scientific reports

Check for updates

OPEN

Contrastive representation learning with transformers for robust auditory EEG decoding

Lies Bollens[1,2], Bernd Accou[1,2]✉, Hugo Van hamme[2] & Tom Francart[1]✉

Decoding of continuous speech from electroencephalography (EEG) presents a promising avenue for understanding neural mechanisms of auditory processing and developing applications in hearing diagnostics. Recent advances in deep learning have improved decoding accuracy. However, challenges remain due to the low signal-to-noise ratio of the recorded brain signals. This study explores the application of contrastive learning, a self-supervised learning technique, to learn robust latent representations of EEG signals. We introduce a novel model architecture that leverages contrastive learning and transformer networks to capture relationships between auditory stimuli and EEG responses. Our model is evaluated on two tasks from the ICASSP 2023 Auditory EEG Decoding Challenge: a binary stimulus classification task (match-mismatch) and stimulus envelope decoding. We achieve state-of-the-art performance on both tasks, significantly outperforming previous winners with 87% accuracy in match-mismatch classification and a 0.176 Pearson correlation in envelope regression. Furthermore, we investigate the impact of model architecture, training set size, and finetuning on decoding performance, providing insights into the factors influencing model generalizability and accuracy. Our findings underscore the potential of contrastive learning for advancing the field of auditory EEG decoding and its potential applications in clinical settings.

Electroencephalography (EEG), a non-invasive method for measuring brain electrical activity, reveals insights into neural processes during speech perception. By analyzing EEG signals alongside features of presented speech, researchers investigate how the brain processes continuous speech[1–3]. This line of inquiry has implications not only for understanding the neural underpinnings of speech processing but also for developing objective measures for hearing diagnostics[4,5] and designing advanced auditory prostheses. EEG, with its high temporal resolution and ease of use, is a valuable tool for auditory neuroscience research and clinical applications. EEG-based measures can aid in the detection and diagnosis of hearing loss[6–9].

The endeavor to relate EEG signals to continuous speech features has primarily relied on linear regression for predicting EEG from stimuli or decoding stimuli from EEG[1]. While successful in linking EEG to speech intelligibility[10,11], linear methods face limitations due to the low signal-to-noise ratio inherent in EEG data and due to the inherent non-linearity of speech encoding in the brain[12,13]. This challenge has led to the exploration of artificial neural networks (ANNs), including deep learning models, as potential solutions for improving auditory decoding performance[14–20]. The application of ANNs extends further, encompassing speech denoising[21], EEG signal enhancement[22,23], and predicting EEG from acoustic features[24].

In parallel to regression, the match-mismatch task, a classification paradigm that determines which of two speech segments corresponds to a given EEG segment, has gained traction[25,26]. Recent deep learning models have shown promise in tackling this task, surpassing linear methods in accuracy[26–29].

However, the need for a large, standardized public auditory EEG dataset, coupled with variations in evaluation metrics, hinders comprehensive comparisons between different models (e.g., see the comprehensive review by Puffay et al., 2023[20]). Although recently, more public EEG datasets have become available[30–36], comparison of different models has remained difficult. In an attempt to create benchmarks for evaluation, two auditory EEG decoding challenges have been launched[37–40]. These challenges have prompted many researchers to tackle this problem and have brought various innovations. The challenge brought two tasks: a match-mismatch task and a regression task. For the match-mismatch task, the top teams used several different speech representations, such as the mel spectrogram[41], high-frequency envelope modulations (as explained in[42])[43], high-frequency gamma-band features[44] as well more complex representations such as wav2vec[45] features[46] and gpt2[47] features[46], indicating the advantage of using richer, more complex speech representations when trying to relate speech to EEG.

[1]Dept. Neurosciences, ExpORL, KU Leuven, Leuven, Belgium. [2]Dept. of Electrical engineering (ESAT), PSI, KU Leuven, Leuven, Belgium. ✉email: bernd.accou@kuleuven.be; tom.francart@kuleuven.be

As for architectures, many teams incorporated attention mechanisms in their solution[41,48,49]. In addition, the usage of ensemble learning has been shown to yield significant increases in the results. Thirdly, many teams propose bigger networks, using stacks of convolutional blocks and skip connections[46,49,50], or based on the wavenet[51] architecture[52,53]. These models and the ICASSP challenge show the potential of the usage of recent deep learning techniques and their application to auditory EEG decoding. However, the models mentioned above focus on one specific downstream task, for which the models are tailored, which might limit their applicability and hinder the capacity to learn general representations for speech decoding.

