Received 8 August 2023; revised 30 October 2023; accepted 2 November 2023. Date of publication 12 March 2024; date of current version 18 June 2024. The review of this article was arranged by Associate Editor Sergio Benini.

Digital Object Identifier 10.1109/OJSP.2024.3376296

Auditory EEG Decoding Challenge for ICASSP 2023

MOHAMMAD JALILPOUR MONESI [1,2], LIES BOLLENS [1,2], BERND ACCOU [1,2], JONAS VANTHORNHOUT [2], HUGO VAN HAMME [1] (Senior Member, IEEE), AND TOM FRANCART [2]

[1] Department of Electrical engineering (ESAT), KU Leuven, PSI, 3000 Leuven, Belgium
[2] Department of Neurosciences, KU Leuven, ExpORL, 3000 Leuven, Belgium

CORRESPONDING AUTHORS: LIES BOLLENS; MOHAMMAD JALILPOUR MONESI; TOM FRANCART (e-mail: lies.bollens@kuleuven.be; mohammad.jalilpourmonesi@esat.kuleuven.be; tom.francart@kuleuven.be).

(Mohammad Jalilpour Monesi and Lies Bollens contributed equally to this work.)

This work was supported in part by KU Leuven Special Research Fund under Grant C24/18/099 through C2 Project to Tom Francart and Hugo Van hamme, in part by Research Foundation Flanders (FWO) under Grant 1S89622 N, Grant 1SB1421 N, and Grant 1290821 N, and in part by European Research Council (ERC) through European Union's Horizon 2020 Research and Innovation Program under Grant 637424, ERC Starting Grant to Tom Francart.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Medical Ethics Committee UZ KU Leuven/Research (KU Leuven, Belgium) under Application No. S57102, and performed in line with the guidelines for good clinical practices (ICH/GCP) and followed the most recent version of the declaration of Helsinki, which serves to protect people participating in a clinical study.

ABSTRACT This paper describes the auditory EEG challenge, organized as one of the Signal Processing Grand Challenges at ICASSP 2023. The challenge provides EEG recordings of 85 subjects who listened to continuous speech, as audiobooks or podcasts, while their brain activity was recorded. EEG recordings of 71 subjects were provided as a training set such that challenge participants could train their models on a relatively large dataset. The remaining 14 subjects were used as held-out subjects in evaluating the challenge. The challenge consists of two tasks that relate electroencephalogram (EEG) signals to the presented speech stimulus. The first task, match-mismatch, aims to determine which of two speech segments induced a given EEG segment. In the second regression task, the goal is to reconstruct the speech envelope from the EEG. For the match-mismatch task, the performance of different teams was close to the baseline model, and the models did generalize well to unseen subjects. In contrast, For the regression task, the top teams significantly improved over the baseline models in the held-out stories test set while failing to generalize to unseen subjects.

INDEX TERMS Backward modeling, EEG, match-mismatch, neural tracking, speech decoding.

I. INTRODUCTION

To investigate how the brain processes sound, various neuroimaging techniques can be used. Electroencephalography (EEG) is popular because it is relatively easy to conduct and has a high temporal resolution. Besides fundamental neuroscience research, EEG-based measures of auditory processing in the brain are also useful to detect or diagnose a potential hearing loss [1], [2]. They enable differential diagnosis of populations that can otherwise not be tested, such as young children or people with mental disabilities. In addition, there is a growing field of research in which auditory attention is decoded from the brain, with potential applications in smart hearing aids.

An increasingly popular method in these fields is to relate a person's electroencephalogram (EEG) to a feature of the natural speech signal they were listening to. This is typically done using linear regression to predict the EEG signal from the stimulus or to decode the stimulus from the EEG [3], [4], [5]. While these approaches have been widely used in auditory neuroscience and have successfully been linked to speech intelligibility, [4], reconstruction scores remain low and are prone to large inter-subject variability. Given the very low signal-to-noise ratio of the EEG, auditory decoding is a challenging problem, and several alternative methods based on artificial neural networks (ANNs) have been proposed to improve upon the linear methods [6], [7], [8], [9], [10], [11].

Instead of directly decoding a speech feature from the EEG, which is a challenging *regression* problem, an alternative *classification* paradigm, referred to as the match-mismatch task, has been recently proposed [12]. Given an EEG segment and two speech segments, the task is to determine which of the two speech segments corresponds to the EEG segment [13]. Recently, methods based on deep learning models have obtained promising results on this task, outperforming the linear methods [13], [14], [15], [16]. In addition to the above-mentioned regression and match-mismatch task, deep ANNs have recently been a popular choice to relate EEG to speech in various paradigms, such as speech denoising [17], denoising and normalizing EEG [18], [19] and predicting EEG from acoustic features [20]. However, a drawback to neural networks is that they typically require a large amount of data for training. Unfortunately, a large public auditory EEG dataset together with well-defined tasks to relate EEG to speech is not available, which makes it difficult to compare the performance of different models as people use different datasets as well as different metrics to evaluate their models (e.g., see [21]).

In the ICASSP 2023 Auditory-EEG challenge, we provide a large auditory EEG dataset containing data from 85 subjects who listen on average to 110 minutes of single-speaker stimuli for 157 hours of data. Teams compete to build the best model to relate speech to EEG in two tasks: 1) **match-mismatch**; given two segments of speech and a segment of EEG, which of the speech segments matches the EEG segment? and 2) **regression**; reconstruct the speech envelope from the EEG.

II. DATASET

Electroencephalography (EEG) is a non-invasive method to record electrical activity in the brain, which is generated by ionic currents that flow within and across neuron cells. When a large population of thousands or millions of neurons with a similar orientation in a specific brain region synchronizes its electrical activity, the produced electrical field is large enough to be observable on the scalp. When we attach an array of electrodes on the scalp, these electrical fields can be recorded by measuring the electrical potential (typically 10–100 μV) between pairs of electrodes in the array.

