a pre-trained natural language model to decode the most likely sentence given the current neural activity.

Results

In this study, we recorded the neural activity of five native Mandarin-speaking participants who underwent awake surgery to treat brain tumors. Each participant was guided by sequential visual cues to produce 10 sentences consisted of 5-8 Chinese characters from a corpus of 40 Chinese characters with varied tones **(Fig. 1a)**. Among 5 participants, PA1-4 completed the speech production task using normal articulating, while PA5 completed the speech production task using whispering. Participants' brain activity was recorded by temporally placed high-density ECoG grids. Subsequently, we assessed the efficacy of our proposed brain-to-text speech decoder across these 5 participants.

Decoder overview

Our brain-to-text decoder comprises interconnected modules: a speech detector, tone decoder, syllable decoder, and a language model, functioning in a sequential stream. To decode Mandarin tonal sentences from high-density ECoG neural recordings, our brain-to-text decoder starts with a speech detector, which takes in sequence of neural activity recordings of pre-selected speech-responsive electrodes and predicts whether each timepoint belongs to speech production or resting-state. The output of the speech detector was used to identify speech production epochs, where each epoch corresponding to the time interval of one Chinese character. Based on the decoded speech production epochs, our tone and syllable decoders then decode tone label and syllable label from neural signals of pre-selected tone and syllable discriminative electrodes during each epoch.

For PA1 to PA4, we identified 108,144,199, and129 speech responsive electrodes, 30,28,68, and 75 syllable discriminative electrodes and 10,12,21, and 48 tone discriminative electrodes, respectively. For PA5, we found 118 speech responsive electrodes, 47 syllable discriminative electrodes and 14 tone discriminative electrodes. The coverage and overlapping relationship of these electrodes were shown in **Fig. 2a-e**.

For onset predicting, the high gamma activity across all speech-responsive electrodes was processed time point by time point by an artificial neural network (ANN) consisting of Convolutional Neural Network (CNN) structure following a stack of Gated Recurrent Unit (GRU) layers and a single dense (fully connected) layer, capturing both forward and backward temporal dependencies in neural signals, which was designated for inference on dynamic temporal processes. The dense layer projected the latent dimensions of the last GRU layer into probability space for two event classes: speech and rest. **(Fig. 2f)** For simplify, all the hyperparameters in this step was fixed. After that, the curve of the predicted probabilities along the dimension of time was smoothed and binarized according to undefined smoothing window ($\boldsymbol{S}$) and probability threshold ($\boldsymbol{P_t}$), the

onset of each utterance was then predicted if a lasting silence states ($\boldsymbol{T_{off}}$) before the onset and a lasting speech state ($\boldsymbol{T_{on}}$) after it, while slightly errors labels within an undefined error permissive rate (***EPR***) was also allowed **(Fig. 3b)**. These five undefined hyperparameters need further tuning through grid search for each participant.

As for tone and syllable decoder, a 1.2-second time window for syllable and 0.8-second time window for tone of high gamma activity was processed by an ensemble of 5 pairs (10 total) ensemble ANN models. ANNs in tone and syllable decoder share the same architecture but different in their own hyperparameters. Within each ANN, the high gamma activity was processed by an initial convolution with initial filter length ($\boldsymbol{FL_{ini}}$), stride of convolution ($\boldsymbol{ST_{conv}}$). This initial layer was followed by undefined number ($\boldsymbol{C_{layer}}$) of CNN unit[12]. Each CNN unit constituted of a temporal convolution with undefined kernel length ($\boldsymbol{FL_{conv}}$), aforementioned stride length and dimension ($\boldsymbol{C_{dim}}$), a batch normalization, ELU activating function[23], a dropout layer with dropout rate ($\boldsymbol{D}$) and maxpooling layer with max-pooling kernel length ($\boldsymbol{L_{pool}}$) and max-pooling stride ($\boldsymbol{ST_{pool}}$). After that, data was processed by undefined number ($\boldsymbol{R_{layer}}$) of stacked bidirectional gated recurrent unit (GRU) layers with undefined number of dimension ($\boldsymbol{R_{dim}}$) [24]. A dense layer projected the latent dimension from the final GRU layer into probability space, which contained the probability of syllable and tone of each of the words from the 10-syllable set or 4-tone classes, being the feature of target word during the speech production attempt associated with the neural time window. The 10 or 4 probability distributions from the ensembled ANN models were averaged together to obtain the final vector of predicted word probabilities. **(Fig. 2g)** It is worth noting that all the aforementioned undefined hyperparameters required further tuning for each participant.

Subsequently, a language model integrates these decoded tone and syllable labels, along with prior information of their transitional probabilities, to compute the Bayesian likelihood of entire word sequences.

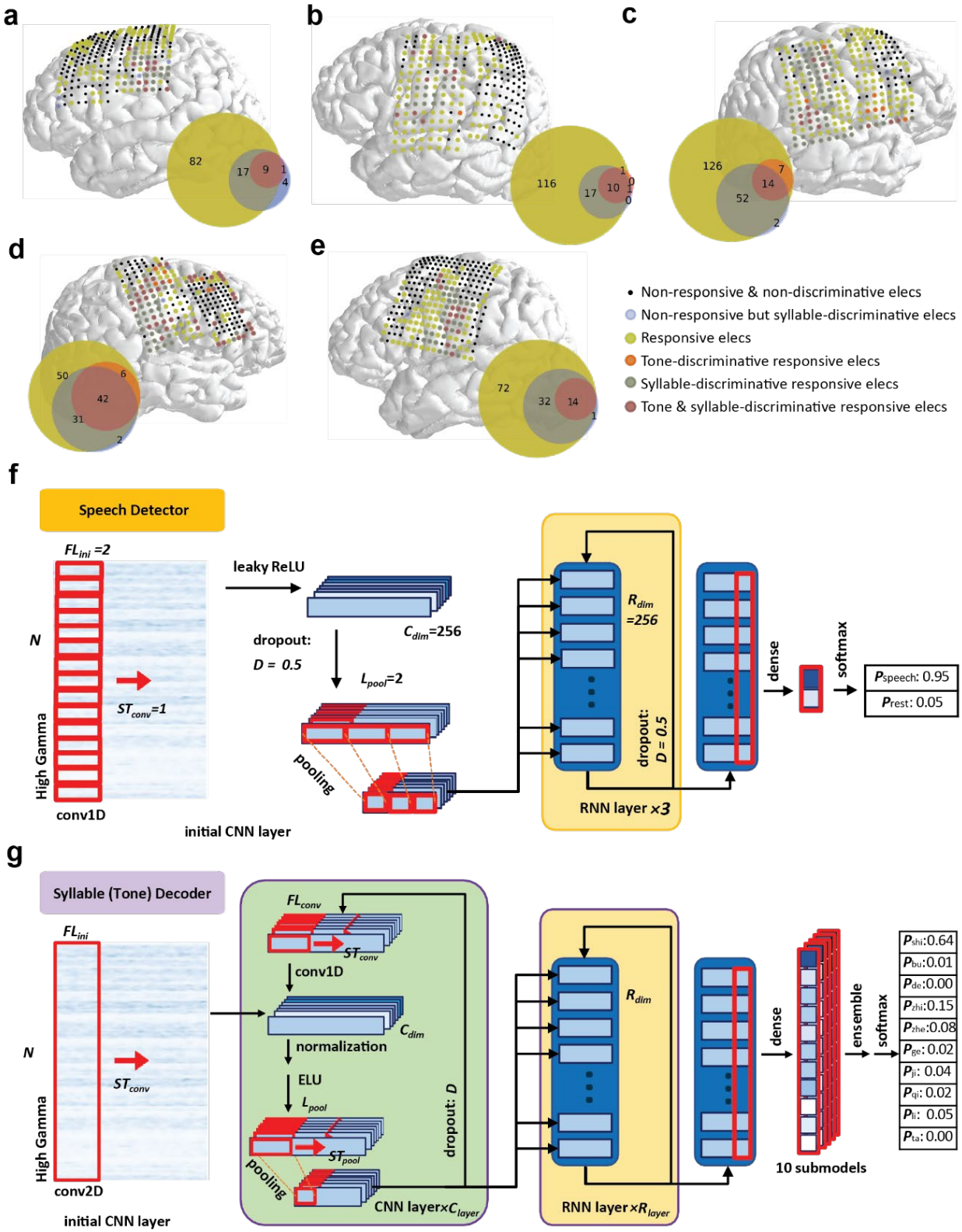


Fig 2 Electrode coverage, category and decoding model schematics

a-e. Anatomical reconstructions of PA1-PA5 (from a to e). The locations of the ECoG electrodes were plotted with colored discs. The colors indicated the electrode categories.

Yellow: responsive electrodes; red: tone-discriminative electrodes; blue: syllable-discriminative electrodes. Electrodes with combined feature were plotted with mixed colors, nonresponsive electrodes were plotted as small black dot. Venn diagrams showed the number of electrodes in each category for each participant.

f. Speech detection model schematic. Predefined hyperparameters in ANN and their values were shown in italic.

g. Syllable classification model schematic. CNN unit shown in green frame while RNN unit shown in yellow. All of the undefined hyperparameters were shown in Italic, which can be roughly divided into time-dimension-related group and size-related group. Former group includes initial convolution filter length ($\boldsymbol{FL_{ini}}$), stride of convolution ($\boldsymbol{ST_{conv}}$), filter length of following convolutional blocks ($\boldsymbol{FL_{conv}}$), max-pooling kernel length ($\boldsymbol{L_{pool}}$) and max-pooling stride ($\boldsymbol{ST_{pool}}$). Later one including number of sequential convolutional blocks ($\boldsymbol{C_{layer}}$), number of layers of RNN ($\boldsymbol{R_{layer}}$), number of filters in each convolutional process ($\boldsymbol{C_{dim}}$), number of dimensions in each RNN process ($\boldsymbol{R_{dim}}$), and dropout value ($\boldsymbol{D}$)) The tone classification model shared the same architecture with syllable classification model but different hyperparameters.

Independent performance of speech detector, tone decoder and syllable decoder

First, we evaluated the performance of each individual decoder module. Since these decoder modules work in sequential order, the performance of the tone and syllable decoders would rely on the output of the speech detector. To evaluate the independent performance of these decoder modules, we first calculate the decoding accuracies of syllable, tone, tonal syllable, and Chinese characters on manually aligned speech onsets rather than onsets predicted by the speech detector.

Each of the participants completed 158 to 160 sentences in the speech production task. Using nested cross-validation, we trained the brain-to-text decoder for each participant and evaluated the decoding performance. Among five participants, our speech detector reached an area under curve (AUC) of 0.96 to 0.99 **(Fig. 3a)**. The tone decoder reached an accuracy of 57%±12% (mean [± SD] classification accuracy across the 4 target tones in 5 participants, chance 25%) on manually aligned neural signals **(Fig. 3c)**. Syllable decoder reached an accuracy of 69%±5% (mean [±SD] classification accuracy across the 10 target syllables in 5 participants, chance 10%) on manually

aligned neural signals **(Fig. 3d)**. When multiply raw predicted probability of tone and syllable, we acquired raw decoding accuracy of all the predicted tonal-syllables, which reaches 46%±8% (mean [±SD] classification accuracy across the 40 tonal syllables on 5 participants, chance 3.0%) at the level of tonal syllable **(Fig. 3e)**. After applying language model to calculate corresponding Chinese words, the accuracy at the level of Chinese words reaches 87%±4% (mean [±SD] classification accuracy across the 40 Chinese words on 5 participants, chance 2.5%) **(Fig. 3f)**.

Overall decoding performance of tonal-sentences

After evaluated the independent performance, we then tested the performance of all the modules when they were interconnected and worked in a stream. Aligned with speech onsets detected by speech detector, tone decoder achieved an accuracy of 56%±13% **(**mean [±SD], **Fig. 3g)**, while syllable decoder achieved an accuracy of 64%±6% **(**mean [±SD] , **Fig. 3h)**. Accuracy at the level of raw tonal syllable reached 44%±7% **(**mean [±SD], **Fig. 3i)**, while the final accuracy at the level of Chinese words reaches 82%±4% **(**mean [±SD], **Fig. 3j)**. These results were consistent with the independent module performance, when actual speech onsets were manually aligned. We also calculated the word error rate (WER) of each decoded sentence. Decoding accuracy was consistent across participants (one way-ANOVA, $F(4, 45) = 0.65$, $p=0.63$), as the overall WER was 25%± 6%, 21%±5%, 21%±6%, 14%±3%, 18%±4% (mean [±SE]) from PA1 to PA5. For each individual sentence, WER ranged from 1% to 52%. 47.74% of all the sentences (95/199) were decoded correctly (WER=0) (**Fig. 3k**). There was no significant difference (Pearson's correlation, r = -0.03, $P = 0.61$) between the decoding accuracy and the complexity (number of words or

phrases) of the sentence (**Fig. 3m**), suggesting that our proposed method worked for both short and long sentences.

