Auditory EEG decoding challenge for ICASSP 2024

Lies Bollens[1,2], Corentin Puffay[1,2], Bernd Accou[1,2], Jonas Vanthornhout[2], Hugo Van Hamme[1], and Tom Francart[1,2]

[1]Dept. of Electrical engineering (ESAT), KU Leuven, PSI
[2]Dept. Neurosciences

August 30, 2024

Posted on 30 Aug 2024 — CC-BY 4.0 — https://doi.org/10.36227/techrxiv.172504102.27360831/v1 — e-Prints posted on TechRxiv are preliminary reports that are not peer reviewed. They should not b...

; revised .

Auditory EEG decoding challenge for ICASSP 2024

Lies Bollens[1,2], Corentin Puffay[1,2], Bernd Accou[1,2], Jonas Vanthornhout[2], Hugo Van Hamme[1], Tom Francart [2]

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

Corresponding authors: Lies Bollens (email: lies.bollens@kuleuven.be), Tom Francart (email: tom.francart@kuleuven.be).

The research conducted in this paper is funded by KU Leuven Special Research Fund C24/18/099 (C2 project to Tom Francart and Hugo Van hamme), by three grants (1S89622N, 1SB1421N and 1290821N, 1S49823N) of the Research Foundation Flanders (FWO) and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 637424, ERC Starting Grant to Tom Francart).

ABSTRACT This paper describes the auditory EEG challenge, organized as one of the Signal Processing Grand Challenges at ICASSP 2024. The challenge provides electroencephalogram (EEG) recordings of 105 subjects who listened to continuous speech, as audiobooks or podcasts, while their brain activity was recorded. The challenge consists of two tasks that relate EEG signals to the presented speech stimulus. The first task, called match-mismatch, is to determine which of five speech segments induced a given EEG segment. The second task, called regression, is to reconstruct the Mel spectrogram from the EEG. EEG recordings of 85 subjects were provided as a training set so that challenge participants could train their models on a relatively large dataset. The remaining 20 subjects were used as held-out subjects for the evaluation step of the challenge.

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

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 [3].

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, e.g., [4]–[6]. While these approaches have been widely used in auditory neuroscience and have successfully been linked to speech intelligibility, [5], 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 [7]–[12].

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 proposed [13]. Given an EEG segment and two speech segments, the task is to determine which of the two speech segments corresponds to the EEG segment [14]. Recently, methods based on deep learning models have obtained promising results on this task, outperforming the linear methods [14]–[16]. In addition to the above-mentioned regression and match-mismatch tasks, 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 forward modeling, in which the task is to predict EEG from acoustic features [20].

However, a drawback of neural networks is that they typically require a large amount of data for training. Until recently, a large public auditory EEG dataset together with well-defined tasks to relate EEG to speech was not available, which made it difficult to compare the performance of

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

different models as people use different datasets as well as different metrics to evaluate their models (see [21]).

In order to solve these problems and provide a large dataset as well as establish benchmarks to compare future models against, in 2023, we organized the first edition of the Auditory EEG Decoding Challenge, presented at ICASSP 2023 [22] and published a large dataset of EEG recordings [23], which provided data from 85 subjects. The challenge used these data and included two alternative tasks. The first task, match-mismatch, aimed to determine which of two speech segments induced a given EEG segment. In the second regression task, the goal was to reconstruct the speech envelope from the EEG. For the match-mismatch task, 21 teams submitted a model. 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, 13 teams submitted a model, and the top teams significantly improved over the baseline models in the held-out stories test set. However, they failed to generalize to unseen subjects.

The positive results and success of the challenge led to the proposal of a second edition with a larger dataset comprising 20 new subjects and slightly more challenging tasks. In the ICASSP 2024 Auditory EEG Challenge, we provided a substantial auditory EEG dataset containing data from 105 subjects, each listening to an average of 110 minutes of single-speaker stimuli, totaling 188 hours of data. Teams competed to build the best model to relate speech to EEG in two tasks: 1) **match-mismatch**; given five segments of speech and a segment of EEG, which of the speech segments matches the EEG segment? and 2) **regression**; reconstruct the Mel spectrogram from the EEG.

II. DATASET

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 $10 - 100\mu$V) 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. The dataset contains data from 105 young, normal-hearing subjects (91 female/14 male, 21.4 ±1.9 years (sd), all hearing thresholds ≤ 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 [23]. 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 [24] and diagnoses of hearing loss and neurological pathologies). Subjects indicating any neurological or hearing-related diagnosis were excluded from the study. We measured air conduction hearing thresholds using the Hughson-Westlake method [25] for octave frequencies between 125 and 8000 Hz. Subjects with any hearing threshold > 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 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-3A 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 59 distinct speakers (32 female/27 male).

B. Training set

Both tasks share the training set. The training set contains EEG responses from 85 subjects. These subjects are numbered from sub-001 to sub-085. As shown in figure 1, each subject listens to between 6 and 10 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 666 trials from 85 subjects using 57 different stimuli, totaling 157 hours. The 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 the subject identifier. The auditory stimuli are provided in a separate folder *stimuli*.

C. Test set

The test set consists of data from 20 held-out subjects. The data for these subjects were acquired using the same protocol as for the other 85 subjects. The test set contains a total of 127 recordings, amounting to 31 hours of test data. This set is 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, subsections A and B).

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 Hz 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 [26]. Afterward, the EEG was re-referenced to a common average, and finally, the EEG was downsampled to 64 Hz.

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 (5 seconds for task 1 A match-mismatch, 30 seconds for task 2 B regression), performing many preprocessing steps could introduce artificial edge effects in the data, which could influence performance.

