ICASSP 2023 AUDITORY EEG DECODING CHALLENGE

Lies Bollens^{1,2,†}, *Mohammad Jalilpour Monesi*^{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
† These authors contributed equally to this work

ABSTRACT

This paper describes the auditory EEG challenge which was organized as one of the Signal Processing Grand Challenges of ICASSP 2023. This challenge consists of two tasks in which the goal is to relate electroencephalogram (EEG) signals to the presented speech stimulus. In the first task, named match-mismatch, the goal is to determine which of the two speech segments matches with a given EEG segment. In the second task, a regression task, the goal is to reconstruct the speech envelope from the EEG.

Index Terms— EEG, match-mismatch, speech decoding

1. INTRODUCTION

Electroencephalography (EEG) is a popular neuroimaging technique to study how the brain processes speech. It has applications in fundamental neuroscience research, as well as in the diagnosis of potential hearing loss. A popular approach is to relate the EEG of a person to a representation of the speech signal they were listening to. Typically, linear regression is used to predict the EEG signal from the stimulus or to reconstruct a representation of the stimulus from the EEG [1]. However, these linear models have low reconstruction scores and high inter-subject variability. As an alternative, several methods based on artificial neural networks (ANNs) have been proposed to improve upon linear models [2, 3].

Instead of directly decoding a speech feature from the EEG, which is a challenging *regression* problem, an alternative *classification* paradigm, named match-mismatch, has been recently proposed. Given an EEG segment, the task is to determine whether it matches with a given speech [4]. Recently, methods based on deep learning models have obtained promising results on this task, outperforming the linear methods [5, 6].

However, a drawback to neural networks is that they typically require a large amount of data to train. Also, no large public auditory EEG dataset exists together with well-defined tasks to relate EEG to speech, making it difficult to compare the performance of different models. In the 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 have competed to build the best model to relate speech to EEG in the following two tasks:

1. **match-mismatch**; given two segments of speech and a segment of EEG, which of the speech segments matches the EEG segment?
2. **regression**; reconstruct the speech envelope from the EEG.

More details can be found on the challenge website (https://exporl.github.io/auditory-eeg-challenge-2023). Code to get started is provided in a public GitHub repository (https://github.com/exporl/auditory-eeg-challenge-2023-code).

2. DATASET

We collected EEG data from 85 young, normal-hearing Dutch-speaking participants in a well-controlled lab environment (soundproof and electromagnetically shielded booth) using a 64-channel Biosemi ActiveTwo EEG recording system with 64 active Ag-AgCl electrodes. During the experiment, participants listened to between 8 and 10 randomized single-speaker stories in Flemish, either podcasts or audiobooks. All participants provided explicit consent for their anonymized data to be included in a publicly accessible dataset [7].

The training set contains EEG responses from 71 subjects, for 508 recordings, each approximately 15 minutes long, amounting to 120h of training data. All subjects listened to a reference story, *Audiobook 1*. The test set has two parts: *held-out stories* for the 71 subjects in the training set and *held-out subjects* for 14 new subjects, totaling 36h.

Two dataset versions are provided: raw EEG data downsampled from 8192 Hz to 1024 Hz, and a preprocessed version in MATLAB. The preprocessing includes downsampling the signal from 8192 Hz to 1024 Hz, removing artifacts with a multichannel Wiener filter, re-referencing to a common average, and further downsampling to 64 Hz. For task 2 (see section 4), a specific envelope version is defined using a gammatone filter bank, as defined in [8].

3. TASK 1: MATCH-MISMATCH

3.1. Description

Task 1 is a classification problem in a match-mismatch paradigm. In this paradigm, three inputs are presented to the model: (1) a segment of EEG, (2) the time-aligned speech stimulus (match), and (3) an imposter stimulus (mismatch). The task of the model is to determine which of the input speech segments corresponds to the EEG segment. The performance metric is the classification accuracy (%). Figure 1 illustrates the problem. The input length of all (EEG, speech) pairs is 3 s. We define the mismatched stimulus to be temporally close to the matched one by taking the segment starting either one second after the end of the matched segment or 4 seconds before the start of the matched segment.

