EEG-Based Speech Envelope Decoding: Structured State Space and U-Net Model Integration

Yueqian Lin and Ming Li*

Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Duke Kunshan University, Kunshan JS 215316, China
ming.li369@dukekunshan.edu.cn

Abstract. In the advancing domain of EEG-based speech envelope decoding, this paper presents a novel architecture that strategically integrates a structured state space model (S4) with a subsequent U-Net denoising block. The S4 layer is a specialized decoder that is particularly adept at handling long-sequence modeling tasks. Following this, the U-Net denoising block functions to further refine the decoded output. The proposed architecture also includes a subject embedding layer with an embedding strength modulator (ESM), to enhance within subject performance. Experimental evaluations indicate that our hybrid model surpasses the existing top-performing model from the ICASSP 2023 Auditory EEG Challenge Task 2. Specific performance metrics, such as Pearson correlation coefficients for within subjects and held-out-subjects tests, demonstrate the effectiveness of the proposed approach.

Keywords: EEG · Structured State Space Model · U-Net Model.

1 Introduction

In the field of auditory neuroscience, while invasive techniques offer high spatial resolution and signal-to-noise ratio, non-invasive methods, notably the electroencephalogram (EEG), are gaining prominence due to their broader applicability and cost-effectiveness, particularly in clinical contexts [2]. These non-invasive methods enrich a methodological framework reflecting the complex nature of auditory perception and neural responses. Central to this investigation is the mechanism of auditory-evoked EEG responses, a pivotal measure capturing the brain's electrical activity following auditory stimulation. According to seminal research, these responses not only delineate the challenges of distinguishing spontaneous brain activities from those induced by auditory stimuli but also emphasize the influence of background spectral activity on the observed waveform [15]. The utility of EEG extends beyond foundational research to the diagnosis of hearing impairments, particularly in challenging cohorts like children or individuals with cognitive impairments. However, in the current landscape, the scope of EEG extends beyond simple diagnostic applications. A burgeoning field focuses

*Corresponding author.

on decoding auditory attention from the brain, a pursuit with transformative implications for developing intelligent hearing aids. The challenge lies in establishing a correlation between an individual's EEG and the natural speech signals they perceive, with traditional linear regression approaches facing complexities due to the characteristic low signal-to-noise ratio of EEG. Recent advances propose a shift towards non-linear methods, particularly incorporating deep neural networks (DNNs), to surmount these challenges. Specifically, Thornton et al. employ a straightforward Convolutional Neural Network (CNN) architecture to decode the speech envelope [16]; however, it is reported that increasing the number of weights does not necessarily increase the decoding performance [2], while Piao et al. use a transformer-like architecture [13], which may not be ideal for modeling long sequences [7] and shows a significant performance gap between seen and unseen subjects.

In this paper, we aim to make three key explorations into electroencephalogram (EEG) based speech envelope decoding. First, we introduce the Structured State Space model (S4) layer, designed to capture long-term dependencies in EEG data, thereby addressing a crucial limitation of existing methods. Second, we incorporate a U-Net denoising block following the S4 blocks to improve signal fidelity and reduce noise. Third, we introduce a subject conditioner with an embedding strength modulator (ESM) to enhance within subject performance. Our experimental evaluations indicate that the proposed hybrid architecture outperforms the current leading model from the ICASSP 2023 EEG Challenge Task 2, as evidenced by superior Pearson correlation coefficients for both within subjects and held-out subjects tests. Consequently, the proposed research holds the potential for advancing both theoretical understanding and practical applications in decoding auditory attention from brain signals.

2 Methods

2.1 Formulation of the Task

The Auditory EEG Challenge Task 2 [1] aims to develop a computational model that reconstructs auditory stimulus features based solely on EEG recordings. Given an auditory stimulus, denoted as S (e.g., an audiobook or podcast), the subject listens to S while their electrical brain activities E are recorded using a 64-channel EEG system with a sampling rate of 8192 Hz. Despite the high temporal resolution, the spatial resolution is limited by the 64-electrode configuration, which is positioned according to the international 10-20 system. The primary objective is to identify a model, represented as M, such that it can predict the stimulus features F_{pred} from E as $F_{\text{pred}} = M(E)$.

