Received 08 August, 2023; revised 20 November, 2023; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version 20 November, 2023.

Digital Object Identifier 10.1109/XXXX.XXXX.XXXXXXX

Attention and Sequence Modeling for Match-Mismatch Classification of Speech Stimulus and EEG Response

Marvin Borsdorf[1], Siqi Cai[2], Member, IEEE, Saurav Pahuja[1], Dashanka De Silva[1], Haizhou Li[3,1,2], Fellow, IEEE, and Tanja Schultz[4], Fellow, IEEE

[1]Machine Listening Lab (MLL), University of Bremen, Germany
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[3]Shenzhen Research Institute of Big Data, School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China
[4]Cognitive Systems Lab (CSL), University of Bremen, Germany

(Invited Paper)

Corresponding author: Marvin Borsdorf (email: marvin.borsdorf@uni-bremen.de).

This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen)

ABSTRACT For the development of neuro-steered hearing aids, it is important to study the relationship between a speech stimulus and the elicited EEG response of a human listener. The recent Auditory EEG Decoding Challenge 2023 (Signal Processing Grand Challenge, IEEE International Conference on Acoustics, Speech and Signal Processing) dealt with this relationship in the context of a match-mismatch classification task. The challenge's task was to find the speech stimulus that elicited a specific EEG response from two given speech stimuli. Participating in the challenge, we adopted the challenge's baseline model and explored an attention encoder to replace the spatial convolution in the EEG processing pipeline, as well as additional sequence modeling methods based on RNN, LSTM, and GRU to preprocess the speech stimuli. We compared speech envelopes and mel-spectrograms as two different types of input speech stimulus and evaluated our models on a test set as well as held-out stories and held-out subjects benchmark sets. In this work, we show that the mel-spectrograms generally offer better results. Replacing the spatial convolution with an attention encoder helps to capture better spatial and temporal information in the EEG response. Additionally, the sequence modeling methods can further enhance the performance, when mel-spectrograms are used. Consequently, both lead to higher performances on the test set and held-out stories benchmark set. Our best model outperforms the baseline by 1.91 % on the test set and 1.35 % on the total ranking score. We ranked second in the challenge.

INDEX TERMS Auditory system, EEG decoding, match-mismatch classification, speech envelope, speech stimulus

I. INTRODUCTION

THE human brain possesses the extraordinary ability to selectively focus on a specific speaker's voice in a crowded scenario, commonly called "the cocktail party effect" [1], [2]. Recent advancements made in fields such as psychoacoustics, biophysiology, and neurosciences have provided valuable insights into the mechanisms underlying auditory attention in the human brain. Notably, recent studies have demonstrated that auditory attention can be decoded from recordings of brain activity, such as electrocorticography (ECoG) [3], magnetoencephalography (MEG) [4], and electroencephalography (EEG) [5], [6] in a cocktail party scenario.

While the selective listening ability can be performed easily and without much effort, people suffering from hearing loss experience severe difficulties in performing this task.

Over the past decade, there has been remarkable progress in equipping machines with this ability, paving the way for being integrated into hearing aids. Those algorithms usually rely on a reference cue of the to-be-extracted target speech signal, commonly given as speech signal [7]–[12], or based on a different modality such as face [13]–[15], text [16], [17], or even gesture [18] information. However, in real-world conversational situations, it might be hard to acquire those cues and the quality may also vary due to occlusion, changes in the light setting, body movement, or interfering signals. This makes it really hard to select the speaker of interest. The EEG signal of a human listener, on the other hand, is connected to the auditory system and directly responds when attention is paid to incoming sounds. This relationship makes the EEG signal quite suitable for being used as a reference cue to steer hearing aids. Furthermore, in the realm of brain-computer interfaces (BCIs), EEG has garnered significant attention due to its non-invasive nature, cost-effectiveness, ease of use, and ability to provide high temporal resolution [19]. These advantages have positioned EEG as a highly suitable modality [5], facilitating its integration into various domains.

For the development of neuro-steered hearing aids, it is crucial to understand the fundamental relationship between a speech stimulus and the EEG response of a human listener attending to the speech signal. The recent Auditory EEG Decoding Challenge 2023[1] (Signal Processing Grand Challenge, IEEE International Conference on Acoustics, Speech and Signal Processing, 2023) dealt with this relationship. One of the challenge's tasks was designed as a match-mismatch classification task for speech stimulus and EEG response in a single-talker scenario. To be more precise, given two speech stimuli and one EEG response, the task was to classify which stimulus elicited the EEG response (see Fig. 1). The match-mismatch task holds great promise for advancements in hearing aid technology through the implementation of cognitive control [20], [21].

The match-mismatch classification task can be considered as an important aspect to be addressed in the emerging field of auditory attention detection (AAD). Recently, AAD has opened up new opportunities for the cognitive control of hearing aids, known as neuro-steered hearing aids [21]. To address EEG-based AAD tasks, various approaches, and techniques have been developed, which can be explored in detail in a comprehensive review by Geirnaert et al. [22].

In this paper, we focus on the match-mismatch classification task in the single-talker scenario, where we match an EEG response with given speech stimuli, while AAD aims at identifying attention when listening to multiple speech stimuli simultaneously. We present our approach to solve the match-mismatch classification task offered by the Auditory EEG Decoding Challenge, 2023. We start from the challenge's baseline model [23] in which the feature extraction for both speech data and EEG data is purely based on convolutional methods (see Section IV).

