Studying Language Processing in the Human Brain with Speech and Language Models

Chao Zhang
Tsinghua University
University College London
cz277@tsinghua.edu.cn

Andrew Thwaites
University College London
University of Cambridge
acgt2@cam.ac.uk

Cai Wingfield
University of Cambridge
cw417@cam.ac.uk

Abstract

Speech and language computational models have been instrumental in advancing Artificial Intelligence in recent years. However, it remains an open question whether the human brain is employing similar approaches to these models. This tutorial aims to provide an accessible introduction to the extensive research on this topic, specifically focusing on studies that seek to establish quantitative correlations between neuroimaging data from human subjects and the output of language models or automatic speech recognition systems. The tutorial covers various aspects of this research, including a brief overview of brain-computer interfaces and neuroscience, common techniques for data processing and pattern analysis, and representative research examples. Finally, the tutorial addresses the main limitations and technical challenges encountered in this field, as well as the relationship between brain mechanism research and brain-inspired artificial intelligence.

1 Motivation and Objectives

The ability to engage in complex verbal communication is a defining capacity that distinguishes humans from other animals, and speech is a primary medium through which this is achieved. The brain recognises individual words and understands the semantics of sentences through the progressive processing of speech signals, a process known as language comprehension. Understanding the mechanisms underlying this process is an important research topic in computational cognitive neuroscience (Chomsky, 1965; Goldberg, 2006). Computational cognitive neuroscience is a field that connects the brain's complex biological neural system with computational models that can describe and simulate its cognitive functions. Technological developments over the last three decades have aided this field. First, modern neuroimaging technology allows us to collect brain activity patterns ("brain responses"), from human subjects while they receive natural language input. Second, human-level automatic speech recognition (ASR) systems and their accompanying language models can provide human-level model responses to any arbitrary inputs of speech or text. By employing spatiotemporal pattern analysis techniques, we can quantitatively correlate the brain responses and model responses given the same input, providing insights into the brain's spoken language recognition and comprehension mechanisms (Wingfield et al., 2017; Wingfield et al., 2022; Tuckute et al., 2022; Vaidya et al., 2022; Vaidya et al., 2022; Caucheteux et al., 2023). This research method not only provides new means to uncover the secrets of the brain's language comprehension mechanism but also provides a basis for measuring the brain-likeness of models (Wingfield et al., 2022), which can be used both for investigating the intelligence characteristics of the brain and in interpretable artificial intelligence (AI) research. Moreover, the development of connectionism/deep learning AI technology, including large language models (LLMs), continues to draw inspiration from neuroscience, reflecting the close correlation between the two.

This tutorial will mainly introduce the following:

- Brain-computer interface and basics of neuroscience.
- Common data processing and pattern analysis techniques.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 17-23, Harbin, China, August 3 - 5, 2023.

- Language models and automatic speech recognition technologies.
- Representative research work and its research achievements.
- Major technical challenges and prospects for future work.

2 Tutorial Overview and Structure

2.1 Brain-computer interface and basics of neuroscience (15 min)

While several technologies allow the direct or indirect measurement of human brain responses through non-invasive means, electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) are three of the most common. EEG and MEG directly measure the accompanying electric field and magnetic field of neuronal dendrites during the generation of action potentials, while fMRI indirectly measures the energy consumption of neurons performing computations by measuring blood oxygen levels. Functional near-infrared spectroscopy (fNIRS) also measures blood oxygen levels, but via local spectroscopy rather than magnetic resonance imaging, which restricts its utility.

Intracranial EEG (iEEG) is less common, due to its requirement for invasive surgery. iEEG technologies include stent EEG (where EEG electrodes are placed inside arteries or veins adjacent to the cortex) and electrocorticography (ECoG), where electrodes are placed directly into the cortex. Intracranial electrode types can range from single-pin electrodes to complex arrays comprised of thousands of miniaturised microelectrodes (Steinmetz et al., 2021).

2.2 Common data processing and pattern analysis techniques (15 min)

Before conducting model-brain comparisons, it is often necessary to clean brain activity data, especially for electrophysiological measurements from EEG and MEG. This involves steps such as averaging, filtering, and removing signal components deemed to arise from non-neuronal activity. For EEG and MEG, the data may also need to be transformed from "sensor space" (where each measurement represents sensor readings over time) to an estimation of "source space" activity (where each measurement corresponds to neural activity in a specific brain location). This transformation is known as "source localisation"(Hämäläinen and Ilmoniemi, 1994).