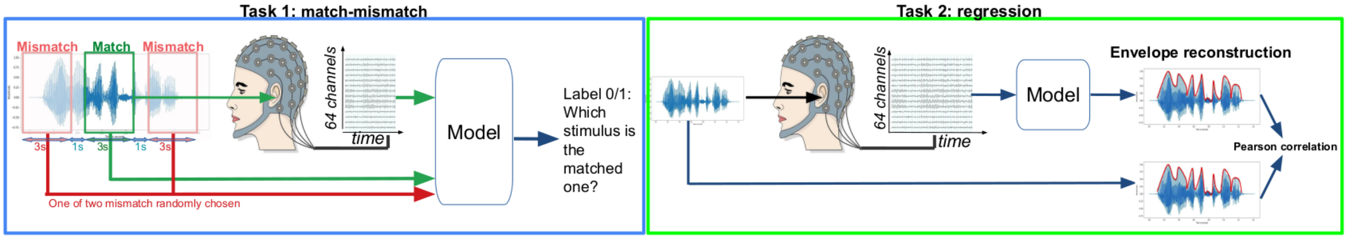

Fig. 1. Schematic overview of the two tasks. Left: Task 1 (match-mismatch). The model gets three inputs: an EEG segment, the matched (in time) speech segment, and a mismatched segment. The task is to determine which of the two segments is matched. Right: Task 2 (regression). The task is to decode the speech envelope from the EEG brain response.

3.2. Baseline method

The dilated convolutional network is used as a baseline for task 1 [6]. The model uses 3 dilated convolutional layers, followed by a rectified linear unit (ReLU), to project EEG and speech inputs to a common embedded space. Then, the cosine similarities between the embedded representation of EEG and speech inputs are fed to a single neuron with sigmoid non-linearity to create the final prediction. The model reaches 77% accuracy on the test set of the challenge.

3.3. Evaluation criteria

The test set consists of held-out stories (test set 1) and held-out subjects (test set 2). For both test sets, we provide pairs of (EEG, stimulus 1, and stimulus 2), with a length of 3 seconds. For evaluation, the mean accuracy per subject is calculated. Then, we calculate the mean accuracy over test subjects from test set 1 (S_1) and test set 2 (S_2) and add them to obtain the final score: $Score = 2/3S_1 + 1/3S_2$.

4. TASK 2: REGRESSION

4.1. Description

Task 2 is a regression problem in which the stimulus envelope is reconstructed from the EEG. Pearson correlation between the original speech and reconstructed envelopes is used as a metric. Figure 1 shows a high-level setup of this task.

4.2. Baseline method

We include a linear baseline decoder [1] as a baseline for task 1. The linear decoder reconstructs the speech envelope from EEG by using a linear transformation across all channels and an integration window of 500 ms. The linear decoder here is trained subject-independently with negative Pearson correlation as a loss function. When applied to the training and test sets of the challenge, an average correlation score of 0.10 is obtained.

4.3. Evaluation criteria

The test set consists of held-out stories (test set 1) and held-out subjects (test set 2). We have split up the stimuli into several smaller segments of 60 seconds. Pearson correlation is calculated for each segment based on original and reconstructed envelopes. For evaluation, the mean correlation value per subject is calculated. Then, we calculate the mean correlation value over test subjects from test set 1 (S_1) and test set 2 (S_2) and add them to obtain the final score: $Score = 2/3S_1 + 1/3S_2$.

5. RESULTS

In Task 1, the performance of 10 out of 21 competing teams exceeded the baseline, with the top-performing team achieving a combined accuracy of 82.13% compared to the baseline score of 77.51%. In Task 2, the performance of 9 out of 13 competing teams exceeded the baseline, with the top-performing team achieving a combined correlation of 0.1589 compared to the baseline score of 0.1023.

6. REFERENCES

[1] Michael J. Crosse, Adam Di Liberto, and Edmund 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.

[2] Jaswanth Reddy and S. Ganapathy, “Deep correlation analysis for audio-eeg decoding,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2742–2753, 2021.

[3] Bernd Accou, Jonas Vanthornhout, Hugo Van Hamme, and Tom Francart, “Decoding of the speech envelope from EEG using the VLAAI deep neural network,” Sept. 2022, Pages: 2022.09.28.509945 Section: New Results.

[4] Alain de Cheveigné, Daniel D.E. Wong, Giovanni M. Di Liberto, Jens Hjortkjær, Malcolm Slaney, and Edmund Lalor, “Decoding the auditory brain with canonical component analysis,” *NeuroImage*, vol. 172, pp. 206–216, 2018.

[5] Mohammad Jalilpour Monesi, Bernd Accou, et al., “Extracting different levels of speech information from eeg using an lstm-based model,” *Proceedings Interspeech 2021*, pp. 526–530, 2021.

[6] Bernd Accou and Mohammad Jalilpour-Monesi et al., “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.

[7] Lies Bollens, Bernd Accou, Hugo Van hamme, and Tom Francart, “A Large Auditory EEG decoding dataset,” 2023, doi: 10.48804/K3VSND.

[8] Jonas Vanthornhout, Jan Wouters, Jonathan Z. Simon, and Tom Francart, “Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope,” *Journal of the Association for Research in Otolaryngology*, vol. 19, pp. 181–191, Apr. 2018.