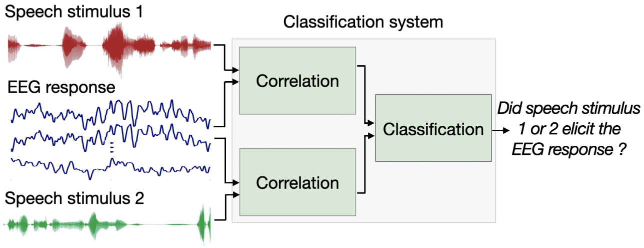

FIGURE 1. **Block diagram of the match-mismatch classification task of the Auditory EEG Decoding Challenge, 2023. The classification system receives an EEG response as well as two speech stimuli and predicts whether speech stimulus 1 or 2 has elicited the EEG response.**

For speech processing tasks, sequence modeling approaches based on recurrent neural networks (RNNs), long short-term memory (LSTM), and gated recurrent units (GRUs) have shown to be effective [24]–[33]. Recent works which studied the relationship between speech stimulus and EEG response successfully applied LSTM-based methods to process the speech data [34], [35]. Therefore, we study the integration of RNN, LSTM, and GRU into the processing pipeline of the speech stimulus. In addition, we consider the bidirectional variants of those, referred to as BiRNN, BiLSTM, and BiGRU. Due to the success of attention-based decoding of EEG signals [36]–[41], we study to replace the spatial convolution in the EEG processing pipeline with an attention encoder (AE). In addition, the baseline model processes the speech stimulus given as a speech envelope. The speech envelope lacks crucial features of the original speech stimulus, such as formants, prosody, and pitch [42]. Recently, the mel-spectrogram representation has been shown to be beneficial for the match-mismatch classification of speech stimulus and EEG response in the single-talker task [35]. Consequently, we compare the classification performance for speech envelope and mel-spectrogram as input speech stimulus types (SSTs). In a comprehensive study, we train and evaluate nine different models, respectively for the speech envelope and for the mel-spectrogram as input SST, leading to eighteen models in total. We evaluate the performance on a test set and two benchmark sets (held-out stories and held-out subjects).

Our experimental results show that using mel-spectrograms instead of speech envelopes as input SST consistently improves the performance on the test set and mostly on both benchmark sets. The application of additional sequence modeling methods such as RNN, LSTM, and GRU benefits from the use of mel-spectrograms and shows enhanced performance on both the test set and the held-out stories benchmark set. Applying an AE to extract the spatial and temporal features of the EEG data boosts the performance on the test set and the held-out stories benchmark set. Our best model shows improvements of 1.91 % accuracy on the test set and 1.35 % accuracy on the total ranking score (weighted sum of both benchmark sets) compared to the baseline model. This work represents

[1] https://exporl.github.io/auditory-eeg-challenge-2023/

an extension of our previously published Auditory EEG Decoding Challenge conference paper [43]. We study additional sequence modeling methods for processing the speech stimulus signal and delve into our results in a more detailed analysis to gain better insights.

The rest of the paper is organized as follows: Section II provides an overview of the related work. Section III describes the data set used in this study and provides details on the applied preprocessing techniques. Section IV explains the baseline architecture and our extension. Section V presents the experimental setup. Section VI illustrates and discusses the results. Finally, Section VII concludes the study and highlights some future work. We make all information and scripts available[2].

II. RELATED WORK

As highlighted in the introduction, relating single-talker speech to the elicited EEG response of a human listener plays a critical role in developing BCI applications. This has been comprehensively discussed in a recent review paper by Puffay et al. [44]. Notably, convolutional-based methods [23], [45] as well as sequence modelling techniques [34], [35] have been proposed to solve this task. Alongside the match-mismatch classification, decoding of the speech is a related task that aims at estimating a single-talker speech signal based on a neural response [46]–[48]. As already described, the single-talker match-mismatch classification task represents an aspect of the widely studied AAD task.

In the field of EEG-based AAD, various methods have been developed to interpret EEG signals and determine auditory attention. Traditional AAD approaches primarily focused on decoding the speech envelope of the attended speaker. These methods relied on clean speech signals and involved comparing reconstructed speech with individual sources in multi-speaker environments [5]. In line with the concept of stimulus reconstruction, various linear decoders have been devised for AAD tasks [49], [50]. However, the correlation between the reconstructed speech envelope and the attended speech is generally weak, which may be attributed to the oversimplified linear computational model.

Considering the inherent non-linear processing of acoustic signals along the auditory pathway [51], Taillez et al. [6] firstly studied a non-linear neural network to map EEG signals to speech envelopes in a cocktail party scenario, that outperforms the linear model baseline. Recently, convolutional neural network (CNN) models [52]–[54] were studied to detect the attended speakers directly from the EEG and audio signals. However, these AAD models mentioned above predominantly utilize CNN architectures, which may not be the most suitable option for capturing the temporal dynamics of brain signals.

Given that speech stimuli and neural responses are inherently time-varying processes, RNNs have demonstrated their effectiveness in capturing the temporal dynamics of EEG signals in numerous studies [55]. Therefore, it is worthwhile to explore the potential benefits of employing RNN-based architectures for AAD.

The recent success of Transformers [56] in various research areas has also made its way into the field of EEG processing. They have been used to decode the EEG signal, e.g., for motor imagery [36]–[38], [40], human brain-visual image classification [37], [39], steady-state visual evoked potential analysis [41], and emotion recognition [40].

To sum up, we propose a model that combines attention and sequence modeling with dilation-based convolutional layers to capture the spatial and temporal dynamics in both the speech stimulus and the EEG response to improve the match-mismatch classification performance.

III. DATA SET

The data set [57], [58] used in the Auditory EEG Decoding Challenge comprises EEG data that was obtained from 85 young adults who are Dutch native speakers and have normal hearing capabilities. Throughout the experiment, each participant engaged in approximately 8 to 10 trials, with each trial lasting around 15 minutes. To avoid any potential biases and promote diversity, the order of the trials was randomized for each participant. The stimuli utilized in this experiment consisted of single-speaker stories presented in Flemish (Belgian Dutch), narrated by a native Flemish speaker. The stimuli comprised podcasts and audiobooks, with some audiobooks exceeding the 15-minute duration. In such cases, the longer audiobooks were divided into two consecutive trials, ensuring continuity for the participants. The data set provides the narrated stories that were presented to the subjects, referred to as speech stimuli, and the EEG responses of the subjects acquired while listening. In total, the data set comprises 157 hours of parallel data, i.e., speech stimuli and EEG responses. In the following, we will describe the preprocessing of the data as well as the construction of both the development set and the benchmark set.

A. SPEECH STIMULUS

Each speech stimulus is given as a story (podcast or audiobook), narrated by a native Flemish speaker, and recorded with a sampling rate of 48 kHz. The match-mismatch classification task initially uses the speech envelope of the original speech stimulus, according to the following processing steps [48]: First, a gammatone filterbank with 28 subbands is applied to the raw audio signal. The bandwidth is equally spaced with center frequencies of 50 Hz to 5 kHz. Second, the absolute value is calculated for each sample and exponentiated with 0.6. Third, the speech envelope is obtained by averaging all subbands. Fourth, the speech envelope is downsampled to 64 Hz.

