communications biology

A Nature Portfolio journal

Review article

https://doi.org/10.1038/s42003-025-08511-z

Progress, challenges and future of linguistic neural decoding with deep learning

Yu Wang [1,2,3], Heyang Liu [1,3], Yuhao Wang [1], Chuan Xuan[1], Yixuan Hou[1], Sheng Feng[1], Hongcheng Liu[1], Yusheng Liao[1,2] & Yanfeng Wang [1,2] ✉

Language is the primary medium through which humans achieve information transfer and exchange. It enables the conveyance of ideas, concepts, and messages, thereby playing an indispensable role in social interaction and knowledge dissemination. Linguistic neural decoding aims to obtain outstanding language information from the evoked human brain during information interaction of both textual and spoken formats. In this work, we present a taxonomy of recent neural decoding progress, focusing on deep learning architectures and strategies, especially those implementing large language models (LLMs) for their powerful information understanding, processing, and generation capacity. We conclude with a concise observation of the challenges and potential future directions. This article aims to provide brain scientists and deep learning researchers with an overarching viewpoint of the significant correlations observed in the human brain during language perception and production from a methodological perspective, and thus facilitate their further investigation.

Language facilitates human communication and information exchange, in which the human brain plays a crucial role, enabling complex cognitive processes supporting the linguistic perception, comprehension and production. Despite advancing insights into the human brain, the linguistic representation and processing are still underexplored, making it challenging to explain intuitively the internal mechanisms[1]. Deep learning presents a viable solution for understanding language in the brain by utilizing large-scale trainable parameters to map the correlation between external stimuli and neural activity. This paper summarizes representative solutions and current progress on linguistic neural decoding, addressing the potential promotion by leveraging large language models (LLMs). Progress in this field involves the joint efforts of neuroscientists and artificial intelligence researchers. We introduce the neurological foundations supporting neural decoding with deep networks and illustrate multiple model architectures. We classify task forms into multiple standardized paradigms, facilitating researchers to further progress their work, and conclude by discussing the challenges faced by related fields and proposing directions for potential applications. It is important to note that the language discussed in this paper is a synthesis of semantic and syntactic information, featuring specific content presented in a defined format, primarily including text and speech forms. Visual image reconstruction is excluded, as it contains semantic content but lacks linguistic syntactic presentation. Similarly, motions such as handwriting are not considered, given their involvement with bodily movements and minimal relevance to language.

Supplementary Note 1 summarizes the main content of this survey. We begin by discussing the neurological basis of linguistic decoding. Neural tracking ensures the temporal alignment of brain responses with linguistic properties, while continuous neural prediction supports the integration of contextual information. Stimuli recognition is the simplest form of neural decoding, involving the differentiation of linguistic stimuli by analyzing the subject's evoked brain responses. For text stimuli reconstruction, decoding is performed at the word or sentence level using classifiers, embedding models, and custom network modules. Considering the dynamics of speech flow, restoring the speech envelope, mel-frequency cepstral coefficient (MFCC), and speech waves present broader challenges. Brain recording translation paradigms are applied in natural reading and listening scenarios, where the decoding system generates the stimulus sequence in textual or speech form based on the evoked brain response. This task is analogous to machine translation, treating brain activity as the source language and translating it into human-understandable text. Speech neuroprosthesis focuses on decoding inner or vocal speech based on human intentions. The field has progressed from phoneme-level recognition to open-vocabulary sentence decoding. Brain-to-speech technology is a promising direction, with spectrograms generated through matching algorithms or by considering speech properties, synthesizer parameters, and articulator

[1]Shanghai Jiao Tong University, Shanghai, China. [2]Shanghai Artificial Intelligence Laboratory, Shanghai, China. [3]These authors contributed equally: Yu Wang, Heyang Liu. ✉e-mail: wangyanfeng622@sjtu.edu.cn

movements. Additionally, to assist neuroscientists and artificial intelligence researchers in better developing decoding systems, we provide evaluation metrics introduced from deep learning tasks before the introduction of brain decoding solutions (Section 2) and a concise summary of the machine learning models and algorithms discussed in this review (Supplementary Note 1). Compared to previous reviews of neural decoding[2,3], our article includes recent advances and expands the task formats to a larger scope. Furthermore, our work focuses on the specification of task paradigms and methodology, which complements the work on the internal mechanisms of language models and human language systems[4].

Brain-network alignment

Brain signal recordings measure and quantify the biometric neural response from the human brain, which can be divided into two categories: invasive and non-invasive. The latter, including functional magnetic resonance imaging (fMRI), electroencephalography (EEG), magnetoencephalography (MEG), etc., are affected by transcranial attenuation with a lower signal-to-noise ratio (SNR)[5]. On the other hand, invasive methods such as electrocorticography (ECoG) are hampered by the limited public availability due to the necessity of neurosurgery. The essence of deep learning lies in leveraging the inherent correlations within data to complete prediction, regression and generation, and the alignment of neural activities and linguistic representations is crucial for enabling these capabilities. Specifically, neural tracking enables the theoretical possibility of achieving temporally continuous decoding from evoked brain activities, while the neural prediction process underscores the benefit of contextual information integration, which is commonly used in current neural decoding approaches.

In this paper, the process by which the brain receives external language stimuli and transforms it into specific neural representations is referred to as perception, which primarily involves neural encoding. During this process, external stimuli are transformed into specific neural response patterns, with neural tracking ensuring the association between language and neural representations[6], as shown in Fig. 1a. The cortical activity automatically tracks the dynamics of speech as well as various linguistic properties, including surprisal, phonetic sequences, word sequences, and other linguistic representations[7–10]. A minor time shift has been observed for information transfer and neural response. It ensures the temporal alignment of brain recordings with linguistic representations, facilitating the serialized and temporal modeling of cortical activities. As shown in Fig. 1b, language stimuli are encoded into regular evoked brain responses. In contrast, linguistic neural decoding aims to reconstruct the stimuli perceived or the intention expressed from high-dimensional brain responses. In Fig. 1c, the brain undergoes the following processes in communications: perception converts external linguistic stimuli into specific neural patterns; comprehension involves steps such as semantic extraction, understanding, and reasoning; generation (production) entails outputting responses in a specific form, for example, by guiding the vocal organs to produce speech. In natural listening settings, the human brain encodes a wide range of acoustic features and processes external language stimuli temporally through prediction, highlighting the importance of contextual information in cortical perception, even at the level of single neurons[11,12]. Predictive processing fundamentally forms the comprehension mechanisms, occurring hierarchically in both acoustic and linguistic levels[13–16]. This phenomenon underscores the profound impact of context on the forecasting and tracking of ongoing speech streams, necessitating the use of contextual representations to investigate cortical responses[17–19]. This characteristic is similar to language models constructed by neural networks, where the same stimuli presented in varying contexts are mapped onto diverse semantic features. Despite ample evidence supporting the predictive characteristics of human language processing, it has recently been suggested that the benefits actually come from the capacities of models to predict brain responses[20]. Regardless of the mechanisms in the perception process, language models have the potential to understand and infer neural responses.

When processing natural language, artificial neural networks exhibit patterns of functional specialization similar to those of cortical language networks[21]. Research, particularly focused on Transformers and LLMs, shows that the representations in these models account for a significant portion of the variance observed in the human brain[22,23]. To further this analogy, it has been verified that the brain encoding models and pre-trained LLMs follow the scaling laws, where the model performance increases as the number of parameters grows, indicating the necessity to develop larger systems to bridge the brain activity patterns and human linguistic representations if given sufficient data and other necessary conditions[24,25]. A recent study[26] has indicated that, in addition to model scaling, the amount of data utilized during the training process positively influences the similarity of representations between the brain and neural networks. Furthermore, alignment training is deemed an effective approach to enhancing this similarity.

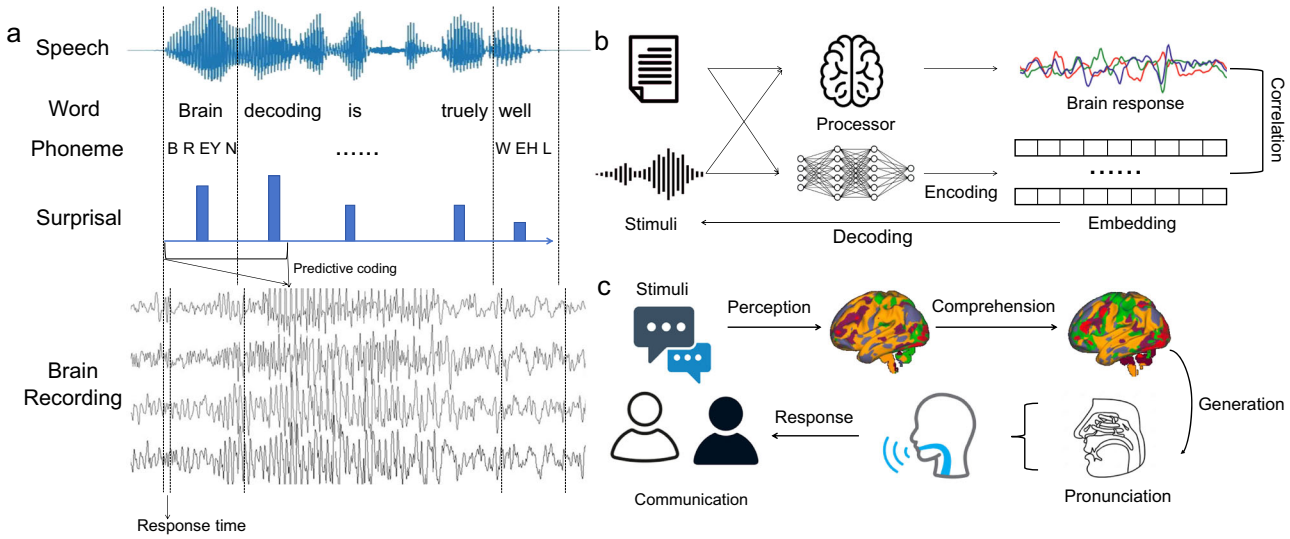

Fig. 1 | The formation of linguistic representation in the human brain. a The human brain tracks the dynamic flow of speech and linguistic properties with minor response delay, and the neural response is performed in a continuous predictive manner. **b** The human brain and the neural networks can both encode textual or verbal stimuli into specific representations, and the decoding process aims to reconstruct the linguistic information. **c** In vocal communications, the brain processes the perception, comprehension, and generation of language. The processor and communication icons are from Vecteezy and Dreamstime.