To evaluate the model's performance, the Pearson correlation coefficient r is employed. It is calculated for individual segments of the stimuli as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where x_i and y_i correspond to the actual and reconstructed feature values of a stimulus segment, respectively. The terms $\bar{x}$ and $\bar{y}$ represent the means of the respective sequences x and y.

The model's performance is summarized using C_s, the average Pearson correlation across all segments for a given subject. Two separate test sets are considered: one where the subject is known to the model (within subject), and one where the subject is not (held-out). These yield mean correlation scores across subjects in a given dataset S_1 and S_2, respectively. The final ranking score is a weighted combination of S_1 and S_2:

$$\text{Score} = \alpha \times S_1 + (1-\alpha) \times S_2 = \alpha \times \frac{1}{N_1} \sum_{i=1}^{N_1} C_{i1} + (1-\alpha) \times \frac{1}{N_2} \sum_{j=1}^{N_2} C_{j2} \quad (2)$$

where α is the weight of S_1 in the final score, N_1 and N_2 represent the number of subjects in the within subject and held-out test sets, respectively.

2.2 Model Overview

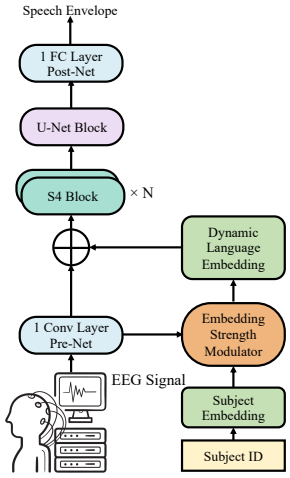

Fig. 1. Model Overview.

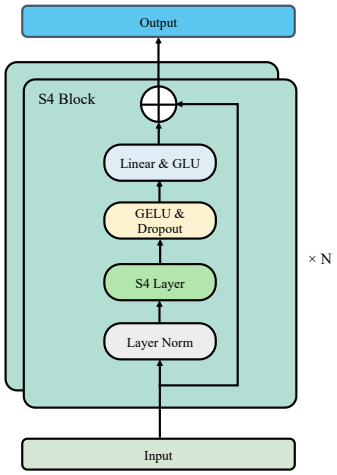

Fig. 2. S4 block Illustration.

Fig. 1 illustrates the architecture of the proposed model. Our EEG-based speech envelope decoding model synergistically combines the S4 model with a U-Net denoising block for refined output processing. Given an EEG signal as input, the model first extracts features via a Pre-Net convolutional layer. A dynamic subject embedding, influenced by an ESM, contextualizes the features for subjects within the test dataset to optimize the performance. These features are then passed through the S4 block for long-sequence modeling, a process that is iteratively repeated N times. The U-Net denoising block further refines the decoded output and forwards it to a single Fully Connected (FC) layer, serving as the post-net. Subsequent sections will offer a detailed explanation of each component of the model.

2.3 Embedding Strength Modulator

To generate the speech envelope specifically tailored for within subject applications, Piao et al. leverage an auxiliary global conditioner, aiming to retain more contextual information [13]. It is imperative to acknowledge that the effectiveness of the subject embedding should not remain static, considering the inherent transferability of diverse subjects' EEG signals [8] and their generalizability for classification tasks [11]. Drawing inspiration from advances in bilingual TTS tasks, as seen in [19] and [20], where an ESM dynamically modulates the strength of language and phonology embeddings, our model integrates an ESM layer following the subject embedding. Yang et al. detail the ESM's mechanism [19], where it is described as an attention-based modulator resembling the framework in [18]. The ESM comprises two primary sub-networks: multi-head attention and a feed-forward network, both of which leverage layer normalization and residual connections. Formally, our ESM processes the Pre-Net outputs with scaled positional encoding, PN_o, alongside subject embedding, SE. This mechanism allows the post-Pre-Net EEG signal data to be modulated by dynamic weights, which are influenced by both the data and the subject embedding. The equations governing the output F_o in our ESM are as follows:

$$Mo = \mathrm{MH}(PN_o, \mathrm{LN}(SE), \mathrm{LN}(SE)) + SE \tag{3}$$

$$Fo = \mathrm{FFN}(LN(Mo)) + Mo \tag{4}$$

where MH(query, key, value), FFN$(\cdot)$ and LN$(\cdot)$ are multi-head attention, feed-forward network and layer normalization, respectively.

