2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) | 979-8-3503-7451-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSPW62465.2024.10626977

BMMSNET: BIDIRECTIONAL MAPPING AND MULTILEVEL SIMILARITY COMPARISON FOR EEG-SPEECH MATCH-MISMATCH PROBLEM

Zelin Qiu[1,2], *Jianjun Gu*[1], *Dingding Yao*[1], *Junfeng Li*[1,*] , *Yonghong Yan*[1]

[1]Institute of Acoustics, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China

ABSTRACT

In this report, we present our approach for the task 1 (match-misatch problem) of the Auditory EEG Decoding Challenge at ICASSP 2024. Existing methods for this problem usually only consider the global similarity between EEG and speech stimulus. To overcome the shortcoming, we propose a novel neural network which mainly consists of a **B**idirectional **M**apping between EEG and speech stimulus and a **M**ultilevel **S**imilarity comparison (BMMSNet). Moreover, we employ a data augmentation-based ensemble learning method to further increase the robustness of our system. Our method achieves a discrimination accuracy of 58.673% on the test data provided by the challenge, ranking third among more than 50 teams. We release the code for our model online.[1]

***Index Terms*—** EEG, auditory EEG decoding, convolutional neural network

1. INTRODUCTION

The Auditory EEG Decoding Challenge at ICASSP 2024 Signal Processing Grand Challenge calls on researchers to explore the relationship between EEG and speech stimulus more deeply by setting two tasks: match-mismatch and regression. We concentrate on task 1, which involves determining the specific speech stimulus that elicits the subject's electroencephalography (EEG) response among five candidates.

Existing methods use neural networks to map EEG and speech stimulus into a hidden space, where their similarity is measured. However, many of these methods only consider the global similarity. To address the drawback, in this report, we propose a novel architecture called BMMSNet. Our approach consists of two main components: (1) a bidirectional mapping between EEG and speech stimulus, which simulates the forward and backward processes of transforming speech into neural signals and vice versa; and (2) a multilevel similarity comparison, which calculates the similarity between EEG and speech stimulus at each layer of the bidirectional mapping, and aggregates them to obtain the final matching result. It is noteworthy that we use the shallow (deep) features of the EEG branch and the deep (shallow) features of the speech branch to calculate the similarity, because these two branches have opposite directions, simulating the human auditory perception process of converting speech to EEG and its inverse process, respectively. Thus, EEG shallow (deep) feature is corresponding to deep (shallow) feature of speech. Our approach has the following advantages: (1) it provides a more natural and intuitive way of modeling the EEG-speech stimulus relationship, which can be interpreted as an auditory perception and an inverse model; (2) it captures the similarity between EEG and speech stimulus at different levels of abstraction, which can reveal more fine-grained and meaningful features for the matching. Furthermore, we adopt an ensemble learning strategy based on data augmentation to further decrease the decoding error rate. Our method achieves a discrimination accuracy of 58.673% on the official test set.

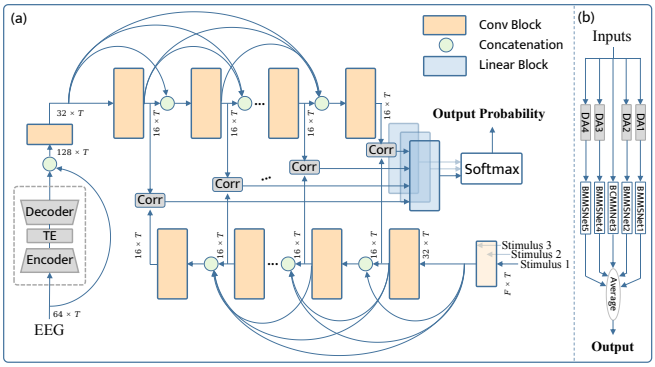


Fig. 1. (a) An overview of the proposed BMMSNet (only three stimuli are shown in the figure for simplicity). (b) The schematic diagram of the ensemble BMMSNet based on data augmentation.

2. METHODOLOGY

2.1. Model Structure

As shown in Figure 1, our proposed model comprises two branches: EEG and stimulus. Each branch consists of a convolutional block (Conv Block) and a densely connected structure of ten Conv blocks (only four are shown in Figure 1 for

*Corresponding author

This research is partially supported by STI 2030—Major Projects (No. 2021ZD0201503), the National Natural Science Foundation of China (No. 12104483).