A. DATA COLLECTION

We measured EEG data in a well-controlled lab environment (soundproof and electromagnetically shielded booth) using a high-quality 64-channel Biosemi (Amsterdam, the Netherlands) ActiveTwo EEG recording system with 64 active Ag-AgCl electrodes and two extra electrodes, which serve as the common electrode (CMS) and current return path (DRL). These two electrodes are responsible for establishing the electrical reference or “ground” for the EEG system. The CMS is the reference channel against which all EEG channels are compared. Meanwhile, the DRL’s role is to minimize the subject’s electrical potential deviation from the system’s “zero” point. The BioSemi head caps were used, which contain electrode holders placed according to the 10–20 electrode system. The data was measured at a sampling rate of 8192 Hz. While the temporal resolution is high, the spatial resolution is low, with only 64 electrodes. All 64 electrodes were placed according to the international 10–20 standard. The dataset contains data from 85 young, normal-hearing subjects (74 female/11 male, 21.4 $\pm$ 1.9 years (sd), all hearing thresholds $\leq$ 30 dB SPL), with Dutch/Flemish as their native language. This study was approved by the Medical Ethics Committee UZ / KU Leuven (EC Research) with reference S57102. Before commencing the EEG experiments, all subjects read and signed an informed consent form. All subjects in the dataset gave explicit consent for their anonymized data to be shared in a publicly accessible dataset [22]. All identifiable subject information has been removed from this dataset.

Before commencing the EEG experiments, subjects completed a questionnaire requesting general demographic information (age, sex, education level, handedness [23] and diagnoses of hearing loss and neurological pathologies. Subjects indicating any neurological or hearing-related diagnosis were excluded from the study. Then, we measured the air conduction threshold using the Hughson-Westlake method [24] for octave frequencies between 125 and 8000 Hz. Subjects with hearing thresholds $\geq$ 30 dB SPL were excluded.

Each subject listened to between 6 and 10 trials, each of approximately 15 minutes in length. The order of the trials was randomized between subjects. After each trial, we asked a question about the stimulus’s content to motivate subjects to pay attention during the recording. All the stimuli are single-speaker stories spoken in Flemish (Belgian Dutch) by a native Flemish speaker. We vary the stimuli between subjects to have a wide range of unique speech material. The stimuli are either podcasts(37 in total) or audiobooks (15 in total). As some audiobooks are longer than 15 minutes, they are split into two trials presented consecutively to the subject. The stimuli were presented to the subjects using electromagnetically shielded Etymotic ER-3 A insert phones, binaurally at 62 dBA for each ear. Distinct speakers narrate all podcasts. The same speaker narrated audiobooks 2, 5,6, and 15, and distinct speakers narrated all other audiobooks. In total, there are 49 distinct speakers (27 female/22 male).

B. TRAINING SET

Both tasks share the training set. The training set contains EEG responses from 71 subjects. These subjects are numbered from sub-001 to sub-071. As shown in Fig. 2, each subject listens to between 6 and 9 trials, each of around 15 minutes in length. Due to measuring errors, not all trials for all subjects have been included in the training set. Subjects are divided into groups, depending on which stimuli they listen to. Each such group contains between 2 and 26 subjects. All subjects of all groups listen to a reference story, *Audiobook 1*.

The training set contains 508 trials from 71 subjects, using 57 different stimuli. The total data duration amounts to 7216 minutes (120 hours). Data is structured in a folder per subject, and the trials are named according to the subject and stimulus. Each EEG trial file contains a pointer to the stimulus used to generate the specific brain EEG response and a reference to

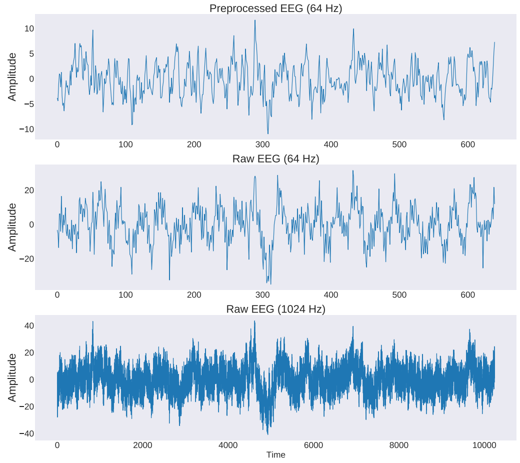
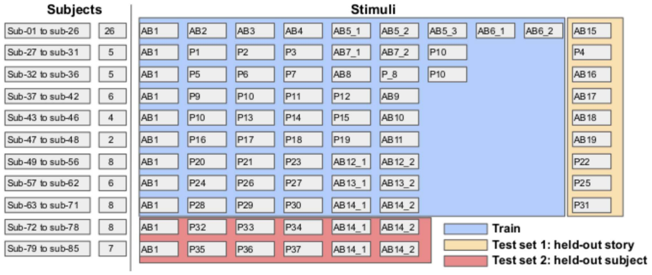


FIGURE 1. Figure displays preprocessed and raw EEG signals recorded from the 'Cz' channel over a 10-second duration for one of the subjects.



Subjects		Stimuli									
Sub-01 to sub-26	26	AB1	AB2	AB3	AB4	AB5_1	AB5_2	AB5_3	AB6_1	AB6_2	AB15
Sub-27 to sub-31	5	AB1	P1	P2	P3	AB7_1	AB7_2	P10			P4
Sub-32 to sub-36	5	AB1	P5	P6	P7	AB8	P_8	P10			AB16
Sub-37 to sub-42	6	AB1	P9	P10	P11	P12	AB9				AB17
Sub-43 to sub-46	4	AB1	P10	P13	P14	P15	AB10				AB18
Sub-47 to sub-48	2	AB1	P16	P17	P18	P19	AB11				AB19
Sub-49 to sub-56	8	AB1	P20	P21	P23	AB12_1	AB12_2				P22
Sub-57 to sub-62	6	AB1	P24	P26	P27	AB13_1	AB13_2				P25
Sub-63 to sub-71	8	AB1	P28	P29	P30	AB14_1	AB14_2				P31
Sub-72 to sub-78	8	AB1	P32	P33	P34	AB14_1	AB14_2				
Sub-79 to sub-85	7	AB1	P35	P36	P37	AB14_1	AB14_2				

FIGURE 2. Overview of the different stimuli and their division into training and testing set. Left side: subjects are divided into different groups. Each group of subjects listens to the same stimuli, which are depicted on the right side. Each horizontal line defines one group and the corresponding stimuli. Short names for the stimuli are given, with AB = audiobook and P = podcast.

the subject identifier. The auditory stimuli are provided in a separate folder *stimuli*.

C. TEST SET

The test set consists of two parts: *held-out stories* and *held-out subjects*. These sets are split into two parts, ensuring that the test sets of the two tasks do not overlap. More details about the task-specific test sets can be found in the description of each task (see Section III, Section III-A and III-B).

Held-out stories: contains data from the 71 subjects seen in the training set. We held out one story for each group of subjects, which never occurred in the training set, amounting to 944 minutes in total.