Table 1 | Divided categories and their corresponding characteristics

Category	Task/Experiment	Data	Stimuli	Decoding Target
Stimuli recognition	Textual stimuli classification	N	Text	Text id (word, sentence, ...)
	Speech stimuli reconstruction	N	Speech	Text ID, speech features, and waves
Brain recording translation	Nature reading	N	Text	Text sentences
	Nature listening	N	Speech	Text sentences
Speech neuroprosthesis	Inner speech recognition	I	Speak	Text (phoneme, word, sentence)
	Brain-to-speech	I	Speak	Speech wave

'N' and 'I' in the Data column represent non-invasive and invasive data, respectively.

Table 2 | Evaluation metrics for linguistic neural decoding

Target	Metric	Application	Origin	Methods
Text	Accuracy	Textual stimuli classification	Classification	Percentage of correct output
	BLUE			Precision of n-grams
	ROUGE	Brain recording translation	MT	Recall of n-grams
	BERTScore			Semantic similarity
	PER			Phoneme accuracy
	CER	Inner speech recognition	ASR	Character accuracy
	WER			Word accuracy
Speech	PCC		Statistics	Linear correlation of variables
	STOI		Intelligibility	Human intelligibility correlation
	FFE	Speech reconstruction		Accuracy of pitch (F0)
	MCD		TTS	Accuracy of MFCC
	MOS			Subjective human evaluation

Neural decoding division and evaluation

Linguistic neural decoding aims to generate the corresponding external stimuli or inner intention from the activated brain signals. This field has lacked a fine-grained division, preventing researchers from systematically conducting their work. In this review, previous research has been categorized according to the experiment design, stimulus type and decoding target (Table 1). Stimuli recognition is the simplest form and usually requires a modest candidate set and limited sequence length. For speech stimuli, in addition to identifying the textual content, some work considers reconstructing simple speech features and waveforms. These tasks are typically treated as simple classification or regression. Brain recording translation differs in its ability to handle open-vocabulary continuous decoding, which means a sharp increase in the search space and results in a deterioration in the accuracy without introducing intrusive signals. This task focuses more on semantic consistency rather than the absolute identity of the text. Speech neuroprosthesis aims to generate inner speech from spontaneous neural activation patterns. The subjects do not receive external stimuli but perform pronunciation tasks of imagined speech or attempted speech. Researchers have achieved word-level high-precision continuous decoding with invasive recordings.

As an interdisciplinary field of neuroscience and artificial intelligence, early work on neural decoding mainly follows the paradigm of classification, recognition and sequence decoding. Similar experiments are closely related to machine translation (MT), text-to-speech (TTS), and automatic speech recognition (ASR). Table 2 summarizes the evaluation metrics. In the textual stimuli classification paradigm, accuracy is widely used to measure the percentage of correct instances. As for sequential decoding, ASR and MT tasks generate text sequences with distinct accuracy requirements. The evaluation metrics for the latter focus on semantic consistency, which is extensively employed in brain recording translation. To be more specific, BLEU (bilingual evaluation understudy)[27] calculates the precision of n-grams compared to reference translations, and ROUGE (recall-oriented understudy for gisting evaluation) pays more attention to recall. BERTScore[28] is a recent metric leveraging deep contextualized embeddings from BERT[29] to capture semantic similarity instead of matching exact n-grams. When invasive data is used, ASR metrics become more applicable, such as inner speech recognition in speech neuroprosthesis. WER (word error rate) is a common metric of ASR systems. It measures the accuracy of decoded hypotheses word by word. In addition to the word-level calculation, CER (character error rate) and PER (phoneme error rate) are carried out on character- and phoneme-level, respectively.

In natural listening and speaking scenarios, the metrics derived from TTS are mainly used in speech reconstruction tasks, for the decoding outputs of both are speech waves. The simplest method is to calculate the statistical correlation between the generated and reference speech, with the PCC (Pearson correlation coefficient) showing the most preference. It measures the linear relationship between two continuous variables. STOI (short-time objective intelligibility)[30] is used to evaluate the speech intelligibility. It is designed to provide an objective measure that correlates well with human subjective intelligibility ratings. FFE (F0 frame error)[30] and MCD (mel-cepstral distortion)[31] aim to evaluate the accuracy of pitch and MFCC, respectively, which have been widely used in TTS. MOS (mean opinion score) is commonly used to estimate the perceived quality of audio, video, and multimedia content. It provides a subjective measure of quality based on human judgments and typically uses a five-point scale where participants rate the quality of the synthesized speech slices.

Stimuli recognition

As shown in Fig. 2a, compared with fine-grained decoding, a moderate set of candidates is necessary for stimulus recognition. The subjects passively receive external information by reading text or listening to podcasts, and

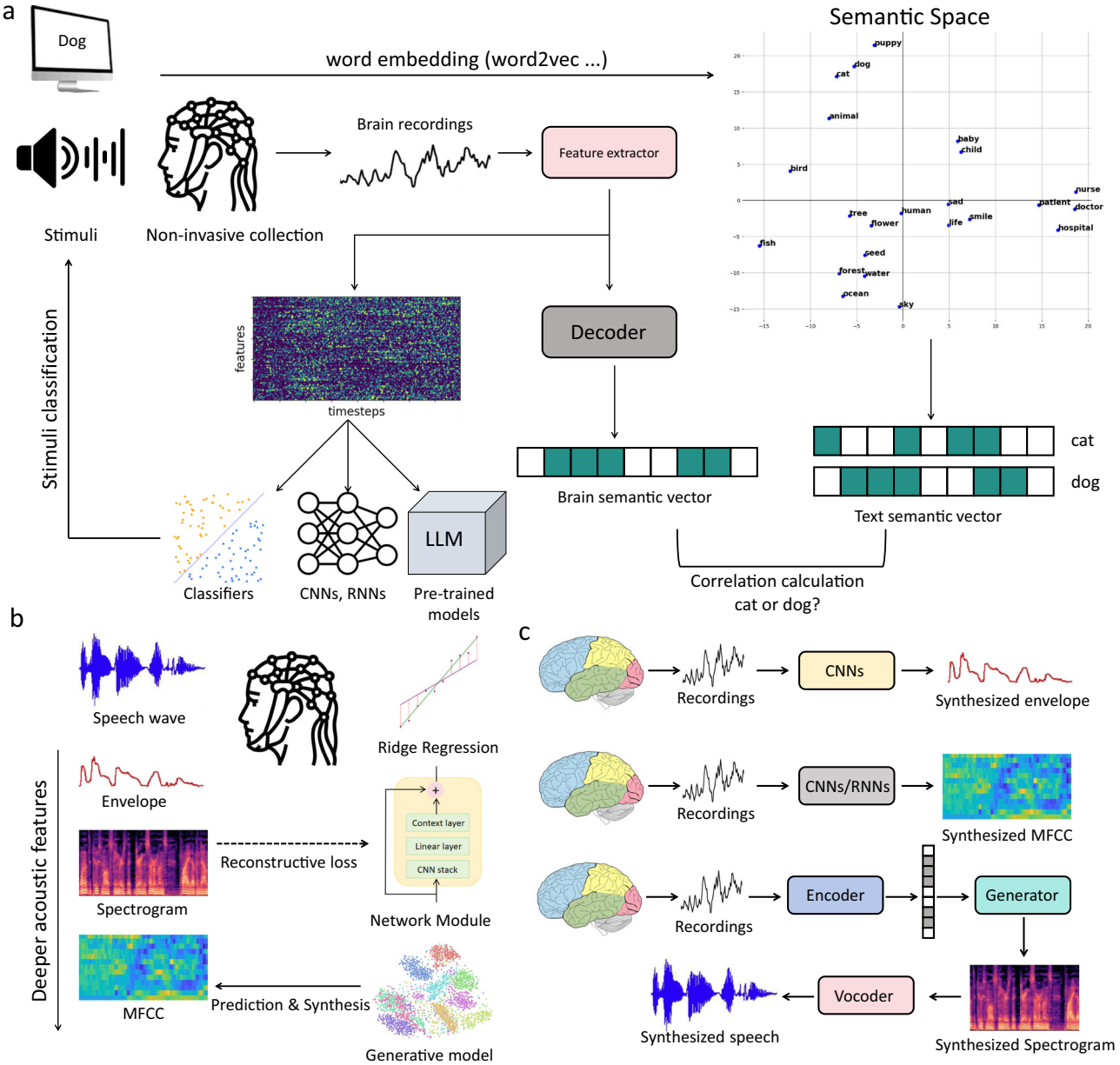

Fig. 2 | Stimuli recognition of evoked brain activity. a An overview of the stimuli recognition task. The subject receives textual or vocal information while the active brain signals are collected. The raw brain recordings are processed into feature space, followed by classifiers, networks or pre-trained models to distinguish the original stimuli based on the complexity and candidate size. Several approaches adopted word embeddings (i.e., word2vec[87]) to compare the decoded vector in a semantic space. **b** In natural listening scenarios, restoring the original speech features and waveform is a more complex task. Regression models (i.e., ridge regression), CNN and RNN-based network modules, and paramount generation models (i.e., GAN) are widely used. **c** The decoding architecture for various speech-related targets. The speech envelope can be easily reconstructed with CNNs, while more complex networks are necessary for the decoding of MFCC[61,65]. The most difficult task is to synthesize the stimulus wave, where an encoder-generator-vocoder architecture has been verified effective[70]. The non-invasive collection icon is from Vecteezy.

deep learning methods are adopted to classify the original stimuli based on evoked brain signals.

Textual stimuli classification

Language presented in text highly condenses information and avoids the temporal variability of corresponding speech signals. Early work focused on recovering language information from text stimuli. This paradigm distinguishes the original information provided to the subject from several candidates. The previous approach defined a word set of concrete nouns to avoid neural representations of abstract concepts[32,33]. Classifiers were adopted to distinguish which word had been perceived by the subject. Following this, other studies extended to abstract nouns, proving the superiority of text-based models over visually grounded approaches[34], resulting in the evaluation of 8 different word embedding models for predicting another given either the neural activation patterns or word representations[35]. In ref. 36, the researchers presented a neural decoding system based on a semantic space trained on massive text corpora. The decoded representations were detailed enough to differentiate between sentences with similar meanings. Larger vocabularies bring greater difficulties. In ref. 37, a network module with dense layers and a regression-based decoder was implemented to directly classify an fMRI scan over a 180-word vocabulary. The recognition effect far exceeded the chance probability (5.22% Top-1 and 13.59% Top-5 accuracy). Following these achievements, researchers predicted masked words and phrases[38]. The proposed approach

utilized an encoder-decoder paradigm and achieved 18.20% and 7.95% top-1 accuracy over a 2000-word vocabulary on the two tasks, respectively.

