ICASSP 2023 AUDITORY EEG DECODING CHALLENGE

Lies Bollens[1,2,†], *Mohammad Jalilpour Monesi*[1,2,†], *Bernd Accou*[1,2], *Jonas Vanthornhout*[1], Hugo Van Hamme[2], Tom Francart[1]

[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

1. INTRODUCTION

To investigate how the brain processes sounds, various neuroimaging techniques can be used. Electroencephalograpy (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. 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 the electroencephalogram (EEG) of a person to a feature of the natural speech signal they were listening too. This is typically done using linear regression to predict the EEG signal from the stimulus or to decode the stimulus from the EEG [1]. While these approaches have been widely used in auditory neuroscience and have successfully been linked to speech intelligibility, [2], 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 [3, 4, 5, 6].

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 [7]. Recently, methods based on deep learning models have obtained promising results on this task, outperforming the linear methods [8, 9, 10, 11]. 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 [12], denoising and normalizing EEG [13, 14] and predicting EEG from acoustic features [15]. However, a drawback to neural networks is that they typically require a large amount of data to be trained. Also, there exists no large public auditory EEG dataset together with well defined tasks to relate EEG to speech, which makes 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 a total of 157 hours of data. Teams will compete to build the best model to relate speech to EEG in the following two tasks:

Task 1 : match-mismatch; given two segments of speech and a segment of EEG, which of the speech segments matches the EEG segment?

Task 2 : regression; reconstruct the speech envelope from the EEG.

2. 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 synchronises 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.

2.1. Data collection

We measure EEG data in a well-controlled lab environment (soundproof and electromagnetically shielded booth), using a high-quality 64-channel Biosemi 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). The data is measured at a sampling rate of 8192 Hz. While the temporal resolution is high, the spatial resolution is low, with only 64 electrodes for billions of neurons. All 64 electrodes are placed according to international 10-20 standards. The dataset contains data from 85 young, normal-hearing subjects (all hearing thresholds $<=$ 25 dB Hl), with Dutch as their native language. This study was approved by the Medical Ethics Committee with reference S57102. Before commencing the EEG experiments, all particpants read and signal an informed consent form, aproved by the Medical Ethics committe UZ KU Leuven/Research. All participants in the dataset gave explicit consent for their anonymized data to be shared in a publicly accessible dataset. All identifiable subject information has been removed from this dataset.

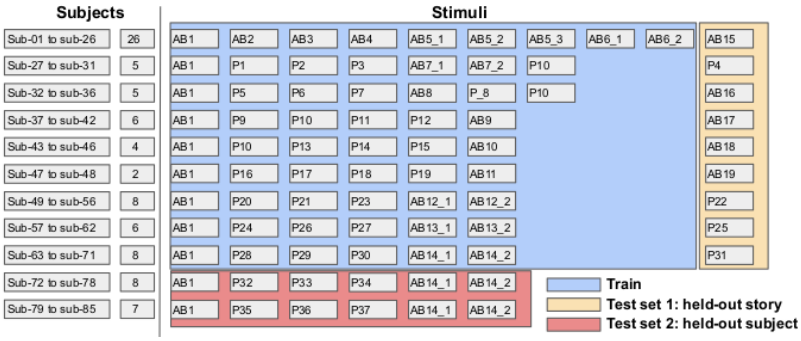

Fig. 1. Overview of the different stimuli and their division into training and testing set. Left side: subjects are divided into different groups. Each group of subjects listens to the same stimuli, which are depicted on the right side. Each horizontal line defines one group and the corresponding stimuli. Short names for the stimuli are given, with AB=audiobook and P=podcast.

Each subject listened to between 8 and 10 trials, each of approximately 15 minutes in length. The order of the trials is randomized between participants. 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 podcast or audiobooks. Some audiobooks are longer than 15 minutes. In this case, they are split into two trials presented consecutively to the subject.

2.2. Training set

The training set contains EEG responses from 71 subjects. These subjects are numbered from sub-01 to sub-071. As shown in figure 1, 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 listened to. Each such group contains between 2 and 27 subjects. All subjects of all groups listen to a reference story, *Audiobook 1*.