During the past few years, self-supervised learning techniques, particularly contrastive learning, have become very popular and have revolutionized representation learning in fields like speech and image processing. This approach has demonstrated impressive results in various domains, including natural language processing[54] and computer vision[55]. In the context of auditory EEG decoding, contrastive learning offers the potential to learn meaningful representations of EEG signals by contrasting examples of similar and dissimilar auditory stimuli, leading to improved decoding performance[46,56,57]. The core principle of contrastive learning involves training models to distinguish between positive pairs (similar examples) and negative pairs (dissimilar examples). In this way, it can be seen as an extension or generalization of the match-mismatch paradigm, with multiple negative samples instead of one. An advantage of contrastive losses is that it encourages the model to find latent representations that find the best discrimination between different features between all the segments of the batch, thus forcing the model to find and retain actual informative data for discrimination and discard the rest. The success of contrastive learning in bridging modalities, exemplified by CLIP (Contrastive Language-Image Pre-Training)[58] and its application to brain imaging[56], demonstrates its potential for relating complex, disparate data.

In this paper, we extend this paradigm to auditory EEG decoding, exploring the use of contrastive learning to create effective latent representations of EEG signals. We hypothesize that a model trained with contrastive learning will outperform models trained without contrastive learning on downstream tasks that utilize these latent EEG representations. To test this hypothesis, we introduce a novel model that leverages contrastive learning to capture the relationship between auditory stimuli and their corresponding EEG responses. We evaluate the quality of the learned EEG representations by assessing their performance on two well-established tasks[37]: match-mismatch classification and stimulus feature regression. We show that using contrastive learning, our model learns to generate useful, discriminative latent representations. When evaluating two downstream tasks, we obtain state-of-the-art performance without retraining the latent representations and using simple linear classification or regression methods.

Methods

Preprocessing

EEG

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

Audio

We extract different speech representations.

Mel Spectrogram

The initial stage of our analysis involved extracting speech features using Mel spectrograms. A Mel-scale filter bank encompassing 28 equally spaced frequency bands within the range of 50 Hz to 5000 Hz was applied to the short-time Fourier transform (STFT) representation of the audio signals. This process yielded Mel spectrograms that capture the spectral characteristics of the speech data in a perceptually relevant manner.

Wav2vec

In addition to mel spectrograms, we investigated the utilization of self-supervised neural network models for speech representation, specifically, the wav2vec 2.0 architecture[45]. This approach has demonstrated promising results in automatic speech recognition (ASR) tasks, owing to its ability to learn general-purpose representations from vast amounts of unlabeled audio data. We employed a wav2vec 2.0 model finetuned for the Dutch language[60]. This model consists of 24 transformer layers, the intermediate outputs of which have been shown to effectively capture linguistic features and exhibit linear mapping to brain activations[61,62]. We utilized the outputs of the 19th layer, resampled to 64 Hz, to align with EEG data as speech representations. These representations, each with a dimensionality of 1024, are hypothesized to contain rich phonetic and linguistic information pertinent to the Dutch language[63].

Speech envelope

For the regression task (see section Experiments), a speech envelope is computed with a gammatone filterbank as specified in the SparrKULee paper[30] and the ICASSP 2023 challenge[37].

Dataset

The experiments and the experimental protocol that have been carried out have been approved by the Medical Ethics Committee UZ KU Leuven/Research (KU Leuven Belgium) with reference S57102. All experiments were performed in accordance with their relevant guidelines and regulations. All participants provided informed consent, with a subset explicitly consenting to the sharing of their pseudo-anonymized data[64], as per the

guidelines of the Medical Ethics Committee UZ KU Leuven/Research (KU Leuven Belgium) with reference S57102.

The study utilized a dataset comprising EEG recordings from 194 normal-hearing ($\leq$ 30 dB SPL, 125-8000 Hz) native Dutch/Flemish-speaking participants aged 18-45 years. A subset of this dataset (105 participants) is publicly available and has been previously employed in the ICASSP 2023/2024 Auditory EEG decoding challenge[30,37].

Participants completed a questionnaire detailing demographic information (age, sex, education level, handedness[65]) and any diagnoses of hearing loss or neurological pathologies. Hearing acuity was assessed via pure-tone audiometry and the Flemish MATRIX test.

EEG data were acquired in a sound-attenuated and electromagnetically shielded booth using a 64-channel Biosemi ActiveTwo system with Ag-AgCl electrodes. Signals were sampled at 8192 Hz using ER3A insert earphones (62 dBA intensity per ear), an RME Multiface II sound card, and APEX software[66]. During the experiment, participants listened to 8-10 randomized single-speaker Flemish stories (podcasts/audiobooks), segmented into 15-minute parts. To ensure engagement, participants were informed of post-story comprehension questions.