For task 2 B: regression, we defined a specific version of the Mel spectrogram. We calculate a Mel spectrogram using Librosa (CITE), using 10 subbands, which are spaced between 50 and 5000 Hz. The resulting Mel spectrogram is downsampled to 64 Hz. We provide code to create these Mel spectrogram representations, as well as to perform the described preprocessing steps.



FIGURE 1. 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.

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 [27]. 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 speech segment was not presented while the EEG segment was recorded.

There are many possibilities to define a match-mismatch task based on the concept of matched/mismatched segments [14], [15], [28].

In the first edition of the Auditory EEG decoding challenge, we opted for the method described in [14]. The task was to find the matched segment among two speech segments. The mismatched speech segment was chosen to be temporarily close to the matched segment, either ending one second before the start of the matched segment or starting one second after the end of the matched segment. While the results were promising, there was relatively little variance to be observed among teams, and most teams did not significantly outperform the baseline. Therefore, we opted to extend the task in the second edition of the challenge by providing four mismatched segments instead of one. The task of the model is now to determine which of the five speech segments corresponds to the EEG segment. The length of the segments was extended from last year's 3 seconds to 5 seconds this year in order to avoid increasing the difficulty of the task too much.

A schematic of the chosen match-mismatch task is illustrated in the left plot of Figure 2. The model is provided with six inputs of length 5 seconds: (1) a segment of EEG, (2) the time-aligned speech stimulus (match), and (3-6) four unaligned stimuli (mismatch). The task of the model in this paradigm is to determine which of the five input stimulus segments correspond to the EEG segment. We randomly

chose four non-overlapping mismatched segments from the same recording for every matched segment.

We implement and suggest the following recommendations during training to use the data in the classification paradigm. First, we present each EEG segment multiple times to the model, with the matched speech segment appearing at different places and the output label corresponding to the input position of the matched segment: (EEG, matched stimulus, mismatched stimulus 1, mismatched stimulus 2, mismatched stimulus 3, mismatched stimulus 4, output label 0), ..., (EEG, mismatched stimulus 1, mismatched stimulus 2, mismatched stimulus 3, mismatched stimulus 4, matched stimulus, output label 4). This ensures that matched and mismatched candidates occur equally in all input positions. 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 windowing (1s in our code) is dividable by the spacing between the matched and mismatched segment (1s). Failing to follow these suggestions will likely result in models simply remembering the segment-label associations learned during training and thus failing to generalize to the test set, as shown in more detail in [29].

2) Baseline method

An adapted version of the dilated convolutional network [14] is used as a baseline for task 1. The model uses three dilated convolutional layers, followed by a rectified linear unit (ReLU), to project EEG and speech inputs to a common embedded space. All stimulus segments share the weights for the convolutional layers. After these non-linear transformations, the embedded EEG representation is compared to each embedded stimulus representation using cosine similarity. Finally, all obtained similarity matrices are fed to an individual neuron with a linear activation function. The resulting outputs (corresponding to the number of stimulus segments) are fed into a softmax layer, which selects the matching segment.

When trained on the training set and tested on the test set of the challenge for 5 s segments, a classification accuracy of 50.74% is obtained.

3) Evaluation criteria

The test set for the match-mismatch task contains half of the EEG recordings from the 20 held-out subjects. It is not possible to use overlapping test sets, as in the regression task, the stimulus is the target of the regression, whereas in the match-mismatch task, the stimulus is one of the inputs.

We provide pairs of (EEG, stimulus 1, stimulus 2, stimulus 3, stimulus 4, and stimulus 5), with a length of 5 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 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 0, 1, 2, 3 or 4.

For evaluation, we calculate the mean accuracy per subject and the mean accuracy across test subjects to obtain the final score.

B. Task 2: Regression

1) Description

Task 2 is a regression problem: reconstructing the speech Mel spectrogram from the EEG. After reconstruction, the Pearson correlation measures the similarity between the reconstructed and original stimuli by measuring the correlation per band and then averaging over bands. The right plot in Figure 2 illustrates the regression task. The stimulus representation is the Mel spectrogram, as described in section D.

2) Baseline method

We include a simple linear model as the baseline for task 2. The linear decoder reconstructs the Mel spectrogram from EEG using a linear transformation across all channels and an integration window of 400 ms. The linear decoder is trained with the negative Pearson correlation as a loss function. When trained using the training set and tested on the challenge's test set, an average correlation score of 0.0214 is obtained.

3) Evaluation criteria

The test set consists of 20 held-out subjects listening to held-out stories. All stimuli are held out, i.e., they do not appear in the training set. We split the stimuli into several smaller segments of 30 seconds and made these available with a segment ID and a subject ID for each segment.

For each segment of 30 seconds, we expect a reconstructed Mel spectrogram, which is then compared to the original Mel spectrogram, as defined in section D, using Pearson correlation. We use the `scipy.stats.pearsonr` function to calculate the correlation c_i for each segment i, per band. Afterward, we average these values. For evaluation, the mean correlation value per subject is calculated, and the final score is obtained by averaging over subjects.

C. Provided online code

The dataset is available from the KU Leuven RDR platform [30]. Detailed information about the dataset can be found in the corresponding dataset paper [23]. 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 sparrkulee@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.