Comparing neural activity with computational models can be achieved in various ways. The goal is to assess the similarity between a model's behaviour and brain activity, either explicitly or implicitly. The most straightforward techniques simply compare the model's outputs directly with the activity in each brain location. This comparison can be done using similarity metrics like Pearson's Rho, Euclidean distance, or mutual information (*e.g.* (Thwaites et al., 2017; Pérez et al., 2022)).

Such direct comparison requires the model to make specific predictions about each region's activity. If the model is unable to make such a prediction, then the researcher might choose to train a wrapper model that transforms the output of the original model to match the neural data (*e.g.*, (Caucheteux et al., 2023; Oota et al., 2023)). Alternatively, they might choose to relax the constraint that the model's output must match a single location, either by fitting a wrapper model that tries to learn the relationship between multiple locations of activity and the model (using classification or linear regression (Millan et al., 2002)) or by indirectly comparing the patterns of how both the models and brain regions reacted under certain conditions, an approach known as representation similarity analysis (RSA) (Kriegeskorte et al., 2008).[1]

Relaxing the constraint of direct comparison between model and activity has numerous drawbacks, however, including a decrease in statistical power and interpretability.

2.3 Language models and automatic speech recognition technologies (20 min)

Two broad families of computational models are usually considered relevant to the study of the brain's speech recognition and understanding mechanisms: text-based language models and speech-based ASR

[1]The family of approaches that relaxes the "single region constraint" is known as "multivariate pattern analysis" (MVPA) (Haxby, 2012).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 17-23, Harbin, China, August 3 - 5, 2023.

models. Text-based language models have been found to be useful models for both human language syntax and semantics. GloVe (Pennington et al., 2014) and word2vec (Mikolov et al., 1997) are commonly used word vectors. Pre-trained language models based on the Transformer structure (Vaswani et al., 2017) have profoundly influenced the development of AI, which include embeddings from language models (ELMo) (Peters et al., 2018), Generative pretrained Transformer (GPT) (Radford et al., 2018), bidirectional encoder representations from Transformers (BERT) (Devlin et al., 2019), GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020), *etc.* In particular, LLMs represented by ChatGPT (OpenAI, 2022) demonstrate powerful multi-task capabilities in language and are considered a milestone in general AI.

ASR is the AI task that most resemble the speech recognition ability of the brain. The development of ASR technology has gone through two stages: systems based on hidden Markov models (HMM) and end-to-end systems. Modular systems include acoustic models, language models, pronunciation dictionaries, and decoding programs (Jelinek, 1998), commonly using Gaussian mixture models (GMM) and ANN to model the observation probability of HMM (Young et al., 2015; Bourlard and Morgan, 1994; Hinton et al., 2012). ASR models that only use ANN models without HMM appeared as early as the late 1980s (Robinson and Fallside, 1987), but it has only recently gradually and completely replaced HMM models as the mainstream method. Connectionist temporal classification (CTC) is a popular pure neural network ASR method (Graves et al., 2006) equivalent to an ANN-HMM (Li et al., 2019). End-to-end ASR uses an ANN model to directly convert the input speech sequence to the output text sequence (Graves, 2012; Graves et al., 2013; Chorowski et al., 2015; Lu et al., 2015; Chan et al., 2016). The end-to-end ASR have added a memory mechanism for elements in the output word sequence to model at the sentence and semantic levels like text language models, equivalent to audio-grounded language models (Li et al., 2019), and can fully utilize audio information for speech understanding (Sun et al., 2023). However, even large end-to-end ASR models trained with a massive amount of speech data still have a significant gap in accuracy and robustness in real applications compared to humans (Zhang et al., 2022; Radford et al., 2022). Consequently, studying the brain's speech recognition and understanding mechanisms is of great value in inspiring improvements in ASR (Wingfield et al., 2022).

2.4 Representative research work and its research achievements (30 min)

Humans consume language via different means, and the study of the brain's language comprehension mechanism is consequently also wide-ranging. For example, reading text is a common experimental method in neuroscience research for language cognition. By using fMRI to collect blood oxygen level signals when subjects read words and pictures, Mitchell et al. in 2008 demonstrated that word vectors constructed based on co-occurrence frequency are able to predict brain responses related to isolated words (Mitchell et al., 2008). Wehbe and colleagues built on this idea and used more natural narrative text as stimuli to study lexical features (Wehbe et al., 2014) and syntactic features (Reddy and Wehbe, 2021). Wehbe et al. also used MEG data with high temporal precision and established a correspondence between the MEG data and word vectors of the RNN language model by learning a linear mapping function (Wehbe et al., 2014). More recent studies have increasingly used neuroimaging data such as EEG and MEG with high temporal precision (Toneva et al., 2020; Hollenstein N et al., 2021; Schrimpf et al., 2021; Chehab et al., Data Analysis; Toneva et al., 2022; Caucheteux and King, 2022; Murphy et al., 2022), as well as vector representations derived from text models like word2vec, GloVe, ELMo, GPT, BERT, and GPT-2. These studies show that language models can be used to interpret the brain's language-processing mechanisms, thereby enhancing our understanding of human language cognition.

