Independent Feature Enhanced Crossmodal Fusion for Match-Mismatch Classification of Speech Stimulus and EEG Response

Shitong Fan[1,†], Wenbo Wang[2,†], Feiyang Xiao[1], Shiheng Zhang[1], Qiaoxi Zhu[3], and Jian Guan[1,]*

[1]Group of Intelligent Signal Processing (GISP), College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]Faculty of Computing, Harbin Institute of Technology, Harbin, China
[3]University of Technology Sydney, Ultimo, Australia

fanshitong@hrbeu.edu.cn, wwb1325864697@outlook.com, xiaofeiyang128@gmail.com, zhangshiheng@hrbeu.edu.cn, qiaoxi.zhu@gmail.com, j.guan@hrbeu.edu.cn

Abstract

It is crucial for auditory attention decoding to classify matched and mismatched speech stimuli with corresponding EEG responses by exploring their relationship. However, existing methods often adopt two independent networks to encode speech stimulus and EEG response, which neglect the relationship between these signals from the two modalities. In this paper, we propose an independent feature enhanced crossmodal fusion model (IFE-CF) for match-mismatch classification, which leverages the fusion feature of the speech stimulus and the EEG response to achieve auditory EEG decoding. Specifically, our IFE-CF contains a crossmodal encoder to encode the speech stimulus and the EEG response with a two-branch structure connected via crossmodal attention mechanism in the encoding process, a multi-channel fusion module to fuse features of two modalities by aggregating the interaction feature obtained from the crossmodal encoder and the independent feature obtained from the speech stimulus and EEG response, and a predictor to give the matching result. In addition, the causal mask is introduced to consider the time delay of the speech-EEG pair in the crossmodal encoder, which further enhances the feature representation for match-mismatch classification. Experiments demonstrate our method's effectiveness with better classification accuracy, as compared with the baseline of the Auditory EEG Decoding Challenge 2023.

Index Terms: Auditory EEG decoding, multi-modal learning, feature fusion, cross-attention

1. Introduction

Electroencephalogram (EEG) is widely used to explore the mechanism by which the brain processes external information [1–4]. To analyze and interpret the brain's response to speech stimulus, auditory EEG decoding has recently attracted increasing attention due to its importance for the development of neuro-steered hearing aids [5], such as the study on the match-mismatch classification to determine whether a given pair containing speech stimulus and EEG response correspond [6].

To address the match-mismatch problem, existing methods adopt various deep-learning-based structures to improve the representation of speech stimulus and EEG response for a better match and mismatch prediction. For example, the study in [2] presents a model based on long short-term memory (LSTM) [7], which improves the accuracy of match-mismatch classification by considering variable delays in the EEG response to speech stimulus. The study in [8] further explores the influence of using different levels of speech stimulus (i.e., envelope, voice activity, phoneme identity, and word embeddings) as the input for LSTM-based models in the match-mismatch task, and finds that mel-spectrogram can provide better classification performance. Whereas the study [9] introduces word boundary information to LSTM to account for the brain's discrete processing of speech stimulus, which further improves the classification performance.

Meanwhile, there are also some studies [3, 5, 6] that employ dilated convolutional neural networks [10, 11] to capture the contextual information of both speech stimulus and EEG response, expanding the receptive field while maintaining a low parameter count [6]. Based on [6], the study in [3] introduces the fundamental frequency of speech and combines it with speech envelope, which improves the match-mismatch classification performance. Subsequently, [5] utilizes an attention encoder to replace the spatial convolution in [6], allowing for better exploration of the spatial and temporal feature of EEG response. However, all these methods employ two separate networks to encode speech stimulus and EEG response, which neglects the relationship between these two modalities, leading to limited match-mismatch classification performance.