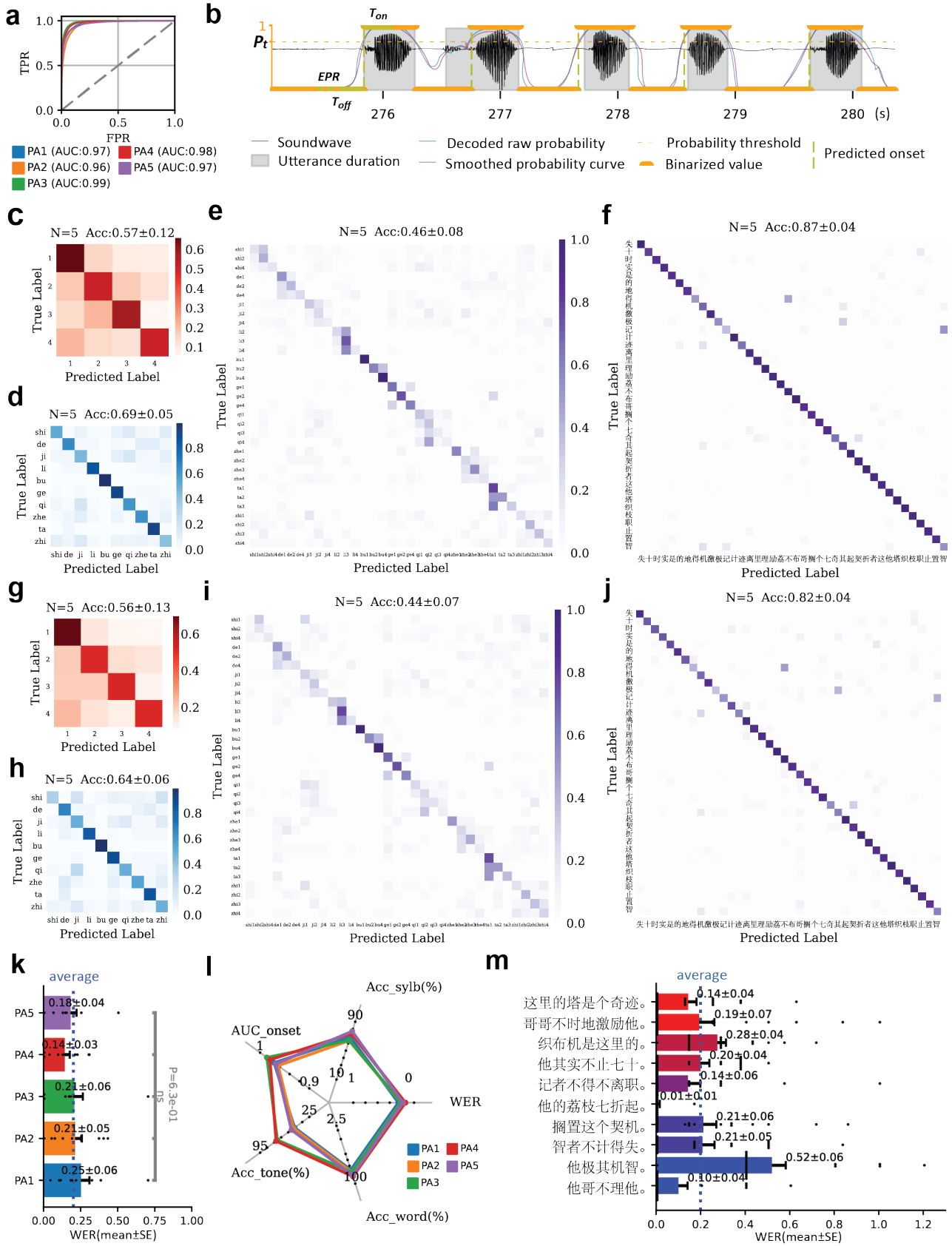


Fig 3. Evaluation of the overall neural-to-text decoding performance of the decoder.

a. Receiver Operating Characteristic (ROC) curves and corresponding area under curve (AUC) of speech detectors in each participant.

b. One example trial, in which the participant produced the sentence "□ □ □ □ □ " (His elder brother ignores him). The raw speech sound waveform was plotted in black. The time course of the predicted speech probability was plotted blue, while speech probability after smoothing and binarized was plotted red and orange. Finally, the detected speech event onsets from the neural decoder was plotted as yellow dotted vertical lines. All of the undefined hyperparameters were shown in Italic.

c-f. Confusion matrices of c) the tone labels, d) the syllable labels, e) the tonal syllables, and f) the words (with language model), evaluated on the test set, using manually aligned actual speech onsets.

g-j. Confusion matrices of g) the tone labels, h) the syllable labels, i) the tonal syllables, and j) the words (with language model), evaluated on the test set, using decoded speech onsets from the onset decoder.

k. The averaged word error rate (WER) of decoded sentences in 5 participants, mean [±sem], the vertical dotted line indicated the overall averaged performance across all 5 participants.

l. Performances of the speech detector, the tone decoder, the syllable decoder and the overall performances of each participant (in the same color keys as panel **k**), shown by AUC of speech-silent classifier (AUC_onset), tone accuracy (Acc_tone), syllable accuracy (Acc_sylb), word accuracy (Acc_word) and WER respectively.

m. The averaged WER of each individual sentence across all 5 participants, mean [±SE], dotted line indicated the average performance across all sentences.

Robustness of the speech decoder under tonal variance in natural speech

In natural speech, the actual pitch trajectories of lexical tones often deviate from their canonical forms, due to accent, emotions, and other contextual effect such as coarticulation and tone sandhi[25]. For example, all the patient articulated the tone of "不" in "不计得失" with a pitch trajectory more similar to tone 2 rather than tone 4 in single syllable form or in other phrases such as "不得不", due to the rule of tone sandhi (Fig 4a). Even the same tonal syllable with different phonological context would result in different pitch trajectories (Fig 4b). This suggests that the underlying neural code commanding tone articulation may also encounter great variance. Therefore, the decoding algorithm should not only consider the stereotypical canonical monosyllable instances of lexical

242 tones, but also able to account for such significant variance during natural speech and robustly
243 decode lexical tones regardless of the variance.

244 First, we quantify the variance caused by tonal sentence context and tone sandhi behaviorally. We
245 performed tone intelligibility assessment (IA) test through which native mandarin speaking
246 participants listening to audio of each tonal Chinese word clipped from the natural tonal sentences
247 and judging the tone. We find the accuracy of the tone IA test only reached 70%±2% (mean [±SE]),
248 significantly lower (t-test, $t(38) = 7.38$, $p = 7.6e-9$) than behavioral performance under full natural
249 context (tona IA accuracy 91%±1%, mean [±SE], Fig 4c). Therefore, using contextual information
250 is important for listeners to overcome tonal variances in natural speech production.

251 Similarly, given only monosyllabic information, the tone decoder would perform suboptimal. To
252 show this, we adopted a baseline monosyllable decoder model previously used in Liu et al[15]. The
253 monosyllable decoder only took in the neural activity aligned to the current syllable utterance and
254 did not consider contextual syllables. In our test dataset, the monosyllable baseline decoder model
255 achieved an averaged tone decoding accuracy of 35%±3% (mean [±SE], Fig. 4**c**).

256 Finally, we tested the performance our proposed sentence decoder. We compared our tone-decoder
257 (without the language model) with our previous published tone decoder designed for
258 monosyllable[15]. Our framework (decoding accuracy 58%±9%, mean [±SE]) outperformed (t-test,
259 $t(30) = -4.62$, $p = 6.9\times10-5$) previous monosyllable tone-decoder (decoding accuracy 32%±3%,
260 mean [±SE]) on tonal sentence decoding task, which suggests our framework is more robust on
261 decoding tone components from natural language settings with multiple and unbalanced syllables.
262 Furthermore, after introducing language model to our decoding framework, a tone decoding
263 accuracy of 93%±3% (mean [±SE]) was achieved, which significantly exceed (t-test, $t(34) = 7.12$,

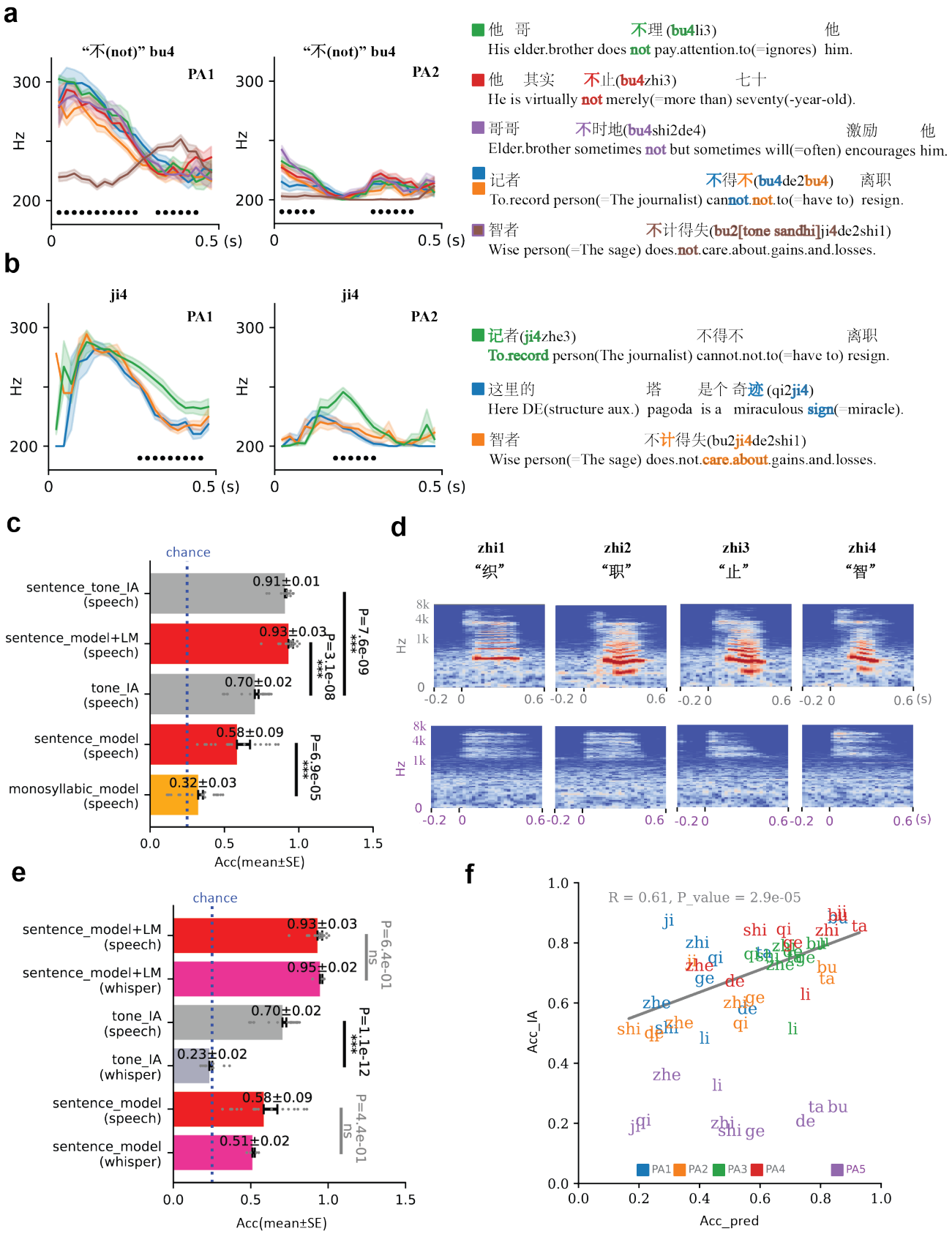
$P = 3.1\times10^{-8}$, **Fig. 4c**) the accuracy of monosyllable tone IA results (70%±2%, mean [±SE]), and approximate the actual behavioral performance of native speakers under natural context (91%±1%,). Hence, applying language model will largely eliminate the tone variance in natural language by introducing contextual relationships, which is a promising solution of decoding tone in natural tonal sentences accurately.

Speech decoding during non-intelligible tonal speech production

To further evaluate the robustness of our proposed decoder under different speech production scenario, we tested if the decoder could decode from patient who did not produce intelligible speech. In particular, PA5 completed the speech production task using whispering and the produced tonal speech was largely non-intelligible. We first plot the Mel-spectrograms of Chinese words clipped from whispering natural tonal sentences, finding no obvious base frequency (F0) and consonant peak (**Fig. 4b**). Furthermore, the tone IA score of PA5 was 23%±2% (mean [±SE]), which was significantly lower than tone IA scores of normal articulating audios (t-test, $t(28) = -12.18$, $p = 1.1\times10^{-12}$) and within the range of chance level (**Fig. 4c**). Therefore, behaviorally PA5 was not able to produce intelligible tonal speech during the task.