The **training set** on which we train our models consists of a combination of the official ICASSP 2023 auditory EEG challenge training set (comprising 71 subjects) and the remaining 109 subjects who are not part of the challenge. This amounts to a total of around 350 hours of training data. We produce our tests on the official ICASSP 2023 auditory EEG challenge 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. **Held-out stories** contains data from 70 subjects seen in the training set. We held out one story for each group of subjects, which never occurred in the training set, amounting to around 16 hours in total. These 70 subjects are part of the data that is publicly available. **Held-out subjects** contains data for 14 subjects that are not in the training set, further referred to as held-out subjects, for a total of around 21 hours. The data for these subjects were acquired using the same protocol as for the other 70 subjects.

Model

Code for the model and pre-trained models can be found at https://github.com/exporl/transformer-clip-eeg

Figure 1 provides a high-level overview of the network. It consists of multiple subparts, which are explained below.

EEG encoder

The EEG encoder consists of three components that are repeated *N* times. Each repetition contains approximately 340,000 parameters. Figure 1(c) provides a visual overview of the EEG encoder. The general approach for this model is inspired by the approach taken by[67]. More specifically, the self-attention architecture and implementation are the same as their implementation. The convolutional layer is a different implementation, as explained in more detail below. Another key difference between our proposed model and that proposed by[67] is that in our model, all three components are repeated. In contrast, in the former architecture, only the self-attention module is repeated.

(1) Convolution

The convolutional module proposed by[67] is inspired by the convolutional architecture that was originally proposed by[68]. In this architecture, the convolution is split into two one-dimensional convolutions: one over the temporal and one over the spatial dimension. Then, they use batch normalization and Exponential linear units (ELUs) as the activation function. After this, an average pooling among the time dimension follows. Accou et al.[15] recently proposed a model for envelope regression and obtained high correlation values, prompting us to base our convolutional model upon this model. They show that bigger convolutional models can effectively be used for envelope regression. We opt to only use one convolutional layer, as opposed to 5 in the original model, as there are multiple repetitions of the whole model, and we want to keep the computational complexity feasible.

In our convolutional module, first, there is a convolutional layer using 64 filters, each with a temporal kernel size of 64. This is followed by layer normalization[69] and the GELU activation[70]. In parallel, we add a skip connection. While skip connections yielded no significant benefit in reconstruction score, by the model proposed by Accou et al[15], it has been shown that they promote stability in model training[71].

(2) Self-Attention

Second, we implement an attention network following the approach by[67]. The output tokens from the convolutional layer are transformed into queries (Q), keys (K), and values (V). The dot product of Q and K models the correlation between different time tokens, scaled by $\sqrt{k}$ (where k is the token length) to avoid vanishing gradients. This is passed through a Softmax function to obtain a weighted matrix serving as the attention score, which is then weighted using the input matrix V. This can be expressed as: $Att(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{k}})V$

Instead of a single attention module, we employ a multi-head attention strategy, allowing the model to have diverse representations. We divide the tokens into $h = 8$ segments and feed each segment into different self-attention modules. The results are concatenated and passed to the next layer. $MultiHeadAtt(Q, K, V) = [Att_0 \quad Att_1 \quad \cdots \quad Att_{h-1}]$ and $Att_l = Att(Q_l, K_l, V_l)$.

With Q_l, K_l, V_l, the tokens partitioned into h different parts. We add another skip connection between the input and output of this attention mechanism.

(3) Feedforward

The attention mechanism is followed by two fully connected layers, which expand the output to four times its original size and then reduce it back to its original size, aiding the model's fitting capability. Again, a skip connection is added between the input and output of the feedforward part.