Having subjects listen to narrative speech is more natural than reading text, as speech inherently has temporality, making it easier to determine the order and duration of responses to individual words. In 2014, Mesgarani et al. had epilepsy patients listen to natural continuous speech and used intracranial electrodes to record brain electrical signals with high spatiotemporal precision, finding that the superior temporal cortex of the brain concentrated speech features (Mesgarani et al., 2014). Later, Wingfield and others found commonalities between humans and ASR based on GMM-HMM acoustic models in simultaneously collected EEG and MEG data, including significant correlations between both phonetic

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 17-23, Harbin, China, August 3 - 5, 2023.

features and the hidden layers of the DNN-HMM acoustic model with those areas of the brain related to auditory and phonetic processing (Wingfield et al., 2016; Wingfield et al., 2017; Wingfield et al., 2022). Défossez and others found that using comparative learning of brain responses collected by EEG or MEG when listening to speech and the pre-trained wav2vec 2.0 model's responses could achieve top-1 and top-10 classification accuracy rates of up to 44.1% and 72.5% respectively from 1,594 3-second speech segments heard by the subject (Défossez et al., 2022). The increased interest in audio-based ASR as a computational model has partially dampened the interest in fMRI recordings (which have a relatively poor temporal resolution compared with EEG and MEG), and ASR-related studies in EEG and MEG are gradually increasing (Wang et al., 2022).

In China, many universities and research institutions have made significant research achievements in the neural mechanisms of language cognition (Wang et al., 2022; Lu et al., 2019; Zou et al., 2022; Wang et al., 2020; Zhang et al., 2022; Fu et al., 2022; Qian et al., 2016; Liu et al., 2022; Jin et al., 2018; Sheng et al., 2019). Although the development of ASR and LLMs in China is broadly synchronized with the world's leading edge, there is relatively less research on using these AI computational models to parse brain language cognition mechanisms, most of which use reading text rather than listening to speech as stimuli (Zou et al., 2022; Wang et al., 2020). The team of researchers Shaonan Wang and Chengqing Zong from the Institute of Automation of the Chinese Academy of Sciences used open-source fMRI data based on listening to English narrative speech, and used ELMo and BERT language models to study the representation of different syntactic features in the brain (Zhang et al., 2022), and in 2022, they released Chinese data collected with fMRI and MEG respectively (Wang et al., 2022).

2.5 Major technical challenges and prospects for future work (10 min)

Research on human brain language cognition mechanisms using ASR and language models is a complex interdisciplinary field of neuroscience and AI, involving various fundamental sciences such as physics, statistics, physiology, neuroscience, psychology, and linguistics, as well as advanced hardware and software technologies such as brain-computer interfaces, signal processing, machine learning, and multi-voxel pattern analysis. The diversity of the disciplines involved is high, and many technologies (such as AI and brain-computer interfaces) are not yet fully mature, resulting in the facing of many technical challenges. However, such disciplinary characteristics also bring important scientific opportunities and many new application opportunities. In the medical field, relevant research methods and results can help understand and treat brain diseases related to speech (such as autism and dementia); in the field of human-computer interaction, it can be significant to the development of brain-computer interfaces based on imagined speech.

3 The Presenter and Co-Authors

Chao Zhang (presenter, presentation in Mandarin) is a tenure-track Assistant Professor at the Department of Electronic Engineering at Tsinghua University and holds an Honorary Professorship at University College London. He obtained both his BEng and MSc degrees from Tsinghua University and his PhD degree from Cambridge University. He was a research scientist at Google.

Andrew Thwaites (co-author) is a Senior Research Fellow at UCL and Affiliated Lecturer in Statistics at the University of Cambridge's Department of Psychology. His research focuses on computational neuroscience, in particular speech and auditory processing. Dr Thwaites received his PhD in Computational Neuroscience from the University of Cambridge.

Cai Wingfield (co-author) is a Visiting Scientist at the University of Cambridge (MRC Cognition and Brain Sciences Unit). His research focuses on speech and language processing, as well as on the dual roles of language and simulation in conceptual cognition. He received his PhD in the mathematical foundations of computer science at the University of Bath.