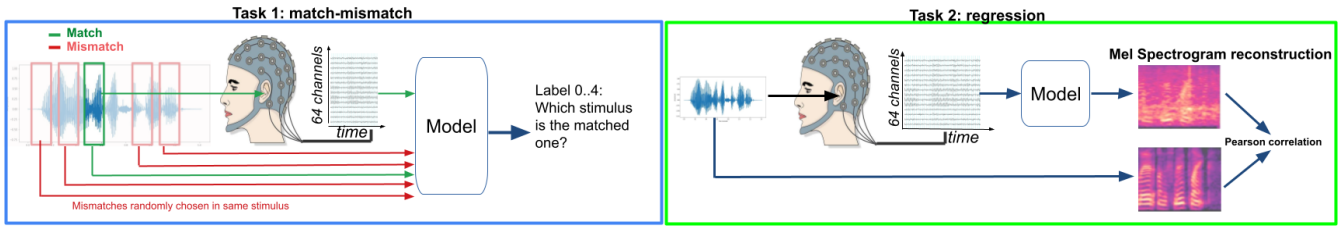


FIGURE 2. **Schematic Overview of the two tasks. Left: Task 1 (match-mismatch). The model gets six inputs: an EEG segment, the matched (in time) speech segment, and four mismatched segments. The task is to determine which of the five segments is matched. Right: Task 2 (regression). The task is to decode the Mel spectrogram from the EEG brain response. Reconstructed and original Mel spectrograms are compared by computing the Pearson correlation.**

We provided a simple Python codebase in our Github 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, please refer to the repository's README file.

IV. RESULTS

THERE were 59 submissions for the match-mismatch task and 49 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 *Fröhliche Würstchen (Thornton et al.)*, with their solution "Detecting gamma-band responses to the speech envelope for the ICASSP 2024 auditory EEG decoding signal processing grand challenge". They obtained a final accuracy score of 62.8%. This team also won the match-mismatch task in the previous edition of the challenge [31]. Their approach builds on the architecture from their previous year's submission. Their solution involves combining two decoders: a low-frequency envelope-tracking decoder and a high-frequency, gamma-band EEG decoder, which relates the speech envelope to EEG signals within the broad gamma range (70 Hz to 220 Hz). The gamma-band EEG signals were derived by filtering the EEG between 35 Hz and 150 Hz and resampling the original signal to 512 Hz. In addition, they used the low-frequency signals, which are downsampled to 64 Hz. A composite decoder was formed by combining the output logits of both decoders using linear discriminant analysis (LDA). The LDA classifier estimates the probability that a pair is matched, with the highest probability pair considered the match. Additionally, they used an averaging technique, averaging the output logits of 16 trained models of each decoder. This method is similar to majority voting in machine learning, aggregating multiple outputs to make the final decision.

The second place was secured by the team *PKU_SHRC (Wang et al.)* with their solution titled "Self-supervised speech representation and contextual text embedding for match-mismatch classification with EEG recording." Their approach introduced several novel aspects. Firstly, instead of relying solely on low-level speech representations like the speech envelope or Mel spectrogram, they integrated more complex speech representations. Their final solution combined the speech envelope and Mel spectrogram with representations extracted from the self-supervised speech model wav2vec 2.0 [32]. They used the output of the 14th layer of wav2vec 2.0, reducing its dimensionality from 1024 to 64 using PCA. Additionally, they extracted contextual text embeddings using the pre-trained language model GPT-2 [33]. Transcripts were obtained via Whisper [34], and the ninth layer of GPT-2 was used as input, with dimensionality reduced from 768 to 4 using PCA. Secondly, they introduced a contrastive learning objective with the InfoNCE loss [35], enhancing the possibility for discrimination between different stimuli. Thirdly, their proposed architecture featured deep convolutional neural networks with skip connections for both EEG and speech encoders. These encoders comprised stacks of convolutional blocks with GELU activation functions and residual connections. Lastly, they employed ensemble learning to improve model performance. A total of 136

TABLE 1. **Summary table of the top 5 teams. For the match-mismatch task (the more popular), there are three winners; for the regression task, there are two. We report the final mean score and standard deviation (classification accuracy (%) for match-mismatch, Pearson correlation (ρ) for the regression task). P-values are calculated with respect to the baseline using a Wilcoxon signed-rank test ($\alpha = 0.05$) with Holm-Bonferroni correction.**

	Team	Mean	Std	P-value
Task 1: Match-mismatch		(%)		
	Baseline (dilation)	**50.7**	12.7	
1	Fröhliche Würstchen	**62.8**	14.7	$8.20 * 10^{-5}$
2	PKU_SHRC	**60.3**	15.7	$8.22 * 10^{-5}$
3	KLSACUQ	**58.7**	15.8	$1.34 * 10^{-4}$
Task 2: Regression		(ρ)		
	Baseline (linear)	**0.0214**	0.00955	
1	PKU_SHRC	**0.0420**	0.0177	$8.58 * 10^{-5}$
2	Bigbrain	**0.0413**	0.0170	$8.58 * 10^{-5}$

models were trained, and the final classification output was determined through majority voting.

The team *KLSACUQ (Qiu et al.)* ranked third in the match-mismatch task with their solution titled "BMMSNET: Bidirectional Mapping and Multilevel Similarity Comparison for EEG-Speech Match-Mismatch Problem." Their approach proposed an architecture where both EEG and speech are mapped to different latent representations. They introduced a multilevel similarity comparison, which calculates and incorporates the similarity between EEG and speech at various network depths for the final classification. Since EEG is mapped to speech and vice versa, the early layers of the EEG network are compared to the late layers of the speech network and vice versa. Both the EEG and speech encoders utilized a stack of convolutional blocks. Additionally, before the EEG encoder, the EEG data is processed through an EEG enhancement module comprising a convolutional encoder, a Transformer encoder, and a convolutional decoder. They also employed various data augmentation methods, including adding parts of another EEG segment, masking several channels for a set duration, adding white noise, and multiplying each channel's data by different coefficients. The final output was the average of the outputs from all augmented models.