In total, the training set contains 508 trials, from 71 subjects, using 57 different stimuli. The total amount of minutes of data amounts to 7216 minutes (120 hours). Both tasks share the training set. 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 reference to the subject identifier. The auditory stimuli are provided in a separate folder *stimuli*.

2.3. 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 task description 3.1 and 3.2. Both test sets will be released to the participants on January 6, 2023. However, the ground truth labels will only be available to the public after the competition is over.

Test Set 1 **(held-out stories)** contains data for the 71 subjects seen in training. We held out one story for each group of subjects, which never occurs in the training set, amounting to a total of 944 minutes.

Test Set 2 **(held-out subjects)** contains data for 14 subjects (sub-72 to sub-85) 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.

2.4. Preprocessing

We provide two versions of the dataset. The first data version is the raw EEG data, which has been downsampled from 8192 Hz to 1024 Hz. The second version of the dataset has been preprocessed in MATLAB. First, the EEG signal was downsampled from 8192 Hz to 1024 Hz, and artefacts were removed using a multichannel Wiener filter. Then, the EEG signal was re-referenced to a common average. Finally, the EEG signal was downsampled to 64 Hz. These steps are commonly used in EEG signal processing, and the preprocessed version can be used directly in machine learning models. However, challenge participants are free to perform their own preprocessing on both versions of the datasets.

For task 2 3.2, we define a specific version of the envelope, which will be used for evaluating the final outputs. We estimate the envelope using a gammatone filter bank with 28 subbands, spaced by equivalent bandwidth with center frequencies of 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.

3. SIGNAL PROCESSING GRAND CHALLENGE DESCRIPTION

The Audio-EEG challenge features two separate tasks. Participants can submit to either one track or both.

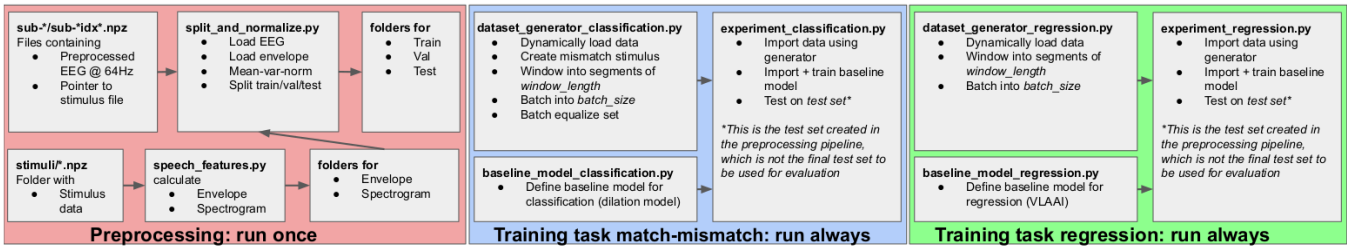

Fig. 2. Overview of the provided code and the recommended flow to get started. Left (red): Preprocessing code to calculate speech representations, normalize and split into different sets. Middle (blue): dataset generator and experiment file to be used for task 1: match-mismatch. Right (green): dataset generator and experiment file to be run for task 2: regression.

3.1. Task 1: Match-mismatch

3.1.1. 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 envelope segments correspond. . . s to the EEG. The performance metric is the classification accuracy (%). Figure 3 illustrates the problem. The input length of all (EEG, envelope) 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.

Figure 2 contains an overview of the provided code and our recommended starting flow. We provide code to generate an envelope and mel spectrogram representation from the raw audio files. However, participants can freely perform audio processing and use their preferred representations.

In order to use the data in the classification paradigm, we implement and suggest the following methods. 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). 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 (1s in our code) is dividable by the spacing between the matched and mismatched segment (1s). Failing to follow these suggestions is likely to result in models simply remembering the training samples and thus failing to generalize to the test set.

