DOI: 10.54254/2755-2721/92/20241716

Exploration on the improvement of EEG dataset collection and preprocessing methods based on neuroscience theory

Yucheng Yang

Institute of Software, Tianjin Normal University, Extension of Bin Shui West Road, Tianjin, China

tjyyc0319@163.com

Abstract. This paper analyzes current methods of EEG dataset collection and preprocessing within the framework of neuroscientific theories and proposes potential improvements. Existing datasets are limited because they primarily capture specific stages of neural activity during continuous recitation, missing other crucial stages of language generation. This limitation impacts the performance of deep learning models used for EEG data classification. By examining the multi-stage process of language formation, the paper highlights the need for data collection methods that encompass all stages—conceptualization, grammatical encoding, lexicalization, morphological and phonological encoding, and articulation. The review also discusses the benefits of using nonlinear dynamic system analysis methods over traditional covariance or cross-correlation matrices to better capture the complexities of EEG data. Finally, it suggests treating EEG data at various stages of language formation as multimodal data, as a strategy to improve the comprehensiveness and practicality of EEG datasets, thereby enhancing the accuracy and practical application of EEG-based brain-computer interfaces.

Keywords: EEG dataset collection, Language generation, Deep learning, Multimodal data processing.

1. Introduction

Electroencephalography (EEG) is a non-invasive technique used to record neural activity in the brain. It is widely employed to monitor the activation status of various brain regions during neural processes [1]. With the continuous advancements in brain-computer interfaces and deep learning technology, deep learning has had a huge impact on BCI based on motor imagery EEG [2]. The goal of unvoiced speech classification based on EEG data is to translate EEG signals into corresponding speech or commands, enabling silent communication with others or devices. This technology holds promise for helping language-impaired patients acquire the ability to communicate effectively.

The formation of language is a complex, multi-stage process that includes conceptualization, grammatical encoding, lexicalization, morphological and phonological encoding, articulation, and self-monitoring. Each stage follows a clear sequence [3]. However, the two existing publicly available datasets, Zhao et al. [4] and Coretto et al. [5] have released datasets focusing on imaginative speech. Both datasets require participants to continuously recite the same pronunciation or word during data collection. This approach often results in data that capture specific neural activity stages while missing potential information from other stages of language generation. Such a data collection method diverges

DOI: 10.54254/2755-2721/92/20241716

from practical applications and limits the performance of algorithms, such as deep learning, in EEG data classification. To address this issue, it is necessary to design new data collection methods that encompass all stages of language generation, thereby improving the accuracy and practical application value of EEG classification models.

This review aims to summarize the shortcomings of the current EEG dataset collection and preprocessing methods for imaginative speech, based on relevant neuroscience theories. It will also explore potential improvements for future EEG dataset collection and preprocessing techniques.

2. Neural Mechanism of Language Generation

The various stages of language formation, although somewhat parallel and interdependent, can be roughly viewed as proceeding in sequence. Previous studies [6,7,8] have outlined the process of language generation as follows: First is conceptualization, which typically occurs within 200-400 milliseconds after the stimulus appears. During this phase, the prefrontal cortex rapidly activates to form basic intentions and concepts for communication. Next is the Lexical Selection stage, where neural activity predominantly occurs in the left inferior temporal gyrus, left inferior frontal gyrus, and left inferior parietal lobule [6,9]. This stage, which transforms abstract concepts into concrete vocabulary, generally happens within 200-300 milliseconds after the stimulus. The subsequent 100-200 milliseconds encompass the Grammatical Encoding stage, primarily involving the left inferior frontal gyrus, left superior temporal gyrus, and left middle frontal gyrus [10,11,12]. Here, speech is constructed according to grammatical rules. Following this, the Morphological and Phonological Encoding stage involves planning the pronunciation of speech through collaboration among the left inferior frontal gyrus, left superior temporal gyrus, and other regions [7,13], lasting approximately 200-300 milliseconds. Finally, during the articulation stage, the primary motor cortex, Broca's area, basal ganglia, and cerebellum coordinate and control the movement of articulators to produce actual speech. [14,10,6]

3. Existing EEG Classification Research

Electroencephalogram (EEG) data exhibits characteristics such as time-varying dynamics, nonlinearity, multi-channel acquisition, multi-frequency domains, and spatial distribution, resulting in high Kirkpatrick complexity. Artificial neural networks, known for their strong fitting ability, excel at automatically discovering and understanding nonlinear relationships and abstract features in data, making them particularly effective for extracting complex data features. In deep learning, training data processing involves two main stages: data preprocessing and model training. Data preprocessing transforms raw data into a format and range suitable for model training. Common methods for classifying EEG data include cross-correlation matrices, covariance matrices, and other techniques.