We then compared the tone-decoding accuracy of our framework on both participant PA5 and other normal articulating participant, the intended speech was consistently decoded, the average accuracy was 93%±3% and 95%±2% (mean [±SE]) with language model, 58%±9% and 51%±2% (mean [±SE]) with language model, no significant difference was observed (t-test, with language model: $t(18) = 0.48$, $p = 6.4\times10^{-1}$, without language model: $t(18) = -0.79$, $p = 4.4\times10^{-1}$).

Finally, tone decoding accuracy and the tone IA scores of each syllable in the 4 articulating participants were significantly correlated (Pearson's correlation $r = 0.61$, $p = 2.9\times10–5$), indicating our tone decoder learned from behaviorally-relevant tone-related neural features, rather than other co-variants (**Fig. 4d**).



290 **Fig 4. Evaluation of tonal decoding under natural contextual variances and whispering.**

a-b) Averaged pitch contours (mean [±SE]) of different Chinese characters of the same tonal syllables (a: bu4, also the same Chinese character "不" which represents negative meanings such as "not", b: ji4), black dots indicated time points with significant mean difference (one-way ANOVA, $p < 0.05$). The text transcriptions were shown in the bottom with the specific words highlighted in the corresponding colors. Corresponding Chinese characters, *Pinyin*, and English translation of the choosen syllables were shown in the same color as the pitch contour at the right part of the figures. In the last sentence of the subplot **a**, the tone 4 of "ji" (also shown in color) leads to the change in the tone of previous "bu", due to the rule of tone sandhi.

c) Tone decoding performances (mean [±SE]) of previous monosyllabic neural decoding model (monosyllabic_model), our sentence-based neural decoding model without language model (sentence_model), our sentence-based neural decoding model with language model (speech_sentence+LM), the accuracy of tone intelligible analysis (tone IA score) by 20 volunteers using the corresponding speech audio syllable clips (tone_IA), and speech audio of each full sentences (speech_tone_IA). Blue dotted line indicated chance level (25%); *** $p < 0.001$, t-test, two-sided.

d) The mel-spectrograms of participant (PA1) spoke aloud four tones of "zhi" (upper row), compared to the mel-spectrograms of participant (PA5) whispering four tones of "zhi" (bottom).

e) Tone decoding performances using our proposed sentence decoder without language model (sentence_model) and with language model (sentence_model + LM), and the tone IA scores (mean [±SE]) evaluated on the speech and whisper participants.

f) Scatterplot showing the correlation between of tone decoding accuracy (Acc_pred) and tone IA score (Acc_IA) of articulating participants (Pearson's correlation, $r = 0.61$, $p = 2.9\times10–5$). Whispering data shown in purple, which was not included in the linear regression due to randomly distributed tone IA scores.

Universal decoding framework with predefined hyperparameters is applicable on participants with great variations

ANN-based BCI decoders often encounter model hyperparameters such as depth and width of the network, kernel size and strides, dropout rate, etc. Training these models usually require optimization of these hyperparameters for each individual subject. A key factor of the generalizability of the BCI decoder is how robust the model performance is regarding different sets of hyper-parameters across participants. To get an overall understanding of heterogeneity

across participants, we computed their best hyperparameter combinations of speech detector, tone decoder and syllable decoder chosen by the optimization process. These hyperparameters include smoothing window (***S***), the probability threshold value ($\boldsymbol{P_t}$), the off-time threshold ($\boldsymbol{T_{off}}$) and on-time threshold ($\boldsymbol{T_{on}}$), and the error permissive rate (***EPR***) which are related to the thresholding utterrance onsets; filter lengths ($\boldsymbol{FL_{ini}}$ for the first convolutional layer, $\boldsymbol{FL_{conv}}$ for the rest convolutional layers, and $\boldsymbol{L_{pool}}$ for max-pooling kernel) and strides ($\boldsymbol{ST_{conv}}$ for convolutional layers and $\boldsymbol{ST_{pool}}$ for max-pooling layers) which determined the temporal feature of neuro-decoders; as well as depths ($\boldsymbol{C_{layer}}$ for sequential convolutional blocks and $\boldsymbol{R_{layer}}$ for stacked recurrent layers) and widths ($\boldsymbol{C_{dim}}$ for number of filters for convolutional process, $\boldsymbol{R_{dim}}$ for number of dimensions in each RNN process, ***D*** for dropout value) which determined the overall model architecture.. Very few optimized hyperparameter remained the same across all participants, while most optimal hyperparameters varied across participants with variations of age, gender, speech behaviors and electrodes coverage **(Fig. 5a-e)**. To figure out the impact of such heterogeneity on decoding performance in our framework, we picked the medium value of each hyperparameter and constructed a fixed pre-defined hyperparameter set. We then applied a universal decoder (UNI) on all the participants using this fixed pre-defined hyperparameter set. We found this universal decoder performed similar to (pair-wise t-test,PA1: $t(9) = -0.62$, $p = 5.5\times10\text{-}1$, PA2: $t(9) = 0.05$, $p = 9.6\times10\text{-}1$, PA3: $t(9) = -2.00$, $p = 7.6\times10\text{-}2$, PA4: $t(9) = 0.23$, $p = 8.3\times10\text{-}1$, PA5: $t(9) = -0.55$, $p = 5.9\times10\text{-}1$) individually-optimized frameworks with hyperparameters optimized on individual participants (**Fig. 5f**). Also, the decoding performance of universal framework across participants was stable (oneway ANOVA, $F(4, 45) = 1.11$, $p = 3.6\times10\text{-}1$). The UNI decoder even outperformed four control models only contained RNN and CNN part of the framework (CNN and RNN) with optimized hyperparameters. Mean WERs of decoding sentences through UNI remained lowest in

all participants (**Fig. 5g**). Though the advantage was not significant in each individual participant due to relatively small sample size, it was significant after combining data of all the participants (paired t-test, UNI with CNN, $t(98) = -3.29$, $p = 1.8\times10\text{-}3$, UNI with RNN, $t(98) = -4.97$, $p = 8.3\times10\text{-}6$).

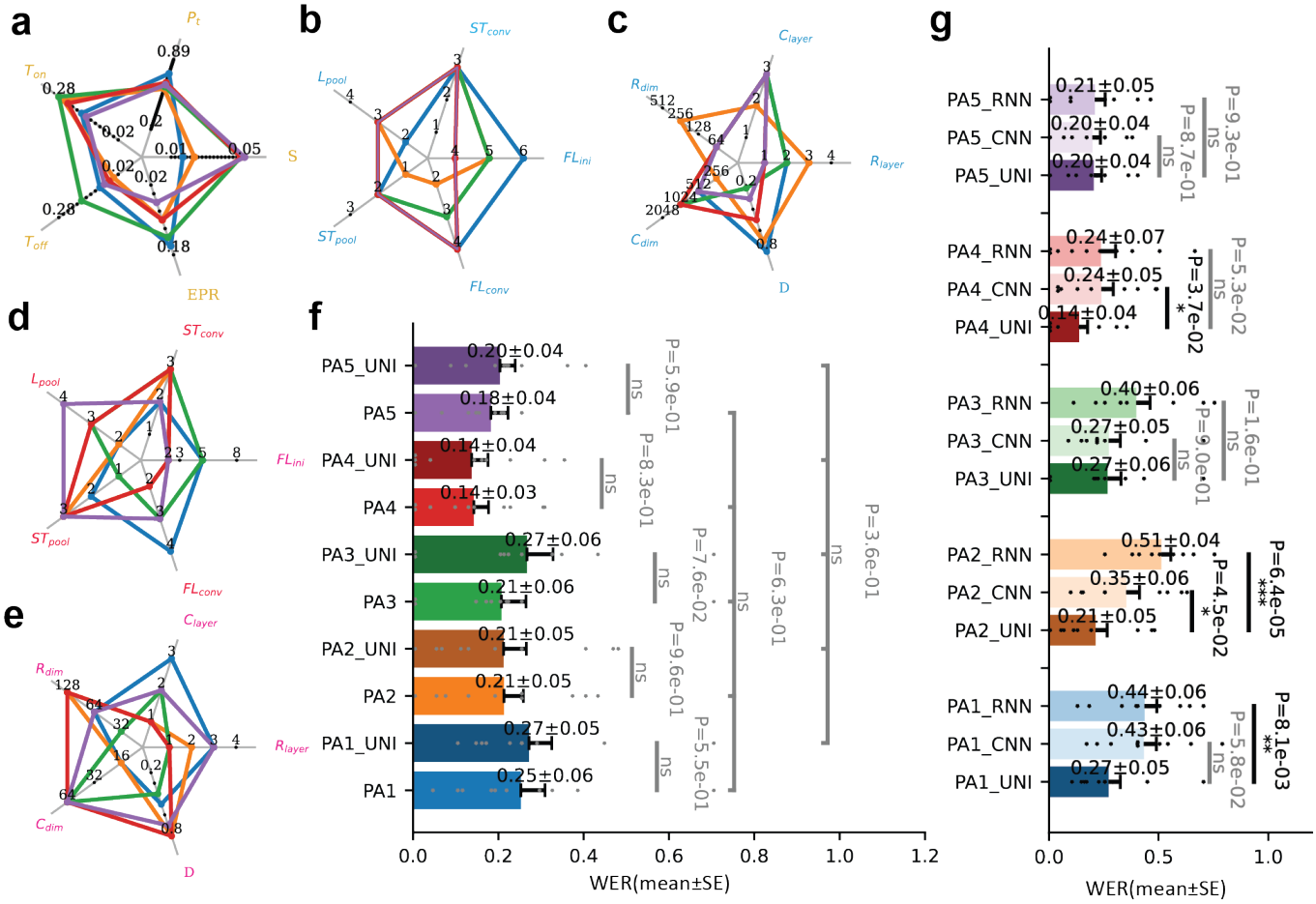


Fig 5. Evaluation of the neural decoder model hyperparameters across different participants.

a) The optimal combinations of hyperparameters for speech detector in 5 different subjects (colors consistent in **f** and **g**): ***S*** represents the smoothing size, $\boldsymbol{P_t}$ represents probability threshold, $\boldsymbol{T_{off}}$ represents off-time threshold and $\boldsymbol{T_{on}}$ for on-time threshold, ***EPR*** represents error permissive rate (detailed description of these hyperparameters see **Methods Section**.

b-c) The optimal combinations of hyperparameters for the tone decoder: **b)** $\boldsymbol{FL_{ini}}$ and $\boldsymbol{FL_{conv}}$ represents the filter length of the initial convolutional layer and all following convolutional layers, $\boldsymbol{ST_{conv}}$ represents for stride of all convolutional layers. $\boldsymbol{L_{pool}}$ and $\boldsymbol{ST_{pool}}$ represent for pooling length and stride of all pooling layers; **c)** $\boldsymbol{R_{layer}}$ and $\boldsymbol{R_{dim}}$ represent numbers of layers and hidden units of bidirectional Gated Recurrent Unit (GRU). $\boldsymbol{C_{layer}}$ and $\boldsymbol{C_{dim}}$ represent

number of convolutional-pooling blocks and number of filters in each convolutional layer. ***D*** represents dropout value of each drop-out layers.

d-e) The optimal combinations of hyperparameters for syllable decoder, similar to **b, c**.

f) Comparison of decoding performances on each individual participant between individually-optimized decoders and the universal decoder with shared hyperparameters. In all the participants, universal decoder showed no significant difference from individually-optimized decoder (pair-wise t-test,PA1: $t(9) = -0.62$, $p = 5.5\times10\text{-}1$, PA2: $t(9) = 0.05$, $p = 9.6\times10\text{-}1$, PA3: $t(9) = -2.00$, $p = 7.6\times10\text{-}2$, PA4: $t(9) = 0.23$, $p = 8.3\times10\text{-}1$, PA5: $t(9) = -0.55$, $p = 5.9\times10\text{-}1$).

g) Comparison of decoding performances on each individual participant between universal frameworks (UNI) and customized control frameworks with hyperparameters optimized. In all the participants, UNI show lower WER than control models. In PA1, PA2 and PA4, these is significance difference (pair-wise t-test, PA1_UNI with PA1_RNN: $t(9) = -3.38$, $p = 8.1\times10\text{-}3$, PA2_UNI with PA2_RNN: $t(9) = -7.00$, $p = 6.4\times10\text{-}5$, PA2_UNI with PA2_CNN: $t(9) = -2.33$, $p = 4.5\times10\text{-}2$, PA4_UNI with PA4_CNN: $t(9) = -2.44$, $p = 3.7\times10\text{-}2$)between decoding performances of UNI and control models.