In this paper, we propose an independent feature enhanced crossmodal fusion model (IFE-CF) for match-mismatch classification of auditory EEG decoding, which leverages the fusion feature of the speech stimulus and the EEG response to achieve speech-EEG pair classification. Our IFE-CF consists of a crossmodal encoder, a multi-channel module and a predictor. Specifically, we introduce a crossmodal attention mechanism [12] to build the crossmodal encoder, which can establish the connection between the latent features of speech stimulus and EEG response from two separate networks, to explore the relationship between the speech stimulus and EEG response. Furthermore, since EEG responses occur after a given moment of speech stimulus, to ensure that crossmodal attention focuses on the reasonable relationships when interacting with the latent features of both modalities, a causal mask strategy is introduced in crossmodal attention, which can filter out the unreasonable relationship in the interaction feature from the crossmodal encoder, thus improving the feature representation.

Then, a multi-channel fusion module is designed to fuse both the interaction feature from crossmodal attention encoder and the independent features from the separate networks. It provides a fine-grained feature that can reflect both the relationship

This work was partly supported by the project of the Ministry of Industry and Information Technology under Grant No.CBZ3N21-2

† These authors contributed equally to this work.

* Corresponding author.

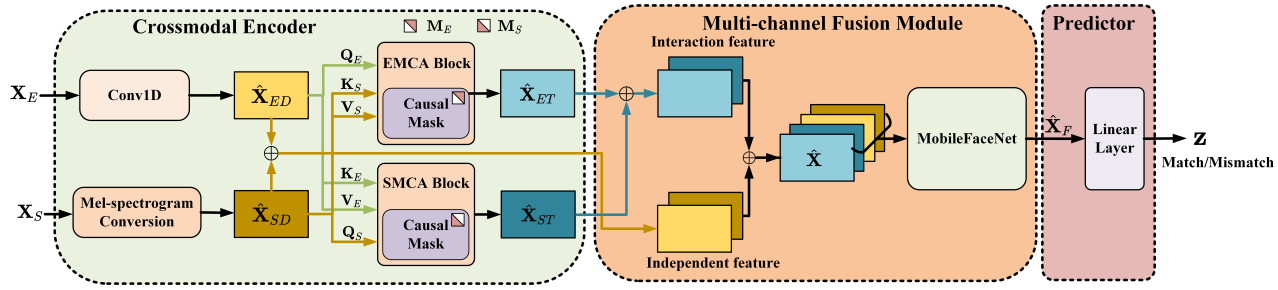

Figure 1: *The model structure of our proposed Independent Feature Enhanced Crossmodal Fusion method (IFE-CF). "$\bigoplus$" indicates the feature concatenate operation. $\mathbf{X}_E \in \mathbb{R}^{D\times T}$ and $\mathbf{X}_S \in \mathbb{R}^{t}$ represent the EEG response and the speech stimulus, respectively. Here, D denotes the number of EEG response channels, T denotes the number of time frames in the EEG response, and t represents the number of sampling points in the speech stimulus waveform. $\mathbf{M}_E$ and $\mathbf{M}_S$ are the causal mask matrices for EEG causal mask crossmodal attention (EMCA) block and speech causal mask crossmodal attention (SMCA) block, respectively.*

between speech-EEG modalities and the specific characteristics of each individual modality for the predictor module, resulting in improved match-mismatch classification performance.

Experiments are performed on the Auditory EEG Decoding Challenge 2023 dataset [13], and the results show that our proposed IFE-CF can outperform the official baseline method and surpass the Top-2 ranking system in Auditory EEG Decoding Challenge 2023. Ablation study and performance analysis are also conducted in our experiments, which validate the effectiveness of each component of our proposed method.

The remainder of the paper is organized as follows: Section 2 presents our proposed method in detail; Section 3 shows the experimental results and analysis; and Section 4 concludes the paper and discusses potential future work.

2. Proposed Method

In this section, we present our proposed independent feature enhanced crossmodal fusion model (IFE-CF) in detail, which consists of a crossmodal encoder, a multi-channel fusion module, and a predictor for match-mismatch classification. The overall framework is given in Figure 1.