References

N. Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

A. Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

C. Wingfield, L. Su, B. Devereux, X. Liu, C. Zhang, P. Woodland, E. Fonteneau, A. Thwaites, W. Marslen-Wilson. 2016. Multi-level representations in speech processing in brain and machine: Evidence from EMEG and RSA. in *Society for the Neurobiology of Language*.

C. Wingfield, L. Su, X. Liu, et al. 2017. Relating dynamic brain states to dynamic machine states: Human and machine solutions to the speech recognition problem. *PloS Computational Biology*, 13:e1005617.

C. Wingfield, C. Zhang, B. Devereux, et al. 2022. On the similarities of representations in artificial and brain neural networks for speech recognition. *Frontiers in Computational Neuroscience*:16.

G. Tuckute, J. Feather, D. Boebinger, et al. 2022. Many but not all deep neural network audio models capture brain responses and exhibit hierarchical region correspondence. *bioRxiv:2022.09.06.506680*.

A.R. Vaidya, S. Jain, A. Huth. 2022. Self-supervised models of audio effectively explain human cortical responses to speech. in *Proc. ICML*.

C. Caucheteux, A. Gramfort, J.R. King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Natural Human Behaviour*:10.1038/s41562-022-01516-2.

N. A. Steinmetz, C. Aydin, A. Lebedeva, M. Okun et al. 2021. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372:abf4588.

J. Pennington, R. Socher, C.D. Manning. 2014. GloVe: Global vectors for word representation. in *Proc. EMNLP*.

T. Mikolov, I. Sutskever, K. Chen, et al. 2013. Distributed representations of words and phrases and their compositionality. in *Proc. NIPS*.

S. Hochreiter, J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735-1780.

M.E. Peters, M. Neumann, M. Iyyer, et al. 2018. Deep contextualized word representations. in *Proc. NAACL-HLT*.

R. Bommasani, D.A. Hudson, E. Adeli, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

A. Radford, K. Narasimhan, T. Salimans, et al. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. in *Proc. NAACL-HLT*.

A. Radford, J. Wu, R. Child, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

T. Brown, B. Mann, N. Ryder, et al. 2020. Language models are few-shot learners. in *Proc. NeurIPS*.

A. Vaswani, N. Shazeer, N. Parmar, et al. 2017. Attention is all you need. in *Proc. NIPS*.

OpenAI. 2022. Introducing ChatGPT. *OpenAI Blog*.

F. Jelinek. 1998. *Statistical Methods for Speech Recognition*. MIT Press.

S. Young, G. Evermann, M. Gales. 2015. *The HTK Book (for HTK version 3.5)*. Cambridge University Engineering Department.

H. Bourlard, N. Morgan. 1994. *Connectionist Speech Recognition: A Hybrid Approach*. Springer Science & Business Media.

G. Hinton, L. Deng, D. Yu, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29:82-97.

A.J. Robinson, F. Fallside. 1987. The Utility Driven Dynamic Error Propagation Network. *Cambridge University Engineering Department*.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 17-23, Harbin, China, August 3 - 5, 2023.

A. Graves, S. Fernández, F. Gomez, et al. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. in *Proc. ICML*.

Q. Li, C. Zhang, P.C. Woodland. 2019. Integrating source-channel and attention-based sequence-to-sequence models for speech recognition. in *Proc. ASRU*.

A. Graves. 2012. Sequence transduction with recurrent neural networks. in *Proc. ICML*.

A. Graves, A. Mohamed, G. Hinton. 2013. Speech recognition with deep recurrent neural networks. in *Proc. ICASSP*.

J.K. Chorowski, D. Bahdanau, D. Serdyuk, et al. 2015. Attention-based models for speech recognition. in *Proc. NIPS*.

L. Lu, X. Zhang, K. Cho, et al. 2015. A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition. in *Proc. Interspeech*.

W. Chan, N. Jaitly, Q. Le, et al. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. in *Proc. ICASSP*.

G. Sun, C. Zhang, I. Vulić, P. Budzianowski, P.C. Woodland 2023. Knowledge-Aware Audio-Grounded Generative Slot Filling for Limited Annotated Data. *arXiv preprint arXiv:2307.01764*.

Y. Zhang, D.S. Park, W. Han, et al. 2022. BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 16:1519-1532.

A. Radford, J.W. Kim, T. Xu, et al. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

T.M. Mitchell, S.V. Shinkareva, A. Carlson, et al. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320:1191-1195.

L. Wehbe, B. Murphy, P. Talukdar, et al. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9:e112575.

A.J. Reddy, L. Wehbe. 2021. Can fMRI reveal the representation of syntactic structure in the brain?. in *Proc. NeurIPS*.