The speech envelope is a less complex representation of the raw speech signal and, due to its slow rate of change, it is easy to handle and to process. However, the speech

[2] https://github.com/mborsdorf/ICASSP2023SPGC_AuditoryEEG

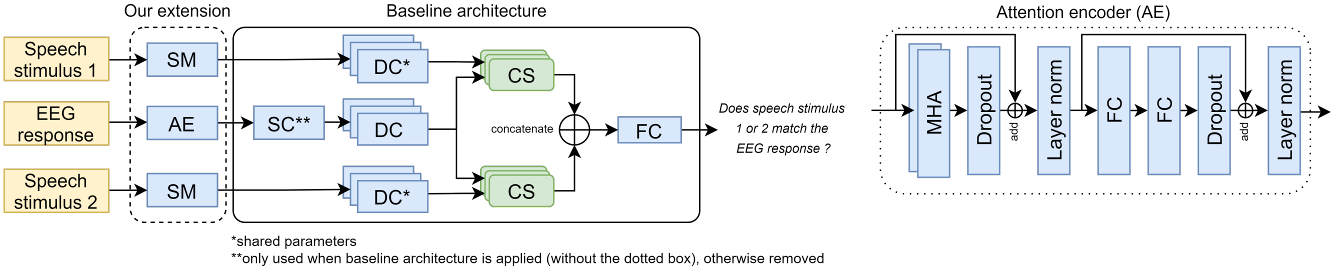

FIGURE 2. The model architecture used in our experiments. The box with the solid line shows the baseline architecture [23]. The spatial convolution (SC) is only used in the baseline architecture and removed in all other models that adopt the baseline architecture. The dashed box shows our model extension. We add sequence modeling (SM) methods to preprocess the speech stimuli. The SM is based on RNN, LSTM, GRU, BiRNN, BiLSTM, or BiGRU, respectively. The construction of the AE is shown in the dotted box.

envelope lacks of important speech characteristics such as formants, prosody, and pitch [42]. Therefore, we use the mel-spectrogram representation of the speech stimulus as the input signal. The mel-spectrogram displays the speech signal in the time-frequency domain using a non-linear scale based on human auditory perception. This provides a more informative signal representation to the model. The calculation of the mel-spectrogram follows the preprocessing carried out by the challenge organizers to ensure compatibility between the data sets. Accordingly, the following preprocessing steps are performed: First, the DC component of the speech stimulus signal is removed. Second, a fast Fourier transform (FFT) is applied with a hop length of 750 samples (15.625 ms) and a Hamming window of 1,200 samples (25 ms) length. To match a total FFT length of 2,048 samples, the Hamming window is padded with zeros. In total, 28 mel-frequency bands are applied with a minimum center frequency of -4.2735 Hz and a maximum center frequency of 5,444 Hz. Third, the resulting mel-spectrogram is exponentiated with 0.6.

B. EEG RESPONSE

The EEG data was recorded with a sampling frequency of 8,192 Hz in a soundproof and electromagnetically shielded booth, using a high-quality 64-channel Biosemi ActiveTwo EEG recording system. All 64 active Ag-AgCl electrodes were placed according to international 10-20 standards. Before utilization, the EEG data is preprocessed as follows [48]: First, a 1st order Butterworth high-pass filter with a cut-off frequency of 0.5 Hz is applied to the data. The filtering is done in both forward direction and backward direction to enable zero-phase filtering. Second, the data is downsampled to 1,024 Hz. Third, artifacts due to eye-blink are removed by using a multi-channel Wiener filter [59]. Fourth, the EEG data is re-referenced to a common average. Fifth, the data is downsampled to a frequency of 64 Hz.

C. DATA SPLITTING

The data set is split up into development data and benchmark data. The development data comprises 71 out of the 85 subjects. The remaining 14 subjects are reserved for benchmarking only. The development data is split up into training, validation, and test sets with 80 %, 10 %, and 10 % of the data, respectively. All sets share the same 71 subjects.

During the Auditory EEG Decoding Challenge, the benchmark data was used to compare the models of the different competing teams. We use this data to compare our models with each other. The benchmark data is split up into two sets. Benchmark set 1 contains 70 subjects that are also part of the development data, but with different stories to determine the performance for held-out stories. We only have 70 subjects in the benchmark set 1, as no benchmark data is available for subject 1. As the test set of the development data contains 10 % of each story, it is also considered as held-out stories condition, similar to benchmark set 1. Even though both study the same condition, the stories are different. Therefore, we keep both, as we have more data for our analysis. Benchmark set 2 consists of the 14 separated subjects but with the same stories as in the development data to determine the performance for held-out subjects.

IV. MODEL ARCHITECTURE

Our solution starts from the challenge's baseline architecture [23], illustrated in Fig. 2. The baseline model receives the EEG response as well as two speech stimuli of which one is the matching signal, that elicited the EEG response, and one is the mismatching signal. The EEG signal is first processed by spatial convolution (SC) to process temporal and spatial features. The SC is implemented as a convolutional 1D layer with 8 filters and a kernel size of 1. Subsequently, the signal is fed into a stacked block of three dilated convolution (DC) layers. Each layer has 16 filters, a kernel size of 3, a stride of 1, and a dilation rate that grows with the kernel size exponentiated by the respective layer index. The speech stimuli are directly processed by stacked DC blocks that share the parameters. The DC blocks have the same settings as the DC block which processes the EEG signal. The processed EEG signal is compared with each processed speech stimulus by calculating the respective cosine similarity (CS) as follows:

$$cos(\theta) = \frac{E \cdot S}{||E|| \, ||S||} \tag{1}$$

with E being the EEG signal representation, S being one of the speech stimulus representations, " $\cdot$ " being the dot product, and "$||\ ||$" being the norm. The cosine similarity value ranges in the interval of $[-1, 1]$. The closer the value is to 1, the more similar E and S are. Both results are concatenated and fed into a fully connected (FC) layer with a single neuron and Sigmoid activation function. The FC layer's output yields the prediction about which of the speech stimuli has elicited the EEG response.