The winner of the regression task was the team *PKU_SHRC (Xu et al.)* with their solution titled "Convconcatnet: A Deep Convolutional Neural Network to Reconstruct Mel Spectrogram from the EEG." Their approach featured six blocks, each comprising a convolutional stack followed by fully connected linear layers, an output context layer inspired by the proposed baseline VLAAI model [8], and a sample spatial attention layer. In their architecture, after each block, the EEG, the output of the context layer, and the output of the spatial attention layer were concatenated and fed into the next layer. During training, both the envelope and the Mel spectrogram were decoded, which enhanced the Mel spectrogram prediction. Additionally, they employed ensemble learning, training a total of 520 models. For the final Mel spectrogram prediction, the outputs from all models were normalized and averaged.

The second place for the regression task was won by the team *Bigbrain (Fang et al.)* with their solution titled "Cross-Attention-Guided Wavenet for Mel Spectrogram Reconstruction in the ICASSP 2024 Auditory EEG Challenge." Their approach utilized several iterations of Wavenet [36], sequentially reconstructing the envelope, the 10-band Mel spectrogram, the 80-band Mel spectrogram, and the magnitude spectrogram in a coarse-to-fine order. They introduced a cross-attention mechanism to account for correlations across modalities. This mechanism fused the EEG with the output of each block (envelope, 10-band Mel spectrogram, or 80-band spectrogram) before feeding it into the Wavenet block, aiming to capture dependencies and relationships between different features and modalities. As a loss function, they employed a combination of L1 loss, negative Pearson correlation, and Kullback-Leibler divergence. Additionally, they used data augmentation with the Mixup technique, which creates artificial segments of data by mixing two existing segments.

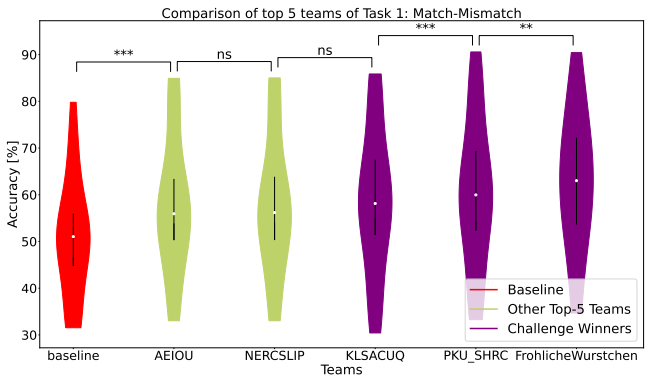


FIGURE 3. accuracy of task 1: Match-mismatch of top 5 models. The violin plots are shown per team over subjects. Each point in the violin plots corresponds to one team's average accuracy for one subject. Red: Baseline model. Purple: Challenge winners. Green: For comparison. Models in top 5, but not in top 3. Pairwise significance is calculated using single-sided Wilcoxon signed-rank tests with Bonferroni-Holm corrections ($* : \rho < 0.05$, $** : \rho < 0.01$, $*** : \rho < 0.001$.).

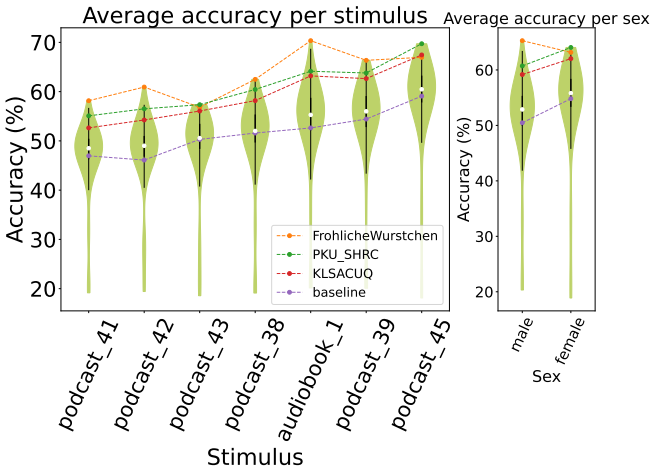


FIGURE 4. Match-mismatch accuracy of top models per stimuli and sex of the speaker of the stimuli. Each point in the violin plots corresponds to the average value of one team over all subjects. Violin plots are shown over teams. Left: Average accuracy per unique stimulus over all the stimuli included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

V. SUMMARY OF RESULTS

A. Task 1: match-mismatch

In total, 59 teams submitted to task 1. Of these 59 teams, 43 performed better than the baseline. We tested the significance of these teams' results with respect to the baseline using a single-sided Wilcoxon signed-rank test with Bonferroni-Holm corrections and an alpha value of 0.001. Nineteen teams significantly outperformed the baseline.

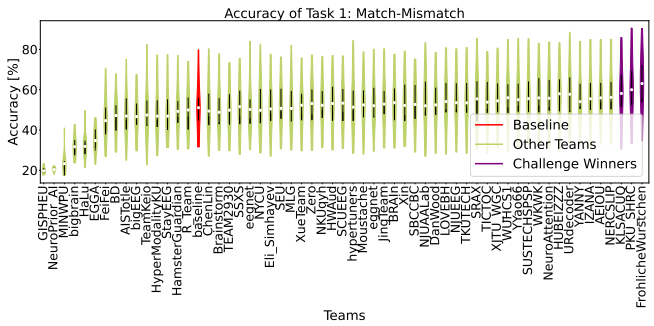


FIGURE 5. Comparing the match-mismatch accuracy of different teams on the test sets. The violin plots are shown per team over subjects. Each point in the violin plots corresponds to one team's average accuracy for one subject.