Starting from these approaches, some work treated the sentence-level responses as a combination of latent word effects, bridging the relationship between the neural process when receiving words and a whole sentence[39–41]. Following these approaches, the holistic encoding of sentence stimuli was proposed[42,43]. Studies further evaluated various distributed semantic models to predict or decipher brain response to textual sentences, with the Transformer-based model achieving the best performance[44]. Another classification task was performed on the passage level. The researchers predicted the evoked brain response during natural reading and classified the corresponding brain activity by distance to the synthesized brain image[45]. In ref. 46, the approach bridged the textual stimuli pattern and MEG recordings using multiple network architectures, with BERT showing the best performance.

Textual stimulus classification is greatly limited by the decoding range and is almost performed on dozens or hundreds of candidates, which is separate from real-world applications. As an initial attempt, this task illustrates the possibility of obtaining textual information from the evoked cortex, gradually developing into open vocabulary sequence decoding.

Speech stimuli reconstruction

Speech perception entails processes that convert acoustic signals into neural representations. In neuroscience, this includes the complete pathway from the cochlear nerve to the auditory cortex areas. Previous research has demonstrated that the hierarchical structure in neural networks trained on speech representations aligns with that of the ascending auditory pathway, supporting the feasibility of deep learning approaches[47].

The speech stimuli reconstruction aims at forming semantic information, acoustic features, and synthesized perceived speech from evoked brain activity (Fig. 2). Classifiers had been used to distinguish perceived stimuli before the deep learning methods were applied. The logistic regression was applied to classify the speech stimuli perceived by an unseen subject during training[48]. Inspired by the ASR systems, the phoneme-level Viterbi decoding was introduced to recognize the heard utterance in a question-answering setting[49]. Another work introduced a contrastive learning model inspired by CLIP[50] to predict the correct segment out of 1000 possibilities[51]. It leveraged the correlation between speech waves and EEG/MEG time series with wav2vec 2.0[52] and convolutional neural networks (CNNs) as the speech and brain modules, respectively. The research on content and subject recognition is not separated, considering that the speech flow can be identified in both spaces. One attempt was to adopt variational autoencoders to transform the EEG space into disentangled latent spaces, representing the content and subject distribution, respectively[53].

The speech envelope refers to the variations in amplitude and intensity of a speech signal over time. It plays a crucial role in speech perception and understanding, for our brains are tuned to these variations, helping recognize speech sounds, syllables, and words[54,55]. Earlier work focused on the signal processing and linear model to align the envelope representation with brain activity[56]. After that, some other research implemented convolutional models[57,58] or based on mutual information analysis[59]. In ref. 60, the researchers evaluated the envelope construction performance of ridge regression, convolution and fully connected layers. The more in-depth research led to the development of the VLAAI, a convolution-based architecture to achieve more precise reconstruction[61]. Considering the highly robust correlation between envelope and linguistic information, some extended to a cocktail party setting, where the attended speech envelope was predicted with a context-aware neural network[62]. A recent work adopted a transformer-based encoder-decoder architecture[63]. Compared with the speech envelope, MFCC is a widely used feature in speech recognition that represents the short-term power spectrum of sound. The parallels between speech recognition and brain-to-text technologies inspired the prediction of MFCC from brain recordings using custom networks, regression and generative models[64,65]. Subsequent research extended this approach to various acoustic features, predicting 16 different types using an attention-based regression model[66].

Instead of reconstructing the acoustic features, synthesizing speech directly from brain recordings is more challenging, yet it holds greater practical significance and application prospects. In ref. 67, the researchers opened up the possibility of speech restoration with evoked brain recordings. This approach implemented a linear spectrogram model with strict recording quality and word selection requirements. The following studies investigated the reconstruction performance of linear and non-linear models based on speech spectrogram and vocoder parameters of the synthesizer[68]. The result demonstrated the significance of non-linear neural networks. Other studies leveraged Wasserstein GAN (wGAN)[69] for generator pre-training to obtain the spectrogram representation[70], and dual generative adversarial network (DualGAN)[71] for cross-domain mapping between EEG signals and speech waves[72]. In this field, network optimization contributes to performance improvement, with the self-attention module demonstrating its superiority to multi-layer perceptrons (MLPs) and CNNs to restore the spectrogram[73].

Compared with text, speech varies more and contains richer information, which brings more challenges to restoring the speech stimuli. From the current perspective, reconstructing recognizable speech waveforms requires multiple rounds of iterations of recording quality and network architecture.

Brain recording translation

Decoding natural sentences from brain signals remains a significant challenge. Unlike simpler tasks that convert brain signals into categorical labels, brain recording translation directly decodes linguistic stimuli into word sequences (Fig. 3). This process borrows concepts from machine translation, as both tasks aim to map representations between two different units of analysis. Brain recording translation involves open-vocabulary decoding based on neural patterns, which implies a vast search space. However, it fundamentally differs from machine translation, for the stimulus text or speech is deterministic, while the potential targets can be numerous for the latter. Given the resolution limitations of non-invasive neuroimaging, this task demonstrates the balance between the brain recording quality and the recognition granularity.

The brain recordings are typically collected during natural reading and listening scenarios (Fig. 4a). Researchers reconstruct text stimuli through deep learning solutions. In ref. 74, the authors first introduced the concept of machine translation into neural decoding. Although this work decoded word sequences during attempted speech, the serialization of text generation provided new insights for subsequent work. The neural network architecture contained temporal convolution to model contextual relations and encoder–decoder recurrent neural networks (RNNs) to generate predicted text. The experiment was conducted with ECoG recordings and carried out on a vocabulary list of several hundred words. The following work turned to the BLEU and ROUGE scores[75]. This work largely expanded the decoding vocabulary (~50,000) by fully leveraging the inference capabilities of pre-trained LLMs. Specifically, a multi-layer Transformer encoder is used to map non-invasive EEG features to the embedding space of the BART tokenizer[76], and the decoded sentence is generated through its decoder. Following these achievements, this paradigm was progressed by directly interpreting raw brain signals with contrastive learning methods and introducing discrete encoding into the EEG recording representation borrowed from VQ-VAE[77,78]. However, their models were highly estimated with teacher-forcing schema during evaluation[79], which means instead of feeding the model's previous predictions for the next time step, the actual target values were used during inference. This prevented them from generating meaningful sentences in real-life applications.

Alternatively, a solution with implementation potential was proposed to generate text directly from MEG recordings without teacher forcing[80]. The proposed architecture, NeuSpeech, utilized MEG instead of EEG or fMRI and incorporated a Whisper model. During the training process, only a small portion of parameters within the encoder were fine-tuned, while the

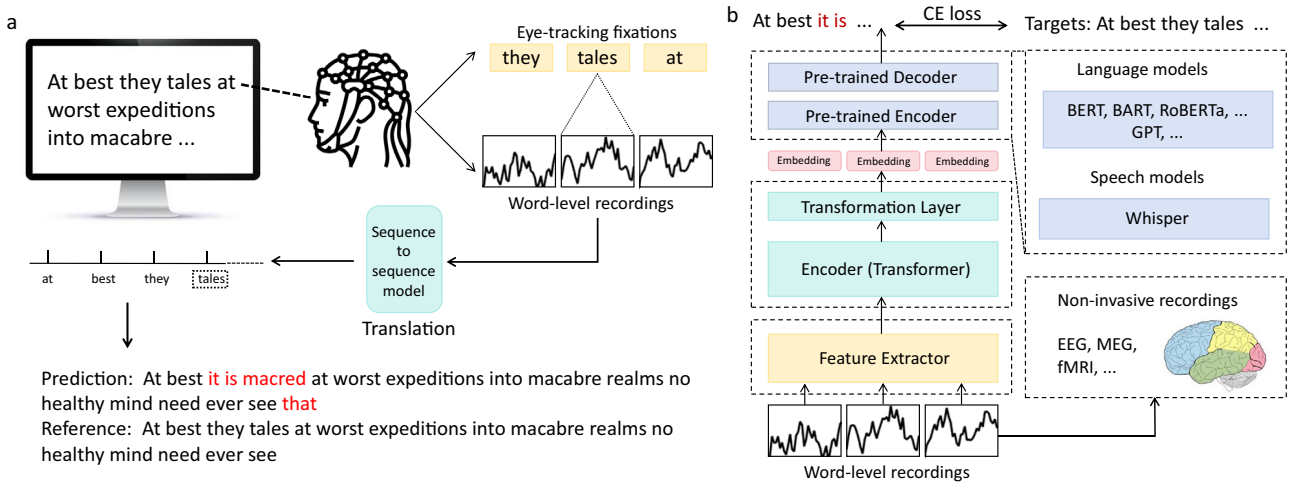

Fig. 3 | The experiment setting and model architecture of brain recording translation. a For natural reading, the subjects are exposed to text while the active brain signals are collected. The eye movements are typically recorded to determine the text transcription corresponding to the brain data at each time step. A sequence-to-sequence model processes the evoked brain recordings to determine the related word and then forms the decoded sentences. **b** A feasible translation model architecture, including feature extraction, feature transformation and a pre-trained encoder–decoder to generate the decoding sentence. Both the pre-trained language models (i.e., BART) and speech models (i.e., Whisper) have been verified to be effective. The non-invasive collection icon is from Vecteezy.

Transformer layers in the encoder and the entire decoder remained frozen. The advanced solution contributed to an open-vocabulary MEG-to-text translation model capable of generating unseen text[81], where multiple alignments were conducted between the MEG recordings and speech audio. The brain module was mapped to Whisper representations in three aspects: the Mel spectrogram, hidden state and decoded text. Another work proposed simultaneously leveraging the inferring ability of LLMs and implemented an fMRI encoder to learn a suitable prompt in an auditory-decoding setting. The prompt of text and fMRI modalities were aligned through a contrastive loss[82]. In ref. 83, the researchers directly used the representation decoded from fMRI as the input for LLMs and found a closer alignment with content deemed surprising for the LLM backbone. As for the improvement from modeling strategies, PREDFT utilized the predictive encoding with a side network to generate predictive representation with a multi-head self-attention module[84].