Discussion

In this study, we present a brain-to-text framework capable of decoding natural tonal sentences from high-density ECoG recordings. We adopted a modular approach to delineate speech onset, base syllables, lexical tones, and leveraged contextual information through Bayesian likelihood and the Viterbi algorithm to enhance the decoding process. For natural speech, our proposed method achieved a tone decoding accuracy of 93%, similar to the behavioral performance of native speakers. The overall word error rate of decoded natural speech was as low as 14% in the best participant. Notably, we proposed a generalized spatiotemporal decoding framework for syllable and tone decoders. The robustness of our framework was evident across diverse participant profiles, including variations in gender, age, education, pronunciation behaviors, and electrode coverage, indicating that our model hyperparameters possess a high degree of generalizability. Significantly, the framework's standardization of hyperparameters negates the need for extensive individual optimization, a step forward in practical application and scalability. Furthermore, our system adeptly managed the inherent language heterogeneity encountered in Mandarin, effectively

393 handling the variances introduced by tone sandhi, regional dialects and individual speech patterns.
394 Our decoder even showed robust performance in whispering conditions. These results underscore
395 the potential for clinical applications in aiding patients with anarthria and broadening the
396 communicative efficiency of BCIs.

397 This work extends the scope of ECoG-based speech brain-computer interface to natural speech of
398 tonal languages. The past decade has witnessed significant breakthroughs in speech decoding and
399 brain-computer interfaces using intracranial neural recordings. Previous works have also used
400 ECoG to record local field potential directly from the sensorimotor cortex responsible for speech
401 production. These works diverge in their decoding targets: some directly decode neural activity
402 into speech sound or acoustics like spectrogram[2,26,27]; others map neural activity into discrete
403 linguistic units such as words or phonemes[5,10,28]. Here we adopted a brain-to-text framework
404 similar to the latter strategy where neural activity is first decoded into discrete syllables and lexical
405 tones. Although directly decoding speech sound allows for continuous and infinite speech output,
406 the quality of reconstructed speech is limited by the noisy neural signal. On the other hand,
407 decoding into a finite set of syllables and tones extracts the invariant information from the noisy
408 neural recordings. Furthermore, for cases like whispering or even completely covert speech, there
409 is no ground truth of explicit speech output. As a result, brain-to-text may be feasible to such silent
410 speech cases.

411 Our work underscores the importance of ventral sensory-motor cortex (vSMC) in speech
412 production and decoding, particularly for tonal languages. Similar to our previous work[15], we show
413 that tonal speech production can be reliably decoded from neural activity in vSMC. Previous
414 intracranial neurophysiology studies have investigated the spatiotemporal coding of the
415 articulatory movements responsible of pitch control and phonetics[20-22]. A theoretical foundation of

these studies is that there exist spatially distinct and distributed neural populations in the vSMC, representing different articulatory gestures corresponding to phoneme and pitch articulation. High-density ECoG recordings have proven to reliably cover the distributed network and dissociate these fine-grained neural coding[15,29].

Our results demonstrate the excellent efficacy of our proposed light-weighted models in decoding articulatory movements, which is in line with a recent discovery that shallow feedforward networks achieve better performance on motor control than deeper and complex ones[30]. Articulatory movement is innervated via motor nerves which locates only one or two (considering inhibitory interneurons) synapses downstreaming the large pyramidal neurons (corticonuclear tract) originated from the motor cortex[31]. Based on intracranial neural recordings of higher signal-noise ratio (SNR), it is quite reasonable that light-weighted ANN is capable enough to replace the signal-processing function of such a few layers of synapses (including the lateral connections within the layers of cortex) [32,33]. In our UNI frameworks, the numbers of trainable parameters in speech detectors, tone decoders and syllable decoders range from 2.3M to 6.7M. Such light-weighted frameworks not only showed less sensitivity to variations in hyperparameters, but also achieved better performance when trained on very limited amount of training data[34]. Such frameworks also reduce the responding-time and energy-consumption of computational infrastructures, which is a promising candidate for practical neuroprosthetic systems.

Our framework also provides insights into critical design considerations essential for speech BCI models. Previous studies of speech BCI typically relied on extensive hyperparameter optimization for individual participants who undergo chronic implantation[5,9] . Such decoding models may not be directly generalizable across different patients, resulting in repeated hyperparameters optimizing procedures for each individual. In contrast, our study reveals the feasibility of a

universal, hyperparameter-optimization-free framework to five individuals, demonstrating its robustness across a spectrum of ages, genders, educational backgrounds, pronunciation habits, and variations in brain electrode coverage. Furthermore, as oppose to prior work where specific CNN and Recurrent Neural Network (RNN) models were designed for tone and syllable decoding respectively[15], we proposed a unified CNN-RNN framework for both the tone and the syllable decoders in this study. Our discovery further validates that this combined CNN-RNN model achieved better decoding performance compared with baseline frameworks that employ only one of these network types. The benefits of our combined approach cannot be replicated through hyperparameter optimization alone. Our research highlights the potential for developing a broadly applicable, hyperparameter-independent framework for neural decoding. Although our universal framework has yielded stable performance across five distinct participants, future works remain to be done to consolidate its generalizability in patients with anarthria, across different Chinese dialects, and potentially other tonal languages.

This brain-to-text framework represents a pioneering effort in language BCI, designed to decode the full spectrum of Mandarin characters. Mandarin's linguistic complexity is reflected in its use of over 6,000 commonly utilized characters, each a single syllable word. Yet, within this vast lexicon, there are only 416 unique segmental combinations of consonants and vowels. These unique combinations, combined with suprasegmental pitch features (4 different lexical tones), define 1664 unique tonal syllables[35]. Although there are some important phonetic features to the distinction standard 4 lexical tones such as turning point and ΔF0, amplitude, and speaker F0 range[36]. Other factors that determine the pitch contour in Chinese include the prosodic structure of the language, the interaction between syntax and phonology, and paralinguistic factors such as speech rate and tempo[18]. These factors can influence the pitch contour and contribute to the overall

tonal patterns in Chinese, confusing native listeners listening to audio of each syllable clipped from natural sentences. Such bias obviously added challenges to speech BCI decoding tone exclusively from clips of neural signals while producing each syllable in natural sentences. The present decoding strategy inherently integrates the critical tonal features of Mandarin, effectively handling the variances introduced by tone sandhi, regional dialects and individual speech patterns, which are paramount for accurate speech communication, offering advantages over phoneme-level decoding. Besides, the current findings also highlight a pivotal insight: the decoding performance is not hindered by the length or complexity of the sentences, indicating the scalability of our framework for broader applications and more general settings in tonal language communication. Also, Mandarin characters are single syllable words usually consisted of only two or three phonemes, which is far less discriminative than most English. For example, "ji" versus "qi", or "zhi" versus "shi", as demonstrated in our study[2]. This structural intricacy of Mandarin, with its concise phonemic diversity, renders the phoneme-level decoding strategies used in recent speech BCI studies for non-tonal languages ineffective[7,8]. In response, our framework targets the decoding of Mandarin speech at the mono-syllable word level, aligning more closely with the language's inherent structure and providing a practical blueprint for decoding its entirety—potentially expanding our current focus from the most frequently used 4×10 tonal syllables to all 4×416 tonal syllables.

In addition to the established use of high-density ECoG in speech decoding, various neural recording techniques offer distinct advantages in terms of coverage and temporal resolution. Examples include the Utah array[9,37], stereoelectroencephalography (SEEG) [38], and neuropixels[39,40]. When choosing among these methods, a crucial consideration involves striking a balance between obtaining high-resolution neural signals, such as investigating fine-grained spiking properties of

multiple single units or microcircuits underlying speech production within a limited area of the cortex[9,40], and achieving broad coverage of cortical networks, such as collecting neural signals across the entire vSMC, which depicts a comprehensive view of neurodynamics of the functional regions[21,22,29]. Future works remain to be done to investigate the decoding capabilities of natural tonal languages using signals of varying coverage and resolution scales. This entails identifying the optimal trade-off point where high decoding performance aligns with decoding robustness across patients with high heterogeneity.

Acknowledgements

Dr. Junfeng Lu is supported by STl 2030-Major Projects (2022ZD0212300) and the National Natural Science Foundation of China General Program (32371146). Dr. Yuanning Li is supported by the National Natural Science Foundation of China General Program (32371154) and Shanghai Pujiang Program (22PJ1410500). Dr. Jinsong Wu receives funding from Innovation Program of Shanghai Municipal Education Commission (2023ZKZD13) and National Social Science Fund of China Major Program (No. 22&ZD299).

Competing interests

The authors report no competing interests.

Data availability

Data relevant to this study are accessible from the authors under restricted access according to our clinical trial protocol, which enables us to share de-identified information with researchers from other institutions but prohibits us from making it publicly available. Access can be granted upon reasonable request. Any data provided must be kept confidential and cannot be shared with others unless approval is obtained. To protect the participants' anonymity, any information that could identify him or her will not be part of the shared data. Source data and code to recreate the figures in the manuscript will be publicly released with code upon publication of the manuscript.

Code availability

Code and source data to replicate the main findings of this study can be found on GitHub at https://github.com/yuanningli/tonal_BCI_decoding.

Author contributions

J.W., J.L. and Y.Li. conceived and supervised the project. J.L., Y.Li., D.Z., Y.Liu., Z.Z., and S.L. designed the experiment. J.L, D.Z., Y.Q., and Z.Z. collected the data. D.Z., H.Z., X.H. finished the phonetic and phonological transcription. Y.Li., D.Z. designed the neural network. Y.Li., Z.W.,

D.Z., and W.L. designed the language model. D.Z. and Z.W. analyzed the data. J.L., Y.Li., L.C., K.X., and D.Z., interpreted the data. D.Z., Z.W., Y.Li., J.L. and J.W. wrote and revised the manuscript. All authors reviewed and approved the manuscript.

Methods

Participants

Five participants (a 41-year-old female, a 44-year-old female, a 54-year-old male, a 46-year-old male and a 30-year-old female) were all patients with brain lesions who underwent awake surgery and temporarily placed high-density, subdural electrode arrays covering the lateral surface of the brain to collect the neural signal in speech task. All participants are native speakers of Mandarin and perform daily communication in Mandarin without any difficulty. All protocols approved by the Institutional Review Board of Huashan Hospital (KY2019-538), and the conducting of research complied with all relevant ethical regulations. Participants were asked to participate in this study only if there was clinical necessity to perform awake surgery for safe resection of occupying lesions and protection of the eloquent area.

Design of the sentence corpus

To have a representative set of phonological features in Mandarin and maximize the representativeness of our speech task, we selected the top 10 most frequently used open syllables with monophthong, which cover the pronunciation of nearly 25.9% of all Chinese characters[41]. Using these 10 syllables, we obtained 40 distinct Chinese characters with 4 lexical tones and constructed 29 Chinese words and phrases from these 40 characters. Finally, 10 sentences were constructed with these 29 phrases, which eventually consisted the sentence corpus used in our decoding task.

The 10 syllables chosen for this work is:

1. 'shi', /ʂɚ/

2. 'de', /tɤ/
3. 'ji', /tɕi/
4. 'li', /li/
5. 'bu', /pu/
6. 'ge', /kɤ/
7. 'qi', /tɕ^{h}i/
8. 'zhe', /ʈʂɤ/
9. 'ta', /t^{h}a/
10. 'zhi' /ʈʂɚ/

The 40 Chinese characters used in this work and their corresponding tonal syllables (five-level tone marks):

1. 失: shi55
2. 十: shi35
3. 时: shi35
4. 实: shi35
5. 是: shi51
6. 的: de
7. 地: de
8. 得: de35
9. 机: ji55

10. 激: ji55

11. 极: ji35

12. 记: ji51

13. 计: ji51

14. 迹: ji51

15. 离: li35

16. 里: li214

17. 理: li214

18. 励: li51

19. 荔: li51

20. 不: bu35/bu51

21. 布: bu51

22. 哥: ge55

23. 搁: ge55

24. 个: ge51

25. 七: qi55

26. 奇: qi35

27. 其: qi35

28. 起: qi214

29. 契: qi51

30. 折: zhe35

31. 者: zhe214

32. 这: zhe51

33. 他: ta55

34. 塔: ta214

35. 织: zhi55

36. 枝: zhi55

37. 职: zhi35

38. 止: zhi214

39. 置: zhi51

40. 智: zhi51

*to simplify the decoding process, we use T1 (-) to denote 55, T2 (/) to denote 35, T3 (\/) to denote 214, and T4(\) to denote 51. The neutral tone syllables “的:de” and “地:de” were marked as T1 or T4 according to the actual pronunciation of each individual patient.