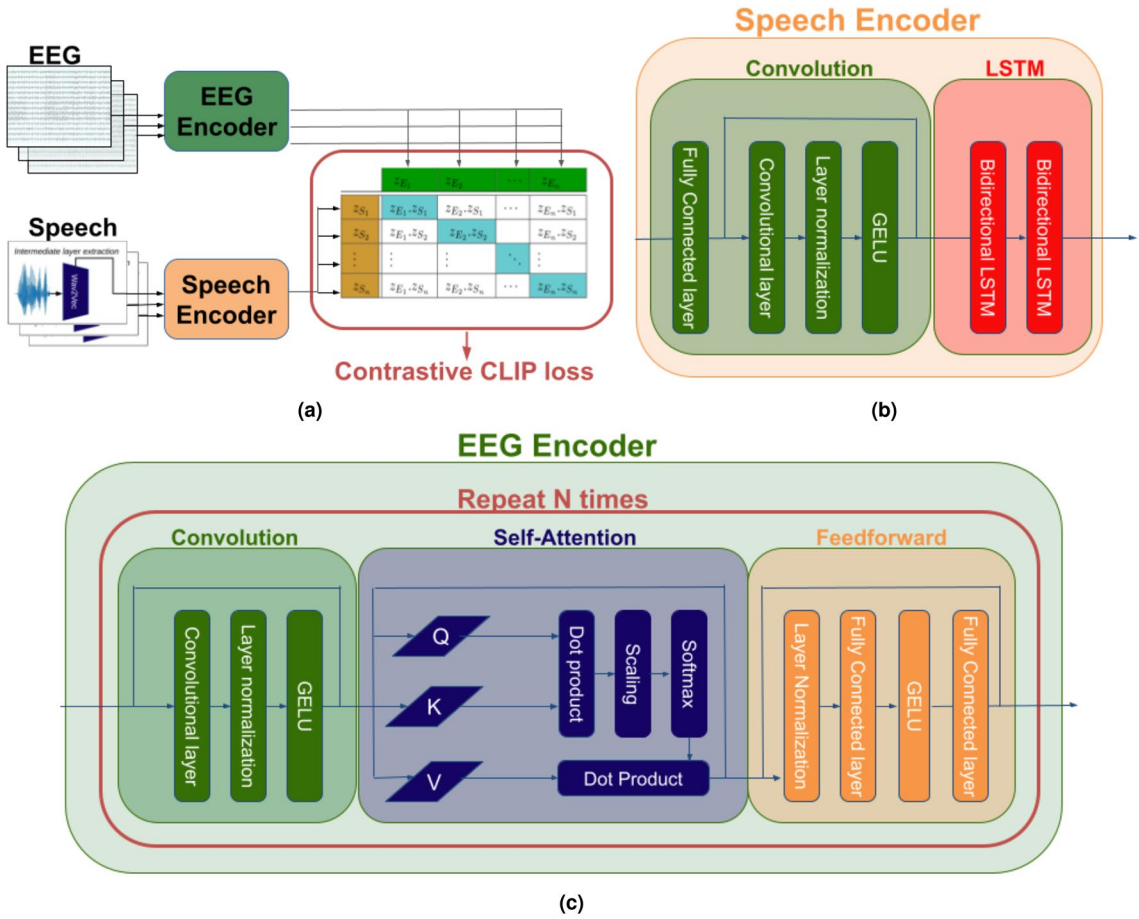

Fig. 1. Overview of the network. **(a)**: High-level overview of the different encoders. **(b)**: The Speech encoder in more detail. Two bidirectional LSTM layers follow a convolutional block. **(c)**: The EEG encoder in more detail. A Convolutional block is followed by self-attention and a Feedforward network. This sequence is repeated N times

Speech encoder

The speech module begins with a fully connected layer, which recombines the input channels into 64 channels. This is followed by a convolutional layer, similar to the convolutional layer in the EEG encoder, with 64 filters and a temporal kernel size of 64. Layer normalization and a GELU activation function follow. We add a skip connection over the convolutional layer. To facilitate alignment of the speech with the EEG, the convolutional layer is followed by two bidirectional LSTM layers of dimensions 64 and 4, respectively. The speech module's output is 8-dimensional.

Contrastive loss

Match-mismatch loss, which is often used in auditory EEG decoding tasks, can be seen as a simplified contrastive loss, where there is always just one negative example. Given the promising results of extended contrastive losses, we generalize this loss by adding more negative examples, potentially enhancing model performance. A popular method for relating different modalities is the CLIP loss (Contrastive Language-Image Pre-Training)[58], designed to match latent representations in text and images.

By generalizing this loss, the model is encouraged to identify discriminative features in the EEG/Stimulus features that distinguish them from others. In addition, similar segments will be encoded similarly, defining an ordering in the latent space, which can be useful for downstream tasks. The CLIP loss consists of two contrastive losses. First, the latent EEG representation is compared against K latent stimulus representations, with K the batch size, with a similarity metric, to find the most fitting latent speech segment. Secondly, the latent speech representation is compared against K latent EEG representations, using the same similarity metric to find the most fitting latent EEG segments.