The setting of the brain recording translation is reasonable. Under this paradigm, more work emerged that implements LLMs to translate brain signals in large vocabularies, including schemes using contrastive learning and curriculum learning[85]. By constructing positive and negative sample pairs from the EEG of different subjects exposed to identical or different sentence stimuli, the method aimed to pull closer the representation distances of semantically similar inputs, while pushing apart dissimilar ones. More similar sample pairs are considered challenging, and the strategy followed a progression from easy to difficult. A similar approach was also used for decoding fMRI signals, which used an encoder-decoder architecture with BART as a text generator[86]. The reconstruction loss of fMRI signals was used to train a better encoder, and the discretized EEG signal and the text vector after word2vec[87] were fed to the contrastive learning module in EEG-text pairs, in which the EEG representation aligned with pre-trained language models. Another method of experiment was to collect brain recordings while participants listened to narrative stories[88]. The fMRI data was sent into GPT after the feature extractor to complete the sequence generation task. Under the same experimental context,[89] employs encoders and projectors to align the distributions between fMRI and text. An external large model, GPT, samples candidate words before selecting the option with the closest distribution to the predicted fMRI signal. This process completes the sequence decoding in an autoregressive manner.

The models of brain recording translation, especially the structures proposed in the past year, and their performance on various datasets are shown in Supplementary Table 1. The word sequence decoded from the non-invasive brain signal shows great disparity with the original textual signal, as reflected in the high WER, while they are consistent with semantic correlation, achieving a promising BERTScore. Considering the promotion prospects of non-invasive signal acquisition equipment, this is a feasible experiment design, which does not require accurate decoding of text information but focuses more on semantics reconstruction.

Speech neuroprosthesis

Some neurological diseases can result in the loss of communication abilities. Many patients rely on a brain–computer interface (BCI) to spell words[90,91], move the computer cursor[92], and direct handwriting[93]. Although these systems can improve the quality of life for patients, communication efficiency is a concern. A major challenge is to overcome the limitations of current spelling-based methods to achieve a natural rate of communication. The goal of speech neuroprosthesis (SN) is to directly decode the words or speech waves that the experiment participants intend to speak from their brain signals (Fig. 4). This represents a hopeful path for creating devices that assist in voice communication.

Inner speech recognition

The inner speech was first called imagined speech in a two-phoneme classification task[94]. The subjects have lost their ability to produce recognizable sounds, and the brain signals are recorded as they try to speak. In some experiments, the brain signals during vocal speech are also collected. Unlike brain recording translation, inner speech recognition demands high-quality brain waves, as high-resolution neural recordings improve the accuracy of speech decoding[95]. This task is highly correlated with ASR, for they both: (1) model the relation between diverse temporal features and deterministic textual information; (2) correlate with pronunciation and acoustics; (3) aim to generate language-compliant text. A recent study shows that even at the level of a single neuron, there are significant neural representations related to inner and vocalized speech that are sufficient to discriminate between words from a small vocabulary[96].

Phonemes, recognized as the foundational elements of speech pronunciation, have historically been the focus of initial studies aiming to decipher human articulatory patterns through brain activities. Previous studies have provided evidence for the neural representation of phonemes and other acoustic features during the perception of speech[97,98]. The pioneer attempted to apply instance-based matching algorithms and demonstrated the feasibility of text decoding from brain recordings even without learning

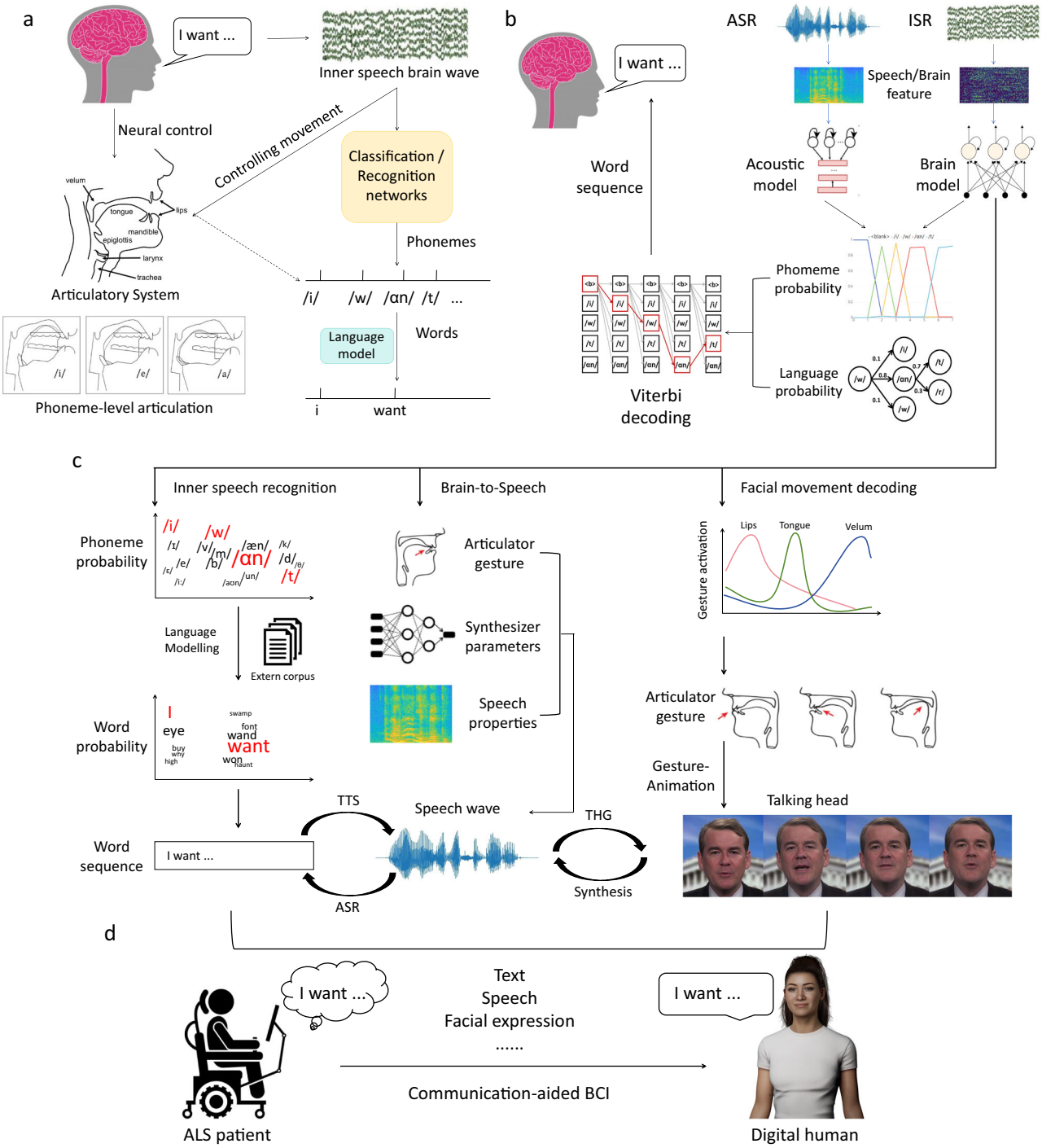

Fig. 4 | Overview of speech neuroprosthesis. a The experimental setting for inner speech recognition. From the neurological perspective, brain waves control the movement of the articulatory system to complete the pronunciation of each phoneme in a series, indicating the mapping from evoked brain signals to movements of the articulators to phonemes. The classification and recognition module is adopted to generate the corresponding phoneme sequences before leveraging the language model to form word sequences. **b** The comparison between ASR and inner speech recognition (ISR). The raw time-series signals are processed for feature extraction and then fed into the acoustic and brain models, respectively. Both models aim to bridge the relationship between learnable features related to acoustics and phoneme sequences. The Viterbi decoding algorithm is performed on the sum of the phoneme probability from the acoustic/brain model and the language probability derived from a language model trained on an extensive corpus to generate the decoded word sequences. **c** The brain model can be implemented to decode various modalities. For inner speech recognition, the phoneme and word sequences are decoded with the aim of language models. For brain-to-speech decoding, the speech waves are synthesized according to the articulator gestures, synthesizer parameters or speech properties. By modeling the articulator gesture probability and adopting a gesture-animation system, the talking head can be generated. Different modalities are associated through TTS, ASR, talking head generation (THG) and synthesis methods. **d** The acoustic-related brain activities show the potential to develop communication-aided BCI for ALS patients, considering the decoding feasibility of text, speech and facial expressions. The articulation and ALS icons are from Oxford Academic, Springer Open, and Iconfinder. The talking head image is from ref. 153.

for features[99–101]. The following research concentrated on identifying these phonetic units, framing the task similarly to classification due to the relatively narrow scope of phoneme varieties. Experiments have been conducted using linear classifiers[102], support vector machine (SVM)[103–105], naive Bayes classifier[106], k-nearest neighbor classifier[94], linear discriminant analysis (LDA) classifier[107–109], flexible discriminant analysis (FDA)[110], and based on brain recording features after principal component analysis (PCA)[111]. The above work was conducted with a few phoneme candidates with clear acoustic boundaries. Following this, the researchers achieved full-set phoneme decoding of American English[112], and implemented a similar approach with brainwave recorded by mobile EEG devices[113].

Progressing from phonemes, researchers achieved advancements toward decoding brain signals into words within a modest vocabulary range. Many investigations were conducted in severely restricted sets with clearly distinguishable pronunciations. Due to the small vocabulary (typically involving several, a dozen, or several dozen candidates), such approaches employed classifiers. In refs. 114–116, the researchers introduced a human-defined lexicon, where a multiclass SVM and relevance vector machine were used for intended speech decoding. Another work based on classification distinguished five words in Spanish and focused on the multiple-modality fusion of text, sound and EEG signals[117]. The most recent achievements conducted the illustration of speech-related representation on a single neuron level recognition[96]. The LDA classifier was adopted to distinguish six words and two pseudowords. Deep learning methods have also been applied to the recognition of imagined speech. The premier attempt implemented several networks to classify imagined words "yes" and "no"[118], followed by research utilizing deep belief neural networks for brain activity feature extraction as well as phoneme and word recognition[119]. The cascade approaches divided the pipeline into convolutional-based modules, including an MFCC prediction module and a word classification model[65]. Network structures with larger parameters are suitable for more complex recognition units, for instance, conducting long word recognition using a mixed network module containing CNNs and RNNs[120]. To test the recognition performance of the network model on longer units, the researchers investigated the decoding performance of five imagined and spoken phrases with fully connected layers and CNNs[121].