2.1. Crossmodal Encoder

The crossmodal encoder adopts a two-branch structure to obtain the independent and interaction features of speech stimulus and EEG response via the independent and interaction feature encoding branches, respectively.

Independent Feature Encoding: For independent feature encoding, speech stimulus and EEG response are leveraged as the input of the crossmodal encoder module to obtain their independent features.

Inspired by work [8], a mel-spectrogram conversion is employed to convert the speech stimulus into the mel-spectrogram as the independent feature $\hat{\mathbf{X}}_{SD} \in \mathbb{R}^{d\times T}$ of speech stimulus, where d and T represent the dimension of extracted feature and the number of time frames, respectively.

Meanwhile, a spatial convolution is utilized to obtain the independent feature of EEG response $\hat{\mathbf{X}}_{ED} \in \mathbb{R}^{d\times T}$, as follows

$$\hat{\mathbf{X}}_{ED} = \mathrm{Conv1D}(\mathbf{X}_E), \tag{1}$$

where $\mathbf{X}_E \in \mathbb{R}^{D\times T}$ and $\mathrm{Conv1D}(\cdot)$ are the original input of EEG response and spatial convolution operation, respectively. D denotes the channel number of the EEG response.

Interaction Feature Encoding: The interaction feature encoding aims to explore the relationship between speech stimulus and EEG response according to their independent feature via two crossmodal attention blocks, i.e., EEG causal mask crossmodal attention (EMCA) block and speech causal mask crossmodal attention (SMCA) block. EEG causal mask crossmodal attention block is used to consider the EEG's relationship with speech in the encoding process of EEG response, while speech crossmodal attention block accounts for speech's relationship with EEG in the encoding process of speech stimulus.

Since these two crossmodal attention blocks perform in a similar way, we take the SMCA block as an example to present our method in detail for the sake of simplicity as follows.

In the speech causal mask crossmodal attention block, to obtain the interaction feature of speech stimulus $\hat{\mathbf{X}}_{ST} \in \mathbb{R}^{d\times T}$, the independent feature $\hat{\mathbf{X}}_{SD}$ and $\hat{\mathbf{X}}_{ED}$ is linearly transformed to obtain speech's query $\mathbf{Q}_S$, key $\mathbf{K}_E$, value $\mathbf{V}_E$, as follows

$$\begin{cases} \mathbf{Q}_S = \mathbf{W}_{Q,S}\hat{\mathbf{X}}_{SD}, \\ \mathbf{K}_E = \mathbf{W}_{K,E}\hat{\mathbf{X}}_{ED}, \\ \mathbf{V}_E = \mathbf{W}_{V,E}\hat{\mathbf{X}}_{ED}, \end{cases} \tag{2}$$

where $\mathbf{W}_{Q,S}$, $\mathbf{W}_{K,E}$, and $\mathbf{W}_{V,E} \in \mathbb{R}^{d_k\times d}$ are weight matrices, and d_k is the dimension of these weight matrices.

Then, $\mathbf{Q}_S$, $\mathbf{K}_E$, and $\mathbf{V}_E$ are split into h heads, and the i-th head's query $\mathbf{Q}_s^i$, key $\mathbf{K}_e^i$, and value $\mathbf{V}_e^i$ can be obtained, as follows

$$\begin{cases} \mathbf{Q}_s^i = \mathbf{W}_{Q,S}^i\mathbf{Q}_s, \\ \mathbf{K}_e^i = \mathbf{W}_{K,E}^i\mathbf{K}_e, \\ \mathbf{V}_e^i = \mathbf{W}_{V,E}^i\mathbf{V}_e, \end{cases} \tag{3}$$

where $\mathbf{W}_{Q,S}^i$, $\mathbf{W}_{K,E}^i$ and $\mathbf{W}_{V,E}^i \in \mathbb{R}^{\frac{d_k}{h}\times d}$ are weight matrices.