The 29-Chinese phrases used in this work (with translation in English):

1. 智者 sage
2. 记者 journalist
3. 荔枝 lichee
4. 七折起 thirty-precent off

5. 他 him/he

6. 他的 his

7. 他哥 his brother

8. 哥哥 elder brother

9. 不时地 often

10. 激励 encourage

11. 极其 extremely

12. 机智 smart

13. 织布机 loom

14. 这里的 here

15. 是 is

16. 的 DE (structure auxiliary)

17. 塔 pagoda

18. 不理 ignore

19. 是个 is a/an

20. 奇迹 miracle

21. 不计得失 do not care about gains and losses (Chinese Idiom)

22. 不得不 have to

23. 离职 resign

24. 搁置 shelve

25. 这个 this

26. 契机 opportunity

27. 其实 actually

28. 不止 more than

29. 七十 seventy

The 10 sentences used in this work:

1. 他哥不理他。

Tā	gē	bù	lǐ	tā
his	elder.brother	not	pay.attention.to	him

‘His elder brother ignores him.’

2. 他的荔枝七折起。

Tā	de	lìzhī	qīzhé	qǐ
his	DE	lichee	30.percent.off	at.least

‘His lichee is at least 30-percent off.’

3. 记者不得不离职。

Jìzhě	bùdébù	lízhí
to.record.person	cannot.not.to	resign

‘The journalist has to resign.’

4. 他其实不止七十。

Tā qíshí bùzhǐ qīshí

he virtually not.merely seventy

‘He is virtually more than seventy(-year-old).’

5. 哥哥不时地激励他。

Gēge bùshíde jīlì tā

elder.brother sometimes.not.but.sometimes will encourage him

‘The elder brother often encourages him.’

6. 他极其机智。

Tā jíqí jīzhì

he extremely smart

‘He is extremely smart.’

7. 织布机是这里的。

Zhībùjī shì zhèlǐ de

loom is here DE

‘The loom belongs here.’

8. 这里的塔是个奇迹。

Zhèlǐ de tǎ shì gè qíjì

here DE pagoda is a miraculous.sign

‘The pagoda here is a miracle.’

9. 智者不计得失。

Zhìzhě bújìdéshī

wise.person do.not.care.about.gains.and.losses

'The sage does not care about gains and losses.'

10. 搁置这个契机。

Gēzhì zhègè qìjī

shelve this opportunity

'Shelve this opportunity

Task design

Experiments and data in this research abided by the relevant ethical regulations. Participants were guided by sequential visual cues to produce one of the ten sentences consisted of 5-8 Chinese characters from a corpus of 10 base syllables and 40 Chinese characters with varied tones **(Fig. 1a)**. In every single trial, the participant was instructed to produce all ten sentences once. The ten sentences were presented in a random order. Ideally, each participant produced 160 sentences (4 blocks × 4 trials/block × 10 sentences/trial), which yielded approximately 30 repetitions per tonal syllable. But not all participant completed the entire reading task (158-160 sentences were completed).

Each participant was guided by visual cues to perform block tasks. Each block started with a black cross at the center of white background on the screen, which lasted for 30 seconds. After that, the cross turned grey for 3 seconds, and one of the sentences in the sentence set was shown in the middle of the screen in grey text. With the individual Chinese characters turning black for 1.2 seconds in a sequential order from the beginning to the end of the sentence, the participant was instructed to pronounce the sentence in a relatively uniform speed following these go cues. The inter-sentence time interval and the inter-trial interval were both 3 seconds.

Audio recordings were obtained synchronously with the ECoG recordings using a mounted microphone. The first four participants (PA1-4) performed the speech task articulating, while the last participants (PA5) performed the speech task whispering. We collected two types of blocks of the sentence task: first 3 optimization blocks (Trial 1-12, containing 118-120 sentences in total for each patient) and one evaluating block (Trial 13-16, containing 38-40 sentences for each patient).

Data acquisition and signal processing

Electrocorticography and audio sound were recorded simultaneously using the Tucker-Davis Technologies ECoG system. For each participant, two 128-channel electrode arrays were placed by an experienced neurosurgeon, and the location was determined based on clinical exposure and avoidance of the tumor. The arrays were connected to an amplifier. ECoG signals were recorded at a sampling rate of 3052 Hz. The researcher visually and quantitatively inspected the channels for artifacts or excessive noise (typically 50 Hz line noise). High-gamma frequency component (70-150 Hz) was extracted via Hilbert transform and down-sampled to 400 Hz. Audio sound waves were recorded at a sampling rate of 24414 Hz.

Phonetic and phonological transcription

For the audio recording, transcriptions and labels, including monosylabic Chinese characters, consonants, vowels, and tones, were manually marked with praat (Version 6.1.01, https://www.fon.hum.uva.nl/praat/) at the syllable level to ensure the label reflected the actual sound the participant produced. Unexpected voicing unrelated to language task (such as communication with clinicians) was excluded from samples used in training models.

Computational modeling infrastructure

705 The training and testing of the decoding models were performed offline using cluster of multiple
706 NVIDIA GPUs.

707 **Data splitting**

708 We split the data for speech detector and the syllable/tone decoder model testing from electrode
709 selection and hyperparameter searching. For responsive and discriminant electrode selection and
710 hyperparameter optimization, only the optimization blocks were used, which contain Trial 1-12
711 with each trial consists of randomly aranged 10 sentences. The evaluation block contains
712 remaining Trial 13-16.

713 During the optimization stage of the speech detector model, we used a six-fold nested cross
714 validation (CV) and each fold consisted of 2 trials. At evaluation stage, we used a similar cross
715 validation process, which leaves each trial in evaluation block for testing model, while the other
716 15 trials for training and validating. For speech detector models, we used 10% of non-testing data
717 as validation set to perform early-stopping, while the left 90% for training, in each CV runs (**Fig.**
718 **S1A&B**). For syllable/tone decoder model, we trained 5 (optimization stage) or 10 (evaluation
719 stage) sub-models in each CV runs, with different potions of data performing early-stopping while
720 left for training. The final model used for evaluation was an ensemble of 5 or 10 sub models. And
721 the overall decoding performance was evaluated by averaging the test performance in all the CV
722 runs (**Fig. S1C&D**).

723 **Speech-responsive electrodes**

724 Speech-responsive electrodes were identified using two-sample t-test. Specifically, each time point
725 in the [-400 ms, 800 ms] time window relative to the consonant onsets in each word was tested
726 against the [-1800ms, -400ms] baseline time window before the onset of each sentence. If the

results were significant (P < 0.01, Bonferroni corrected for the total number of electrodes and all times points) in 40 consecutive time points (100 ms), the electrode would eventually be marked as speech-responsive.

- **Tone discriminant electrodes**

To identify electrodes that were discriminative between lexical tones, we again aligned the high-gamma responses to the onsets of single syllables. We then used the one-way ANOVA to test whether the mean high-gamma responses of the four mandarin tones differed significantly. The time window of the average response was [−500 ms, 500 ms] relative to the onset (400 total time points). The significant time points were determined with a two-sided $P < 0.05$ threshold using Bonferroni correction for the total number of electrodes and time points. To account for electrodes with multiple peaks of high-gamma activity discriminant for tone decoding, we set the following criterion: if there was a continous 200ms time window in which more than half of the time points were significant, the electrode would be marked as tone-discriminant.

- **Syllable discriminant electrodes**

Similar to tone discriminant electrodes, the syllable discriminant electrodes were defined using one-way ANOVA. We tested whether the mean high-gamma responses of the ten syllables were significantly different. The time window of the average response was [−400 ms, 800 ms] relative to the onset (480 total time points). The significant time points were determined with a two-sided $P < 0.05$ threshold using Bonferroni correction for the total number of electrodes and time points. Similarly, if there were a 200ms time window in which more than half of the time points were significant, the electrode would be marked as syllable-discriminant.

- **Cortical surface extraction and electrode visualization**

Electrodes on each individual brain were marked by the BrainLab® neuronavigation system via preoperative T1 MRI and validated via intro-operative photos by a neurosurgeon. The cerebral surface was reconstructed using Freesurfer while anatomical labeling and plotting were performed via customized python code[42].

Decoding framework

We built different neural network modules of speech detector, tone decoder and syllable decoder to decode speech onsets and offsets, syllable labels and tone labels from the neural activity, respectively. Inspired by previously published decoding models such as the EEGNet[43,44], our decoders had separable CNN layers extracting within- and cross-electrodes spatiotemporal features, followed by GRU layers extracting sequential information. Finally, we also used a natural language model and a Viterbi decoder to combine the sequential outputs of the tone and syllable decoders and generate the entire sentence using maximum a posteriori probability **(Fig. 1b)**.

Speech utterance detection model (speech detector)

The speech utterance detection model processed each time point of neural activity on speech responsive electrodes using the time window spanning from 0.25 second before the time point to 0.25 seconds after the timepoint. The speech/silent state of each time point was decoded based on the time-windowed neural activity.

We used the Torch 1.13.1+cu117 Python package to create and train the event classification model[45]. The event classification architecture was a CNN structure followed by a stack of three GRU layers with a latent dimension size of 256 and a dropout of 0.5 applied at each layer.

The initial CNN structure contains an 1D convolutional layer with a kernel size ($\boldsymbol{FL_{ini}}$) of 3 and a stride($\boldsymbol{ST_{conv}}$) of 1 followed by a Leaky ReLU activation function to introduce non-linearity,

and a max-pooling layer with kernel size ($\boldsymbol{L_{pool}}$) of 2 for temporal downsampling[46]. This initial processing using convolution and pooling allows the model to extract local features from the input data effectively, while following recurrent layers are capable of maintaining an internal state through time that can be updated with new individual time samples of input data, making them well suited for real-time inference with temporally dynamic processes. Also, the bidirectional nature of the GRU enables the model to capture both forward and backward temporal dependencies in the ECoG data, making it well-suited for event classification tasks that require understanding the sequential context. The GRUs are followed by a fully connected layer to project the last latent dimensions to probabilities across the two events (rest and speech). During training, the model is optimized to minimize weighted crossentrophy. The weighted crossentrophy loss was calculated using the ratio of the number of all the speech time points and the number of all the rest time points, to offset the bias caused by imbalanced samples. The batch size is 1024 and an Adam optimizer was used with a learning rate of 0.001. The training process stops after the validation loss no longer decreases for 10 epochs or after 50 epochs, ensuring that the model has sufficiently learned the underlying patterns in the data but has not yet overfitted. A schematic depiction of this architecture is given in **Fig. 2f**.

During testing, the neural network predicted probabilities for each speech event label (rest, speech) given the input neural data. To convert these predicted probabilities into the time of detected onsets, we need to apply thresholding to the predicted speech probabilities. First, we smoothed the probabilities using a sliding window average with a smoothing size ($\boldsymbol{S}$). Next, we applied a probability threshold ($\boldsymbol{P_t}$) to the smoothed probabilities to binarize each frame (with a value of 1 for speech and 0 otherwise). Afterwards, we "debounced" these binarized values by applying time thresholds for both a minimum continuous duration of silent period before the predicted onset, and

794 a minimum duration of continuous speech period after the predicted onset (off-time threshold (T_{off})
795 and on-time threshold (T_{on})). Detailed examples of thresholding process see **Fig. 3b**.

796 But the required continuous state is not strict, we set an error permissive rate (***EPR***) to allow a
797 small portion of timepoints with uncorrectly predicted speech event labels presented during both
798 continuous silent period before the predicted onset as well as the speech period after the predicted
799 onset. Besides, we will not predict two or more onsets within 0.5s, since normal speaking rate of
800 our task do not allow such short interval and this update in algorithm speed up the iteration process.

801 In all, this process of obtaining onset events from the predicted probabilities was parameterized by
802 five thresholding hyperparameters: the size of the smoothing window (***S***), the probability threshold
803 value (P_t), the off-time threshold (T_{off}) and on-time threshold (T_{on}), and the error permissive rate
804 (***EPR***). We used hyperparameter optimization to determine values for these parameters, which will
805 be described in later sections.

806 **Tone decoder**

807 For each speech utterance onset that was detected or manually aligned, tone decoder computed
808 tone likelihood by processing the neural activity in tone- discriminant electrodes spanning from
809 0.2 second before to 0.6 seconds after the detected onset of speech.

810 The tone decoder consists of ten ensumble neural networks with the same hyper-parameters but
811 different validation sets and training sets (see **Fig. S1D**). The input was high gamma ECoG
812 snippets with dimension $N \times T$, N is the number of tone-discriminated electrodes, while $T = 320$,
813 which is calculated by the time duration (0.8seconds) multiplies downsampled EcoG frequency
814 (400 Hz). Each network among 10 networks consists of the following stages:

815 - **Stage 1**

In the first stage, one-layer 2D convolution with the kernel size ($\boldsymbol{FL_{ini}}$, $\boldsymbol{N}$) and a stride of ($\boldsymbol{ST_{conv}}$, **0**), followed by a batch normalization layer (momentum=0.1, affine=True, and eps=1e-5) applied to the output of the spatial convolution. After an Exponential Linear Unit (ELU) activation function. Max-pooling is then performed along the temporal axis with a kernel size of ($\boldsymbol{L_{pool}}$, **1**) and a stride of ($\boldsymbol{ST_{pool}}$, **1**). During training, we applied a dropout layer a dropout value of $\boldsymbol{D}$ in the end of each CNN block for regularization to prevent overfitting.