3.1.2. Baseline method

We include a dilated convolutional network [11] as a baseline for task 1. The dilated convolutional network consists of four steps. First, the EEG channels are 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 N dilated convolutional layers with a kernel size of K and 16 filters. These N convolutional layers are applied to both EEG and envelope stimulus segments. After each convolutional layer, a rectified linear unit (ReLU) is applied. Both stimulus envelope segments share the weights for the convolutional layers. After these non-linear transformations, the EEG is compared to both stimulus envelopes, 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 of approximately 77% correct is obtained.

3.1.3. Evaluation criteria

The test set for the match-mismatch task contains half the data from test set 1 and half from test set 2. For test set 1, all test stimuli are held-out stimuli. For test set 2, some stimuli have already appeared in the training set, but the subjects are unseen.

For these test sets, we will 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 should 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 should be of the form (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 will then be 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 add them to obtain the final score, which will serve as the ranking value $Score = 2/3S_1 + 1/3S_2$. This value will be updated in the online leaderboard.

3.2. Task 2: Regression

3.2.1. Description

Task 2 is a regression problem: To reconstruct the stimulus from the EEG. After reconstruction, a metric to measure the similarity is used between the reconstructed stimulus and the original stimulus. In this task, we use the Pearson correlation. Figure 3 shows a high-level setup of this task. For this task, the stimulus representation is defined as the envelope, as described in section 2.4. Figure 2 gives a schematic overview

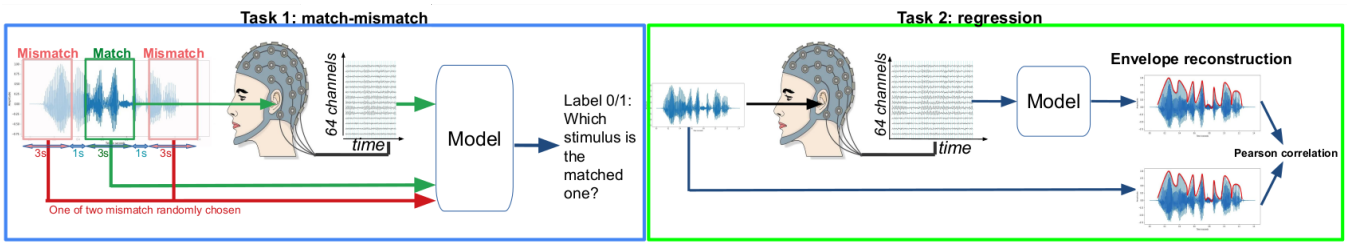

Fig. 3. Schematic overview of the two tasks. Left: Task 1 (match-mismatch). The model gets three inputs: an EEG segment, the matched (in time) speech segment, and a mismatched segment. The task is to determine which of the two segments is matched. Right: Task 2 (regression). The task is to decode the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

of the provided code and our recommended starting flow. Participants are free to create their methods. However, remember that the stimulus objective is fixed, as defined by the python file `envelope.py`.

3.2.2. Baseline method

We include the Very Large Augmented Auditory Inference (VLAAI) network [4] as a baseline for task 1. The VLAAI network consists of multiple (N) blocks, consisting of 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 special layers enhances the predictions made by the model up until that point, by taking the previous context into account and combining it with the current sample. At the end of each block except the last, a skip connection is present with the original EEG input. 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 0.19 is obtained.

3.2.3. Evaluation criteria

The test set for the regression task contains half the data from test set 1 and half from test set 2. All stimuli are held-out stimuli, i.e., they do not appear in the training set. We have split up 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 will then be compared to the original envelope, as defined in section 2.4, using Pearson correlation. We will use the `scipy.stats.pearsonr` function to calculate the correlation c_i for each segment i. Afterwards, 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 add them to obtain the final score, which will serve as the ranking value: $Score = 2/3S_1 + 1/3S_2$

Participants should submit a NumPy dictionary file for the test set to an online form on our website, which contains the reconstructed envelopes for all EEG segments. Afterward, we will calculate the score mentioned above and update this in the online leaderboard. Each entry in the submitted dictionary should be of the form (EEG ID) : (Reconstructed Envelope). A correlation value of 0 will be taken in case of absent EEG ID entries. The reconstructed envelope should be of dimensions 1 x 3840 (i.e., 60 seconds of data at the prescribed sample rate of 64 Hz).