Held-out subjects: contains data for 14 subjects (sub-072 to sub-085) that are not in the training set, further referred to as held-out subjects, for a total of 1260 minutes. The data for these subjects were acquired using the same protocol as for the other 71 subjects.

D. PREPROCESSING

We provide two versions of the dataset. The first data version is the raw EEG data, downsampled from 8192 Hz to 1024 Hz. In addition to the raw EEG recordings, we also provide preprocessed EEG and speech stimuli, which have undergone commonly used preprocessing steps. All steps were conducted in Python 3.7, and the code for preprocessing is available on our GitHub repository (https://github.com/exporl/auditory-eeg-dataset).

First, EEG data was downsampled from 8192 Hz to 1024 H and high-pass filtered, using a 1st-order Butterworth filter with a cut-off frequency of 0.5 Hz. Zero-phase filtering was conducted by filtering the data forward and backward. Subsequently, eyeblink artifact removal was applied to the EEG, using a multichannel Wiener filter [25]. Afterward, the EEG was re-referenced to a common average, and finally, the EEG was downsampled to 64 Hz. Fig. 1 shows 10 seconds of preprocessed and raw EEG signals for channel 'Cz' for one of the subjects in the dataset.

The steps in our preprocessing are commonly used in EEG signal processing, and the preprocessed version can be used directly in machine learning models. Challenge participants are free to perform their preprocessing on both versions of the datasets. However, since the test set is already split up into segments of EEG and stimuli (3 seconds for task 1 III-A match-mismatch, 60 seconds for task 2 III-B regression), performing many preprocessing steps could introduce artificial edge effects in the data, which could influence performance.

For task 2 III-B: regression, we defined a specific version of the envelope, which has been defined and validated in [26]. We estimated the envelope using a gammatone filter bank with 28 subbands spaced by one equivalent rectangular bandwidth with center frequencies from 50 Hz to 5 kHz. Subsequently, the absolute value of each sample in the filters is taken, followed by exponentiation with 0.6. Then, all subbands are averaged to obtain one speech envelope. Finally, the resulting envelope is downsampled to 64 Hz. We provide code to create these envelope representations, as well as to perform the described preprocessing steps.

III. AUDITORY EEG DECODING CHALLENGE

A. TASK 1: MATCH-MISMATCH

1) DESCRIPTION

Task 1 is a classification task based on the concept of matched and mismatched (EEG, speech) pairs [12]. A matched (EEG, speech) pair means that the EEG segment was recorded while the speech segment was presented to the subject. A mismatched (EEG, speech) pair indicates that the EEG segment is not a response to the speech segment.

There are many possibilities to define a match-mismatch task based on the concept of matched/mismatched segments [14], [27], [28]. For this challenge, we opted for the method described in [28]. A schematic of the chosen match-mismatch task is illustrated in the left plot of Fig. 3. The model is provided with three inputs of length 3 seconds: (1) a segment of EEG, (2) the time-aligned speech stimulus (match), and (3) an unaligned stimulus (mismatch). The task of the model in this paradigm is to determine which of the two input stimulus segments correspond to the EEG segment.

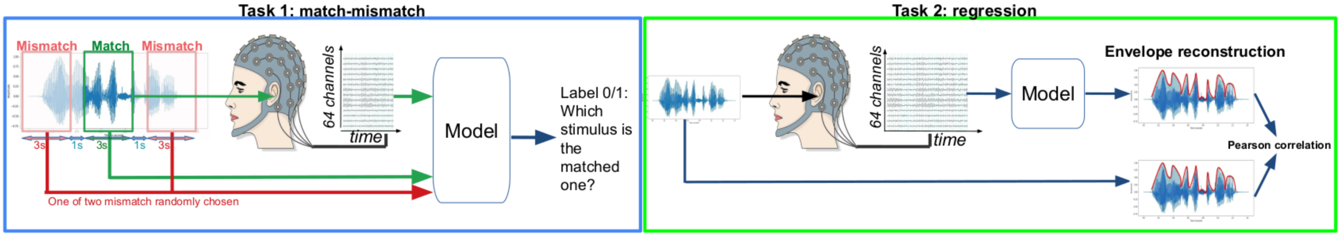


FIGURE 3. Schematic Overview of the two tasks. Left: Task 1 (match-mismatch). The model gets three inputs: an EEG segment, the matched (in time) speech segment, and a mismatched segment. The task is to determine which of the two segments is matched. Right: Task 2 (regression). The task is to decode the speech envelope from the EEG brain response. Reconstructed and original envelopes will be compared by computing the Pearson correlation.

More specifically, we define the mismatched stimulus as temporally close to the matched one by randomly taking the segment starting either one second after the end or 4 seconds before the start of the matched segment. A mismatched segment close to the matched segment makes the training process more challenging for the model (see [21], section 3.6).

We implement and suggest the following recommendations during training to use the data in the classification paradigm. First, we present each EEG segment twice to the model: (EEG, matched stimulus, mismatched stimulus, output label 0) and (EEG, mismatched stimulus, matched stimulus, output label 1). This ensures that matched and mismatched candidates occur equally in both input positions (first or second). Second, we make sure that a mismatched stimulus segment is also a matched segment with another EEG segment. A way to do this is to ensure that the shift when windowing (1 s in our code) is dividable by the spacing between the matched and mismatched segment (1 s). Failing to follow these suggestions will likely result in models simply remembering the training samples and thus failing to generalize to the test set, as shown in more detail in [21].

2) BASELINE METHOD

We include a dilated convolutional network [13] as a baseline for task 1. The dilated convolutional network transforms both EEG and stimuli to a latent presentation, after which these latent representations are compared to make a final decision. There is a separate EEG and stimulus path. For the EEG path, the EEG channels are first combined, from 64 to 8, using a 1D convolutional layer with a kernel size of 1 and a filter size of 8. Second, there are three dilated convolutional layers with a kernel size of 3 and 16 filters. After each convolutional layer, a rectified linear unit (ReLU) is applied.

Similarly, the stimulus path contains three dilated convolutional layers with a kernel size of 3 and 16 filters. Both stimulus segments share the weights for the convolutional layers.

After these non-linear transformations, the latent EEG representation is compared to both latent stimuli representations, using cosine similarity. Finally, the similarity scores are fed to a single neuron, with sigmoid non-linearity, to create a prediction of the matching stimulus segment. When applied to the training and test sets of the challenge, a performance accuracy of approximately 78% is obtained.