The challenge of low SNR in brain signal recordings, primarily from non-invasive techniques, is a significant obstacle to expanding the decoding space[5]. In ref. 74, the authors achieved word sequence decoding on a vocabulary of 250 words using an RNN-based encoder-decoder architecture with invasive ECoG recordings. The most promising approach to generating sentences originates from speech recognition tasks. Specifically, the hybrid model ASR includes an acoustic model, a language model, and a lexicon. The acoustic model calculates the scores of recognition units and then adds them to the language model scores to generate the decoding hypothesis. The cascade speech neuroprosthesis replaces the acoustic model with a brain model and decodes the corresponding phoneme or small-vocabulary word hypothesis before generating the sentences[122,123]. These works typically adopted the Gaussian mixture model (GMM) to fit the data distribution of invasive brain activities. Such approaches did not make a groundbreaking impact until the replacement of GMM with artificial neural networks contributed to a steady improvement[124,125]. This groundbreaking work used RNNs to model the mapping relationship between invasive brain activity and phonemes. The phoneme scores, in conjunction with an n-gram language model trained on a large amount of external text, worked together through the Viterbi search algorithm to decode sentence hypotheses, and a lexicon established the connection from phonemes to words. Through this work, researchers achieved a 25.8% WER on a vocabulary of 125,000 words within the acceptable bounds of performance[126], with a recognition rate of 62 words per minute. A similar previous work was proposed[127], in which an encoder–decoder architecture with a feature regularization module was used to decode character sequences from ECoG recordings. However, the regularization process consumed acoustic and articulatory kinematic features, which are unavailable for ALS patients. The continuous speech decoding has extended to logosyllabic languages like Mandarin Chinese, designing three CNNs to predict the initials, tones and finals of Pinyin, a phonetic text input system based on the Latin alphabet[128]. The prediction of initials was based on the articulatory feature, including the place and manner of articulation and whether voiced or aspirated. A more convincing result appeared in multilingual recognition, where the participant was presented with the target phrases either in English or Spanish[129]. In ref. 130, encoder–decoder RNNs were implemented to recognize the vocal speech, where the representations generated by revised wav2vec[131] yielded superior decoding performance to the original ECoG data. Another recent approach introduced an end-to-end framework with pre-trained LLMs for decoding invasive brain signals, leveraging the comprehensive inferring capability of GPT-2, OPT, and LLaMA2[132–135]. As an initial attempt, this model achieved comparable performance to the cascade model, demonstrating a promising avenue.

Since cascade inner speech recognition and LLM-augmented approaches have achieved efficient and accurate performance, breakthroughs in this field have been accelerated. However, invasive data collection introduces medical risks, which makes it difficult to promote among patient groups. Additionally, it has been verified that brain patterns vary over time and in subjects[124]. We believe that inner speech recognition is the most promising solution for communication-aided BCI, but there's still a distance from a high-security, high-quality, and low-latency strategy.

Brain-to-speech

Another challenging approach is to directly synthesize speech waves from brain signals. Neuroprostheses using speech synthesis employ deep learning models to convert brain activity records sequentially into synthesizer commands[136,137], kinematic features (e.g., amplitude envelope), or acoustic features (e.g., pitches and MFCC)[138,139], thereby reconstructing the original speech signal. For instance, a study implemented the DenseNet regression model[140] to map ECoG features to the spectrogram[141]. Articulatory-based speech synthesizers generate intelligible speech signals from primary speech articulators using articulator representations[142] or electromagnetic articulography (EMA)[143,144]. EMA measures the position of mouth articulators: the tongue, lips, velum, jaw, and larynx. This method is based on the finding that during speech production, activity in the brain's sensorimotor cortex closely aligns with articulatory characteristics[145]. Additionally, various features related to synthesized speech, such as vocal pitch[146], articulatory kinematic trajectories[147,148], and speech energy[149], can be identified based on brain activity. Speech synthesis without relying on deep learning, such as unit selection, has also been extensively studied[150]. Besides synthesizing intelligible waves, researchers are also focusing on generating spontaneous speech, including speech with accurate lexical tones. A feasible approach involved constructing specific neural networks to separately decode the neural activities of tones and syllables, then using the combined decoded features to synthesize tonal speech[151]. The synthesis delay is an important factor in realizing speech-centric BCI systems. In ref. 152, an online speech synthesis was proposed with a neural voice activity detection to generate speech-sensitive neural pieces, a bidirectional decoding model to estimate acoustic features and a vocoder to obtain the corresponding speech wave.

In addition to speech synthesis, information related to other modalities can be obtained through invasive brain signals. The most intuitive attempt is to leverage articulator gestures for facial movement synthesis[125], which can be achieved by decoding orofacial representations in the speech motor cortex[142,147]. It has been verified that facial movement could be generated using an avatar-animation system, and the progress on talking head generation inspired restoring the patient's own face[153]. In theory, multiple elements of building a digital human can be obtained from invasive brain activity, including textual sentences, speech waves, facial movements, as well as body movements not related to language[154,155]. This may be the future development direction of communication-aided BCIs, which can restore the patient's dignity to the greatest extent possible and communicate with the outside world through a virtual image that is the same as a normal person's (Fig. 4d). For patients who are confined to bed and unable to move, especially ALS patients, this can greatly improve their quality of life.

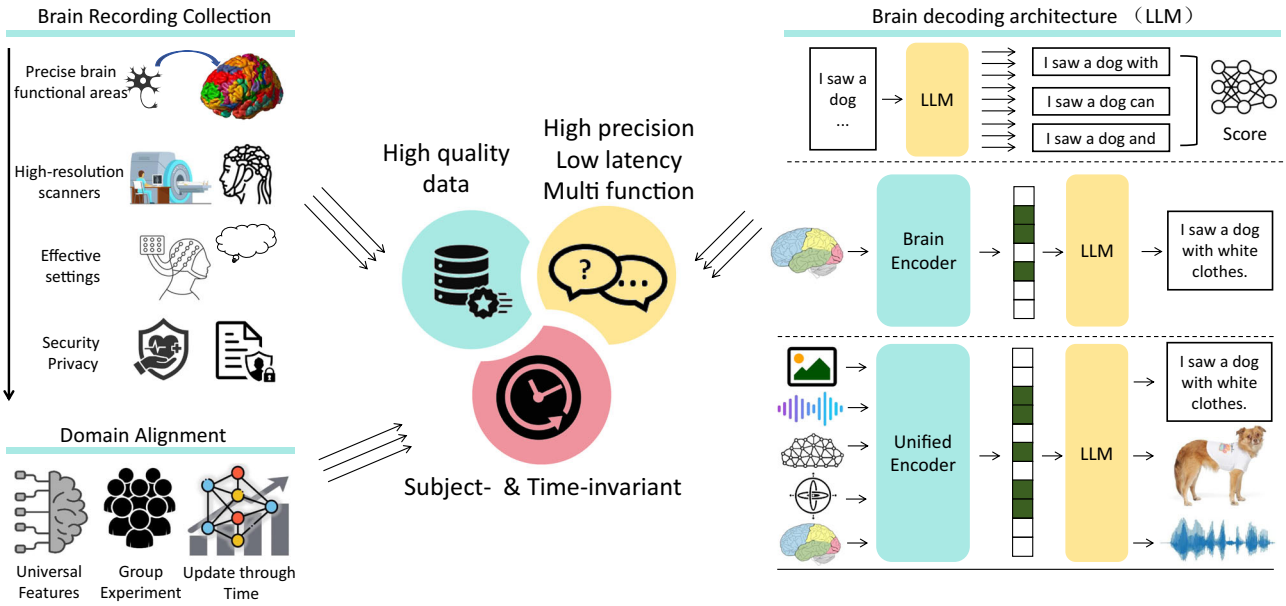

Fig. 5 | Characteristics of an ideal BCI system for communication its achieving solutions. The BCI system requires high-quality brain recordings and addresses the problem of individual and time differences through strategies such as domain alignment. Additionally, the reform of network structure, especially the application of LLM, provides ideas for high-precision, low-latency, and multi-functional interactions. Some icons are from Dreamstime, Vecteezy and Iconfinder.

Progress, challenges, and future

Progress to idea BCI and current challenges

Language is the primary means of human communication, and decoding linguistic information from brain activity is crucial for the development of future BCIs. We summarize the gap between neural decoding systems and ideal BCIs from the following aspects, addressing both progress and challenges (Fig. 5):

- Neural signal collection: Even though the invasive recording outperforms with its superior qualities of brain imaging, the necessary surgery and unbearable medical risks prevent its spread in patients. The collection of high-quality non-invasive data is a prerequisite for word-level fine-grained sequence decoding. Limited by the current level of neural recording collection and the noise resistance of the network architecture, it has not yet been possible to achieve an acceptable level of open-vocabulary continuous decoding with non-invasive data. A feasible alternative, brain recording translation, is to focus on the semantic consistency of the decoded text, not requiring absolute consistency of the corresponding text or high restoration of the speech, but achieving a considerable level of semantic accuracy.
- Subject- and time-invariant: For the same neural stimulation, brain activity varies across subjects and acquisition time[124]. On a small vocabulary, a 3-month clinical trial in an ALS patient showed that speech commands could be accurately detected and decoded without recalibrating or retraining the model[156], and another study showed that the developed decoding system worked successfully in two human patients[96]. However, experiments on a wider population with an open vocabulary have not yet been carried out, and the generalization of models trained on a single data source still needs to be discussed.
- High precision, low latency and multi-function: the upper bound of speech-related BCIs can be viewed as a corresponding ASR system, considering the unified backend of the neural networks and the superposition of noise from the brain response to speech. The development of more sophisticated and responsive BCIs could revolutionize how we interact with machines, offering applications in medical rehabilitation, verbal communication and even entertainment. Furthermore, integrating multiple modalities-such as visual and auditory inputs-can enhance the functionality of BCIs, enabling more comprehensive communication solutions. Current experiments on multiple tasks have shown that text, speech, and visual reconstruction of neural signals have achieved the ability to restore semantic features[80,152,157–159], which indicates a potential solution by modality fusion and system integration. However, it must be emphasized that the detailed restoration performance of the above experiments needs great improvement. There are also discussions to be addressed on striking a balance to avoid error accumulation and promoting the main modalities with auxiliary information.
- Privacy preservation: ethical debates regarding collecting and decoding neural signals from the human brain remain an important limiting factor[160,161]. Invasive data collection is only carried out on a small population due to its surgical risk and usually requires ensuring the necessity of craniotomy for medical treatment. Non-invasive data has much more promotional potential but also carries significant risks of privacy leakage—a more comprehensive data usage convention may be necessary, including the standardization of the collection process, the requirements of experiment subjects, the decoding granularity and vocabulary, and necessary solutions to avoid violation of personal privacy. To have widespread potential, BCI systems must be privacy-preserving and ethically sound. The system should be clearly aware of what information can be accessed, displayed, or made public, and choose to conceal or ignore when it comes to personal privacy or inner thoughts. A responsible stance within society that firmly opposes the misuse of neurodata would serve as the ethical guide for the future advancement of neurotechnology.