[1]https://github.com/QQQQQZL/BMMSNet

licensed use limited to: MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE. Downloaded on November 03,2026 at 12:50:05 UTC from IEEE Xplore. Restrict

simplicity), where each Conv block contains a convolution layer, a batch normalization layer, and a PReLU activation layer. The two branches are bidirectionally oriented, and the intermediate layer outputs can be regarded as intermediate state signals between the EEG and stimulus. We compute the Pearson's r of these intermediate states which are subsequently assigned with different weights by a linear layer to obtain the final output. Moreover, for better feature extraction, we incorporate an EEG feature enhancement module, which includes a convolutional encoder, a transpose convolutional decoder, and a transformer encoder (TE) [1] for temporal modeling. To enrich the information, the stimulus we use is the concatenation of speech envelope and mel-spectrogram. We also apply a dropout of 0.2 after all convolution operations to avoid overfitting.

2.2. Ensemble Model with Data Augmentation

We apply various data augmentation (DA) methods to enhance the model's robustness to EEG distortions, and use an ensemble model based on different DA methods to further boost the system's performance, as illustrated in Figure 1(b). For a given segment of EEG, we use four different DA methods as follows:

- Randomly add another segment of EEG scaled by a small coefficient ($0 \sim 0.2$) to the original EEG segment.
- Randomly mask (set to zero) some ($1 \sim 10$) channels and some time intervals ($0 \sim 1.5$s) of the signal.
- Add white noise (SNR$\in [-5, 10]$ dB).
- Multiply the data of each channel by different positive coefficients ($0.3 \sim 2.0$).

We train different models (DA models) using the original and the four augmented datasets, and the final output is the average of the outputs of these models.

3. EXPERIMENTAL SETUP

We conduct our experiments using the official training dataset [2] provided by the challenge and follow the same preprocessing steps. Since the labels of the official test data are unavailable, in this report, we use our own data splitting method, specifically, we randomly select 80%/10%/10% (68/9/8) of the subjects from the official training dataset as the training/validation/test data. In the subsequent paragraphs, all results are based on such data spliting.

We divide the EEG data and the corresponding speech stimulus into 5-s segments using a sliding window with a stride of 1-s for further processing. For each EEG segment, we use one matching speech segment and four mismatching speech segments as the input, and adopt the same method as described on the official webset to avoid label learning. The models are trained with the Adam optimizer and softmax loss function, using a batch size of 60, a learning rate of 0.001 and a decay factor of 0.98 per epoch. Each DA model is trained for 100 epochs, and the one with the best performance on the validation set is selected for testing.

Table 1. The decoding performance of different models.

Model	Decoding Accuracy (%)	Subject Std (%)
baseline	58.5	11.11
Ensemble BMMSNet	**67.2**	10.99
BMMSNet	65.6	11.83
BMMSNet DA1	65.2	11.33
BMMSNet DA2	65.1	10.98
BMMSNet DA3	64.4	10.56
BMMSNet DA4	65.0	10.61

4. RESULTS

Table 1 shows the results of the official dilated convolution baseline model [3] and our proposed model, as well as the different variants of our proposed model, on our split test set. The proposed ensemble model has a significant improvement ($p < 0.005$, $Wilcoxon$ signed-rank test) in performance compared to the baseline model, which proves the effectiveness of our design. Furthermore, the model ensemble technique proves to be effective in enhancing the system's performance.

5. CONCLUSION

In this report, we present our approach for the Auditory EEG Decoding Challenge at ICASSP 2024. We propose a novel neural network architecture, named BMMSNet, which leverages bidirectional mapping between EEG and speech stimulus to obtain effective intermediate variables and captures the similarity at different levels of abstraction . We also employ an ensemble strategy based on data augmentation to improve the system's robustness. Experiments demonstrate that our proposed method significantly outperforms the baseline.

6. REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[2] Bernd Accou, Lies Bollens, Marlies Gillis, Wendy Verheijen, Hugo Van hamme, and Tom Francart, "Sparrkulee: A speech-evoked auditory response repository of the ku leuven, containing eeg of 85 participants," *bioRxiv*, pp. 2023–07, 2023.

[3] Bernd Accou, Mohammad Jalilpour Monesi, Jair Montoya, Tom Francart, et al., "Modeling the relationship between acoustic stimulus and eeg with a dilated convolutional neural network," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1175–1179.

licensed use limited to: MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE. Downloaded on November 03,2026 at 12:50:05 UTC from IEEE Xplore. Restrict