3) EVALUATION CRITERIA

The test set for the match-mismatch task contains the first half or the second half of each EEG recording from the held-out stories set and half of the recordings from the held-out subjects set (in both test sets, the other half is used in the regression task). It is not possible to use overlapping test sets as for the regression task, the stimulus is the target of the regression, whereas, in the match-mismatch task, the stimulus is one of the inputs.

For both test sets, we provide pairs of (EEG, stimulus 1, and stimulus 2), with a length of 3 seconds, each with a unique identifier and a subject identifier. As an output, participants had to submit a NumPy dictionary file to an online form on our https://exporl.github.io/auditory-eeg-challenge-2023/website, which contains the predicted label for all EEG segments. Each entry in the submitted dictionary must be (EEG ID) : (label). In case of absent EEG ID entries, the sample will be assigned the wrong label. Labels should be either 0 or 1.

The mean classification accuracy per subject is then calculated as $ACC_s = \sum_{i=0}^{n_s}[label_{predicted} = label_{true}]/n_s$. Then, we calculate the mean accuracy over test subjects set 1 ($S_1 = \sum_{s=1}^{71} ACC_s/71$) and test set 2 ($S_2 = \sum_{s=72}^{85} ACC_s/14$) and average them to obtain the final Score, which will serve as the ranking value $Score = 2/3S_1 + 1/3S_2$.

B. TASK 2: REGRESSION

1) DESCRIPTION

Task 2 is a regression problem: reconstructing the EEG speech envelope. After reconstruction, the Pearson correlation measures the similarity between the reconstructed and original stimuli. The right plot in Fig. 3 illustrates the regression task. The stimulus representation is the envelope, as described in Section II-D.

2) BASELINE METHOD

We include a simple linear model as well as the Very Large Augmented Auditory Inference (VLAAI) network [7] as a baseline for task 2. The linear model is implemented using a one-dimensional convolutional layer in TensorFlow with kernel_size of 32 (corresponds to a 500 ms integration window). The VLAAI network consists of multiple ($N = 4$) blocks with 3 different parts. The first part is a CNN stack, a convolutional neural network. This CNN consists of $M = 4$ convolutional layers. The second part is a simple, fully connected layer of 64 units, which recombines the output filters of the CNN stack. The last part is the output context layer. This convolutional layer enhances the predictions made by the model up until that point by taking the previously predicted samples into account and combining them with the current sample. A skip connection is present with the original EEG input at the end of each block except the last. After the last block, the linear layer at the top of the VLAAI model combines the filters of the output context layer into a single speech envelope. When applied to the training and test sets of the challenge, an average correlation score of 0.136 is obtained.

3) EVALUATION CRITERIA

The test set for the regression task contains the first half or the second half of each EEG recording from the held-out stories set and half of the recordings from the held-out subject set. All stimuli are held-out stimuli, i.e., they do not appear in the training set. We split the stimuli into several smaller segments of 60 seconds and made these available with a segment ID and a subject ID for each segment.

For each segment of 60 seconds, we expect a reconstructed envelope, which is then compared to the original envelope, as defined in Section II-D, using Pearson correlation. We use the `scipy.stats.pearsonr` function to calculate the correlation c_i for each segment i. Afterward, the mean correlation value per subject is calculated as $C_s = \sum_{i=1}^{n} c_i/n$. Then, we calculate the mean correlation values over all subjects for test set 1 ($S_1 = \sum_{s=1}^{71} C_s/71$) and test set 2 ($S_2 = \sum_{s=72}^{85} C_s/14$) and average them to obtain the final Score, which will serve as the ranking value: $Score = 2/3S_1 + 1/3S_2$

C. PROVIDED ONLINE CODE

The dataset is available from the KULeuven RDR platform [29]. Detailed information about the dataset can be found in [22]. For most subjects, the data is publicly accessible. However, due to privacy concerns, there are some subjects for which the data is restricted to registered users. Users requesting access should mail to mailto:sparrkulee@kuleuven.besparrkulee@kuleuven.be, stating what they want to use the data for. Access will be granted to non-commercial users, complying with the CC-BY-NC-4.0 license.

We provided a simple Python codebase in our ExpORL https://github.com/exporl/auditory-eeg-challenge-2023-code/tree/mainGithub repository to make it easier for participants to start with the challenge. The repository contains two top folders, each associated with one of the tasks defined in the challenge. Each folder contains code for baseline models and code to preprocess data, create the test set of the challenge and train the baseline models. For more information, refer to the README file of the repository.

IV. RESULTS

There were 21 submissions for the match-mismatch task and 13 for the regression task. Both tracks are separate. In total, there can be five winners. As a result, the top 3 teams from the match-mismatch task (the most popular task) and the top 2 teams from the regression task were accepted as the challenge top 5 teams.

The winner of the match-mismatch task is the team *UnderDawgs (Thornton, Mandic and Reichenbach)*, with their solution “Relating EEG recordings to speech using envelope tracking and the speech-FFR” [30]. The main idea in their solution is combining multiple (50) instances of the baseline dilation model by averaging the outputs of the Sigmoids, the last layer of the baseline model. The idea is similar to the concept of majority voting in machine learning, in which multiple outputs of multiple models are summed or counted to make the final classification decision. They also show that using the high-frequency envelope modulations, as explained in [35], as a speech representation along with the speech envelope improves the baseline model’s performance compared to only using the envelope. They average multiple (30) instances of the f0 dilation model. For the within-subject test set, they further finetuned the subject-independent baseline model on each subject and achieved higher accuracy. They employ a composite model for the unseen subjects, which combines the averaged baseline model with the averaged f0 model via linear discriminant analysis (LDA).

The second place is won by the team *HyperAttention (Borsdorf et al.,)*, with their solution “Multi-head attention and GRU for improved match-mismatch classification of speech stimulus and EEG response” [31]. Their proposed solution is based on the baseline dilation model. However, they use a multi-head attention block, with two attention heads [36] in the EEG path just before the dilated convolutions, as well as a gated recurrent unit (GRU) [37] in the speech path before the dilated convolutions. Moreover, they use the mel spectrogram as a speech representation instead of the envelope. The team *MINWPU (Cui et al.,.)* gets the third place with their solution “Relate auditory speech to EEG by shallow-deep attention-based networks” [32]. In contrast to the baseline model in which the EEG and speech are only connected in the cosine similarity layer, their solution uses an attention-based correlation module (ACM) after each layer to connect the EEG to both speech paths. More specifically, the ACM module tries to capture the global relationship between speech and EEG and consists of a residual attention layer and a feed-forward layer. Finally, a shallow-deep similarity classification module (SDSCM) classifies the sample based on embeddings from different model layers.