To consider the time delay between speech-EEG pair, a causal mask $\mathbf{M}_S \in \mathbb{R}^{T\times T}$ is introduced in the SMCA block to filter out the noise in the interaction feature calculation, which can be defined as

$$\mathbf{M}_S = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ -\infty & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\infty & -\infty & \cdots & 0 \end{bmatrix}, \tag{4}$$

Similar to the construction of the lower triangular mask matrix $\mathbf{M}_S$ in SMCA block, an upper triangular mask matrix $\mathbf{M}_E$ is

also designed in EMCA block to filter out the irrelevant noise in the interaction feature.

Then, the interaction feature $\hat{\mathbf{X}}_{ST}$ can be obtained as follows

$$\hat{\mathbf{X}}_{ST}^{i} = \text{softmax}\left(\frac{\mathbf{Q}_s^i \mathbf{K}_e^{i\top}}{\sqrt{d_k}} + \mathbf{M}_S\right)\mathbf{V}_e^i, \tag{5}$$

$$\hat{\mathbf{X}}_{ST} = \text{Concat}(\hat{\mathbf{X}}_{ST}^1, \ldots, \hat{\mathbf{X}}_{ST}^h)\mathbf{W}_O, \tag{6}$$

where $\text{softmax}(\cdot)$ and $\text{Concat}(\cdot)$ are softmax function and concatenation operation respectively. $\mathbf{W}_O \in \mathbb{R}^{d\times d}$ is a learnable weight matrix.

Finally, we can obtain four features including speech independent feature $\hat{\mathbf{X}}_{SD}$, EEG independent feature $\hat{\mathbf{X}}_{ED}$, speech interaction feature $\hat{\mathbf{X}}_{ST}$ and EEG interaction feature $\hat{\mathbf{X}}_{ET}$ via our two-branch crossmodal encoder.

2.2. Multi-Channel Fusion Module

The multi-channel fusion module is employed to further capture the relationship between speech stimulus and EEG response via deeply fusing the independent and interaction features of speech stimulus and EEG response. Here, the independent feature of speech stimulus and EEG response are introduced in the fusion processing, which is used to ensure the fusion feature obtained by our multi-channel fusion module not only contains the relationship between speech stimulus and EEG response, but also includes the characteristics of these signals in their modality.

Thus, we concatenate $\hat{\mathbf{X}}_{SD}, \hat{\mathbf{X}}_{ED}, \hat{\mathbf{X}}_{ST}$ and $\hat{\mathbf{X}}_{ET}$ to obtain the fusion feature $\hat{\mathbf{X}}$ as follows

$$\hat{\mathbf{X}} = \text{Concat}(\hat{\mathbf{X}}_{SD}, \hat{\mathbf{X}}_{ED}, \hat{\mathbf{X}}_{ST}, \hat{\mathbf{X}}_{ET}), \tag{7}$$

Then $\hat{\mathbf{X}}$ is fed into a MobileFaceNet structure [14] to obtain the embedding feature $\hat{\mathbf{X}}_F$ for mach and mismatch prediction, as follows

$$\hat{\mathbf{X}}_F = \text{MobileFaceNet}(\hat{\mathbf{X}}; \Theta), \tag{8}$$

where Θ denotes the learnable parameters of MobileFaceNet. The detailed architecture of MobileFaceNet used in our method is given in Table 1.

2.3. Predictor

To predict the match-mismatch result $\mathbf{z} \in \mathbb{R}^{1\times 2}$, we apply a simple linear layer as our predictor for match-mismatch prediction, expressed as follows

$$\mathbf{z} = \text{softmax}(\mathbf{W}\hat{\mathbf{X}}_F + \mathbf{b}), \tag{9}$$

where $\mathbf{W}$ and $\mathbf{b}$ denote the weight matrix and bias of the linear layer, respectively.