We adopt the general baseline architecture but replace the SC with an attention encoder (AE). The AE consists of a multi-head attention (MHA) layer with two attention heads and an embedding dimension of 64, followed by a dropout(0.5) layer, layer norm, two FC layers with 32 and 64 neurons, respectively, dropout(0.5), and a final layer norm. The structure is shown in Fig. 2. The AE operates across the EEG channel dimension. As described in Section I, RNNs are shown to be effective for speech processing tasks. Therefore, we add an RNN to extract features from the sequential speech stimulus data. For a comprehensive study, we also investigate variants of the RNN architecture, namely LSTM, GRU, BiRNN, BiLSTM, and BiGRU. This leads to seven new architectures in total. The model architectures are slightly changed according to the number of channels of the respective input speech stimulus type (SST) (1 channel for speech envelope and 28 channels for mel-spectrogram).

V. EXPERIMENTAL SETUP

In our experiments, we develop seven new model architectures and compare them to the challenge baseline architecture [23]. The experiments and models are implemented in Python using the Tensorflow-Keras framework. In the following, we describe the details of how the experiments are conducted as well as the metrics used for testing and benchmarking.

A. IMPLEMENTATION DETAILS

To construct the new models, we adopt the baseline architecture except for the SC. Instead, we add an AE into the processing pipeline of the EEG response. Furthermore, we add a sequence modeling block, based on RNN, LSTM, GRU, BiRNN, BiLSTM, or BiGRU, respectively, to the beginning of the speech stimulus processing pipeline. The input speech stimulus type is given as either a speech envelope or a mel-spectrogram.

The data fed to the models is constructed as tuples. Each tuple consists of an EEG response, two speech stimuli, and the true class label of which stimulus has elicited the EEG response (label "0": speech stimulus 1; label "1": speech stimulus 2). The tuples are constructed as follows: First, a three-second long EEG segment is obtained (192 samples). Second, the respective three-second long matching speech stimulus segment is extracted (either speech envelope or mel-spectrogram). Third, a three-second long mismatching speech stimulus segment is chosen randomly, either starting four seconds before the matching segment starts or one second after the matching segment ends. This shift ensures a one-second long distance between both speech stimuli segments, i.e., they are different but temporally close. The data generator uses each EEG segment twice and the position of the matching and mismatching speech stimuli inside the data tuple permute. Consequently, the true class label changes as well.

1) (EEG response, matching speech stimulus, mismatching speech stimulus, label: "0")
2) (EEG response, mismatching speech stimulus, matching speech stimulus, label: "1")

Following this method naturally doubles the batch size. In addition, the data generation makes sure that mismatching speech stimulus segments are also matching segments to EEG responses in other data tuples. This algorithm helps the model to better generalize to the test and benchmark data.

During training, the model receives the input data and predicts whether speech stimulus 1 or speech stimulus 2 has elicited the EEG response. We apply Adam [60] as the optimizer and use a learning rate scheduler (LRS) which divides the learning rate by ten, if the validation loss does not decrease within two subsequent epochs. Each model is trained for a maximum of 100 epochs, but we stop the training if there is no improvement in the validation loss within six consecutive epochs. The batch size fed to the model is 128 after using the batch duplication method as described above. During training, we apply the binary cross-entropy as loss function:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{n=1}^{N}[Y_n log(\hat{Y}_n) + (1 - Y_n) log(1 - \hat{Y}_n)] \tag{2}$$

with $\hat{Y}$ being the probability for the predicted class, Y being the true class label $\{0,1\}$ (ground truth), and N being the number of samples.

B. EVALUATION

The models are evaluated on the test set and the two benchmark sets (held-out stories and held-out subjects). The results on the test set are reported as mean BCE loss and mean accuracy (%). To obtain the results for both the held-out stories and the held-out subjects benchmark sets, we first calculate the mean accuracy for each subject in both benchmark sets as follows:

$$Acc_s = \frac{1}{n_s}\sum_{i=0}^{n_s}[\hat{Y}_i == Y_i] \tag{3}$$

with s being the subject number, n_s being the number of samples for each subject, and $\hat{Y}_i$ and Y_i being the model's prediction and the true class label, respectively. Next, we determine the mean accuracy for the held-out stories set (Eq. 4) and the held-out subjects (Eq. 5) set, respectively. In addition, we calculate the total ranking score (TRS) as the weighted sum of both benchmark scores (Eq. 6). The TRS is used to rank the different models in the scope of the Auditory EEG Decoding Challenge. Since the EEG responses, speech stimuli, and speech envelopes for benchmarking are already cut to a length of 3 seconds, the speech stimuli are padded with zeros when calculating the mel-spectrograms such that the latter corresponds to a duration of 3 seconds.

$$S_1 = \frac{1}{70}\sum_{s=1}^{70} Acc_s \tag{4}$$

$$S_2 = \frac{1}{14}\sum_{s=72}^{85} Acc_s \tag{5}$$

$$TRS = \frac{2}{3}S_1 + \frac{1}{3}S_2 \tag{6}$$

VI. RESULTS AND DISCUSSION

In total, we trained eighteen models and evaluated their match-mismatch classification performance. Table 1 shows the results on the test set. The results on the benchmark sets are given in Table 2 and, to provide a better comparison, as box plots in Fig. 3 and 4. We conduct statistical analyses using the IBM SPSS statistics software with a significance level of 0.05. Descriptive statistics, including means and standard deviations, are utilized to summarize the data. To evaluate potential significant improvements between different models, we perform non-parametric analyses using the Wilcoxon signed-rank test with Bonferroni correction. In addition, we create topographic maps corresponding to the EEG electrodes based on the EEG response data for some selected models (Fig. 5). We extract the attention weights from the AE. Since the model is trained on 71 subjects, the attention weights can be considered as mean across them. All 64 EEG electrodes are represented as black dots. The attention weighting is illustrated using color gradients, with the red shade indicating a higher weight.