- **Stage 2**

In the second stage, $\boldsymbol{C_{layer}}$ convolutional pool blocks are applied after the aforementioned pooling layer in series, each consisting of a dropout layer, a convolutional layer with kernel size ($\boldsymbol{FL_{conv}}$, **1**) and a stride of ($\boldsymbol{ST_{conv}}$, **1**), batch normalization, ELU activation, and max-pooling identical with aforementioned ones. These blocks further process and extract features from the data. During the first two stages, the filter number was set equal to $\boldsymbol{C_{dim}}$.

- **Stage 3**

In the last stage, the output from the last convolutional pool block is fed into a stack of $\boldsymbol{R_{layer}}$ bidirectional Gated Recurrent Unit (GRU) layers. Each GRU layer has $\boldsymbol{R_{dim}}$ hidden units. The final hidden state of the last GRU layer is passed through a fully connected (dense) layer which projest to 10 output units, representing the number of word classes for classification. **(Fig. 2g)**.

The model is optimized to minimize crossentrophy loss using a batch size of 8 and an Adam optimizer with a learning rate of 0.0005. The training process stops after the validation loss no longer decreases for 50 epochs. The weighted crossentrophy loss was calculated in the similar same way as speech detector to offset the bias caused by imbalanced samples.

Syllable decoder

838 For each speech onset that was detected, the syllable decoder computed syllable likelihood by
839 processing the neural activity in syllable- discriminant electrodes spanning from 0.4 second before
840 to 0.8 seconds after the detected onset of speech utterance. The structure and training process of
841 syllable decoder is identical with the aforemental tone decoder, the only difference between
842 them was the crossentrophy loss was not weighted, since the syllable is almost balanced-
843 distributed in the task.

844 **Hyperparameter optimization and the universal framework**

845 To find optimal values for the model hyperparameters used during performance evaluation, we
846 used hyperparameter optimization procedures to evaluate many possible combinations of
847 hyperparameter values, which were sampled from custom search spaces, with objective functions
848 that we designed to measure model performance. During each hyperparameter optimization
849 procedure, a desired number of combinations were tested, and the combination associated with
850 the best objective function value across all combinations was chosen as the optimal
851 hyperparameter value combination for that model and evaluation type. The data used to measure
852 the associated objective function values were distinct from the data that the optimal
853 hyperparameter values would be used to evaluate (hyperparameter values used during evaluation
854 of a test set were never chosen by optimizing on data in that test set). We used the *hyperopt*
855 Python package, which samples hyperparameter value combinations probabilistically during the
856 optimization procedure. And we used three types of hyperparameter optimization procedures to
857 optimize a total of 25 hyperparameters (see **Fig. 5** for undefined hyperparameters, their searching
858 ranges and their optimal values in each patient). During the optimization process, all the neural
859 networks were tested through the six-fold (2 trials per fold) cross-validation.

860 - **Speech detection optimization**

We used this procedure to optimize the size of the smoothing window (***S***), the probability threshold value (***P_t***), the off-time threshold (***T_off***) and on-time threshold (***T_on***), and the error permissive rate (***EPR***). Because these thresholding hyperparameters were only applied after speech probabilities were predicted, these hyperparameters did not affect training or evaluation of the artificial neural network model driving the speech detector. In each iteration of the optimization procedure, the current hyperparameter value combination was used to generate detected speech events from the existing speech probabilities. We used an objective function to calculate the performance. For time point of real onsets, we use $\hat{X}$ to represent the index of time points within 0.25s range of real onsets ($t_i^*, i \in [1, \dots, n]$). For time point of predicted onsets, we use $\tilde{X}$ to represent the index of time points within 0.25s range of real onsets. Then, we calculated the accuracy (Acc) of $\hat{X}$ and $\tilde{X}$. The formal expression of this objective function was illustrated through following equations:

$$\hat{X}[t] = \begin{cases} 0, & \text{others} \\ 1, & \text{if } \min\limits_{i \in [1,\dots,n]} |t - t_i^*| < 0.25 \end{cases}$$

$$\tilde{X}[t] = \begin{cases} 0, & \text{others} \\ 1, & \text{if } \min\limits_{i \in [1,\dots,n]} |t - t_i^*| < 0.25 \end{cases}$$

$$\mathbb{1}(\mathrm{x}) = \begin{cases} 0, & \mathrm{x} \neq 0 \\ 1, & \mathrm{x} = 0 \end{cases}$$

$$\mathrm{Acc} = \frac{\sum_t \mathbb{1}\left[\hat{X}[t] - \tilde{X}[t]\right]}{\sum t}$$

We use ***Acc*** to measure the model performance with each hyperparameter value combination. In each detection hyperparameter optimization procedure, we evaluated 500 hyperparameter value combinations before stopping.

Once the most optimized combination of hyperparameters was defined, we calculated the predicted onset time list for all the 12 trials via six-fold prediction, we compare them with the

ground truth of onset of the 12 trials. If one onset time point in predicted list falls within 0.25s range of real onsets, the labels (including tone, syllable and Chinese characteristic) belongs to the mannually aligned onset will be given to the predicted time point, which participated in the following hyperparameter optimization. For those time points can not match with any real onset time points, no labels will be given and this part of the data will not participate in the following hyperparameters. The similar match process will be performed in the evaluating stage, calculating confusion matrix for syllable, tone and word on the bases of onset detection. Because some real onset time points do not match with any detected onset time point, the number of syllables and tones that actually used in calculating confusion matrices of decoding performance after automated onset detection is less than the number of onsets detected.

- **Tone and Syllable decoder optimization**

We used this procedure to optimize ten hyperparameters for both tone decoder and syllable decoder. These hyperparameters can be divided into two major groups according to their properties.

The first group includes initial convolution filter length ($\boldsymbol{FL_{ini}}$), stride of convolution ($\boldsymbol{ST_{conv}}$), filter length of following convolutional blocks ($\boldsymbol{FL_{conv}}$), max-pooling kernel length ($\boldsymbol{L_{pool}}$), max-pooling stride ($\boldsymbol{ST_{pool}}$), which are five hyperparameters related to the time dimension.

The second group includes the number of sequential convolutional blocks ($\boldsymbol{C_{layer}}$), number of layers of RNN ($\boldsymbol{R_{layer}}$), number of filters in each convolutional process ($\boldsymbol{C_{dim}}$), number of dimensions in each RNN process ($\boldsymbol{R_{dim}}$), and dropout value ($\boldsymbol{D}$), which are five hyperparameters related to model architectures and sizes.

We performed model testing on mannually-aligned-onset dataset and the dataset generated from predicted onsets by the previously optimized speech detector. And the final loss is the average

value of loss on both datasets, in order to improve the model's robustness when countering automatedly aligned neural signals in our final evaluation. Each optimization evaluated 500 different combinations of hyperparameter values. We used the same cross-entrophy loss functions to calculate the loss during the iteration.

- **The universal framework**

The universal framework shares a predefined set of hyperparameters when testing on all 5 participants. We used the median value of each optimized hyperparameters among 5 participants as the predefined value of hyperparameters.

Conversion from tonal syllables to Chinese words

However, many Chinese characters are homophones that share the same pronunciations (including syllable and tone), which prevented us from directly generating Chinese sentences using decoded tonal syllables. To eventually decoding sentences consists of actual Chinese characters rather than sequences of tonal syllables, we designed a natural language model that computed the next-character transition probabilities given the previous Chinese character in a sequence. We first divided the sentences into aforementional 29 words and phrases consisting of 1 to 4 Chinese characters. We trained this model on a collection of sentences from the CCL corpus[47] that included transfeing pairs (transfers) between those words from the 29-word set. After that, we used a Viterbi decoder to determine the most likely sequence of words given the predicted tonal syllable probabilities from the tone decoder and syllable decoder, as well as the word-sequence probabilities from the natural language model[8]. With the incorporation of the language model, the Viterbi decoder was capable of projecting sets of probability (likelihoods) of tonal syllable to exact Chinese characters, which eventually decoded sentences from neural activity.

Language model

- **Collection of the corpus**

To train a domain-specific language model for the sentence task, we used CCL corpus from Peking University[47]. We first measure the number of transfers between each two phrases (the counts of the previous phrase transferred to the first word of the next phrase in the whole CCL corpus). Due to the unequity of Chinese languge, we set a cut-off of 512 counts, and then performed a ninth root for normalization. Since not all the transfers in our task appeared in CCL corpus, we also incorperate the transfer probability of our own task corpus into our language model.

- **Model fitting and word-sequence probabilities**

We extracted all n-grams with n ∈ {1, 2, 3, 4} from each phrase in our task corpus. Here, an n-gram is a word sequence with a length of n Chinese characters. For example, the n-grams (represented as tuples) extracted from the longest phrases "不计得失" in this approach would be:

1. (不)
2. (不计)
3. (不计得)
4. (不计得失)

Then we expand our transfer probability matrix into n-grams. We add the number of transfering from q_{i-1} (such as (不计)) to q_i (such as (不计得)). Except for full phrases, n-grams as partial components of phrases can only transfer from q_{i-1} to q_i . The calculation of number of each transfering from a full phrase to the beginning of next phrases is aformentioned.

- **Hidden Markov model representation of the sentence decoding procedure**

During a sentence trial, the relationship between the sequence of words that the participant produced and the sequence of neural activity time windows provided by the speech detector can be represented as a hidden Markov model (HMM). In this HMM, each observed state y_i is the time window of neural activity at index i within the sequence of detected time windows for any particular trial, and each hidden state q_i is the n-gram containing the words that the participant had attempted to produce from the first word to the word at index i in the sequence. Here, $q_i = \{w_i, \boldsymbol{c}_i\}$, where w_i is the word at index i in the sequence and $\boldsymbol{c}_i$ is the context of that word. The emission probabilities for this HMM are $p(y_i|q_i)$, which specify the likelihood of observing the neural time window y_i given the n-gram q_i. With the assumption that the time window of neural activity associated with the production of w_i is conditionally independent of all of the other attempted word productions given w_i $\left(y_i \perp w_j \mid w_i \; \forall \, j \neq i\right)$, $p(y_i|q_i)$ simplifies to $p(y_i|w_i)$. The word classifier provided the probabilities $p(w_i|y_i)$, which was used directly as the values for $p(y_i|w_i)$ by applying Bayes' theorem and assuming a flat prior probability distribution. The transition probabilities for this HMM are $p(q_i|q_{i-1})$, which specify the probability that q_i is the n-gram at index i (the sequence of at most n words, containing w_i as the final word, that the participant attempted to produce) given that the n-gram at index $(i-1)$ was q_{i-1}. Here, q_{-1} can be defined as an empty set, indicating that q_0 is the first word in the sequence. Because any elements in $\boldsymbol{c}_i$ will be contained in q_{i-1} and w_i is the only word in q_i that is not contained in q_{i-1}, q_i simplifies to $p(w_i|\boldsymbol{c}_i)$, which were the word sequence prior probabilities provided by the language model. Implicit in this simplification is the assertion that $p(q_i|q_{i-1}) = 0$ if q_i is incompatible with q_{i-1} (for example, if the final word in ci is not equal to the second-to-last word in q_{i-1}).

- **Viterbi decoding implementation**

To predict the words that the participant attempted to produce during the sentence task, we implemented a Viterbi decoding algorithm with this underlying HMM structure. The Viterbi decoding algorithm uses dynamic programming to compute the most likely sequence of hidden states given hidden-state prior transition probabilities and observed-state emission likelihoods[5]. To determine the most likely hidden-state sequence, this algorithm iteratively computes the probabilities of various "paths" through the hidden-state sequence space (various combinations of q_i values). Here, each of these Viterbi paths was parameterized by a particular path through the hidden states (a particular word sequence) and the probability associated with that path given the neural activity. The creation of new Viterbi paths from existing paths can be expressed using the following recursive formula:

$$V_i = \{v_{i-1} + \log p\left(y_i \mid q_{i,v_{i-1}} \right) + L \log p\left(q_{i,v_i-1} \mid q_{i-1,v_{i-1}} \right) | v_{i-1} \in V_{i-1} \wedge w_i \in W\}$$

with the following variable definitions:

• V_j : The set of all Viterbi paths created after the word production attempt at index j within a sentence trial.