Future directions

Even though we are still a long way from efficient and harmless BCIs, some directions have shown bright prospects. A unified brain representation could be the next big breakthrough in neural decoding, which has made a great impact on other modalities. There is a consensus on the individual and temporal variability of neural signals. Performing individualized data collection and model training is not a feasible solution considering the corresponding recording duration and computational resources. Instead, the implementation of a unified neural representation is a promising solution by

fine-tuning with limited user-specific recordings to form personalized decoding systems. This requires expanding the previous experiment to a group population and collecting much more dynamic neural recordings, with self-supervised learning providing a strong relationship with semantic information[130,162,163].

While invasive neural decoding has demonstrated superior performance, the main limitation of non-invasive signals is their significant noise level. It is worth practicing to perform data augmentation and denoising on neural signals, and existing solutions are mainly based on generative models, such as GAN and diffusion[164,165]. Research on the robustness of model architectures is still in its early stages, especially since LLM has recently demonstrated extraordinary reasoning performance. Considering the significant mismatch between the tokenized text and neural space, robust neural networks with a stable training strategy are possible to boost the generation performance.

Large language models preserve powerful understanding, reasoning, and generation capabilities, and previous studies have shown that LLMs trained on vast amounts of textual corpus enable the ability to be aligned with other modalities through smaller-scale fine-tuning, thereby generating content with strong semantic consistency and vivid details. The same phenomenon also applies to neural data, where the most significant trend in linguistic neural decoding has been implementing a textual LLM as the backend decoder for text generation[82,85,135]. As shown in Fig. 5, in the initial attempts, the LLMs were adopted to generate hypothesis candidates with a separate module score for each potential sentence[89,166]. A more promising approach treats the LLM as the inferring core to generate correlated textual information[135], and gradually evolves into a unified decoding system with multi-modality inputs and user-specified output[167]. We believe that the update and iteration of LLMs will promote qualitative changes in neural decoding, thereby achieving application levels in the near future.

Parallel to model improvement, in neuroscience, a pressing issue is the precise collection of neural recordings related to language processing, including acoustic and phonological aspects. This requires identifying specific neuronal populations and brain areas involved in language functions[11,12,168]. High-resolution scanners, wearable neurotechnology devices and advanced equipment are also necessary[169], and more reasonable experimental settings need to be explored. An important aspect is to unify the data collection framework to explore the possibility of developing a massive neural corpus from multiple resources, which means diverse stimuli and subject conditions. The neural recordings from a single experiment trial are typically suitable for small network training, while pre-training and fine-tuning on a larger scale are likely to process data spanning several orders of magnitude.

As for privacy preservation and technology regulation, strict management and supervision need to cover the entire process of data collection, model training and deployment application[161]. The premise is to form clear data usage standards, minimize the dimensions and duration of neural recordings while ensuring decoding performance, and strictly discard potential privacy-aware instances. The dissemination and use of neural data need to ensure that the goal of the corresponding experiment is for human welfare, and data encryption, differential privacy and federated learning are protection measures that need to be considered. As the modality and experimental population of neural decoding expand, we strongly call for the formation of a unified ethical perspective, such as human rights guidelines, which requires neural computing companies and related major researchers to assume corresponding scientific responsibilities.

The interaction between the brain and the environment is bidirectional. This article mainly explains the direction of neural decoding, that is, from the neural recordings to linguistic stimuli or intended messages. Stimuli encoding, by performing tiny simulated currents on the cortex to generate evoked brain activity, might be a solution for sensory loss, including blindness and deafness. Guiding brain cognition through artificial stimulation, commonly known as deep brain stimulation (DBS), is a promising direction for disease treatment and has emerged as an effective treatment for neurological conditions such as Alzheimer's[170] and Parkinson's disease[171]. Another question is whether BCIs can improve the efficiency of information transmission. Information interaction via voice or visual text is limited by the rate of speech flow and vision refresh, while the brain's information reception rate may far exceed both thresholds. When machine operating efficiency reaches a certain level, a large-scale industrial revolution may come from a leap in information transmission efficiency. In general, brain linguistic decoding is a cross-disciplinary collaboration. We expect a further revolution from strengthened cooperation between biology, engineering, and machine intelligence to promote innovation and accelerate the development of brain signal recording technology and its applications.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Received: 3 September 2024; Accepted: 10 July 2025;
Published online: 24 September 2025

References

1. Abnar, S., Beinborn, L., Choenni, R. & Zuidema, W. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp 191–203 (2019).
2. Frisby, S. L., Halai, A. D., Cox, C. R., Ralph, M. A. L. & Rogers, T. T. Decoding semantic representations in mind and brain. *Trends Cogn. Sci.* **27**, 258–281 (2023).
3. Silva, A. B., Littlejohn, K. T., Liu, J. R., Moses, D. A. & Chang, E. F. The speech neuroprosthesis. *Nat. Rev. Neurosci.* **25**, 473–492 (2024).
4. Tuckute, G., Kanwisher, N. & Fedorenko, E. Language in brains, minds, and machines. *Annu. Rev. Neurosci.* **47**, 271–301 (2024).
5. Ball, T., Kern, M., Mutschler, I., Aertsen, A. & Schulze-Bonhage, A. Signal quality of simultaneously recorded invasive and non-invasive eeg. *Neuroimage* **46**, 708–716 (2009).
6. Ahissar, E. et al. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc. Natl Acad. Sci. USA* **98**, 13367–13372 (2001).
7. Brodbeck, C., Hong, L. E. & Simon, J. Z. Rapid transformation from auditory to linguistic representations of continuous speech. *Curr. Biol.* **28**, 3976–3983 (2018).
8. Koskinen, M., Kurimo, M., Gross, J., Hyvärinen, A. & Hari, R. Brain activity reflects the predictability of word sequences in listened continuous speech. *NeuroImage* **219**, 116936 (2020).
9. Donhauser, P. W. & Baillet, S. Two distinct neural timescales for predictive speech processing. *Neuron* **105**, 385–393 (2020).
10. Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T. & Brodbeck, C. Neural markers of speech comprehension: measuring EEG tracking of linguistic speech representations, controlling the speech acoustics. *J. Neurosci.* **41**, 10316–10329 (2021).
11. Leonard, M. K. et al. Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature* **626**, 593–602 (2024).
12. Khanna, A. R. et al. Single-neuronal elements of speech production in humans. *Nature* **626**, 603–610 (2024).
13. Clark, A. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
14. Schrimpf, M. et al. The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl Acad. Sci. USA* **118**, e2105646118 (2021).
15. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P. & De Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl Acad. Sci. USA* **119**, e2201968119 (2022).
16. Caucheteux, C., Gramfort, A. & King, J.-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat. Hum. Behav.* **7**, 430–441 (2023).

17. Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J. & Lalor, E. C. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* **28**, 803–809 (2018).
18. Caucheteux, C., Gramfort, A. & King, J.-R. Disentangling syntax and semantics in the brain with deep networks. In: *International conference on machine learning*, pp 1336–1348 (PMLR, 2021).
19. Toneva, M., Mitchell, T. M. & Wehbe, L. Combining computational controls with natural text reveals aspects of meaning composition. *Nat. Comput. Sci.* **2**, 745–757 (2022).
20. Antonello, R. & Huth, A. Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiol. Lang.* **5**, 64–79 (2024).
21. Alkhamissi, B., Tuckute, G., Bosselut, A., Schrimpf, M.: The llm language network: A neuroscientific approach for identifying causally task-relevant units. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 10887–10911 (2025).
22. Whittington, J. C., Warren, J. & Behrens, T. E. Relating transformers to models and neural representations of the hippocampal formation. In: *International Conference on Learning Representations* (2021).
23. Liu, X. et al. Coupling artificial neurons in Bert and biological neurons in the human brain. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 8888–8896 (2023).
24. Antonello, R., Vaidya, A. & Huth, A. Scaling laws for language encoding models in fMRI. *Adv. Neural Inf. Process. Syst*. **36**, 21895–21907 (2024).
25. Lin, H. et al. Selecting large language model to fine-tune via rectified scaling law. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 30080–30107 (2024).
26. Ren, Y., Jin, R., Zhang, T. & Xiong, D. Do large language models mirror cognitive language processing? In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 2988–3001 (2025).
27. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp 311–318 (2002).
28. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: evaluating text generation with bert. In: *International Conference on Learning Representations* (2019).
29. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186 (2019).
30. Taal, C. H., Hendriks, R. C., Heusdens, R. & Jensen, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp 4214–4217 (IEEE, 2010).
31. Kubichek, R. Mel-cepstral distance measure for objective speech quality assessment. In: *Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, vol. 1, 125–128 (IEEE, 1993).
32. Just, M. A., Cherkassky, V. L., Aryal, S. & Mitchell, T. M. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS ONE* **5**, e8622 (2010).
33. Sudre, G. et al. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage* **62**, 451–463 (2012).
34. Anderson, A. J., Kiela, D., Clark, S. & Poesio, M. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Trans. Assoc. Comput. Linguist.* **5**, 17–30 (2017).
35. Abnar, S., Ahmed, R., Mijnheer, M. & Zuidema, W. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In: *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pp 57–66 (2018).
36. Pereira, F. et al. Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **9**, 963 (2018).
37. Affolter, N., Egressy, B., Pascual, D. & Wattenhofer, R. Brain2word: decoding brain activity for language generation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2009.04765 (2020).
38. Zou, S., Wang, S., Zhang, J. & Zong, C. Towards brain-to-text generation: neural decoding with pre-trained encoder-decoder models. In: *NeurIPS 2021 AI for Science Workshop* (2021).
39. Anderson, A. J. et al. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cereb. Cortex* **27**, 4379–4395 (2017).
40. Wang, J., Cherkassky, V. L. & Just, M. A. Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states. *Hum. Brain Mapp.* **38**, 4865–4881 (2017).
41. Anderson, A. J. et al. An integrated neural decoder of linguistic and experiential meaning. *J. Neurosci.* **39**, 8969–8987 (2019).
42. Sun, J., Wang, S., Zhang, J. & Zong, C. Towards sentence-level brain decoding with distributed representations. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp 7047–7054 (2019).
43. Gauthier, J. & Levy, R. Linking artificial and human neural representations of language. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp 529–539 (2019).
44. Sun, J., Wang, S., Zhang, J. & Zong, C. Neural encoding and decoding with distributed sentence representations. *IEEE Trans. Neural Netw. Learn. Syst*. **32**, 589–603 (2020).
45. Wehbe, L. et al. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS ONE* **9**, e112575 (2014).
46. Jat, S., Tang, H., Talukdar, P. & Mitchell, T. Relating simple sentence representations in deep neural networks and the brain. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5137-5154 (2019).
47. Li, Y. et al. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nat. Neurosci.* **26**, 2213–2225 (2023).
48. Liu, Y. & Ayaz, H. Speech recognition via fNIRS-based brain signals. *Front. Neurosci.* **12**, 395799 (2018).
49. 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 (2019).
50. Radford, A. et al. Learning transferable visual models from natural language supervision. In: *International conference on machine learning*, pp 8748–8763 (PMLR, 2021).
51. Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O. & King, J.-R. Decoding speech perception from non-invasive brain recordings. *Nat. Mach. Intell.* **5**, 1097–1107 (2023).
52. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **33**, 12449–12460 (2020).
53. Bollens, L., Francart, T. & Van Hamme, H. Learning subject-invariant representations from speech-evoked EEG using variational autoencoders. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 1256–1260 (IEEE, 2022).
54. Aiken, S. J. & Picton, T. W. Human cortical responses to the speech envelope. *Ear Hear.* **29**, 139–157 (2008).