The team named *HappyQuoka (Piao et al.,)* won the regression task of the challenge with their solution titled ’HappyQuoka system for ICASSP 2023 auditory EEG challenge’ [33]. The authors propose a model based on feed-forward transformer (FFT) architecture, which uses pre-layer normalization [38]. Furthermore, they use an auxiliary global conditioner [39] that integrates the subject information in the

TABLE 1 **Summary Table of the Top 5 Teams**

Task	Place	Team	Within subjects			Heldout subjects			Final score	
			mean	std	p-value	mean	std	p-value	score	p-value
Match-mismatch		Baseline (dilation)	77.59	7.29		77.34	5.66		**77.51**	
Match-mismatch	1	UnderDawgs [32]	82.71	7.70	$1.07 * 10^{-7}$	80.98	5.27	0.0012	**82.13**	$\mathbf{1.45 * 10^{-09}}$
Match-mismatch	2	HyperAttention [33]	79.61	7.08	0.074	77.93	7.66	0.54	**79.05**	**0.047**
Match-mismatch	3	MINWPU [34]	79.21	7.52	0.14	78.40	5.66	0.52	**78.94**	**0.060**
Regression		Baseline (VLAAI)	0.1557	0.0832		0.0959	0.1557		**0.1358**	
Regression	1	HappyQuokka [35]	0.1895	0.0869	$3.89 * 10^{-8}$	0.0976	0.0444	1	**0.1589**	$\mathbf{3.273 * 10^{-8}}$
Regression	2	TheBrainwaveBandits [36]	0.1741	0.0913	$2.25 * 10^{-6}$	0.1123	0.0447	0.15	**0.1535**	$\mathbf{1.490 * 10^{-7}}$

There are three teams for the match-mismatch task (the more popular task) and two for the regression task. The p-values indicate the significance of these results with respect to the baseline model. We used a two-sided wilcoxon signed-rank test with Bonferonni-Holm corrections.

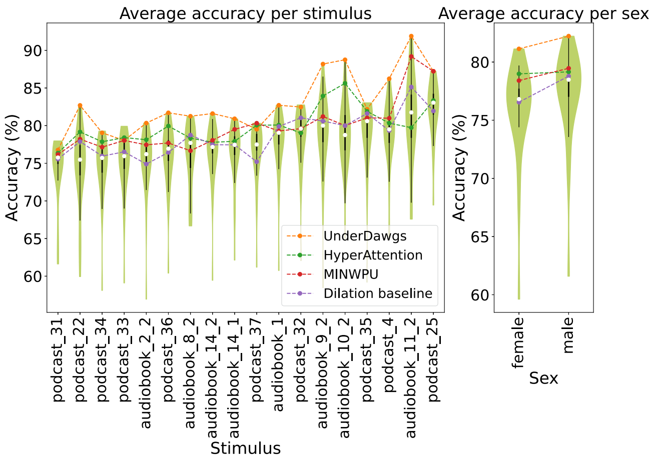


FIGURE 4. **Match-mismatch accuracy of top models per stimuli and sex of the speaker of the stimuli. Each point in the violin plots corresponds to the average value of one team over all subjects. Violin plots are shown over teams. Left: Average accuracy per unique stimulus. Stimuli are ordered based on increasing mean accuracy. Right: Average accuracy per sex of the speaker of stimuli, averaged over stimuli and subjects.**

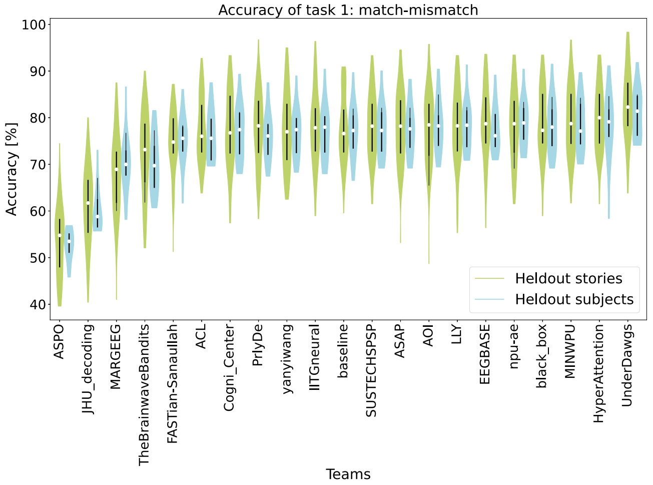


FIGURE 5. **Comparing the match-mismatch accuracy of different teams on the test sets. The violin plots are shown per team over subjects. Each point in the violin plots corresponds to one team's average accuracy for one subject. Green: seen subjects, held-out stories. Blue: held-out subjects, seen stories.**

model. Note that this global conditioner could only be used in the within-subject test set and not in the held-out-subjects test set. Their ablation study shows that using subject information increases the correlation values by 13%.

The team named *The brainwave bandits (Van Dyck et al.,)* got second place in the regression task with their solution titled 'Decoding auditory EEG responses using an adapted wavenet' [34]. Their proposed solution is based on WaveNet [39]. This model has been adapted to use a non-causal dilated 1D convolution and a channel-wise 1D convolution as the first layer that compresses the multichannel EEG input. For the within-subject test set, they further finetuned the model on each subject for better performance.

Since there are only five winners of the challenge, the team that obtained third place in the regression challenge and fourth place in the match-mismatch challenge does not win. However, as mentioned below, their solution for the regression task significantly outperforms both the linear and the VLAAI baseline. The team is called *black box*, and their solution is named "Eeg2vec: Self-Supervised Electroencephalographic Representation Learning" [40]. In their solution, they first use a self-supervised model, inspired by wav2vec 2.0 [41], based on a contrastive loss combined with a reconstruction loss to learn EEG representations. Secondly, the pre-trained model is used as a feature extractor for the downstream tasks (either task 1 or task 2).