The final accuracies from all teams ranged between 19.96% and 62.80%, with a mean score of $50.15 \pm 9.51\%$. The scores of the top three teams are summarized in Table 1. The teams *Fröhliche Würstchen*, *PKU_SHRC*, and *KLSACUQ* won first, second, and third place, with final scores of 62.8%, 60.3%, and 58.7%, respectively.

To measure the significance of these results, we performed a pairwise comparison between the subsequent teams. As shown in figure 3, the first-place team, *Fröhliche Würstchen*, significantly outperformed the second-place team, *PKU_SHRC* ($p = 0.004$). Similarly, the second-place team, *PKU_SHRC*, significantly outperformed the third-place team, *KLSACUQ* ($p = 0.0001$). However, the results of the third-place team, *KLSACUQ*, did not significantly differ from the results of the teams ranked in places 4-8.

In the previous edition (2023) of the challenge, a notable observation was that the sex of the speaker seemed to affect the match-mismatch accuracy of the models. Models obtained higher accuracy on male-spoken stories compared to female-spoken stories, consistent with results obtained using either the fundamental frequency f0 or linguistic features for neural tracking [37]–[39]. However, when we conducted the same analysis for this year's stimuli, this observation did not hold. Only the top team achieved higher accuracy for male-spoken stimuli than for female-spoken stimuli, which is consistent given their use of gamma-band EEG. These results are consistent with previous findings [38], where it is shown that the fundamental frequency f0 cannot be tracked for female speakers using EEG. Therefore, if f0 is included, the match-mismatch accuracy improves for male speakers and remains unchanged for female speakers. However, in the case where the fundamental f0 frequency is not included as a speech feature, no consistent conclusion can be drawn between the sex of the speaker and the match-mismatch performance.

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 figure 4, there is no significant difference in the type of stimulus used as the ordering alternates between audiobooks and podcasts.

The top teams' chosen stimulus representations varied, with the successful use of the Mel spectrogram, fundamental frequency (f0), wav2vec, and GPT-2 representations used in combination with the speech envelope.

Another common strategy among all the top teams was the use of ensemble learning. This involves training multiple instances of a model and averaging their predictions to obtain the final output. This strategy was first introduced in the context of the Auditory EEG challenge by the previous year's winner of task 1 [40], and it improved results for all three top teams this year.

Last year, most teams achieved relatively similar match-mismatch accuracy levels of around 78%. Several hypotheses were proposed to explain this phenomenon. Firstly, many models were based on the same baseline model, limiting the diversity of approaches. Secondly, the match-mismatch task may not have been sufficiently challenging, as it involved a relative decision between two segments of speech. Lastly, a ceiling effect on performance may have been present due to the low signal-to-noise ratio (SNR) of the recorded EEG data.

To address these issues, the match-mismatch task was modified to a paradigm where the model has to find the matched segment out of five speech segments, as opposed to out of two in the previous year. This year, the scores between teams show larger differences. However, upon examining the proposed model architectures, we see a noticeable increase in complexity compared to previous years. Therefore, it is unclear whether the original task was too easy or if the wider range of scores across teams can be attributed to the utilization of more complex models capable of capturing intricate features and relationships between the EEG and speech data.

B. Task 2: regression

Figure 8 provides an overview of the final scores of the competing teams of 2024. Out of 48 teams, 45 scored higher than the linear baseline. We tested the significance of these results with respect to the linear baseline using a single-sided Wilcoxon signed-rank test with Bonferroni-Holm corrections and an alpha value of 0.001. All of these teams significantly outperformed the linear baseline. The top team, *PKU_SHRC*, achieved an average Pearson correlation score of 0.042, while the second team, *Bigbrain*, achieved a score of 0.041, as shown in Table 1.

To assess the significance of these results, we conducted pairwise comparisons between subsequent teams. There was no significant difference between the first team, *PKU_SHRC*, and the second team, *Bigbrain* ($p = 0.004$). However, the results of the second team, *Bigbrain*, significantly differed from those of the third team and all subsequent teams

($p = 0.008$). Significance values were computed using a two-sided Wilcoxon signed-rank test with Bonferroni-Holm corrections.

As previously emphasized in [41] and [29], caution is warranted when interpreting correlation values. While correlation values often serve as metrics for model performance, their significance should be evaluated against the null distribution. Since auditory data exhibits sequential properties, adjacent samples within data are not independent, potentially leading to spuriously high correlations. To address this, we determined the null distribution of the data, following the methodology outlined in [41].

In brief, we conducted a permutation test where each predicted sample of each team was randomly shifted (circularly) ten times, yielding a total of 490 shifted samples per actual sample. Then, the correlation with the actual data was computed. We determined the 95th percentile of this distribution and considered it as the significance threshold. It is important to note that circularly shifting the Mel spectrogram preserves the correlation across different frequency bands, which reflects real speech signals. We hypothesize that maintaining these relationships provides a better estimation of the null distribution compared to randomly shifting all bands separately.

We visualized the comparison between the null distribution and the actual distribution in Figure 6. The 95th percentile value of the null distribution equals 0.056, which surpasses the average correlation value of the displayed team (mean = 0.042), as well as all other teams. However, the lack of statistical significance in any of the results can primarily be attributed to the window length and a potential lack of frequency selectivity of the reconstructed spectrogram. To illustrate this, we construct new predictions by always concatenating four segments of reconstructed Mel, resulting in reconstructed EEG segments of two minutes. Subsequently, when recalculating the null distribution, the 95th percentile is now equal to 0.032, hence rendering the results significant.