A. MATCH-MISMATCH CLASSIFICATION

1) Test set

Models 1 and 10 provide the baseline results on the test set for speech envelope and mel-spectrogram as input SST, respectively (Table 1). Adjusting the learning rate during training based on an LRS helps to train the model better, leading to an improved classification performance (models 2 and 11). Replacing the SC with an AE increases the performance further, showing the benefits of this method to preprocess the temporal and spatial features in the EEG response signal (models 3 and 12).

Applying an additional sequence modeling method to preprocess the speech stimulus does not show to be beneficial when using the speech envelope as input SST because models 4-9 show less performance compared to model 3. Besides model 3, only models 4 and 9 perform better than the baseline (model 2), highlighting that a larger model with respect to the number of parameters does not necessarily lead to better classification performance. Two of the bidirectional models can outperform their counterparts by a small margin (models 8 and 9).

When mel-spectrograms are given as input SST, the number of parameters slightly increases due to a change from 1 channel to 28 channels in the input speech data, and the performance improves for all models (models 10-18). This shows the effectiveness of working on a mel-spectrogram representation for this task. In addition, the sequence modeling methods mostly improve the performance further (models 14-18) compared to only adding the AE (model 12), showing their strength on mel-spectrograms as input SST. The results for the bidirectional models show that only the BiRNN (model 16) can outperform its non-bidirectional counterpart (model 13).

While model 3 performs best when using the speech envelope, model 14 shows the highest performance when using a mel-spectrogram as input SST, leading to an improvement of 1.91 % compared to the baseline model (model 11).

2) Benchmark set

The results on the benchmark set 1 (held-out stories), given in Table 2 and Fig. 3, are in line with the test set results, shown in Table 1. This is as expected, since both sets represent held-out stories conditions. The performance on both input SSTs can be improved if the SC is replaced with an AE (models 3 and 12). Applying additional sequence modeling methods is not beneficial and the models 5 and 8 are even worse than the baseline model (model 2). Similarly to our findings in the results for the test data, we see a clear advantage of using a mel-spectrogram as input SST, since the performance is generally higher. The only exception is given by model 11 which, surprisingly, shows a slightly lower performance compared to its counterpart model (model 2). All sequence modeling methods (models 13-18) outperform the baseline models as well as the model that uses an AE. The BiRNN (model 16) is the only method that shows a higher performance than its non-bidirectional counterpart. The highest performance when using a speech envelope is given by model 3. Model 14 shows the highest performance when working on a mel-spectrogram. Those are also the best performing models on the test set for the respective input SST.

Benchmark set 2 (held-out subjects), see Table 2 and Fig. 4, determines the generalizability of the models to

TABLE 1. Results for the match-mismatch classification task for different model architectures on the test data. We provide information about the input speech stimulus type (SST), the application of a learning rate scheduler (LRS), the composition of both the EEG and speech processing pipelines, and the total number of model parameters (P). We report the mean loss and mean accuracy on the test data for each model. A dagger (†) indicates the baseline architecture [23].

Model #	SST	LRS	EEG Branch	Speech Branch	P	Loss	Accuracy (%)
†1	Envelope	✗	SC+DC	DC	4,633	0.4959	75.25
†2	Envelope	✓	SC+DC	DC	4,633	0.4908	75.90
3	Envelope	✓	AE+DC	DC	44,465	0.4777	**76.78**
4	Envelope	✓	AE+DC	RNN+DC	44,468	0.4791	76.74
5	Envelope	✓	AE+DC	LSTM+DC	44,477	0.5102	74.47
6	Envelope	✓	AE+DC	GRU+DC	44,477	0.4937	75.74
7	Envelope	✓	AE+DC	BiRNN+DC	44,519	0.4892	75.83
8	Envelope	✓	AE+DC	BiLSTM+DC	44,537	0.5019	75.29
9	Envelope	✓	AE+DC	BiGRU+DC	44,537	0.4887	75.98
†10	Spectrogram	✗	SC+DC	DC	5,929	0.4860	76.04
†11	Spectrogram	✓	SC+DC	DC	5,929	0.4771	76.43
12	Spectrogram	✓	AE+DC	DC	45,761	0.4627	77.85
13	Spectrogram	✓	AE+DC	RNN+DC	47,357	0.4685	77.54
14	Spectrogram	✓	AE+DC	LSTM+DC	52,145	0.4577	**78.34**
15	Spectrogram	✓	AE+DC	GRU+DC	50,633	0.4603	78.04
16	Spectrogram	✓	AE+DC	BiRNN+DC	50,297	0.4616	77.87
17	Spectrogram	✓	AE+DC	BiLSTM+DC	59,873	0.4624	78.22
18	Spectrogram	✓	AE+DC	BiGRU+DC	56,849	0.4640	77.92

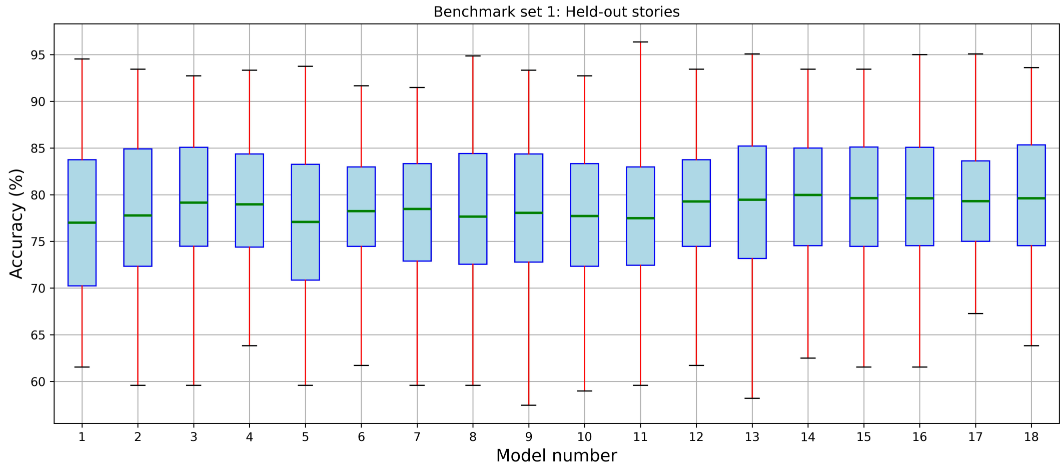

FIGURE 3. Mean accuracy (%) and standard deviation for each model on the held-out stories benchmark set 1, presented as box plots. The outliers are removed from this illustration.

unseen subjects, since this set of subjects is disjoint from the set of subjects in the training. Among all models, the baseline models (models 1, 2, 10, and 11) always perform best, independent of the input SST. All other models, especially the model that applies an AE and that has always outperformed the baseline models on the test set and benchmark set 1, show less performance.