55. Ding, N. & Simon, J. Z. Cortical entrainment to continuous speech: functional roles and interpretations. *Front. Hum. Neurosci.* **8**, 311 (2014).
56. Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z. & Francart, T. Speech intelligibility predicted from neural entrainment of the speech envelope. *J. Assoc. Res. Otolaryngol.* **19**, 181–191 (2018).
57. Crosse, M. J., Di Liberto, G. M., Bednar, A. & Lalor, E. C. The multivariate temporal response function (mtrf) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* **10**, 604 (2016).
58. Accou, B. et al. Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network. In: *2020 28th European Signal Processing Conference (EUSIPCO)*, pp 1175–1179 (IEEE, 2021).
59. De Clercq, P., Vanthornhout, J., Vandermosten, M. & Francart, T. Beyond linear neural envelope tracking: a mutual information approach. *J. Neural Eng.* **20**, 026007 (2023).
60. Thornton, M., Mandic, D. & Reichenbach, T. Robust decoding of the speech envelope from EEG recordings through deep neural networks. *J. Neural Eng.* **19**, 046007 (2022).
61. Accou, B., Vanthornhout, J., Hamme, H. V. & Francart, T. Decoding of the speech envelope from EEG using the VLAAI deep neural network. *Sci. Rep.* **13**, 812 (2023).
62. de Taillez, T., Kollmeier, B. & Meyer, B. T. Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. *Eur. J. Neurosci.* **51**, 1234–1241 (2020).
63. Xu, Z. et al. Decoding selective auditory attention with EEG using a transformer model. *Methods* **204**, 410–417 (2022).
64. Krishna, G., Han, Y., Tran, C., Carnahan, M. & Tewfik, A. H. State-of-the-art speech recognition using eeg and towards decoding of speech spectrum from eeg.Preprint at *arXiv* https://doi.org/10.48550/arXiv.1908.05743 (2019).
65. Petrosyan, A., Voskoboynikov, A. & Ossadtchi, A. Compact and interpretable architecture for speech decoding from stereotactic EEG. In: *2021 Third International Conference Neurotechnologies and Neurointerfaces (CNN)*, pp 79–82 (IEEE, 2021).
66. Krishna, G., Tran, C., Carnahan, M. & Tewfik, A. H. Advancing speech synthesis using EEG. In: *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp 199–204 (IEEE, 2021).
67. Pasley, B. N. et al. Reconstructing speech from human auditory cortex. *PLoS Biol.* **10**, e1001251 (2012).
68. Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D. & Mesgarani, N. Towards reconstructing intelligible speech from the human auditory cortex. *Sci. Rep.* **9**, 874 (2019).
69. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein generative adversarial networks. In: *International Conference on Machine Learning*, pp 214–223 (PMLR, 2017).
70. Wang, R. et al. Stimulus speech decoding from human cortex with generative adversarial network transfer learning. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp 390–394 (IEEE, 2020).
71. Yi, Z., Zhang, H., Tan, P. & Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp 2849–2857 (2017).
72. Guo, Y., Liu, T., Zhang, X., Wang, A. & Wang, W. End-to-end translation of human neural activity to speech with a dual–dual generative adversarial network. *Knowl. Based Syst.* **277**, 110837 (2023).
73. Senda, J. et al. Auditory stimulus reconstruction from ECoG with DNN and self-attention modules. *Biomed. Signal Process. Control* **89**, 105761 (2024).
74. 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 (2020).
75. Wang, Z. & Ji, H. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp 5350–5358 (2022).
76. Lewis, M. et al. Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880 (2020).
77. Van Den Oord, A. et al. Neural discrete representation learning. *Adv. Neural Inf. Process. Syst*. **30**, 6309–6318 (2017).
78. Duan, Y., Chau, C., Wang, Z., Wang, Y.-K. & Lin, C.-t. Dewave: discrete encoding of EEG waves for EEG to text translation. *Adv. Neural Inf. Process. Syst*. **36**, 9907–9918 (2024).
79. Jo, H. et al. Are EEG-to-text models working? Preprint at *arXiv* https://doi.org/10.48550/arXiv.2405.06459 (2024).
80. Yang, Y., Duan, Y., Zhang, Q., Xu, R. & Xiong, H. Neuspeech: Decode neural signal as speech. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2403.01748 (2024).
81. Yang, Y. et al. Mad: Multi-alignment meg-to-text decoding. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2406.01512 (2024).
82. Chen, X., Du, C., Liu, C., Wang, Y. & He, H. Open-vocabulary auditory neural decoding using fMRI-prompted llm. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2405.07840 (2024).
83. Ye, Z. et al. Generative language reconstruction from brain recordings (2024).
84. Yin, C., Ye, Z. & Li, P. Language reconstruction with brain predictive coding from fMRI data. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2405.11597 (2024).
85. Feng, X., Feng, X., Qin, B. & Liu, T. Aligning semantic in brain and language: a curriculum contrastive method for electroencephalography-to-text generation. *IEEE Trans. Neural Syst. Rehabil. Eng*. **31**, 3874–3883 (2023).
86. Xi, N. et al. Unicorn: unified cognitive signal reconstruction bridging cognitive signals and human language. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp 13277–13291 (2023).
87. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1301.3781 (2013).
88. Tang, J., LeBel, A., Jain, S. & Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nat. Neurosci.* **26**, 858–866 (2023).
89. Zhao, X. et al. Mapguide: a simple yet effective method to reconstruct continuous language from brain activities. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 3822–3832 (2024).
90. Chen, X. et al. High-speed spelling with a noninvasive brain–computer interface. *Proc. Natl Acad. Sci. USA* **112**, E6058–E6067 (2015).
91. Metzger, S. L. et al. Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. *Nat. Commun.* **13**, 6510 (2022).
92. Leuthardt, E. C. et al. Using the electrocorticographic speech network to control a brain–computer interface in humans. *J. Neural Eng.* **8**, 036004 (2011).
93. 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 (2021).
94. Brigham, K. & Kumar, B. V. Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy. In: *2010 4th International Conference on Bioinformatics and Biomedical Engineering*, pp 1–4 (IEEE, 2010).
95. Duraivel, S. et al. High-resolution neural recordings improve the accuracy of speech decoding. *Nat. Commun.* **14**, 6938 (2023).