These findings underscore the importance of comparing results against their null distribution. Given the relatively low reconstructed correlations, longer windows should be considered when decoding the Mel spectrogram from the EEG. Future research efforts could focus on enhancing these correlations, potentially enabling statistically significant results with shorter windows.

Lastly, we conducted a similar analysis for the regression task as we did for the match-mismatch task. Specifically, we grouped the results of different teams according to the type of stimulus and the sex of the story's speaker.

We observed an effect based on the sex of the story's speaker, with models performing better on male-spoken stimuli than female-spoken stimuli (see Figure 7). However, since the reverse was true for the match-mismatch case and female spoken stimuli had a higher average score, care should be taken when interpreting these results.

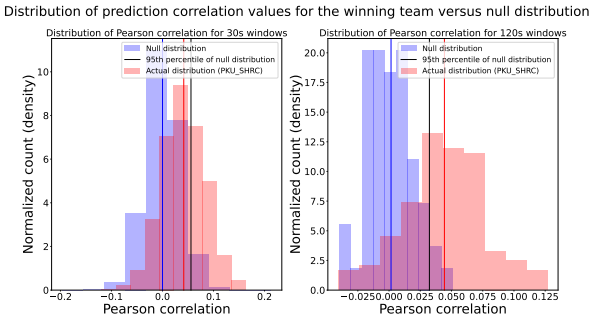


FIGURE 6. comparison of the null distribution versus the actual (predicted) correlation values for the winning team (PKU_SHRC). Left: comparison for segments of 30 seconds, as was the objective of the challenge. p-value=0.056. Right: comparison for concatenated segments of total length 2 minutes. p-value=0.032.

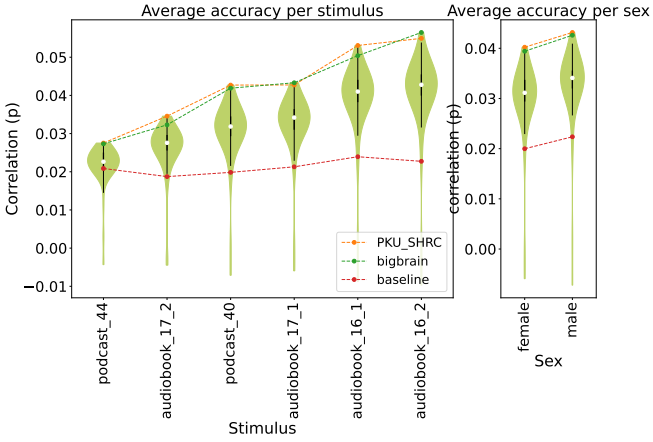


FIGURE 7. Correlation scores of top models per stimulus and sex of the speaker of the stimuli. Each point in the violin plots corresponds to the average value of one team over all subjects. Left: Average correlation per unique stimulus. Stimuli are ordered based on increasing mean correlation. Right: Average correlation per sex of the speaker of stimuli, averaged over stimuli and subjects.

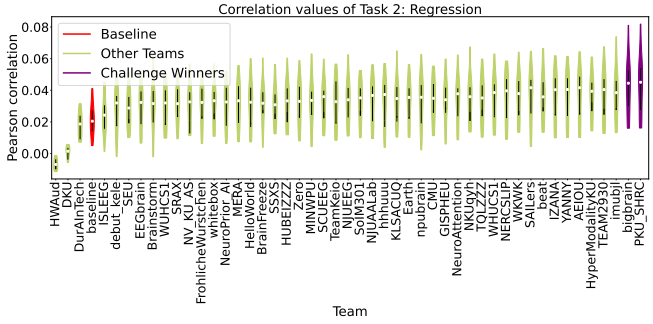


FIGURE 8. Comparing the final correlation scores of different teams on the test sets. Each point in the violin plot corresponds to the average correlation of one subject for one team.

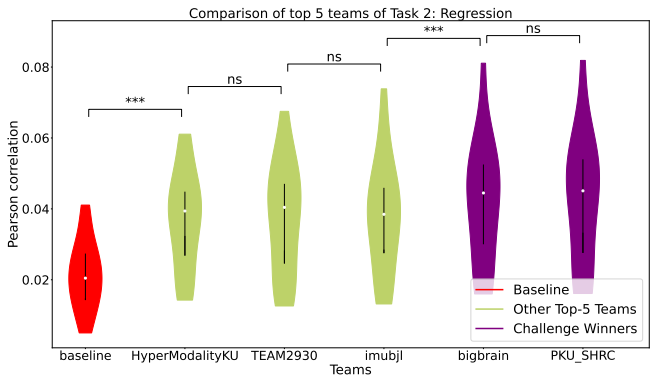


FIGURE 9. correlation scores of task 2: Regression of top 5 models. The violin plots are shown per team over subjects. Each point in the violin plots corresponds to one team's average accuracy for one subject. Red: Baseline model. Purple: Challenge winners. Green: for comparison. Models are in the top 5 but not in the top 2. Pairwise significance is calculated using single-sided Wilcoxon signed-rank tests with Bonferroni-Holm corrections ($* : \rho < 0.05$, $** : \rho < 0.01$, $*** : \rho < 0.001$.).