2.4 Structured State Space Decoding

The S4 block is pivotal in enhancing speech envelope estimation. Originally conceived as an innovative sequence model, the S4 represents an evolution of the continuous-time state space model. It distinguishes itself by adeptly modeling long-term dependencies and maintaining computational efficiency. A primary feature of the S4 is its utilization of linear state space transformations to depict relationships among latent state spaces [7], expressed by the equations

$$\dot{x} = Ax + Bu \tag{5}$$

$$y = Cx + Du \tag{6}$$

where A, B, C, D are trainable matrices that map the input u, hidden state x, and output y.

This framework of linear transformations facilitates a more precise and efficient estimation of latent variables, crucial for tasks such as the auditory EEG Challenge Task 2. Compared to conventional sequential models like Recurrent Neural Networks (RNNs) and Transformers, the S4 provides notable advantages. The S4 mitigates issues related to computational complexity and positional information, often encountered in Transformers, thus showcasing superior performance in modeling long sequences [7].

The S4's versatility is evidenced by its effective application in various tasks, including autoregressive inference tasks like waveform generation [6] and language modeling [7]. Studies have underscored the S4 decoder's capability in enhancing the naturalness of synthetic speech, especially when compared with Transformer-TTS [12]. Additionally, it is noteworthy that S4-based speech enhancement models are highly efficient and achieve commendable results even with a reduced model size [10]. A distinguishing hallmark of the S4 is its capability for parallel training and recurrent generation, characterized by sub-quadratic complexity. This leads to significantly faster generation speeds; for instance, the S4 model demonstrates a performance that is 60 times faster than conventional autoregressive models in language modeling tasks on the WikiText-103 dataset [7]. Additionally, S4-based world models have demonstrated superiority over Transformer-based models in terms of long-term memory and training efficiency [5].

Turning our attention to design specifics, the S4 block's structure (as illustrated in Fig. 2) adapts from the Pre-LN FFT block architecture, previously presented in [13]. Modifications include the elimination of the position-wise feed-forward layer and the replacement of the multi-head self-attention layer with the S4 layer[1]. This is because the S4 model implicitly incorporates positional information within its architecture [7]. In summary, the integration of the S4 block enhances the robustness of our proposed model in handling long-form scenarios.

2.5 U-Net Denoising Block

The integration of the U-Net denoising block into our EEG-based speech envelope decoding architecture is informed by the intrinsic characteristics of U-Net models. These models are particularly adept at denoising and feature integration, which aligns with the objectives of denoising diffusion probabilistic models (DDPMs) [9]. DDPMs operate within the latent variable models and Markov chains framework to generate data resembling real-world distributions. U-Net architectures, known for their encoder-decoder structure with skip connections, are capable of capturing detailed contextual information at multiple scales, which is beneficial for EEG signal processing.

In the realm of EEG, U-Net models have proven effective for artifact removal, enhancing the clarity of brain signals for interfaces and decoding tasks. The IC-U-Net, utilizing a U-Net architecture, has shown promise in removing EEG artifacts and reconstructing signals, a capability that aligns with our U-Net block's objectives [4]. This advancement supports our approach, leveraging U-Net's architecture for improved EEG signal decoding, essential for the high-dimensional and noisy data characteristic of EEG [17].

The U-Net block in our design is grounded in the forward process of DDPMs, accepting and conditioning the output from the S4 blocks. We utilize the U-Net architecture to refine and predict the signal from the S4 blocks, following progressive dilation training strategies to improve training efficiency [14]. Figure 3 illustrates the integration of the U-Net denoising block within our architecture.

[1] A detailed implementation of the S4 model is accessible at https://github.com/espnet/espnet/blob/master/espnet2/asr/state_spaces/s4.py.