4. GUIDELINES FOR PARTICIPANTS

- To register, participants should send a mail to auditoryeegchallenge@kuleuven.be with the name of their team members, emails, affiliations and team name.
- Participants can submit their predictions up to five times. The latest received submission counts as the official score.
- The Audio-EEG challenge features two separate tasks. Participants can submit to either one track or both. Results should be accompanied by a 2-page paper describing the proposed method
- The top 5 ranked teams will be invited to submit a 2-page paper, to be presented at ICASSP-2023, which should be submitted before the camera-ready deadline. In addition, they will receive an invitation to submit a full paper about their work to the IEEE Open Journal of Signal Processing (OJ-SP). Submitted OJ-SP papers will undergo peer review by the OJ-SP Editorial Board in collaboration with the IEEE SPS CDC committee and ICASSP-2023 GC Chairs.
- Winners will be selected according to the best performance for every single task. One winner for each task will be selected
- The top 5 teams will be determined as follows: the top 2 teams for track 1 and the top two teams for track 2. The fifth team will be chosen as the third-ranking team in the task with the most submissions.
- Each participant can only be included in one participating team

- The use of any external data that the organizers do not provide is strictly prohibited.
- Participants can use the full training set to train the models for both tasks. We encourage the participants to explore different preprocessing and data augmentation methods to better train the models. All stimulus information can be used for training, provided that the information can be deduced based on the training set.
- All parameters should be tuned on the training set. The test sets can be used only in an evaluation paradigm to provide output labels/predictions.

5. LIST OF POTENTIAL PARTICIPANTS

Researchers from the following labs have expressed interest to compete in the challenge:

- Jonathan Z Simon, Computational sensorimotor systems lab, University of Maryland
- Nima Mesgarani, Neural acoustic processing lab, Columbia university
- Mounya Elhilali, Laboratory for computational audio perception, Johns Hopkinks
- Emina Alickovic, Linkoping University
- Malcolm Slaney, Machine hearing group, Google research
- Tobias Reichenbach, Sensory neuroengineering, FAU university

More broadly speaking, we expect to attract interest from researchers in related fields, both from engineering (deep learning, automatic speech recognition, EEG signal processing, ...) and from neuroscience (EEG, auditory neuroscience, ...). We will circulate the call for participation widely in our professional networks in both engineering (co-author prof. Van hamme) and neuroscience (co-author prof Francart).

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] 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, no. 2, pp. 181–191, Apr. 2018.

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

[4] 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.

[5] Gautam Krishna, Co Tran, Yan Han, Mason Carnahan, and Ahmed H Tewfik, "Speech synthesis using eeg," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1235–1238.

[6] Madhumitha Sakthi, Ahmed Tewfik, and Bharath 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.

[7] 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.

[8] Mohammad Jalilpour Monesi, Bernd Accou, Jair Montoya-Martinez, Tom Francart, and Hugo 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.

[9] 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.

[10] Corentin Puffay, Jana Van Canneyt, Jonas Vanthornhout, Hugo Van hamme, and Tom Francart, "Relating the fundamental frequency of speech with eeg using a dilated convolutional network," *Proceedings Interspeech 2022*, 2022-09-18.

[11] Bernd Accou, Mohammad Jalilpour-Monesi, Jair Montoya-Martínez, Hugo Van hamme, and Tom 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.

[12] Maryam Hosseini, Luca Celotti, and Éric 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.

[13] Jaswanth Reddy Katthi and Sriram 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.

[14] Lies Bollens, Tom Francart, and Hugo 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.

[15] Gautam Krishna, Co Tran, Mason Carnahan, Yan Han, and Ahmed H Tewfik, "Generating eeg features from acoustic features," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1100–1104.