In general, the attention mechanism has the ability to learn global as well as local patterns, which has leveraged the auditory attention domain already [61], showing substantial improvements in within-subject tasks. However, as the brain behavior between individuals differs, applying learned attention patterns to unseen subjects may face limitations. We believe that training a model on a representative group of subjects may improve the generalizability in this paradigm.

Comparing the sequence modeling methods to each other, we find that the RNN (model 4) and the LSTM (model 14) perform best. The highest classification performance when using a speech envelope and when using a mel-spectrogram as input SSTs are given by models 2 and 11, respectively.

TABLE 2. Results for the match-mismatch classification task for different model architectures on the benchmark data. We provide information about the input speech stimulus type (SST), the application of a learning rate scheduler (LRS), the composition of both the EEG and speech processing pipelines, and the total number of model parameters (P). For each model, we report the mean performance (μ) as well as the standard deviation (σ) over all subjects for both held-out stories and held-out subjects conditions. We denote statistically significant improvements for models 3-9 with respect to model 2 and for models 12-18 with respect to model 11 (with *p$<$0.001, after Bonferroni correction). In addition, we provide the total ranking score (TRS). A dagger (†) indicates the baseline architecture [23].

Model #	SST	LRS	EEG Branch	Speech Branch	P	Held-out stories (%)		Held-out subjects (%)		TRS (%)
						μ	σ	μ	σ	
†1	Envelope	✗	SC+DC	DC	4,633	77.02	8.920	77.56	5.221	77.20
†2	Envelope	✓	SC+DC	DC	4,633	77.78	8.255	**77.61**	5.533	77.73
3	Envelope	✓	AE+DC	DC	44,465	***79.17**	7.497	76.47	6.633	78.27
4	Envelope	✓	AE+DC	RNN+DC	44,468	*78.98	7.261	77.04	5.417	**78.33**
5	Envelope	✓	AE+DC	LSTM+DC	44,477	77.09	7.958	74.92	6.123	76.37
6	Envelope	✓	AE+DC	GRU+DC	44,477	78.25	7.235	75.05	8.097	77.18
7	Envelope	✓	AE+DC	BiRNN+DC	44,519	78.48	6.810	75.83	6.237	77.60
8	Envelope	✓	AE+DC	BiLSTM+DC	44,537	77.66	7.919	74.69	7.201	76.67
9	Envelope	✓	AE+DC	BiGRU+DC	44,537	78.06	7.443	75.09	8.321	77.07
†10	Spectrogram	✗	SC+DC	DC	5,929	77.72	7.929	78.58	5.598	78.01
†11	Spectrogram	✓	SC+DC	DC	5,929	77.49	7.948	**79.09**	5.594	78.03
12	Spectrogram	✓	AE+DC	DC	45,761	*79.28	7.655	77.81	6.859	78.79
13	Spectrogram	✓	AE+DC	RNN+DC	47,357	*79.47	8.041	76.64	7.783	78.53
14	Spectrogram	✓	AE+DC	LSTM+DC	52,145	***79.98**	7.182	78.17	6.534	**79.38**
15	Spectrogram	✓	AE+DC	GRU+DC	50,633	*79.63	7.054	77.95	7.688	79.07
16	Spectrogram	✓	AE+DC	BiRNN+DC	50,297	*79.62	7.685	77.80	7.181	79.01
17	Spectrogram	✓	AE+DC	BiLSTM+DC	59,873	*79.31	6.570	77.40	6.962	78.67
18	Spectrogram	✓	AE+DC	BiGRU+DC	56,849	*79.63	7.758	77.35	7.473	78.87

The TRS is given as a weighted sum of both benchmark sets (Eq. 6). When the speech envelope is used as input SST, the highest performance is attained by model 4 with a TRS of 78.33 %. When working on the mel-spectrogram, model 14 shows the best classification accuracy with a TRS of 79.38 %, outperforming the baseline model (model 11) by 1.35 %.

B. EEG FEATURE EXTRACTION ANALYSIS

We employ models with different processing pipelines for the speech stimulus, i.e., DC, RNN-DC, LSTM-DC, and GRU-DC, while the EEG processing pipeline remains unchanged. In this way, we anticipate to gain insights into the effects of different speech processing techniques and input SSTs on the EEG signal processing. We create topographic attention-based maps (Fig. 5) based on the weights in the attention layer of the AE. The attention weights are averaged over time and over the 71 training subjects. We compare the topographic maps of different models. In this way, we try to visualize and study the interplay of EEG channels in the domain of speech perception and speech processing for the match-mismatch detection.

The salient observation is that the emphasis is consistently on frontal channels such as F3, F1, Fp1, F6, Fpz, and F7. These channels, located in the anterior portions of the brain, play a pivotal role in higher-level cognitive tasks, attention mechanisms, and executive functions, especially during speech perception [62]. Also, the importance of centroparietal channels (such as C4, C5, C6, and C3) in models that process speech envelopes indicates their great importance in sensory information, an important factor for auditory processing. This correlates with previous studies that highlight the centrality of central parietal regions in auditory processing and the conception of continuous speech [63].

Unfortunately, the observed similarities in the topographic maps calculated from the attention weights make it difficult to obtain a general understanding of the underlying cognitive processes.

VII. CONCLUSION AND FUTURE WORK

In this paper, we worked on a solution for the match-mismatch classification of speech stimulus and EEG response. The study was done in the context of participation in the Auditory EEG Decoding Challenge 2023 (Signal Processing Grand Challenge, IEEE International Conference on Acoustics, Speech and Signal Processing, 2023).