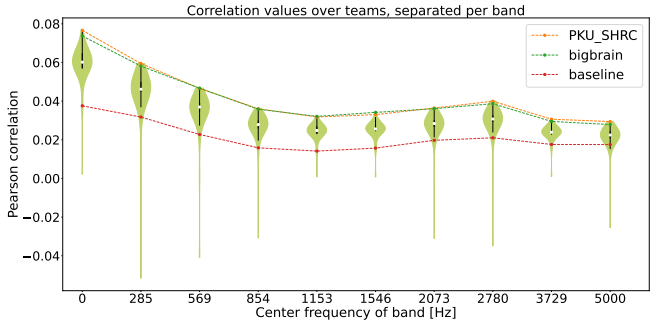


FIGURE 10. Comparing the regression scores per band. Violin plots are shown over teams. One point corresponds to the mean value of one team for all segments of a specific band.

VI. DISCUSSION

MOST top teams employed speech representations such as the Mel spectrogram rather than the traditional speech envelope, aligning with findings from similar studies [16]. This shift to richer, more complex representations, including wav2vec and GPT text embeddings, marked a notable advancement from last year's solutions, which mostly used the Mel spectrogram, along with the speech envelope.

Last year's match-mismatch task winner [40] introduced ensemble learning, training multiple model instances and averaging their outputs, which led to improved performance and reduced variance. This strategy was adopted by 4 of the top 5 teams this year, with the average number of models ranging from 5 to 520, consistently resulting in higher performance.

Two of the top five teams incorporated an attention mechanism in their solutions, achieving good performances. Given the recent success of attention and transformer-like architectures [**?**], this is unsurprising. However, the winner of the match-mismatch challenge used a relatively simple network based on the baseline network, which outperformed the attention-based approach used by the team which placed second in the match-mismatch task. Further research is necessary to determine if larger EEG datasets could enhance the efficacy of transformer-like architectures. We encourage researchers to explore this potential avenue.

Four of the five teams utilized architectures with a substantial amount of parameters, which featured stacked convolutional blocks and residual connections. Interestingly, the winner of the match-mismatch task employed a relatively simple architecture closely aligned with the baseline model. This prompts the question of whether complex architectures offer a true advantage or if focusing on feature representations and preprocessing methods might be more beneficial. Too complex architectures might overfit on the data and thus yield a lower total performance. Notably, the winning team used EEG data sampled at 512 Hz and incorporated gamma-band responses. Further research is required to draw more definitive conclusions on this matter.

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 data (e.g., the SNR of the recorded data, the attention levels or the general mental state of the subjects during the experiment).

Compared to last year, the dataset increased by 20 participants, adding a total of 36 extra hours of data. With the current dataset size at approximately 200 hours, it is pertinent to question whether further increasing the data volume would lead to better results. A larger dataset could enable more complex models to capture diverse patterns and achieve improved generalization. Notably, the architectures proposed this year exhibit increased complexity, suggesting that an even larger dataset might be beneficial despite the substantial size of the current dataset. Several teams employed various data augmentation techniques, including adding parts of another EEG segment, masking several channels for a set duration, adding white noise, or creating artificial segments of data by mixing two existing segments. These techniques facilitated model training, indicating that expanding the dataset further could enhance model performance.

Investigating the impact of dataset size on model performance and the benefits of using larger models could provide valuable insights into the scalability and limits of current approaches. A similar analysis has been conducted for the baseline dilation model [14] and shows that the dilation model reaches a ceiling accuracy after including 27 subjects and that adding further subjects does not yield better performance. The same author performed the same experiment for the VLAAI network [8], a model that predicts the envelope and has notably more parameters and showed that there was an increase when adding more subjects, up until the 80 subjects that were included in the dataset. However, the biggest relative increase was found when

adding the first subjects, and adding extra subjects yielded diminishing returns.

In summary, the discussions surrounding the match-mismatch and regression tasks of the Auditory EEG Decoding Challenge underscore potential avenues for further research. Key areas include exploring task modifications and evaluating the benefits of larger datasets and more complex models. These steps are crucial for advancing auditory EEG decoding techniques and improving model accuracy and generalization capabilities.

VII. CONCLUSION

In conclusion, the Auditory EEG Decoding Challenge encompassed two tasks: a classification task (match-mismatch) and a regression task (Mel spectrogram 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 compared to the state-of-the-art models and bringing innovation to the field. Additionally, the task establishes a benchmark to compare future models. Conversely, while all solutions for the regression tasks significantly outperformed the baseline, none of the results were significantly different from the proposed null distribution. These outcomes underscore the necessity for further advancements in regression models to enhance generalization and address the complexities associated with reconstructing Mel spectrograms from EEG signals. This challenge serves as a platform for model evaluation and fosters future research and advancements in auditory EEG decoding techniques.

ACKNOWLEDGMENT

The research conducted in this paper is funded by KU Leuven Special Research Fund C24/18/099 (C2 project to Tom Francart and Hugo Van hamme), by three PhD grants (1S49823N, 1S89622N,1SB1421N) of the Research Foundation Flanders (FWO). The authors thank 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 Systems with Applications*, 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," *Journal of Next Generation Information Technology*, vol. 4, pp. 204–212, May 2013.

[3] 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.

[4] 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," *Frontiers in Human Neuroscience*, vol. 10, no. NOV2016, pp. 1–14, 2016.

[5] J. Vanthornhout, J. Wouters, J. Z. Simon, and T. Francart, "Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope," *Journal of the Association for Research in Otolaryngology*, vol. 19, no. 2, pp. 181–191, Apr. 2018. [Online]. Available: https://doi.org/10.1007/s10162-018-0654-z

[6] I. Iotzov and L. C. Parra, "EEG can predict speech intelligibility," *Journal of Neural Engineering*, vol. 16, no. 3, p. 036008, Mar. 2019. [Online]. Available: https://doi.org/10.1088%2F1741-2552%2Fab07fe

[7] J. Reddy Katthi and S. Ganapathy, "Deep correlation analysis for audio-eeg decoding," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2742–2753, 2021.