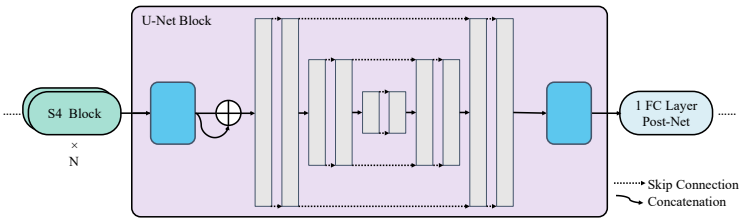

Fig. 3. U-Net Denoising Block Illustration.

3 Experiments

3.1 Dataset

We employ a high-quality dataset derived from the Auditory EEG Challenge at ICASSP 2023 [3], which comprises EEG recordings from 85 normal-hearing, Dutch-speaking young adults. EEG recordings were obtained using a 64-channel Biosemi ActiveTwo system, with a sampling rate of 8192 Hz, in a controlled setting. The training set is composed of 71 subjects and includes 508 trials, totaling 7,216 minutes. The dataset is partitioned into training and test sets, the latter being further split into two subsets for a more nuanced evaluation. Specifically, the test set is divided into two subsets: one for held-out stories and another for held-out subjects, each designed to assess intra-subject and inter-subject generalization, respectively. Test Set 1 features the same 71 subjects as in the training set but exposes them to new auditory stimuli, accounting for a total of 944 minutes of data. Test Set 2 introduces 14 new subjects, not seen in the training set, yielding an additional 1,260 minutes of EEG recordings, with both sets employing the same experimental protocol as the training set.

3.2 Experimental Setup

To verify the performance of our proposed system, we utilize both the same training and test sets from the challenge and adopt the corresponding evaluation metrics: the α is set as 1/3 in Eq. 2. We use PyTorch to train the model. The number of S4 blocks was set to 8, and we used 128 as the dimension of the state in the S4 model and enabled its bi-directionality (convolution kernel will be two-sided). For the U-Net denoising block, the U-Net channels at each layer are [128, 256, 512], downsampling and upsampling factors at each layer are [1, 2, 2], and the number of repeating items at each layer are [2, 2, 2]. We trained our model for 1000 epochs using the Adam optimizer with an initial learning rate of 0.0005. We also applied a StepLR scheduler with a learning rate decay factor of 0.9. During training, we used 5-second segments of signals for stable training, which was randomly cropped from each EEG/speech envelope segment. For inference, input signals are divided into multiple 5-second segments, and their corresponding outputs are concatenated to form the complete envelope. Most of the settings are the same as those in [13].

3.3 Evaluation

We conduct a comprehensive evaluation of our proposed approach, assessing the contributions of each incorporated module. The evaluation results are presented in Table 1. Initially, we assess the efficacy of integrating a single U-Net

denoising block. This integration yields a mean score increase to 0.1281 for held-out subjects, significantly outperforming systems from the 2023 Auditory-EEG Challenge, and a competitive mean score of 0.1835 for within subjects, thereby validating our hypothesis regarding the denoising and robustness-enhancing attributes of the U-Net denoising block. Subsequently, we incorporate the ESM layer, as detailed in Section 2.3. The optimal performance reflected in the held-out subjects' mean scores substantiates the functionality of the ESM layer in dynamically adjusting the strength of the subject embedding, enabling the model to assimilate more universal information. The notable improvement in both the within subject mean and the final score for the U-Net + ESM + S4 system underscores the system's augmented capacity to model long sequences and manage information over extended temporal spans in continuous time series

Table 1. Pearson correlation values of different systems

Method	Within subjects mean ↑	Held-out subjects mean ↑	Score ↑
HappyQuokka (1st in 2023 Auditory-EEG Challenge [1])	0.1895	0.0976	0.1589
TheBrainwaveBandits (2nd in 2023 Auditory-EEG Challenge [1])	0.1741	0.1123	0.1535
U-Net	0.1835	0.1281	0.1650
U-Net + ESM	0.1900	**0.1295**	0.1698
U-Net + ESM + S4	**0.2040**	0.1258	**0.1779**

4 Conclusion

In this study, we present an EEG-based model for speech envelope decoding leveraging the S4 blocks and U-Net model, with a specialized ESM layer that has been demonstrated to be effective for estimating both within subject and heldout subject performance. The S4 blocks contribute to the model's enhanced performance in learning long-term dependencies and the incorporation of a U-Net model further enhances the robustness of our proposed system. Experimental evaluations indicate that our proposed model outperforms the top two models from the ICASSP 2023 Auditory EEG Challenge Task 2. Therefore, we posit that future research could benefit from exploring the utility of S4 blocks and U-Net models in EEG-based studies or other domains involving continuous time series data to address current challenges.