3.1. Models

In most previous studies, besides essential operations such as noise removal and data segmentation, additional mathematical tools have been employed to measure the positive and negative correlations between channels, aiding classification models in better extracting data features. For example, cross-correlation matrices are commonly used to describe the similarity or correlation between two signals and are often applied in data preprocessing for deep learning [14,15,16]. In EEG data processing, cross-correlation matrices are frequently used to assess the similarity between different EEG signal channels and capture spatiotemporal dependencies between them. Covariance matrices are another tool used to measure the degree of linear correlation between two random variables. For instance, they have been utilized to extract the linear correlation between electrode data in various studies [17].

While these methods effectively highlight linear correlations between different brain regions compared to raw EEG data, they may overlook more complex nonlinear relationships. Using covariance or cross-correlation matrices as training data can result in a loss of significant nonlinear relationship information present in the original data. To preserve as much original information as possible while providing advantageous features for deep learning models, nonlinear dynamic system analysis methods can be introduced to calculate the nonlinear dynamic features between channels [18,19,20,21].

DOI: 10.54254/2755-2721/92/20241716

Given that EEG data classification tasks typically involve extracting relatively fixed types of features and do not face the complex language environments encountered by large language models, there is no need for overly complex model structures. Currently, most EEG data classification tasks utilize model architectures such as Convolutional Neural Networks (CNNs) [22], Recurrent Neural Networks (RNNs) [23,24,25,26], and Long Short-Term Memory Networks (LSTMs) [27]. Additionally, small classification models that achieve high accuracy can also be designed and employed effectively.

3.2. Datasets

As the only two publicly available imagined speech datasets [3,4], they provide valuable resources for applying deep learning technology to EEG data. These datasets offer foundational data for developing EEG-based speech imagination decoding technology and contribute to advancing brain-computer interface technology.

However, both datasets have a common feature: they require participants to continuously recite a specific vocabulary or phonetic symbol during data collection. For instance, dataset [5] requires a silent interval of five seconds, while dataset [6] does not specify a duration but involves participants recording 50 imagined pronunciations of each phoneme or vocabulary item. This approach leads to a potential issue: because it differs from the natural language generation process, the data collected tends to be concentrated primarily in the speech planning stage.

For the model to better understand the language generation process, the dataset needs to cover all stages of neural activity such as Conceptualization, Grammatical Encoding, Lexicalization, Morpho Phonological Encoding, and Articulation. This requires experimental design to avoid having participants repeat a single word silently, but to design more complex and diverse language generation tasks to capture comprehensive neural activity data. Of course, this approach will also greatly increase the difficulty of data collection.

4. Data Splitting Method

4.1. Traditional method

Directly slice the raw data into equal length slices, and label each slice with its corresponding pronunciation. Subsequently, the segmented data is preprocessed and input into the model training. This equal length slicing method is commonly used in text data recognition tasks, and cutting by different phases can expand the data volume to a certain extent, providing support for model training.

4.2. Split according to the meaning of the data

According to Pieter M. Levelt's classification of the language generation process, before a sentence is spoken, the brain regions involved in each stage are specifically activated alternately. In other words, different brain regions express the same content with different features at different stages. Author S. Oviatt pointed out that well-designed multimodal systems integrate complementary modalities to create highly collaborative mixtures, where the advantages of each mode are fully utilized to overcome the weaknesses of others [28]. This type of system can theoretically be more robust than single-modal systems that rely on only one recognition technique.

5. Conclusion

This paper aims to discuss the shortcomings of current EEG (electroencephalography) dataset collection and preprocessing methods and propose improvements based on neuroscientific theories. Existing public datasets, such as KARA ONE and those from the Universidad Nacional de Entre Ríos, typically require participants to repeatedly recite the same pronunciation or word during data collection. This approach captures neural activity at only specific stages of the language generation process and fails to comprehensively cover all stages. By analyzing the multi-stage process of language generation, this paper emphasizes the importance of designing new data collection methods that encompass all stages, including conceptualization, grammatical encoding, lexical selection, morphological and phonological

DOI: 10.54254/2755-2721/92/20241716

encoding, and articulation. Additionally, the paper explores the advantages of nonlinear dynamic system analysis methods over traditional covariance or cross-correlation matrices in capturing the complexities of EEG data. Finally, it suggests the use of multimodal data acquisition and targeted electrode placement to enhance the comprehensiveness and utility of EEG datasets, thereby improving the accuracy and practical application of EEG-based brain-computer interface classification models.