V. SUMMARY OF RESULTS

A. TASK 1: MATCH-MISMATCH

The baseline dilation model obtained a total score of $\approx 78\%$. The scores of the teams who submitted to the match-mismatch task ranged from 53 % to 82 %, with an average of $\approx 75.23\%$ $\pm$ 6.55 (std). Fig. 5 overviews the competing teams' final scores for both test sets. Of the 21 submitted teams, ten scored higher than the baseline. We tested the significance of the results of these ten teams with respect to the baseline, using a two-sided Wilcoxon signed-rank test using Bonferroni-holm corrections and an alpha value of 0.05. Only the top two teams, *Underdawgs* and *Hyperattention* have a score that differs significantly from the baseline (p-values of $1.45 * 10^{-9}$ and 0.047, respectively).

Some trends were observed amongst most of the submissions. Most of the models performed equally well on the held-out subjects as they did on held-out stories. This indicates that the models generalize well to new, unseen subjects and might suggest that the models can extract a subject-independent representation of the stimuli from the EEG signals. Another observation is that the sex of the speaker seems to affect the match-mismatch accuracy of the models. As shown in Fig. 4, models obtain higher accuracy on male-spoken stories compared to female-spoken stories, consistent with results obtained using either the fundamental frequency f0 or linguistics features for neural tracking [21], [42], [43].

We also grouped the results of the different teams per stimulus type (i.e., audiobook vs. podcast) to see if the stimulus type affects the match-mismatch accuracy of the models. To this end, we ordered the performance of the different stimuli based on the mean accuracy of all the teams. As seen in Fig. 4, there is no significant difference in the type of stimulus used as the ordering alternates between audiobooks and podcasts.

A common theme across teams is to employ finetuning strategies for the seen subjects, which improves classification accuracy, consistent with the previously published results [14], [28], [44]. Among the top teams, the chosen stimulus representation differs from the speech envelope, as the mel spectrogram and the fundamental frequency (f0) representations were also successfully used.

Most of the teams have relatively similar match-mismatch accuracy of around 78%. This might be due to several reasons. First, most models were based on the baseline model, which already works well ($\approx$ 78%) on the match-mismatch task. It makes sense to see similar results from all these models as they made some adjustments to improve the baseline model. Another possible explanation is that the proposed match-mismatch task is not difficult enough since it is a relative decision between two segments of speech, which might not entice the model to develop very robust latent speech and EEG representations. Hence, the models perform relatively well using simple speech representations such as the envelope. There might also be a ceiling effect on the performance which can be obtained. In other words, the SNR of the recorded EEG is very low, rendering it impossible to achieve higher accuracy than a certain point. However, more research is needed before any conclusions can be drawn.

The baseline linear, backward model obtained a total score of 0.102 on the regression task, whereas the baseline VLAAI model obtained a total score of 0.136. The scores of the teams that submitted ranged from 0.097 to 0.159, with an average Pearson correlation of 0.126 $\pm$ 0.025 (std). Fig. 7 overviews the competing teams' final scores. Of the 13 teams, nine scored higher than the linear baseline, and 4 outperformed the VLAAI baseline. We tested the significance of the results of these four teams with respect to the VLAAI baseline, using a two-sided Wilcoxon signed-rank test using Bonferroni-holm corrections and an alpha value of 0.05. Only the top 3 teams have a total score that differs significantly from the VLAAI baseline (p-values of respectively $3.273 * 10^{-8}$, $1.490 * 10^{-7}$ and $5.964 * 10^{-4}$).

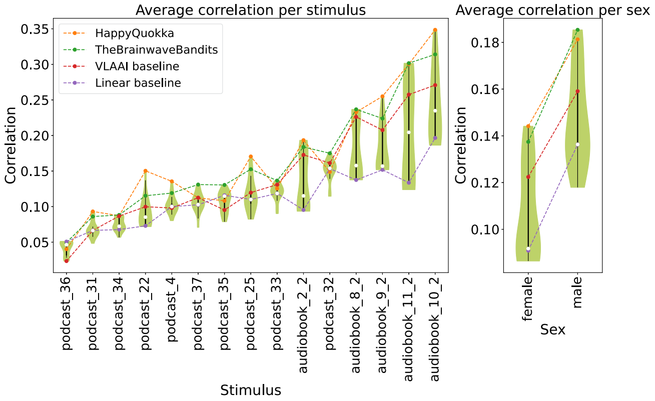


FIGURE 6. Correlation scores of top models per stimuli and sex of the speaker of the stimuli. Each point in the violin plots corresponds to the average value of one team over all subjects. Left: Average correlation per unique stimulus. Stimuli are ordered based on increasing mean correlation. Right: Average correlation per sex of the speaker of stimuli, averaged over stimuli and subjects.

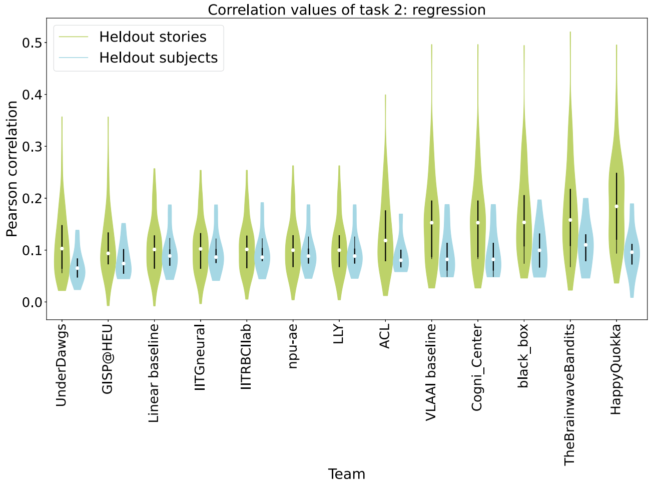


FIGURE 7. Comparing the final correlation scores of different teams on the test sets. Each point in the violin plot corresponds to the average correlation of one subject for one team. Green: seen subjects, held-out stories. Blue: held-out subjects and stories.

One notable observation is the substantial difference in the models' performance between the within-subject test set and the test set comprising held-out subjects. While the top three teams achieved significantly higher scores than both baseline models on the within-subject test set, none outperformed the two baseline models on the held-out subject test set. While this might suggest that the models fail to generalize to unseen subjects, it is worth looking at these results in more detail.

Therefore, we performed the same analysis that we did for the match-mismatch task. More specifically, we grouped the results of the different teams according to the type of stimulus and the sex of the speaker. Similarly to task 1, we observe a significant effect based on the sex of the story's speaker, with models performing better on male-spoken stimuli than female-spoken stimuli (see Fig. 6). When we compare the performance of the models on stimulus type (audiobook vs. podcast), we observe that models perform significantly better on audiobooks than podcasts. A possible explanation for this might be that the podcasts are recorded using compression techniques, which are employed to generate an evenly loud signal. It might be that the models fail to generalize to different compression techniques rather than to different unseen subjects.