3. Experiments and Results

3.1. Experimental Setup

Dataset: We evaluate our method on the dataset of Auditory EEG Decoding Challenge 2023 [13], which collects EEG response data from 85 young participants with normal auditory systems. All these participants are native Dutch speakers. The speech stimuli are stories narrated by a single speaker whose native language is Flemish (Belgian Dutch). The training set contains the EEG responses of 71 participants, and the EEG responses of the remaining 14 participants are used as the evaluation set named **held-out subjects**. In addition, the EEG responses of these 71 participants, obtained by hearing new stories narrated by the speaker, are leveraged as another evaluation set named **held-out stories**. In our experiments, the evaluation sets for **held-out subjects** and **held-out stories** are consistent with the evaluation sets used in Auditory EEG Decoding Challenge 2023.

Table 1: *MobileFaceNet architecture used for feature embedding, where ef denotes the expansion factor, c is the number of output channels per layer, n represents the number of times each row operation is repeated, st is the stride.*

Input	Operator	ef	c	n	st
28×192 × 4	conv3x3	-	64	1	2
14× 96× 64	depthwise conv3x3	-	64	1	1
14× 96× 64	bottleneck	2	64	5	2
7× 48× 64	bottleneck	4	128	1	2
4× 24× 128	bottleneck	2	128	6	1
4× 24× 128	bottleneck	4	128	1	2
2× 12× 128	bottleneck	2	128	2	1
2× 12× 128	conv1x1	-	512	1	1
2× 12× 512	linear GDC conv7x7	-	512	1	1
1× 1× 512	linear conv1x1	-	128	1	1

Evaluation Metrics: To evaluate our method, the average accuracy Acc_s of the model's correct matched classification for subject s is calculated as follows

$$Acc_s = \textstyle\sum_{i=0}^{n_s}[label_{predicted} = label_{true}]/n_s, \tag{10}$$

where n_s represents the number of samples. Then, the average accuracy of held-out subjects S_1, the average accuracy of held-out stories S_2 and the final overall score $Score$ can be denoted as

$$S_1 = \frac{1}{14}\sum_{s=72}^{85} ACC_s, \tag{11}$$

$$S_2 = \frac{1}{71}\sum_{s=1}^{71} ACC_s, \tag{12}$$

$$Score = \frac{2}{3}S_1 + \frac{1}{3}S_2, \tag{13}$$

where s denotes the subject identity number.

Hyperparameters: All parameters in our model are optimized using the Adam optimizer [15] with a learning rate of 1e-3 and a batch size of 64.

Table 2: *Performance comparison with different methods, where held-out stories(%), held-out subjects(%), and score(%) are used for evaluation.*

Methods	Held-out stories	Held-out subjects	Score
Baseline [6]	77.98	78.55	78.17
A-SM [5]	79.98	78.17	79.38
Top-1 [16]	**82.71**	**80.98**	**82.13**
Top-2 [17]	79.61	77.93	79.05
Top-3 [18]	79.21	78.40	78.94
IFE-CF (Ours)	<u>80.82</u>	<u>80.48</u>	<u>80.71</u>

3.2. Performance Comparison

To show the effectiveness of our IFE-CF, we compare our IFE-CF with official baseline method (i.e., official Baseline [6]), the top three ranking systems of Auditory EEG Decoding Challenge 2023 (i.e., Top-1 [16], Top-2 [17] and Top-3 [18]), and

a recent state-of-the-art method (i.e., A-SM [5]). The results are given in Table 2.

Here, *held-out stories* represents the average accuracy of the model on the evaluation set for known subjects receiving unknown speech stimuli, and *held-out subjects* denotes the average accuracy of the model on the evaluation set for unknown subjects. The *Score* is the overall performance of the model on the evaluation set, which is the weighted sum of the average accuracy of the two parts.

Table 3: *Results of ablation study, where held-out stories, held-out subjects, and score are used for evaluation.*

Methods	Held-out stories (%)	Held-out subjects (%)	Score (%)
IFE-CF(w/o -D)	78.67	77.60	78.31
IFE-CF(w/o -T)	79.60	80.32	79.84
IFE-SF	78.90	79.02	78.94
IFE-CF	**80.82**	**80.48**	**80.71**

Table 2 shows that our method outperforms the official baseline method and the recent state-of-the-art method while surpassing the Top-2 and Top-3 systems in Auditory EEG Decoding Challenge 2023, which is only slightly lower than the Top-1 system. Note that the Top-1 system is an ensemble system that integrates several systems with ensemble learning [19], whereas our proposed method is the only single system that achieved such performance for this task. The results demonstrate the effectiveness and superiority of our proposed method.