• v_j : A Viterbi path within V_j . Each of these Viterbi paths was parameterized by the n-grams ($q_0, \ldots, q_j$) (or, equivalently, the words ($w_0, \ldots, w_j$)) and the log probability of that sequence of words occurring given the neural activity, although these equations only describe the recursive computation of the log probability values (the tracking of the words associated with each Viterbi path is implicitly assumed).

• q_{j,v_k} : The n-gram q_j , containing the word w_j and the context of that word. This context is determined from the most recent words within the hidden state sequence of Viterbi path v_k.

• $p\left(y_i \mid q_{i,v_{i-1}} \right)$: The emission probability specifying the likelihood of the observed neural

activity y_i given the n-gram $q_{i,v_{i-1}}$.

• $p\left(q_{i,v_i-1} \mid q_{i-1,v_{i-1}} \right)$: The transition probability specifying the prior probability of transitioning to the n-gram q_{i,v_i-1} from the n-gram $q_{i-1,v_{i-1}}$.

• L : The language model scaling factor, which is a hyperparameter that we used to control the weight of the transition probabilities from the language model relative to the emission probabilities from the word classifier. In this study, we fixed it as 1.8.

• W : The 29 words set plus the end token.

• log: The natural logarithm.

Using the simplifications described in the previous section, the aforementional recursive formula can be simplified to the following equation:

$$V_i = \{v_{i-1} + \log p(w_i \mid y_i) + L \log p(w_i \mid \boldsymbol{c}_{i-1}) | v_{i-1} \in V_{i-1} \wedge w_i \in W\}$$

where $\boldsymbol{c}_{j,v_k}$, is the context of word w_j determined from the Viterbi path v_k, $p(w_i \mid y_i)$ are the emission probabilities (obtained from the word classifier), and $p(w_i \mid \boldsymbol{c}_{i-1})$are the transition probabilities (obtained from the language model). At the start of each sentence trial, the index i was reset to zero (the first word in each trial was denoted w_0), and any existing Viterbi paths from a previous trial were discarded. To initialize the recursion, we defined V_{-1} as a singleton set containing a single Vitrbi path with the empty set as its hidden statezero. We used log probabilities in practice for sequence and an associated log probability in practice for numerical stability and computational efficiency.

Evaluation

- **Evaluation of independent decoding performances via Receiver Operating Characteristic (ROC) and Confusion Matrices**

To evaluate the performance of our speech detector, we computed the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve for each participant. To evaluate the tone decoder and syllable decoder, we computed the classification accuracy and plotted the confusion matrix, based on both the predicted and the manually aligned speech onsets. We also calculated the classification accuracy and plotted confusion matrices of tonal-syllables and Chinese characters, on both the predicted and the manually aligned speech onsets, in order to further assess the capacity of the entire decoding system.

- **Evaluation of overall decoding performance via Word Error Rates (WERs)**

To evaluate the overall performance of our neural-to-text decoder, we analyzed the decoded sentences using Word Error Rates (WERs) between the target and decoded sentences for each sentence. WER is a commonly used metric

to measure the quality of predicted word sequences, computed by calculating the edit (Levenshtein) distance between a reference (target) and decoded sentence and then dividing the edit distance by the number of words in the reference sentence. Here, the edit distance measurement can be interpreted as the number of word errors in the decoded sentence. It is computed as the minimum number of insertions, deletions, and substitutions required to transform the decoded sentence into the reference sentence. Below we demonstrate each type of edit operation that can be used to transform an example decoded sentence (on the left side of each arrow) into the target sentences. In each case, the example decoded sentence has an edit distance of 1 to the target sentence.

• Insertion: 他哥不理他 → 他哥哥不理他

• Deletion: 他哥不理他 → 他哥理他

• Substitution: 他哥不理他 → 哥哥不理他

Lower edit distances and WERs indicated better performance. We computed WERs of each sentence in the evaluating block using predictions made with the language model and Viterbi decoder.

- **Comparison with published tone decoder**

To further validate the complexity of tone of sentences compared with tone of single syllable, we applied our previously published monosyllabic tone decoder on tone decoding in sentence task. It is a sequential CNN-LSTM structure to generate the syllable label based on manually aligned neural activity when participants produce eight syllables "ma (tone1), ma (tone2), ma (tone3), ma (tone4), mi (tone1), mi (tone2), mi (tone3), mi (tone4)" [15]. **(Fig. S2)**

- **Tone intelligibility assessment (IA)**

Listening tests using crowdsourcing are standard for evaluating the quality in natural language processing. To further validate the variance of lexical tones in natural sentences compared with the canonical tones of single syllables, we extracted the audio clips of each syllable utterance in the evaluation blocks of each participant, using the same timespan as the neural decoder. We shuffled the sequence clips by syllables to avoid any possible of leakage of sementic information. The presented options included all tones. For sentence tone IA, we extracted the audio clip of each sentence each patient pronounced in evaluation blocks. The evaluators were instructed to listen to the audio of the entire and then choose from the options which tone they had just heard for each syllable in the sentence. 20 evaluators (native Mandarin speakers) were instructed to listen to the isolated audio clips and to decide which tone they had just heard. For sentence tone IA, we extracted the audio clip of each sentence each patient pronounced in evaluation blocks. The

evaluators were instructed to listen to the audio of the entire and then choose from the options which tone they had just heard for each syllable in the sentence. We recruited 20 native Mandarin speakers from the Medical School of Fudan University, which do not know the purpose of this test before the evaluation. The tone IA score was defined as the average accuracy of evaluators' judgement.

$$\mathrm{IA} = 1 - \frac{T}{A}$$

Where T is the number of error tone choices; A is the number of total tests.

- **Comparison with baseline models**

We evaluated the performance of our models and compared it against several baselines. The baseline models include (1) CNN models, which do not contain RNN structure, all the other parts remain identical with optimized models. (2) RNN models, which do not contain CNN structure, all the other parts remain identical with original optimized models (**Fig. S3**). WER of decoded sentences were caluculated to evaluate the performance of these baselines.

- **Electrode contributions (saliences)**

To measure how much each electrode contributed to detection and classification performance, we computed electrode contributions (saliences) with the artificial neural networks driving the speech utterance detection model, the syllable and tone classification models, respectively. We used a salience calculation method that has been used for identifing image regions that were useful for image classification[48]. To compute electrode saliences for each type of ANN, we first calculated the gradient of the loss function for the ANN with respect to the input features. The input features were individual time samples of high gamma activity detected time windows for the speech event, tone and syllable classifier. For each input feature, we backpropagated the gradient through the

ANN to the input layer. We then computed the Euclidean norm across time (within each evaluation block) of the resulting gradient values associated with each electrode. Next, we computed the mean across blocks of the Euclidean norm values, yielding a single salience value for each electrode. Finally, we normalized each set of electrode saliences by calculating their root mean square values **(Fig. S4)**.

Supplementary Information

Daohan Zhang[1,2,3#], Zhenjie Wang[4#], Youkun Qian[1,2,3#], Zehao Zhao[1,2,3], Yan Liu[1,2,3], Xiaotao Hao[1,2,3], Wanxin Li[1,2,3], Shuo Lu[5], Honglin Zhu[6], Luyao Chen[7], Kunyu Xu[8], Yuanning Li[4,9*], Jinsong Wu[1,2,3*], Junfeng Lu[1,2,3,8,10*]

Author affiliations:

[1]Department of Neurosurgery, Huashan Hospital, Shanghai Medical College, Fudan University, Shanghai 200040, China

[2]Shanghai Key Laboratory of Brain Function Restoration and Neural Regeneration, Shanghai 200040, China

[3]National Center for Neurological Disorders, Huashan Hospital, Shanghai Medical College, Fudan University, Shanghai 200040, China

[4]School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China

[5]Department of Chinese Language and Literature, Sun Yat-sen University, Guangzhou 510080, China

[6]Faculty of Life Sciences and Medicine, King's College London, London SE1 1UL, UK

[7]School of International Chinese Language Education, Beijing Normal University, Beijing 100875, China

[8]Institute of Modern Languages and Linguistics, Fudan University, Shanghai 200433, China

[9]State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai 201210, China

[10]MOE Frontiers Center for Brain Science, Huashan Hospital, Fudan University, Shanghai 200040, China

[#] These authors contributed equally to this work.

[*] Corresponding authors: Junfeng Lu (junfeng_lu@fudan.edu.cn), Jinsong Wu (wujinsong@huashan.org.cn), Yuanning Li (liyn2@shanghaitech.edu.cn)

Figure S1. Schematic depiction of data organization..... 58

Figure S2. Schematic of control models for speech detector and tone/syllable decoder. 59

Figure S3. Schematic of our previously published tone decoder which successfully decoded tone from intracranial neural signals while participants articulate monosyllable of four target tones. 4

Figure S4. Participants' brains reconstruction overlaid with the locations of the implanted electrodes and their relative contribution of each selected electrodes in decoding onset/syllable and tone. 5

Fig. S1. Schematic depiction of data organization

A. During the optimization stage of the speech detector model, we used a six-fold nested cross validation and each fold consisted of 2 trials, while 10% of data in each trial of the training set

were separated as validation set in order to performing early-stopping.

B. During the evaluation stage of the speech detector model, we used a similar cross validation. Still, 10% of data in each training trial were separated as validation set in order to performing early-stopping.

C. During the optimization stage of the syllable/tone decoder model, we used a six-fold nested cross validation and each fold consisted of 2 trials. In each fold, we trained 5 sub-models with different potions of data performing early-stopping while left for training. The training/validation/test split was 4:1:1.

D. During the evaluation stage of the syllable/tone decoder model, we used a similar cross validation process. In each fold, we trained 10 sub-models with different potions of data performing early-stopping while left for training. The training/validation/test split was 12:3:1.

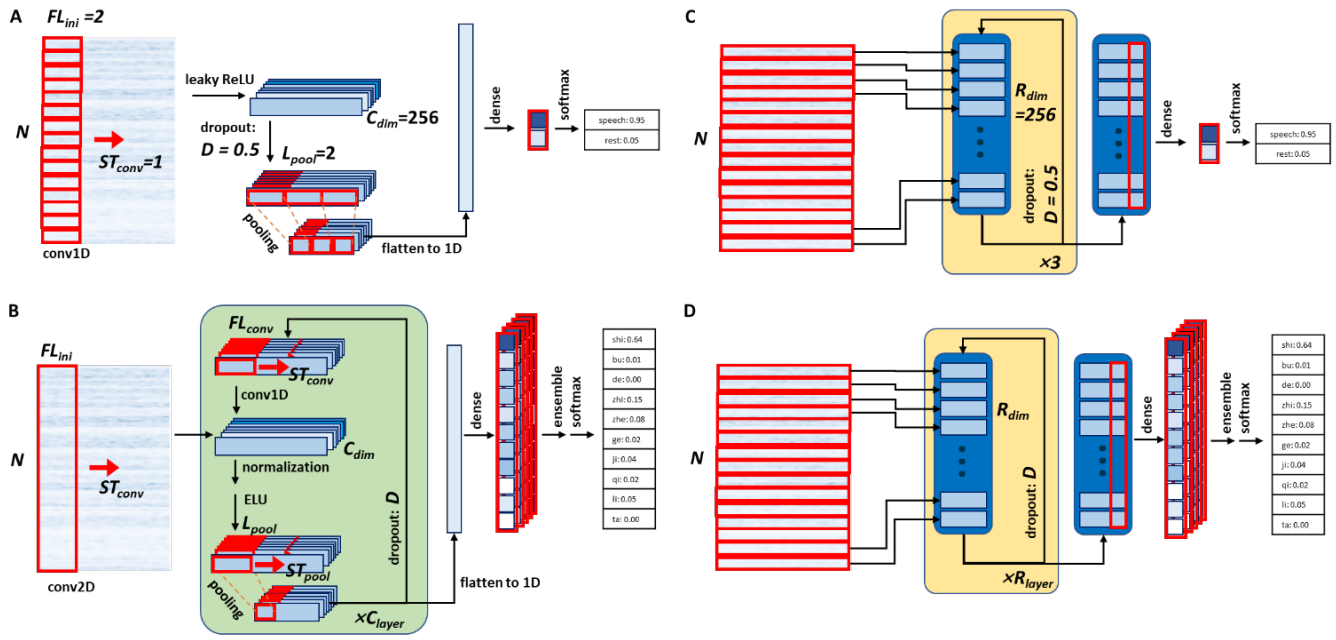


Fig. S2. Schematic of control models for speech detector and tone/syllable decoder.

A. CNN baseline for speech detector, which do not contain RNN structures, and all the other parts remain unchanged.

B. CNN baseline for tone/syllable decoder, which do not contain RNN structures, and all the other parts, including optimized hyperparameters, remain unchanged.

C. RNN baseline for speech detector, which do not contain CNN structures, and all the other parts remain unchanged.

D. RNN baseline for tone/syllable decoder, which do not contain CNN structures, and all the other parts, including optimized hyperparameters, remain unchanged.