VI. DISCUSSION

The results of the Auditory EEG Decoding Challenge provided valuable insights into various aspects of match-mismatch and regression tasks. Most top teams used a different speech representation, such as the mel spectrogram, instead of the speech envelope. This appears to provide an advantage to only using the speech envelope as previously shown in a similar study [15]. Therefore, future work should use richer, more complex speech representations. In addition, finetuning the models for the within-subject test set provides extra accuracy, which is consistent with prior studies [7], [14], [44]. Two of the five teams incorporated an attention mechanism in their solution, achieving good performances. Given the recent success of attention and transformer-like architectures [36], this is not surprising. However, the performance of the attention-based approaches, especially in the match-mismatch task, is not significantly better than other approaches. More research is needed to investigate whether using even larger EEG datasets will eventually make transformer-like architectures superior to others. We encourage researchers to explore this avenue.

In the match-mismatch task, most models generalized well on unseen subjects, which makes the match-mismatch task attractive for relating EEG to speech, as one would not need to train the models on new subjects. Another observation was that most teams had similar results to the provided baseline model. This raises questions about the task's difficulty and potential modifications that could be explored to make it more challenging. For instance, evaluating models on different window lengths and using multiple mismatch candidates could make the task harder, forcing the model to learn better latent representations. Additionally, it is important to consider whether a ceiling effect has been reached, as the accuracy levels achieved by the teams may be limited by the quality of the ground truth data (e.g., the SNR of the recorded data or the attention levels of the subjects during the experiment).

In the regression task, two out of three teams used an attention mechanism in their solution. Finetuning the models or otherwise incorporating subject information is also used to improve the correlations further. Regarding the regression task, it was observed that the models did not perform well on held-out subjects. However, the held-out subjects in the regression task all listened to podcasts. This discrepancy in performance might be attributed to the compression techniques used in podcast audio production, resulting in a more uniform loudness profile, which results in a slightly different, more uniform envelope representation. It is worth investigating whether this compression-induced uniformity challenges the models in reconstructing the envelope. Additionally, one can question whether this difference in correlation may stem from a) compression techniques influencing speech features, particularly the speech envelope, or b) the compressed audio inducing altered neural responses. Understanding the factors contributing to the performance drop on podcast stimuli can provide valuable insights for future improvements in the regression models. In the context of the match-mismatch task, no divergence in accuracy emerges between podcasts and audiobooks. Given the relative nature of the match-mismatch model's decision-making, which may hinge on the presence or absence of specific responses rather than the exact magnitude, the primary concern appears to be a speech feature representation issue. However, further investigation is warranted to elucidate the specific features or cues the models exploit in the speech envelope when confronted with different types of stimuli.

Considering the current dataset size of approximately 150 hours of data, the question arises as to whether increasing the data would lead to better results. With a larger dataset, larger and more complex models could capture more diverse patterns and achieve improved generalization. Investigating the impact of dataset size on model performance and the potential benefits of using larger models could provide valuable insights into the scalability and limits of the current approaches.

In summary, the discussions surrounding the match-mismatch and regression tasks of the Auditory EEG Decoding Challenge highlight the potential avenues for further research. Exploring modifications to the tasks, investigating the impact of compression techniques on model performance, and considering the benefits of larger datasets and models are crucial steps toward advancing auditory EEG decoding techniques and improving the accuracy and generalization capabilities of the models.

VII. CONCLUSION

In conclusion, the Auditory EEG Decoding Challenge encompassed two tasks: a classification task (match-mismatch) and a regression task (speech envelope reconstruction). The participation of numerous teams in each task highlighted the significance of the challenge in evaluating different models and establishing a benchmark for performance comparison. The challenge provided a comprehensive auditory EEG dataset, facilitating research in the field. Findings from the classification task indicated that various deep learning models, such as CNNs and self-attention models, exhibited comparable performances and demonstrated promising generalization capabilities to unseen subjects. Some new architectures were proposed, increasing accuracy for the state-of-the-art models and bringing innovation to the field. Additionally, the task establishes a benchmark to compare future models. Conversely, the regression task revealed the limitations of current models in generalizing to unseen subjects and their dependency on the type of stimuli, specifically Audiobooks versus Podcasts. These outcomes underscore the necessity for further advancements in regression models to enhance generalization and address the complexities associated with reconstructing speech envelopes from EEG signals. Nevertheless, some new ANN models significantly outperform the baseline model. The challenge serves as a platform for model evaluation and fosters future research and advancements in auditory EEG decoding techniques.

ACKNOWLEDGMENT

The authors would like thank to Amelie Algoet, Jolien Smeulders, Lore Kerkhofs, Sara Peeters, Merel Dillen, Ilham Gamgami, Amber Verhoeven, Vitor Vasconselos, Jard Hendrickx, Lore Verbeke and Ana Carbajal Chavez for their help with data collection.

REFERENCES

[1] N. Sriraam, "EEG based automated detection of auditory loss: A pilot study," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 723–731, Jan. 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417411010268

[2] P. M. P, K. Subramaniam, S. Yaccob, A. Hamid, and C. Hema, "EEG based detection of conductive and sensorineural hearing loss using artificial neural networks," *J. Next Gener. Inf. Technol.*, vol. 4, pp. 204–212, May 2013.

[3] M. J. Crosse, A. Di Liberto, and E. C. Lalor, "The multivariate temporal response function (mTRF) toolbox: A. MATLAB toolbox for relating neural signals to continuous stimuli," *Front. Hum. Neurosci.*, vol. 10, 2016, Art. no. 604.

[4] J. Vanthornhout, J. Wouters, J. Z. Simon, and T. Francart, "Speech intelligibility predicted from neural entrainment of the speech envelope," *J. Assoc. Res. Otolaryngol.*, vol. 19, no. 2, pp. 181–191, Apr. 2018, doi: 10.1007/s10162-018-0654-z.

[5] I. Iotzov and L. C. Parra, "EEG can predict speech intelligibility," *J. Neural Eng.*, vol. 16, no. 3, Mar. 2019, Art. no. 036008, doi: 10.1088/2F1741-2552/2Fab07fe.

[6] J. R. Katthi and S. Ganapathy, "Deep correlation analysis for audio-EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2742–2753, 2021.