We adopted the challenge's baseline model and revised three parts as follows: (i) We replaced the spatial convolution in the EEG processing pipeline with an attention encoder. (ii) We applied additional sequence modeling methods based on RNN, LSTM, GRU, BiRNN, BiLSTM, and BiGRU,

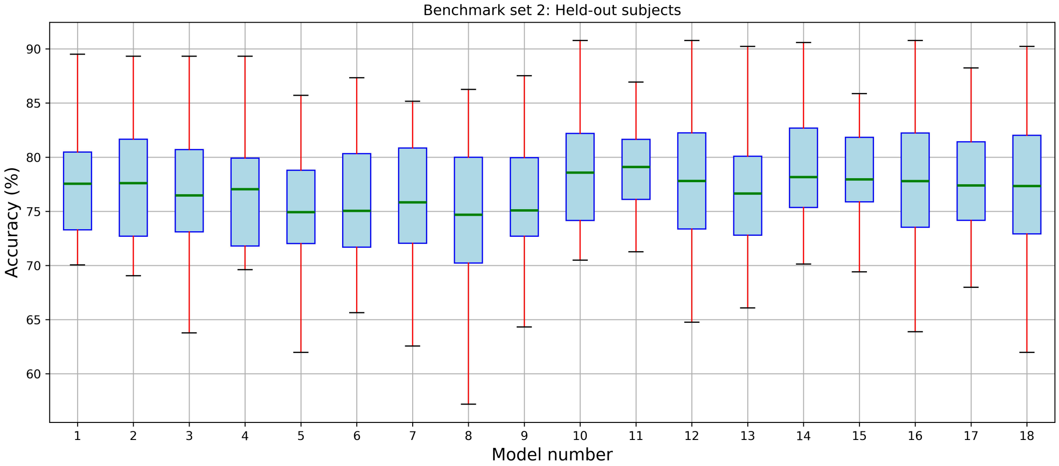

FIGURE 4. Mean accuracy (%) and standard deviation for each model on the held-out subjects benchmark set 2, presented as box plots. The outliers are removed from this illustration.

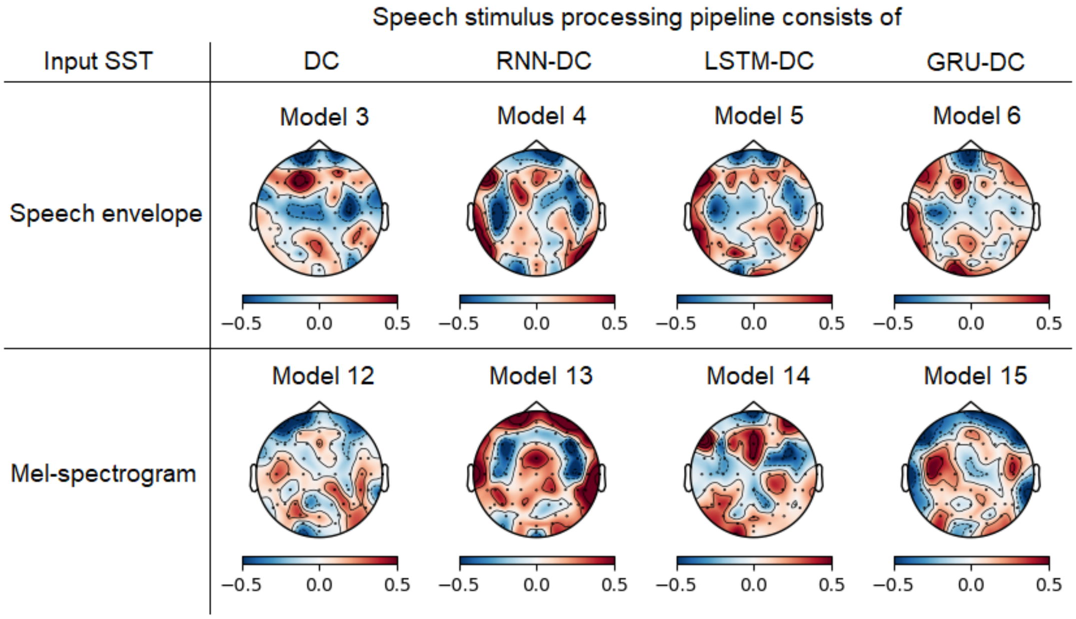

FIGURE 5. Topographic maps for different model architectures to compare the impact of (a) the input speech stimulus type (SST) and (b) the different methods in the speech stimulus processing pipeline. While the models in the first-row process the speech stimulus given as a speech envelope, the models in the second-row process the speech stimulus given as a mel-spectrogram. The results are based on the attention weights of the AE of the respectively trained model, averaged over time and over the 71 training subjects.

respectively, to preprocess the speech stimulus. (iii) We changed the input speech stimulus type from speech envelope to mel-spectrogram.

Our experimental results show that using an attention encoder instead of spatial convolution helps to capture better the temporal and spatial features in the EEG data, leading to an improved classification performance on the test set as well as on the held-out stories benchmark set. The performance on both sets can be enhanced by additional sequence modeling methods in the speech stimulus processing pipeline, if mel-spectrograms are used as input SST. In general, working with mel-spectrograms enhances the classification performance for all models on the test set and for almost all models on the benchmark sets. Our best model shows improvements of 1.91 % on the test set and 1.35 % on the total ranking score (weighted sum of both benchmark sets) compared to the baseline model. Our team reached the second place in the challenge.

In our future work, we plan to study other methods to extract features from the speech stimulus data, e.g., with attention-based algorithms. Secondly, we would like to analyse other attention-based methods in the EEG processing pipeline, as we see great potential here. It would be interesting to study how the attention pattern may fluctuate within subjects during a trial, although this may be more useful for regression tasks than for classification tasks. In addition, we plan to explore how the model size can be reduced in terms of the number of parameters while maintaining the increased classification performance. Finally, we would like to study how the two adjacent tasks of match-mismatch classification and speech envelope reconstruction from EEG can mutually benefit from each other.

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[3] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, pp. 233—-236, 2012.

[4] S. Akram, A. Presacco, J. Z. Simon, S. A. Shamma, and B. Babadi, "Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling," *NeuroImage*, vol. 124, pp. 906–917, 2016.

[5] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[6] T. de Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1234–1241, 2020.

[7] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *ICASSP*, 2018.

[8] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[9] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *INTERSPEECH*, 2017.

[10] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *INTERSPEECH*, 2019.

[11] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1370–1384, 2020.