L. Wehbe, A. Vaswani, K. Knight, et al. 2014. Aligning context-based statistical models of language with brain activity during reading. in *Proc. EMNLP*.

M. Toneva, O. Stretcu, B. Póczos, et al. 2020. Modeling task effects on meaning representation in the brain via zero-shot meg prediction. in *Proc. NeurIPS*.

N. Hollenstein N, C. Renggli, B. Glaus, et al. 2021. Decoding EEG brain activity for multi-modal natural language processing. *Frontiers in Human Neuroscience*:378.

M. Schrimpf, I.A. Blank, G. Tuckute, et al. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118:e2105646118.

O. Chehab, A. Défossez, L. Jean-Christophe, et al. 2022. Deep recurrent encoder: An end-to-end network to model magnetoencephalography at scale. *Neurons, Behavior, Data Analysis, and Theory*

M. Toneva, T.M. Mitchell, L. Wehbe. 2022. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computatonal Science*, 2:745-757.

C. Caucheteux, J.R. King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5:134.

A. Murphy, B. Bohnet, R. McDonald, U. Noppeney. 2022. Decoding part-of-speech from human EEG signals. in *Proc. ACL*.

N. Mesgarani, C. Cheung, K. Johnson, et al. 2014. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343:1006-1010.

A. Défossez, C. Caucheteux, J. Rapin, et al. 2022. Decoding speech from non-invasive brain recordings. *arXiv preprint arXiv:2208.12266*.

S. Wang, X. Zhang, J. Zhang, et al. 2022. A synchronized multimodal neuroimaging dataset for studying brain language processing. *Scientific Data*, 9:590.

L. Lu, Q. Wang, J. Sheng, et al. 2019. Neural tracking of speech mental imagery during rhythmic inner counting. *Elife*, 8:e48971.

S. Zou, S. Wang, J. Zhang, et al. 2022. Cross-modal cloze task: A new task to brain-to-word decoding. *Findings of the ACL*.

S. Wang, J. Zhang, N. Lin, et al. 2020. Probing brain activation patterns by dissociating semantics and syntax in sentences. in *Proc. AAAI*.

X. Zhang, S. Wang, N. Lin, et al. 2022. Probing word syntactic representations in the brain by a feature elimination method. in *Proc. AAAI*.

Z. Fu, X. Wang, X. Wang, et al. 2022. Different computational relations in language are captured by distinct brain systems. *Cerebral Cortex*.

P. Qian, X. Qiu, X. Huang. 2016. Bridging LSTM architecture and the neural dynamics during reading. in *Proc. IJCAI*.

Y. Liu, C. Luo, J. Zheng, et al. 2022. Working memory asymmetrically modulates auditory and linguistic processing of speech. *NeuroImage*, 264:119698.

P. Jin, J. Zou, T. Zhou, et al. 2018. Eye activity tracks task-relevant structures during speech and auditory sequence perception. *Nature Communications*, 9:5374.

J. Sheng, L. Zheng, B. Lyu, et al. 2019. The cortical maps of hierarchical linguistic structures during speech perception. *Cerebral Cortex*, 29:3232-3240.

M.S. Hämäläinen, R.J. Ilmoniemi. 1994. Interpreting magnetic fields of the brain: Minimum norm estimates. *Medical & Biological Engineering & Computing*, 32:35–42.

J. Millan, J. Mouriño, M. Franzé, F. Cincotti, M. Varsta, J. Heikkonen, and F. Babiloni. 2002. A local neural classifier for the recognition of EEG patterns associated to mental tasks. *IEEE Transactions on Neural Networks*, 13:678-686.

N. Kriegeskorte, M. Mur, and P.A. Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*:4.

A. Thwaites, J. Schlittenlacher, I. Nimmo-Smith, W.D. Marslen-Wilson, B.C.J. Moore. 2017. Tonotopic representation of loudness in the human cortex. *Hearing Research*, 344:244–254.

A. Pérez, M.H. Davis, R.A.A. Ince, H. Zhang, Z. Fu, M. Lamarca, M.A. Lambon Ralph, P.J. Monahan. 2022. Timing of brain entrainment to the speech envelope during speaking, listening and self-listening. *Cognition*, 224.

S. R. Oota, K. Pahwa, M. Marreddy, M. Gupta, B. S. Raju. 2023. Neural Architecture of Speech. in *Proc. ICASSP*.

J. V. Haxby. 2012. Multivariate pattern analysis of fMRI: The early beginnings. in *Neuroimage*, 62:852-52012.