[7] B. Accou, J. Vanthornhout, H. Van hamme, and T. Francart, "Decoding of the speech envelope from EEG using the VLAAI deep neural network," *Sci Rep.*, vol. 13, 2023, Art. no. 812.

[8] G. Krishna, C. Tran, Y. Han, M. Carnahan, and A. H. Tewfik, "Speech synthesis using EEG," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 1235–1238.

[9] M. Sakthi, A. Tewfik, and B. Chandrasekaran, "Native language and stimuli signal prediction from EEG," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 3902–3906.

[10] M. Thornton, D. Mandic, and T. Reichenbach, "Robust decoding of the speech envelope from EEG recordings through deep neural networks," *J. Neural Eng.*, vol. 19, no. 4, 2022, Art. no. 046007.

[11] T. de Taillez, B. Kollmeier, and B. Meyer, "Machine learning for decoding listeners' attention from EEG evoked by continuous speech," *Eur. J. Neurosci.*, vol. 51, pp. 1234–1241, 2017.

[12] A. de Cheveigné, D. D. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811918300338

[13] B. Accou, M. Jalilpour-Monesi, J. Montoya-Martínez, H. Van hamme, and T. Francart, "Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network," in *Proc. 28th Eur. Signal Process. Conf.*, 2021, pp. 1175–1179.

[14] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 941–945.

[15] M. J.Monesi, B. Accou, H. Van hamme, and T. Francart, "Extracting different levels of speech information from EEG using an LSTM-based model," in *Proc. Interspeech*, 2021, pp. 526–530.

[16] 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 *Proc. Interspeech*, 2022, pp. 4038–4042.

[17] M. Hosseini, L. Celotti, and Ã. Plourde, "Speaker-independent brain enhanced speech denoising," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 1310–1314.

[18] J. R. Katthi and S. Ganapathy, "Deep multiway canonical correlation analysis for multi-subject EEG normalization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 1245–1249.

[19] L. Bollens, T. Francart, and H. Van hamme, "Learning subject-invariant representations from speech-evoked EEG using variational autoencoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 1256–1260.

[20] G. Krishna, C. Tran, M. Carnahan, Y. Han, and A. H. Tewfik, "Generating EEG features from acoustic features," in *Proc. 28th Eur. Signal Process. Conf.*, 2021, pp. 1100–1104.

[21] C. Puffay et al., "Relating EEG to continuous speech using deep neural networks: A review," *J. Neural Eng.*, vol. 20, 2023, Art. no. 041003 .

[22] 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. https://www.biorxiv.org/content/early/2023/07/26/2023.07.24.550310

[23] S. Coren, "The lateral preference inventory for measurement of handedness, footedness, eyedness, and earedness: Norms for young-adults," *Bull. Psychon. Soc.*, vol. 31, no. 1, pp. 1–3, 1993.

[24] W. Hughson et al., "Manual for program outline for rehabilitation of aural casualties both military and civilian," *Trans. Amer. Acad. Ophthalmol. Otolaryngol.*, vol. 48, no. Suppl, pp. 1–15, 1944.

[25] B. Somers, T. Francart, and A. Bertrand, "A generic EEG artifact removal algorithm based on the multi-channel Wiener filter," *J. Neural Eng.*, vol. 15, no. 3, Jun. 2018, Art. no. 036007. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aaac92

[26] 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 Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 5, pp. 402–412, May 2017.

[27] A. de Cheveigné, M. Slaney, S. A. Fuglsang, and J. Hjortkjaer, "Auditory stimulus-response modeling with a match-mismatch task," *J. Neural Eng.*, vol. 18, no. 4, Aug. 2021, Art. no. 046040. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/abf771

[28] B. Accou, M. J. Monesi, H. Van hamme, and T. Francart, "Predicting speech intelligibility from EEG using a dilated convolutional network," May 2021. [Online]. Available: http://arxiv.org/abs/2105.06844

[29] L. Bollens, B. Accou, H. Van hamme, and T. Francart, "A large auditory EEG decoding dataset," *KU Leuven RDR, V3*, 2023, doi: 10.48804/K3VSND.

[30] M. Thornton, D. Mandic, and T. Reichenbach, "Relating EEG recordings to speech using envelope tracking and the speech-FFR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[31] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[32] F. Cui et al., "Relate auditory speech to EEG by shallow-deep attention-based network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[33] Z. Piao, M. Kim, H. Yoon, and H.-G. Kang, "HappyQuokka system for ICASSP 2023 auditory EEG challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[34] B. Van Dyck, L. Yang, and M. M. Van Hulle, "Decoding auditory EEG responses using an adapted Wavenet," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[35] J. P. Kulasingham, C. Brodbeck, A. Presacco, S. E. Kuchinsky, S. Anderson, and J. Z. Simon, "High gamma cortical processing of continuous speech in younger and older listeners," *NeuroImage*, vol. 222, 2020, Art. no. 117291. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811920307771

[36] A. Vaswani et al., “Attention is all you need,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[37] K. Cho et al., “Learning phrase representations using RNN encoder–Decoder for statistical machine translation,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734. [Online]. Available: https://aclanthology.org/D14-1179

[38] R. Xiong et al., “On layer normalization in the transformer architecture,” in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 10524–10533.

[39] A. V. D. Oord et al., “WaveNet: A generative model for raw audio,” Sep. 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[40] Q. Zhu, X. Zhao, J. Zhang, Y. Gu, C. Weng, and Y. Hu, “EEG2VEC: Self-supervised electroencephalographic representation learning,” 2023, *arXiv:2305.13957*.

[41] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.

[42] J. Van Canneyt, J. Wouters, and T. Francart, “Neural tracking of the fundamental frequency of the voice: The effect of voice characteristics,” *Eur. J. Neurosci.*, vol. 53, no. 11, pp. 3640–3653, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.15229

[43] M. Gillis, J. Vanthornhout, J. Z. Simon, T. Francart, and C. Brodbeck, “Neural markers of speech comprehension: Measuring EEG tracking of linguistic speech representations, controlling the speech acoustics,” *J. Neurosci.*, vol. 41, no. 50, pp. 10316–10329, 2021. [Online]. Available: https://www.jneurosci.org/content/41/50/10316

[44] M. J. Monesi, J. Vanthornhout, H. Van hamme, and T. Francart, “The role of vowel and consonant onsets in neural tracking of natural speech,” *J. Neural Eng.*, vol. 21, 2024, Art. no. 016002.