[12] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A complete time domain speaker extraction network," in *INTERSPEECH*, 2020.

[13] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, 2018.

[14] Z. Pan, R. Tao, C. Xu, and H. Li, "Selective listening by synchronizing speech with lips," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1650–1664, 2022.

[15] Z. Pan, R. Tao, C. Xu, and H. Li, "USEV: Universal speaker extraction with visual cue," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 3032–3045, 2022.

[16] Y. Ohishi, M. Delcroix, T. Ochiai, S. Araki, D. Takeuchi, D. Niizumi, A. Kimura, N. Harada, and K. Kashino, "ConceptBeam: Concept driven target speech extraction," in *ACM International Conference on Multimedia*, 2022, pp. 4252–4260.

[17] J. Li, M. Ge, Z. Pan, L. Wang, and J. Dang, "VCSE: Time-domain visual-contextual speaker extraction network," in *INTERSPEECH*, 2022.

[18] Z. Pan, X. Qian, and H. Li, "Speaker extraction with co-speech gestures cue," *IEEE Signal Processing Letters*, vol. 29, pp. 1467–1471, 2022.

[19] R. Abiri, S. M. 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, 2019.

[20] S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 5, pp. 1045–1056, 2016.

[21] E. Ceolini, J. Hjortkjær, D. D. Wong, J. O'Sullivan, V. S. Raghavan, J. Herrero, A. D. Mehta, S.-C. Liu, and N. Mesgarani, "Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception," *NeuroImage*, vol. 223, p. 117282, 2020.

[22] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, 2021.

[23] B. Accou, M. Jalilpour Monesi, J. Montoya, H. Van hamme, and T. Francart, "Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network," in *EUSIPCO 2020*, 2021.

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

[25] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[26] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *ASRU*, 2013.

[27] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *EMNLP*, 2014.

[28] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *arXiv preprint arXiv:1508.01991v1*, 2015.

[29] R. Rana, "Gated recurrent unit (GRU) for emotion classification from noisy speech," *arXiv preprint arXiv:1612.07778v1*, 2016.

[30] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," *arXiv preprint arXiv:1702.01923v1*, 2017.

[31] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[32] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[33] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, 2020.

[34] 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 *ICASSP*, 2020.

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

[36] Y. Song, X. Jia, L. Yang, and L. Xie, "Transformer-based spatial-temporal feature learning for EEG decoding," *arXiv preprint arXiv:2106.11170v1*, 2021.

[37] Y. Tao, T. Sun, A. Muhamed, S. Genc, D. Jackson, A. Arsanjani, S. Yaddanapudi, L. Li, and P. Kumar, "Gated transformer for decoding human brain EEG signals," in *EMBC*, 2021.

[38] J. Xie, J. Zhang, J. Sun, Z. Ma, L. Qin, G. Li, H. Zhou, and Y. Zhan, "A transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2126–2136, 2022.

[39] S. Bagchi and D. R. Bathula, "EEG-ConvTransformer for single-trial EEG-based visual stimulus classification," *Pattern Recognition*, vol. 129, 2022.

[40] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG conformer: Convolutional transformer for EEG decoding and visualization," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2023.

[41] Z. Wan, M. Li, S. Liu, J. Huang, H. Tan, and W. Duan, "EEGformer: A transformer–based brain activity classification method using EEG signal," *Frontiers in Neuroscience*, vol. 17, 2023.

[42] D. H. Klatt, "Speech perception: A model of acoustic-phonetic analysis and lexical access," *Journal of Phonetics*, vol. 7, no. 3-4, 1979.

[43] 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 *ICASSP*, 2023.

[44] C. Puffay, B. Accou, L. Bollens, M. J. Monesi, J. Vanthornhout, H. Van hamme, and T. Francart, "Relating EEG to continuous speech using deep neural networks: a review," *Journal of Neural Engineering*, vol. 20, no. 4, 2023.

[45] 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," in *INTERSPEECH*, 2022, pp. 4038–4042.

[46] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, 2018.

[47] M. Thornton, D. Mandic, and T. Reichenbach, "Robust decoding of the speech envelope from EEG recordings through deep neural networks," *Journal of Neural Engineering*, vol. 19, no. 4, 2022.

[48] B. Accou, J. Vanthornhout, H. Van hamme, and T. Francart, "Decoding of the speech envelope from EEG using the VLAAI deep neural network," *Scientific Reports*, vol. 13, 2023.

[49] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, 2016.

[50] A. de Cheveigné, D. D. Wong, G. M. Di Liberto, J. Hjortkjaer, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[51] H. Korn and P. Faure, "Is there chaos in the brain? II. Experimental evidence and related models," *Comptes Rendus Biologies*, vol. 326, no. 9, pp. 787–840, 2003.

[52] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'Sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, "Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[53] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *Elife*, vol. 10, p. e56481, 2021.

[54] S. Cai, E. Su, L. Xie, and H. Li, "EEG-based auditory attention detection via frequency and channel neural attention," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 2, pp. 256–266, 2022.

[55] P. Wang, A. Jiang, X. Liu, J. Shang, and L. Zhang, "LSTM-based EEG classification in motor imagery tasks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 11, pp. 2086–2095, 2018.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[57] L. Bollens, B. Accou, H. Van hamme, and T. Francart, "A Large Auditory EEG decoding dataset," *https://doi.org/10.48804/K3VSND, KU Leuven RDR, V1*, 2023.

[58] 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 2023.07.24.550310*, 2023.

[59] B. Somers, T. Francart, and A. Bertrand, "A generic EEG artifact removal algorithm based on the multi-channel Wiener filter," *Journal of Neural Engineering*, vol. 15, no. 3, 2018.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[61] Z. Xu, Y. Bai, R. Zhao, H. Hu, G. Ni, and D. Ming, "Decoding selective auditory attention with EEG using a transformer model," *Methods*, vol. 204, pp. 410–417, 2022.

[62] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Reviews Neuroscience*, vol. 8, 2007.

[63] J. Obleser, R. J. S. Wise, M. A. Dresner, and S. K. Scott, "Functional integration across brain regions improves speech perception under adverse listening conditions," *The Journal of Neuroscience*, vol. 27, pp. 2283–2289, 2007.