[8] B. Accou, J. Vanthornhout, H. V. hamme, and T. Francart, "Decoding of the speech envelope from eeg using the vlaai deep neural network," *Scientific Reports*, vol. 13, no. 1, p. 812, 2023.

[9] G. Krishna, C. Tran, Y. Han, M. Carnahan, and A. H. Tewfik, "Speech synthesis using EEG," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1235–1238.

[10] M. Sakthi, A. Tewfik, and B. Chandrasekaran, "Native language and stimuli signal prediction from EEG," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3902–3906.

[11] 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, p. 046007, 2022.

[12] T. de Taillez, B. Kollmeier, and B. Meyer, "Machine learning for decoding listeners' attention from EEG evoked by continuous speech," *European Journal of Neuroscience*, vol. 51, 12 2017.

[13] 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

[14] 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," *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 1175–1179, 2021.

[15] 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 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 941–945, ISSN: 2379-190X.

[16] M. Jalilpour Monesi, B. Accou, H. Van hamme, and T. Francart, "Extracting different levels of speech information from EEG using an lstm-based model," *Proceedings Interspeech 2021*, pp. 526–530, 2021.

[17] M. Hosseini, L. Celotti, and Plourde, "Speaker-independent brain enhanced speech denoising," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1310–1314.

[18] J. R. Katthi and S. Ganapathy, "Deep multiway canonical correlation analysis for multi-subject EEG normalization," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1245–1249, 2021.

[19] L. Bollens, T. Francart, and H. Van hamme, "Learning subject-invariant representations from speech-evoked EEG using variational autoencoders," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1256–1260.

[20] G. Krishna, C. Tran, M. Carnahan, Y. Han, and A. H. Tewfik, "Generating EEG features from acoustic features," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1100–1104.

[21] C. Puffay, B. Accou, L. Bollens, M. Monesi, J. Vanthornhout, H. Van hamme, and T. Francart, "Relating EEG to continuous speech using deep neural networks: a review," *Journal of Neural Engineering*, Jul. 2023.
[22] M. J. Monesi, L. Bollens, B. Accou, J. Vanthornhout, H. Van Hamme, and T. Francart, "Auditory eeg decoding challenge for icassp 2023," *IEEE Open Journal of Signal Processing*, 2024.
[23] B. Accou, L. Bollens, M. Gillis, W. Verheijen, H. Van hamme, and T. Francart, "Sparrkulee: A speech-evoked auditory response repository from ku leuven, containing the eeg of 85 participants," *Data*, vol. 9, no. 8, 2024. [Online]. Available: https://www.mdpi.com/2306-5729/9/8/94
[24] S. COREN, "The lateral preference inventory for measurement of handedness, footedness, eyedness, and earedness - norms for young-adults," *BULLETIN OF THE PSYCHONOMIC SOCIETY*, vol. 31, no. 1, pp. 1–3, JAN 1993.
[25] W. Hughson, H. Westlake *et al.*, "Manual for program outline for rehabilitation of aural casualties both military and civilian," *Trans Am Acad Ophthalmol Otolaryngol*, vol. 48, no. Suppl, pp. 1–15, 1944.
[26] 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, p. 036007, jun 2018. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aaac92
[27] A. de Cheveigné, D. D. E. 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, May 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811918300338
[28] A. de Cheveigné, M. Slaney, S. A. Fuglsang, and J. Hjortkjaer, "Auditory stimulus-response modeling with a match-mismatch task," *Journal of Neural Engineering*, vol. 18, no. 4, p. 046040, aug 2021. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/abf771
[29] C. Puffay, B. Accou, L. Bollens, M. J. Monesi, J. Vanthornhout, T. Francart *et al.*, "Relating eeg to continuous speech using deep neural networks: a review." *Journal of Neural Engineering*, 2023.
[30] L. Bollens, B. Accou, H. Van hamme, and T. Francart, "A Large Auditory EEG decoding dataset," 2023. [Online]. Available: https://doi.org/10.48804/K3VSND
[31] M. Thornton, D. P. Mandic, and T. Reichenbach, "Decoding envelope and frequency-following eeg responses to continuous speech using deep neural networks," *IEEE Open Journal of Signal Processing*, pp. 1–18, 2024.
[32] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.
[33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
[34] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.
[35] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
[36] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," Sep. 2016, arXiv:1609.03499 [cs]. [Online]. Available: http://arxiv.org/abs/1609.03499
[37] J. Van Canneyt, J. Wouters, and T. Francart, "Neural tracking of the fundamental frequency of the voice: The effect of voice characteristics," *European Journal of Neuroscience*, vol. 53, no. 11, pp. 3640–3653, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.15229
[38] 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," *Proceedings Interspeech 2022*, 2022-09-18.
[39] 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," *Journal of Neuroscience*, vol. 41, no. 50, pp. 10 316–10 329, 2021. [Online]. Available: https://www.jneurosci.org/content/41/50/10316
[40] M. Thornton, D. Mandic, and T. Reichenbach, "Relating EEG recordings to speech using envelope tracking and the speech-FFR," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–2.
[41] M. Crosse, N. Zuk, G. Di Liberto, A. Nidiffer, S. Molholm, and E. Lalor, "Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research," *Frontiers in Neuroscience*, vol. 15, 11 2021.