3.3. Ablation Study

To validate the effectiveness of the two-branch structure to exploit the interaction feature and independent feature from two modalities (i.e., speech stimulus and EEG response), we conduct an ablation study. Here, IFE-CF(w/o -D) and IFE-CF(w/o -T) represent our IFE-CF without independent feature and interaction feature in the fusing processing of multi-channel fusion module, respectively. To better analyze the importance of the interaction and independent features while avoiding the influence of introducing a large number of learnable parameters in the attention calculation process, we replace the crossmodal attention mechanism with a self-attention mechanism in the EEG and speech causal mask crossmodal attention blocks, which is denoted as IFE-SF. Results are given in Table 3.

From Table 3, we can see that the performance of our method IFE-CF without independent feature (i.e., IFE-CF(w/o -D) drops more significantly than that of IFE-CF without interaction feature (i.e., IFE-CF(w/o -T)) on all metrics (i.e., **held-out subjects** and **held-out stories**), which indicates that the independent feature contains more useful information that is not included in the interaction feature. Therefore, the classification performance can be enhanced by introducing the independent feature of speech stimulus and EEG response.

In addition, both IFE-CF without crossmodal attention mechanism (i.e., IFE-SF) and IFE-CF without interaction feature (i.e., IFE-CF(w/o -T)) do not consider the relationship between speech stimulus and EEG response in the encoding process. As a result, they perform worse than IFE-CF on evaluation metrics. Moreover, the performance of IFE-SF is even worse than that of IFE-CF(w/o -T) in terms of all metrics. This indicates that without considering the relationship between speech stimulus and EEG response, simply introducing more learnable parameters or concatenating more independent features decreases the performance due to information redundancy.

Therefore, our IFE-CF can achieve better performance by fusing independent and interaction features in the multi-channel fusion module. This demonstrates the effectiveness of simultaneously exploring the relationship between these two modalities with interaction features and the specific characteristics of independent features from each modality.

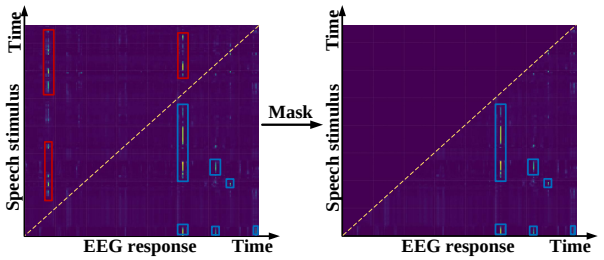

Figure 2: *An example of applying causal mask in speech causal mask crossmodal attention (SMCA) block*

Table 4: *Ablation Study on Causal Mask, where held-out stories, held-out subjects, and score are used for evaluation.*

Methods	Held-out stories (%)	Held-out subjects (%)	Score (%)
IFE-CF(w/o -M)	80.06	**80.53**	80.22
IFE-CF	**80.82**	80.48	**80.71**

3.4. Analysis of Causal Mask

We visually analyze the time delay characteristics of matched speech-EEG pair by providing the heatmaps of the attention maps before and after applying causal mask during crossmodal attention calculation. The results are given in Figure 2.

Due to the time delay that exists between the speech-EEG pair, there is no dependency between the speech stimulus at a given moment and the EEG response preceding this moment. Therefore, by applying our causal mask strategy, the unreasonable parts (i.e., in red rectangles) are masked, while the reasonable parts that conform to the time delay (i.e., in blue rectangles) are retained, as shown in Figure 2.