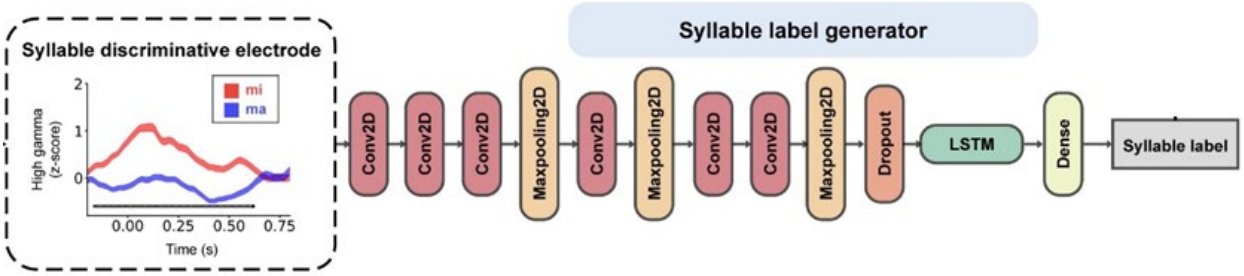


Fig. S3. Schematic of our previously published tone decoder which successfully decoded tone from intracranial neural signals while participants articulate isolated monosyllables of four target tones.

We applied our previously published monosyllabic tone decoder on tone decoding in sentence task. It is a sequential CNN-LSTM structure to generate the syllable label based on manually aligned neural activity when participants produce eight syllables "ma (tone1), ma (tone2), ma (tone3), ma (tone4), mi (tone1), mi (tone2), mi (tone3), mi (tone4)"[15].

1170

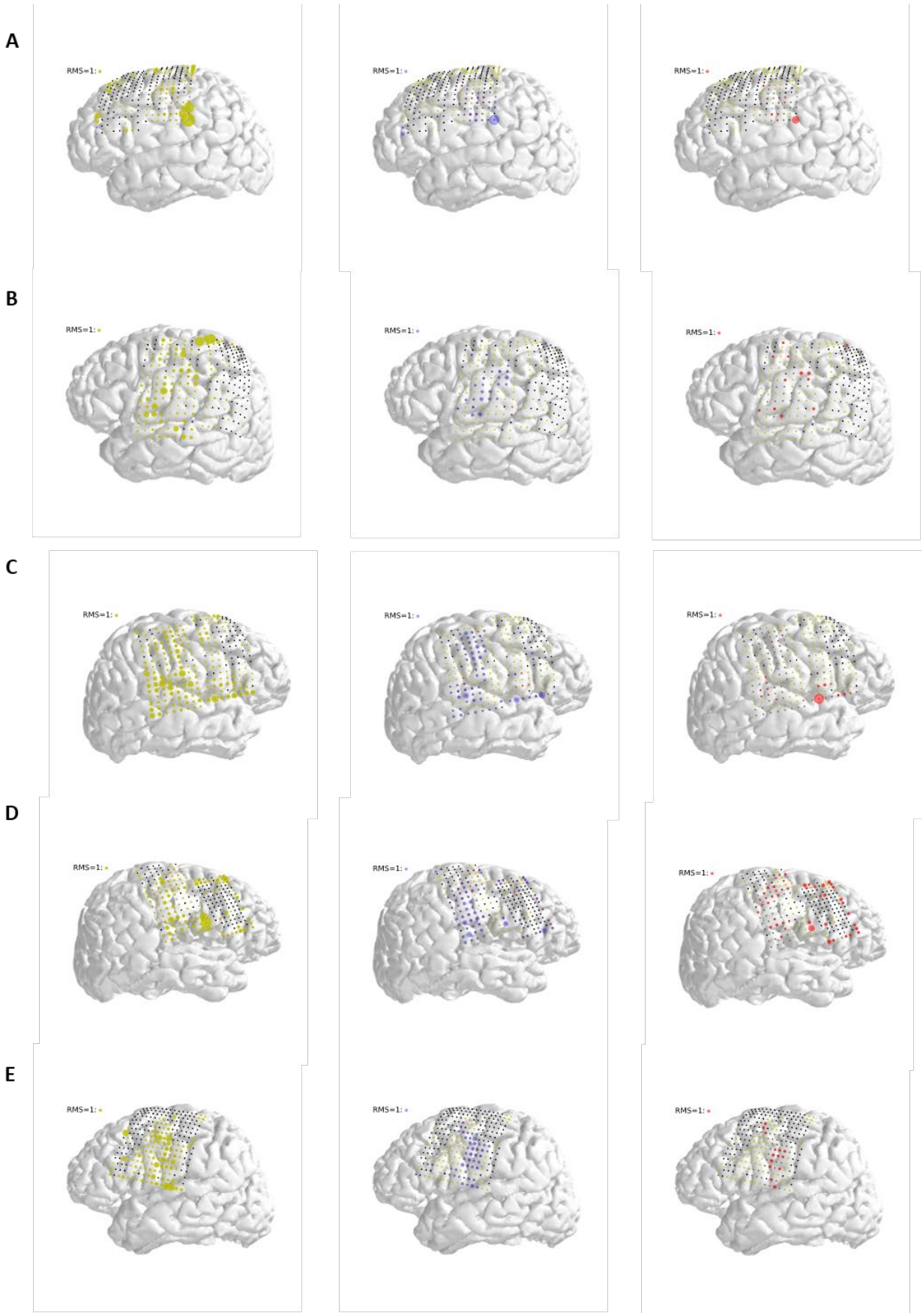


Fig. S4. Participants' brains reconstruction overlaid with the locations of the placed electrodes and their relative contribution of each selected electrodes in decoding onset/syllable and tone.

A-E shown PA1, PA2, PA3, PA4 and PA5. Left part shown speech responsive electrodes in yellow shadows, middle shown syllable discriminative electrodes in blue shadows, while right shown tone discriminative electrodes in red shadows. Radius of shadows around electrodes shown contribution normalized root mean square (RMS), with a scale shown at the upper left of each subplot. The colors of the electrodes themselves represent whether they are speech responsive or syllable/tone discriminative, which is the same as **Fig. 2a-e**.

References:

1 Gilakjani, A. P. & Ahmadi, M. R. A study of factors affecting EFL learners' English listening comprehension and the strategies for improvement. (2011).

2 Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493-498, doi:10.1038/s41586-019-1119-1 (2019).

3 Moses, D. A., Leonard, M. K., Makin, J. G. & Chang, E. F. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nat Commun* **10**, 3096, doi:10.1038/s41467-019-10994-4 (2019).

4 Makin, J. G., Moses, D. A. & Chang, E. F. Machine translation of cortical activity to text with an encoder-decoder framework. *Nat Neurosci* **23**, 575-582, doi:10.1038/s41593-020-0608-8 (2020).

5 Moses, D. A. *et al.* Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. *N Engl J Med* **385**, 217-227, doi:10.1056/NEJMoa2027540 (2021).

6 Angrick, M. *et al.* Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *J Neural Eng* **16**, 036019, doi:10.1088/1741-2552/ab0c59 (2019).

7 Willett, F. R. *et al.* A high-performance speech neuroprosthesis. *Nature*, 1-6 (2023).

8 Metzger, S. L. *et al.* A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 1-10 (2023).

9 Willett, F. R. *et al.* A high-performance speech neuroprosthesis. *Nature* **620**, 1031-1036 (2023).

10 Metzger, S. L. *et al.* A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* **620**, 1037-1046 (2023).

11 Komeiji, S. *et al.* in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 1311-1315 (IEEE).

12 Koch Fager, S., Fried-Oken, M., Jakobs, T. & Beukelman, D. R. New and emerging access technologies for adults with complex communication needs and severe motor impairments: State of the science. *Augmentative and Alternative Communication* **35**, 13-25 (2019).

13 Yip, M. *Tone*. (Cambridge University Press, 2002).

14 Dryer, M. S. & Haspelmath, M. The world atlas of language structures online. (2013).

15 Liu, Y. *et al.* Decoding and synthesizing tonal language speech from brain activity. *Science Advances* **9**, eadh0478 (2023).

16 Fons-ant, H. & Naessentialr, S. A model for the synthesis of pitch contours of connected speech. *annual report of the engineering research institute* **23** (1969).

17 Fujisaki, H., Hirose, K., Hallé, P. A. & Lei, H. in *ICSLP.* 841-844.

18 Chen, M. Y. An overview of tone sandhi phenomena across Chinese dialects. *Journal of Chinese Linguistics Monograph Series*, 111-156 (1991).

19 Ho, D.-a. The characteristics of Mandarin dialects. *The Sino-Tibetan languages*, 126-130 (2003).

20 Dichter, B. K., Breshears, J. D., Leonard, M. K. & Chang, E. F. The Control of Vocal Pitch in Human Laryngeal Motor Cortex. *Cell* **174**, 21-31 e29, doi:10.1016/j.cell.2018.05.016 (2018).
21 Lu, J. *et al.* Neural control of lexical tone production in human laryngeal motor cortex. *Nature Communications* **14**, 6917 (2023).
22 Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327-332, doi:10.1038/nature11911 (2013).
23 Qiumei, Z., Dan, T. & Fenghua, W. Improved convolutional neural network based on fast exponentially linear unit activation function. *Ieee Access* **7**, 151359-151367 (2019).
24 Zargar, S. Introduction to sequence learning models: RNN, LSTM, GRU. *Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, North Carolina* **27606** (2021).
25 Chen, M. Y. *Tone sandhi: Patterns across Chinese dialects*. Vol. 92 (Cambridge University Press, 2000).
26 Pasley, B. N. *et al.* Reconstructing speech from human auditory cortex. *PLoS biology* **10**, e1001251 (2012).
27 Martin, S. *et al.* Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in neuroengineering* **7**, 14 (2014).
28 Herff, C. *et al.* Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices. *Frontiers in neuroscience* **13**, 1267 (2019).
29 Duraivel, S. *et al.* High-resolution neural recordings improve the accuracy of speech decoding. *Nature Communications* **14**, 6938 (2023).
30 Willsey, M. S. *et al.* Real-time brain-machine interface in non-human primates achieves high-velocity prosthetic finger movements using a shallow feedforward neural network decoder. *Nat Commun* **13**, 6899, doi:10.1038/s41467-022-34452-w (2022).
31 Davidoff, M., Robert A. The pyramidal tract. *Neurology* **40**, 332-332 (1990).
32 Yazdan-Shahmorad, A., Kipke, D. R. & Lehmkuhle, M. J. High gamma power in ECoG reflects cortical electrical stimulation effects on unit activity in layers V/VI. *Journal of neural engineering* **10**, 066002 (2013).
33 Sanes, J. N. & Donoghue, J. P. Plasticity and primary motor cortex. *Annual review of neuroscience* **23**, 393-415 (2000).
34 Myung, I. J. The importance of complexity in model selection. *Journal of mathematical psychology* **44**, 190-204 (2000).
35 McCawley, J. D. in *Tone* 113-131 (Elsevier, 1978).
36 Jongman, A., Wang, Y., Moore, C. B. & Sereno, J. A. *Perception and production of Mandarin Chinese tones*. (na, 2006).
37 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).
38 Cuisenier, P. *et al.* Relationship between direct cortical stimulation and induced high-frequency activity for language mapping during SEEG recording. *J Neurosurg* **134**, 1251-1261, doi:10.3171/2020.2.JNS192751 (2020).
39 Leonard, M. K. *et al.* Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature*, 1-10 (2023).
40 Coughlin, B. *et al.* Modified Neuropixels probes for recording human neurophysiology in the operating room. *Nature Protocols* **18**, 2927-2953 (2023).
41 Da, J. in *Proceedings of the fourth international conference on new technologies in teaching and learning Chinese.* 501-511 (Citeseer).

42 Hamilton, L. S., Chang, D. L., Lee, M. B. & Chang, E. F. Semi-automated anatomical labeling and inter-subject warping of high-density intracranial recording electrodes in electrocorticography. *Frontiers in Neuroinformatics* **11**, 62 (2017).

43 Lawhern, V. J. *et al.* EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering* **15**, 056013 (2018).

44 Makin, J. G., Moses, D. A. & Chang, E. F. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience* **23**, 575-582, doi:10.1038/s41593-020-0608-8 (2020).

45 Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).

46 Maas, A. L., Hannun, A. Y. & Ng, A. Y. in *Proc. icml.* 3 (Atlanta, Georgia, USA).

47 Zhan, W., Guo, R., Chang, B., Chen, Y. & Chen, L. in *Corpus Linguistics.* 71-86.

48 Simonyan, K., Vedaldi, A. & Zisserman, A. Visualising image classification models and saliency maps. *Deep Inside Convolutional Networks* **2** (2014).