96. Wandelt, S. K. et al. Representation of internal speech by single neurons in human supramarginal gyrus. *Nat. Hum. Behav.* https://api.semanticscholar.org/CorpusID:269759448 (2024).
97. Chang, E. F. et al. Categorical speech representation in human superior temporal gyrus. *Nat. Neurosci.* **13**, 1428–1432 (2010).
98. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).
99. Suppes, P., Lu, Z.-L. & Han, B. Brain wave recognition of words. *Proc. Natl Acad. Sci. USA* **94**, 14965–14969 (1997).
100. Suppes, P., Han, B. & Lu, Z.-L. Brain-wave recognition of sentences. *Proc. Natl Acad. Sci. USA* **95**, 15861–15866 (1998).
101. D'Zmura, M., Deng, S., Lappas, T., Thorpe, S. & Srinivasan, R. Toward EEG sensing of imagined speech. In *Human-Computer Interaction. New Trends: 13th International Conference, HCI International 2009, San Diego, CA, USA, July 19-24, 2009, Proceedings, Part I 13*, pp 40–48 (Springer, 2009).
102. Tankus, A., Fried, I. & Shoham, S. Structured neuronal encoding and decoding of human speech features. *Nat. Commun.* **3**, 1015 (2012).
103. DaSalla, C. S., Kambara, H., Sato, M. & Koike, Y. Single-trial classification of vowel speech imagery using common spatial patterns. *Neural Netw.* **22**, 1334–1339 (2009).
104. Wang, L., Zhang, X., Zhong, X. & Zhang, Y. Analysis and classification of speech imagery EEG for BCI. *Biomed. signal Process. control* **8**, 901–908 (2013).
105. Stavisky, S. D. et al. Decoding speech from intracortical multielectrode arrays in dorsal "arm/hand areas" of human motor cortex. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp 93–97 (IEEE, 2018).
106. Pei, X., Barbour, D. L., Leuthardt, E. C. & Schalk, G. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *J. Neural Eng.* **8**, 046028 (2011).
107. Deng, S., Srinivasan, R., Lappas, T. & D'Zmura, M. Eeg classification of imagined syllable rhythm using Hilbert spectrum methods. *J. Neural Eng.* **7**, 046006 (2010).
108. Kim, J., Lee, S.-K. & Lee, B. Eeg classification in a single-trial basis for vowel speech perception using multivariate empirical mode decomposition. *J. Neural Eng.* **11**, 036010 (2014).
109. Moses, D. A., Mesgarani, N., Leonard, M. K. & Chang, E. F. Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity. *J. Neural Eng.* **13**, 056004 (2016).
110. Brumberg, J. S., Wright, E. J., Guenther, F. H. & Kennedy, P. R. Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech motor cortex. *Front. Neurosci.* **5**, 7880 (2011).
111. Kellis, S. et al. Decoding spoken words using local field potentials recorded from the cortical surface. *J. Neural Eng.* **7**, 056007 (2010).
112. Mugler, E. M. et al. Direct classification of all American English phonemes using signals from functional speech motor cortex. *J. Neural Eng.* **11**, 035015 (2014).
113. Clayton, J., Wellington, S., Valentini-Botinhao, C. & Watts, O. Decoding imagined, heard, and spoken speech: Classification and regression of eeg using a 14-channel dry-contact mobile headset. In: *INTERSPEECH*, pp 4886–4890 (2020).
114. Mohanchandra, K. & Saha, S. A communication paradigm using subvocalized speech: translating brain signals into speech. *Augmented Hum. Res.* **1**, 3 (2016).
115. Martin, S. et al. Word pair classification during imagined speech using direct brain recordings. *Sci. Rep.* **6**, 25803 (2016).
116. Nguyen, C. H., Karavas, G. K. & Artemiadis, P. Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. *J. Neural Eng.* **15**, 016002 (2017).
117. González-Castañeda, E. F., Torres-García, A. A., Reyes-García, C. A. & Villaseñor-Pineda, L. Sonification and textification: Proposing methods for classifying unspoken words from EEG signals. *Biomed. Signal Process. Control* **37**, 82–91 (2017).
118. Salama, M., ElSherif, L., Lashin, H. & Gamal, T. Recognition of unspoken words using electrode electroencephalograhic signals. In: *The Sixth International Conference on Advanced Cognitive Technologies and Applications*, pp 51–5 (2014).
119. Zhao, S. & Rudzicz, F. Classifying phonological categories in imagined and articulated speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 992–996 (IEEE, 2015).
120. Saha, P. & Fels, S. Hierarchical deep feature learning for decoding imagined speech from EEG. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp 10019–10020 (2019).
121. Dash, D., Ferrari, P. & Wang, J. Decoding imagined and spoken phrases from non-invasive neural (meg) signals. *Front. Neurosci.* **14**, 490970 (2020).
122. Herff, C. et al. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Front. Neurosci.* **8**, 141498 (2015).
123. Moses, D. A. et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *N. Engl. J. Med.* **385**, 217–227 (2021).
124. Willett, F. R. et al. A high-performance speech neuroprosthesis. *Nature* **620**, 1031–1036 (2023).
125. Metzger, S. L. et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* **620**, 1037–1046 (2023).
126. Munteanu, C., Penn, G., Baecker, R., Toms, E. & James, D. Measuring the acceptable word error rate of machine-generated webcast transcripts. In: *Ninth International Conference on Spoken Language Processing* (Citeseer, 2006).
127. Sun, P., Anumanchipalli, G. K. & Chang, E. F. Brain2char: a deep architecture for decoding text from brain recordings. *J. Neural Eng.* **17**, 066015 (2020).
128. Feng, C. et al. A high-performance brain-to-sentence decoder for logosyllabic language (2023).
129. Silva, A. B. et al. A bilingual speech neuroprosthesis driven by cortical articulatory representations shared between languages. *Nat. Biomed. Eng.* **8**, 977–991 (2024).
130. Yuan, B. A. & Makin, J. G. Improving speech decoding from ECOG with self-supervised pretraining. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2405.18639 (2024).
131. Schneider, S., Baevski, A., Collobert, R. & Auli, M. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. Interspeech 2019*, pp. 3465–3469 (2019).
132. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
133. Zhang, S. et al. Opt: Open pre-trained transformer language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2205.01068 (2022).
134. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2307.09288 (2023).
135. Feng, S., Liu, H., Wang, Y. & Wang, Y. Towards an end-to-end framework for invasive brain signal decoding with large language models. In: *Interspeech 2024*, pp 1495–1499 (2024).
136. Guenther, F. H. et al. A wireless brain-machine interface for real-time speech synthesis. *PloS ONE* **4**, e8218 (2009).
137. Chen, X. et al. A neural speech decoding framework leveraging deep learning and speech synthesis. *Nat. Mach. Intell.* **6**, 467–480 (2024).

138. Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498 (2019).
139. Krishna, G., Tran, C., Han, Y., Carnahan, M. & Tewfik, A. H. Speech synthesis using EEG. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 1235–1238 (IEEE, 2020).
140. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 4700–4708 (2017).
141. Angrick, M. et al. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *J. Neural Eng.* **16**, 036019 (2019).
142. Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327–332 (2013).
143. Bocquelet, F., Hueber, T., Girin, L., Badin, P. & Yvert, B. Robust articulatory speech synthesis using deep neural networks for BCI applications. In: *Interspeech 2014-15th Annual Conference of the International Speech Communication Association* (2014).
144. Bocquelet, F., Hueber, T., Girin, L., Savariaux, C. & Yvert, B. Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLoS Comput. Biol.* **12**, e1005119 (2016).
145. Cheung, C., Hamilton, L. S., Johnson, K. & Chang, E. F. The auditory representation of speech sounds in human motor cortex. *elife* **5**, e12577 (2016).
146. 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 (2018).
147. Chartier, J., Anumanchipalli, G. K., Johnson, K. & Chang, E. F. Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron* **98**, 1042–1054 (2018).
148. Mugler, E. M. et al. Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri. *J. Neurosci.* **38**, 9803–9813 (2018).
149. Angrick, M. et al. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Commun. Biol.* **4**, 1055 (2021).
150. Herff, C. et al. Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices. *Front. Neurosci.* **13**, 469935 (2019).
151. Liu, Y. et al. Decoding and synthesizing tonal language speech from brain activity. *Sci. Adv.* **9**, eadh0478 (2023).
152. Angrick, M. et al. Online speech synthesis using a chronically implanted brain–computer interface in an individual with als. *Sci. Rep.* **14**, 9617 (2024).
153. Ling, J., Wang, Y., Xue, H., Xie, R. & Song, L. Posetalk: text-and-audio-based pose control and motion refinement for one-shot talking head generation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2409.02657 (2024).
154. Song, H. et al. Continuous neural control of a bionic limb restores biomimetic gait after amputation. *Nat. Med.* **30**, 2010–2019 (2024).
155. Wang, J. et al. Neural correlate and movement decoding of simultaneous-and-sequential bimanual movements using eeg signals. *IEEE Trans. Neural Syst. Rehabil. Eng.* **32**, 2087–2095 (2024).
156. Luo, S. et al. Stable decoding from a speech BCI enables control for an individual with ALS without recalibration for 3 months. *Adv. Sci.* https://api.semanticscholar.org/CorpusID:264448311 (2023).
157. Wang, S., Liu, S., Tan, Z. & Wang, X. Mindbridge: A cross-subject brain decoding framework. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 11333–11342 (2024).
158. Quan, R., Wang, W., Tian, Z., Ma, F. & Yang, Y. Psychometry: An omnifit model for image reconstruction from human brain activity. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 233–243 (2024).
159. Liu, X. et al. Eeg2video: Towards decoding dynamic visual perception from EEG signals. In: *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*.
160. van Stuijvenberg, O. C., Broekman, M. L., Wolff, S. E., Bredenoord, A. L. & Jongsma, K. R. Developer perspectives on the ethics of AI-driven neural implants: a qualitative study. *Sci. Rep.* **14**, 7880 (2024).
161. Yuste, R. Advocating for neurodata privacy and neurotechnology regulation. *Nat. Protoc.* **18**, 2869–2875 (2023).
162. Wang, C. et al. Brainbert: Self-supervised representation learning for intracranial recordings. In: *The Eleventh International Conference on Learning Representations*.
163. Zheng, H. et al. Du-IN: Discrete units-guided mask modeling for decoding speech from intracranial neural signals. In: *The Thirty-Eighth Annual Conference on Neural Information Processing Systems* https://openreview.net/forum?id=uyLtEFnpQP (2024).
164. Dong, Y. et al. An approach for EEG denoising based on Wasserstein generative adversarial network. *IEEE Trans. Neural Syst. Rehabil. Eng.* **31**, 3524–3534 (2023).
165. Huang, X., Li, C., Liu, A., Qian, R. & Chen, X. Eegdfus: a conditional diffusion model for fine-grained EEG denoising. *IEEE J. Biomed. Health Inform.* **29**, 2557–2569 (2024).
166. Antonello, R., Sarma, N., Tang, J., Song, J. & Huth, A. How many bytes can you take out of brain-to-text decoding? Preprint at *arXiv* https://doi.org/10.48550/arXiv.2405.14055 (2024).
167. Han, J. et al. Onellm: One framework to align all modalities with language. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 26584–26595 (2024).
168. Tankus, A., Rosenberg, N., Ben-Hamo, O., Stern, E. & Strauss, I. Machine learning decoding of single neurons in the thalamus for speech brain-machine interfaces. *J. Neural Eng.* **21**, 036009 (2024).
169. Feinberg, D. A. et al. Next-generation MRI scanner designed for ultra-high-resolution human brain imaging at 7 Tesla. *Nat. Methods* **20**, 2048–2057 (2023).
170. Tatulian, S. A. Challenges and hopes for alzheimer's disease. *Drug Discov. Today* https://api.semanticscholar.org/CorpusID:246553676 (2022).
171. Bucur, M. & Papagno, C. Deep brain stimulation in parkinson disease: a meta-analysis of the long-term neuropsychological outcomes. *Neuropsychol. Rev.* https://api.semanticscholar.org/CorpusID:247615265 (2022).

Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2022ZD0162101).

Author contributions

Yu Wang and Heyang Liu contribute equally to this work, and Yanfeng Wang is the corresponding author. Specifically, Yu Wang, Heyang Liu, Yuhao Wang, Chuan Xuan, Yixuan Hou, Sheng Feng, Hongcheng Liu, Yusheng Liao, and Yanfeng Wang all participate in the discussions and summarize, and make contributions to the review. In writing, Yu Wang and Heyang Liu draft the work, and Yuhao Wang, Chuan Xuan, Yixuan Hou, Sheng Feng, Hongcheng Liu, Yusheng Liao, and Yanfeng Wang review it critically for important intellectual content. All authors approve of the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-08511-z.

Correspondence and requests for materials should be addressed to Yanfeng Wang.

Peer review information *Communications Biology* thanks Ziyi Ye and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Jasmine Pan. A peer review file is available.