Acknowledgements This research is funded in part by the DKU Foundation Project "Interdisciplinary Signal Processing Technologies," National Natural Science Foundation of China (62171207) and the DKU Summer Research Scholars (SRS) Program. We are grateful for the computational resources provided by the Advanced Computing East China Sub-Center, to Zexin Cai for his review and feedback, and to all the reviewers for their contributions to this paper.

References

1. Auditory eeg challenge leaderboard for task 2 - ICASSP 2023. https://exporl.github.io/auditory-eeg-challenge-2023/task2/leaderboard/ (2023)
2. Accou, B., Vanthornhout, J., Hamme, H.V., Francart, T.: Decoding of the speech envelope from EEG using the VLAAI deep neural network. Scientific Reports **13**(1), 812 (2023)
3. Bollens, L., Accou, B., Van hamme, H., Francart, T.: Sparrkulee: A speech-evoked auditory response repository of the ku leuven, containing EEG of 85 participants (2023). https://doi.org/10.48804/K3VSND

4. Chuang, C.H., Chang, K.Y., Huang, C.S., Jung, T.P.: Ic-u-net: A u-net-based denoising autoencoder using mixtures of independent components for automatic eeg artifact removal. NeuroImage **263**, 119586 (2022)
5. Deng, F., Park, J., Ahn, S.: Facing off world model backbones: Rnns, transformers, and S4. arXiv preprint arXiv:2307.02064 (2023)
6. Goel, K., Gu, A., Donahue, C., Ré, C.: It's raw! audio generation with state-space models. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. Proceedings of Machine Learning Research, vol. 162, pp. 7616–7633. PMLR (2022)
7. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022)
8. Hang, W., Feng, W., Du, R., Liang, S., Chen, Y., Wang, Q., Liu, X.: Cross-subject EEG signal recognition using deep domain adaptation network. IEEE Access **7**, 128273–128282 (2019)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
10. Ku, P.J., Yang, C.H.H., Siniscalchi, S., Lee, C.H.: A Multi-dimensional Deep Structured State Space Approach to Speech Enhancement Using Small-footprint Models. In: Proc. INTERSPEECH 2023. pp. 2453–2457 (2023)
11. Lu, Y., Wang, M., Wu, W., Han, Y., Zhang, Q., Chen, S.: Dynamic entropy-based pattern learning to identify emotions from EEG signals across individuals. Measurement **150**, 107003 (2020)
12. Miyazaki, K., Murata, M., Koriyama, T.: Structured state space decoder for speech recognition and synthesis. In: 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
13. Piao, Z., Kim, M., Yoon, H., Kang, H.: Happyquokka system for ICASSP 2023 auditory EEG challenge. CoRR **abs/2305.06806** (2023)
14. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022)
15. Savers, B.M., Beagley, H., Henshall, W.: The mechanism of auditory evoked eeg responses. Nature **247**(5441), 481–483 (1974)
16. Thornton, M., Mandic, D., Reichenbach, T.: Robust decoding of the speech envelope from EEG recordings through deep neural networks. Journal of Neural Engineering **19**(4), 046007 (2022)
17. Vetter, J., Macke, J.H., Gao, R.: Generating realistic neurophysiological time series with denoising diffusion probabilistic models. bioRxiv (2023). https://doi.org/10.1101/2023.08.23.554148
18. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 10524–10533. PMLR (13–18 Jul 2020)
19. Yang, F., Luan, J., Meng, M., Wang, Y.: Improving Bilingual TTS Using Language And Phonology Embedding With Embedding Strength Modulator. In: Proc. INTERSPEECH 2023. pp. 5531–5535 (2023)
20. Zhou, H., Lin, Y., Shi, Y., Sun, P., Li, M.: BiSinger: Bilingual singing voice synthesis. arXiv preprint arXiv:2309.14089 (2023)