References

[1] Roy, Yannick et al. "Deep learning-based electroencephalography analysis: a systematic review.",Journal of neural engineering 16.5 (2019): 051001-051001.

[2] Altaheri, Hamdi et al. "Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: a review",Neural Computing and Applications 35.20 (2021): 14681-14722.

[3] Levelt, Willem J. M. Speaking: From Intention to Articulation. MIT Press, 1989.

[4] Zhao, Shunan, and Frank Rudzicz. "Classifying Phonological Categories In Imagined And Articulated Speech",2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2015): 992-996.

[5] Coretto, German A. Pressel et al. "Open Access Database Of Eeg Signals Recorded During Imagined Speech",12TH INTERNATIONAL SYMPOSIUM ON MEDICAL INFORMATION PROCESSING AND ANALYSIS 10160 (2017)

[6] Indefrey, P, and W J M Levelt. "The spatial and temporal signatures of word production components.",Cognition 92.1 (2004): 101-144.

[7] Indefrey, Peter. "The spatial and temporal signatures of word production components: a critical update.",Frontiers in psychology 2. (2011): 255.

[8] Sahin, Ned T et al. "Sequential Processing Of Lexical, Grammatical, And Phonological Information Within Broca'S Area",Science 326.5951 (2009): 445-449.

[9] Price, Cathy J.. "A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading.",NeuroImage 62.2 (2012): 816-847.

[10] Friederici, Angela D. "The brain basis of language processing: from structure to function.",Physiological reviews 91.4 (2011): 1357-1392.

[11] Hagoort, Peter. "On Broca, brain, and binding: a new framework.",Trends in cognitive sciences 9.9 (2005): 416-423.

[12] Grodzinsky, Yosef, and Andrea Santi. "The battle for Broca’s region",Trends in cognitive sciences 12.12 (2008): 474-480.

[13] Hickok, Gregory, and David Poeppel. "The cortical organization of speech processing",Nature Reviews Neuroscience 8.5 (2007): 393.0-402.

[14] Barachant, Alexandre et al. "Multiclass Brain–Computer Interface Classification by Riemannian Geometry", IEEE Transactions on Biomedical Engineering 59.4 (2012): 920-928.

[15] Blankertz, Benjamin et al. "Optimizing Spatial filters for Robust EEG Single-Trial Analysis", IEEE Signal Processing Magazine 25.1 (2008): 41-56.

[16] Lotte, F et al. "A review of classification algorithms for EEG-based brain-computer interfaces.", Journal of neural engineering 4.2 (2007): R1-R13.

[17] Saha, Pramit et al. "Deep learning the eeg manifold for phonological categorization from active thoughts",2019 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP) abs/1904.04358 (2019): 2762-2766.

[18] Ding, Mingzhou et al. "Granger Causality: Basic Theory and Application to Neuroscience",arXiv: Quantitative Methods (2006)

[19] Stam, C.J.. "Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field", Clinical Neurophysiology 116.10 (2005): 2266-2301.

[20] Rubinov, Mikail, and Olaf Sporns. "Complex network measures of brain connectivity: uses and interpretations.",NeuroImage 52.3 (2010): 1059-1069.

[21] Pereda, Ernesto et al. "Nonlinear multivariate analysis of neurophysiological signals.", Progress in Neurobiology 77.1 (2005): 1-37.

[22] Cun, Le et al. "Back-Propagation Applied to Handwritten Zip-code Recognition", Neural Computation 1.4 (1992): 541-551.

[23] Hopfield, J J. "Neural networks and physical systems with emergent collective computational abilities.", Proceedings of the National Academy of Sciences of the United States of America 79.8 (1982): 2554-2558.

[24] Rumelhart, David E. et al. "Learning representations by back-propagating errors", Nature 323.6088 (1988): 696-699.

[25] Rumelhart, David e et al. "Learning internal representations by error propagation", Parallel distributed processing: explorations in the microstructure of cognition, vol. 1 (1988): 318-362.

[26] Williams, Ronald J., and David Zipser. "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks.", Neural Computation 1.2 (1989): 270-280.

[27] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long Short-Term Memory", Neural Computation 9.8 (1997): 1735-1780.

[28] Oviatt, S. "Ten Myths Of Multimodal Interaction",COMMUNICATIONS OF THE ACM 42.11 (1999): 74-81.