To further demonstrate the effectiveness of our causal mask strategy, we conduct the another ablative experiment. The results are given in Table 4. Here, IFE-CF(w/o -M) denotes our IFE-CF without using causal mask strategy in the crossmodel blocks. We can see that the performance of IFE-CF(w/o -M) is worse than that of IFE-CF, which illustrates that highlighting the meaningful part of the crossmodal attention map by considering the time delay between EEG response and speech stimulus with causal mask strategy can further improve the match-mismatch classification performance.

4. Conclusion

In this paper, we have presented an independent feature enhanced crossmodal fusion model for match-mismatch classification. The crossmodal attention mechanism is utilized to build the connection between speech and EEG encoding process to capture the relationship between these two modalities. Furthermore, these modalities' independent and interaction features are fused in a multi-channel fusion module to explore the relationship between the two modalities and the specific characteristics of independent features from each modality, simultaneously. In addition, a causal mask strategy is introduced to account for the time delay of the matching speech-EEG pair, further improving the classification performance. Experimental results demonstrate the effectiveness and the superiority of our method.

5. References

[1] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based brain–computer interface paradigms," *Journal of Neural Engineering*, vol. 16, no. 1, p. 011001, 2019.

[2] M. J. Monesi, B. Accou, J. Montoya-Martinez, T. Francart, and H. Van Hamme, "An LSTM based architecture to relate speech stimulus to EEG," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 941–945.

[3] C. Puffay, J. Van Canneyt, J. Vanthornhout, H. Van Hamme, and T. Francart, "Relating the fundamental frequency of speech with EEG using a dilated convolutional network," *arXiv preprint arXiv:2207.01963*, 2022.

[4] Y. Bai, X. Wang, Y.-p. Cao, Y. Ge, C. Yuan, and Y. Shan, "Dreamdiffusion: Generating high-quality images from brain EEG signals," *arXiv preprint arXiv:2306.16934*, 2023.

[5] M. Borsdorf, S. Cai, S. Pahuja, D. De Silva, H. Li, and T. Schultz, "Attention and Sequence Modeling for match-mismatch classification of speech stimulus and EEG response," *IEEE Open Journal of Signal Processing*, 2023.

[6] B. Accou, M. J. Monesi, J. Montoya, T. Francart *et al.*, "Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network," in *Proceedings of European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1175–1179.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] M. J. Monesi, B. Accou, T. Francart, and H. Van Hamme, "Extracting different levels of speech information from EEG using an LSTM-based model," *arXiv preprint arXiv:2106.09622*, 2021.

[9] A. Soman, V. Sinha, and S. Ganapathy, "Enhancing the EEG speech match mismatch tasks with word boundaries," *arXiv preprint arXiv:2307.00366*, 2023.

[10] F. Yu and V. Koltun, "Multi-Scale context aggregation by Dilated Convolutions," in *Proceedings of International Conference on Learning Representations (ICLR)*, Nov 2016.

[11] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[12] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for unaligned multimodal language sequences," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 2019. NIH Public Access, 2019, p. 6558.

[13] B. Accou, L. Bollens, M. Gillis, W. Verheijen, H. Van hamme, and T. Francart, "Sparrkulee: A Speech-evoked auditory response repository of the KU Leuven, containing EEG of 85 participants," *bioRxiv*, pp. 2023–07, 2023.

[14] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proceedings of Chinese Conference on Biometric Recognition (CCBR)*. Springer, 2018, pp. 428–438.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2014.

[16] M. Thornton, D. Mandic, and T. Reichenbach, "Relating EEG recordings to speech using envelope tracking and the speech-FFR," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–2.

[17] M. Borsdorf, S. Pahuja, G. Ivucic, S. Cai, H. Li, and T. Schultz, "Multi-Head Attention and GRU for improved Match-Mismatch classification of speech stimulus and EEG response," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–2.

[18] F. Cui, L. Guo, L. He, J. Liu, E. Pei, Y. Wang, and D. Jiang, "Relate auditory speech to EEG by shallow-deep attention-based network," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–2.

[19] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.