More formally, we implement the CLIP loss in the following way. Let E_i be a preprocessed downsampled segment of T samples of EEG (seconds $\times$ sampling rate), with F_E channels, and S_i the downsampled stimulus representation of the stimulus of T samples, with dimension F_S. First, we calculate the latent representations of these segments, denoted as respectively $z_{E_i} \in \mathbb{R}^{F_E \times T}$ and $z_{S_i} \in \mathbb{R}^{F_S \times T}$. For each batch, we sample $K-1$ negative samples z_{E_j} $(j \in 1, ..., K-1)$ and z_{S_j} $(j \in 1, ..., K-1)$. Then, we calculate the probabilities for each EEG segment corresponding with any of the stimulus segments and vice versa. We estimate the probabilities

by taking the dot product of the EEG latent representation and the candidate stimulus latent representations, followed by a softmax. In order to regulate the probabilities better, a scaling factor e^t is added, which is a learnable parameter. The learned probabilities are

$$p_{z_{S_i}} = \frac{e^{<z_{S_i}, z_{E_i}> e^t}}{\sum_{j=1}^{K} e^{<z_{S_i}, z_{E_j}> e^t}} \text{ and } p_{z_{E_i}} = \frac{e^{<z_{S_i}, z_{E_i}> e^t}}{\sum_{j=1}^{K} e^{<z_{S_j}, z_{E_i}> e^t}}$$

with $< z_S, z_E >$ the inner product over the time dimension of z_S and z_E. Then, the model is trained with a cross-entropy loss between the probabilities. The final loss is then calculated as a sum of both probabilities.

$$L_{CLIP} = -\frac{1}{2} \sum_{i=1}^{K} \left(\log(p_{z_{S_i}}) + \log(p_{z_{E_i}}) \right)$$

Ensemble

In addition to the general model, we employed an ensemble model approach, which averages the logits (the unnormalized output scores) from multiple models to enhance prediction accuracy. This approach was first introduced in the context of auditory EEG decoding by the winning team of the auditory EEG decoding task[73] and has been used by four out of the five top teams in the auditory EEG decoding 2024 challenge[40].

Training details

All experiments were carried out using PyTorch within a Python 3.10 environment. Model training was facilitated by the Adam optimizer, employing weight decay ($\lambda = 0.01$), a learning rate of $1e-3$, and β_1 and β_2 values of 0.9 and 0.999, respectively. A step learning rate scheduler with a fixed decay rate ($\gamma = 0.1$) every 10 epochs was implemented. Empirically, we found that using the Adam optimizer with weight decay, together with using a step learning rate scheduler with a fixed decay rate, led to the most stable training process (see Supplementary Information, Section A for more details).

The attention block count (N) was set to 10, with each block containing $h = 8$ attention heads. Training data consisted of 3-second non-overlapping windows extracted from EEG recordings and corresponding stimuli. Batches of 128 were constructed by segmenting each stimulus-EEG pair into 3-second intervals, ensuring negative examples within a batch were temporally proximal to their corresponding positive example. Thus, the stimulus is segmented into separate segments of $[0-3s], [3-6s], [6-9s], \ldots$ The positive pair is always equal to the time-matched [EEG, stimulus] pair, the negative pairs are chosen as the other segments within the same stimulus/EEG recording.

A shuffling mechanism was employed to mitigate overfitting and promote the generalizability of latent representations. After each epoch, batches were reconstructed by randomly permuting the EEG segments associated with a given audio stimulus, leveraging the availability of multiple participants exposed to the same stimulus.

Baseline methods

We compare our model's performance to that of established baseline methods and the top-performing submissions from the 2023 Auditory EEG Decoding Challenge.

Dilated convolutional network

The dilated convolutional model[72] employs separate pathways for processing EEG and stimulus data. In the EEG pathway, the 64 input channels are first reduced to 8 using a 1D convolutional layer. Subsequently, three layers of dilated convolutions with ReLU activations are applied to capture temporal patterns at different scales. The stimulus pathway mirrors this structure, with shared weights for the convolutional layers across both stimulus segments. Cosine similarity is then used to compare the latent EEG representation with the two latent stimulus representations. The resulting similarity scores are passed through a single sigmoid neuron to predict the matching stimulus segment.

Thornton et al. (2023)

The winning approach for the match-mismatch task by Thornton et al.[73] leverages an ensemble of 50 dilated convolutional models. The final prediction is obtained by averaging the sigmoid outputs of these models, akin to a majority voting mechanism. Additionally, they demonstrate that incorporating high-frequency envelope modulations as an additional speech representation alongside the speech envelope enhances model performance. For the within-subject test set, the authors further finetune the subject-independent model for each individual, achieving improved accuracy. A composite model combining the averaged baseline model with a separate model trained on f0 features using linear discriminant analysis is employed for the held-out subjects.

Linear decoder

A simple linear decoder is used as a baseline for the regression task. It is implemented as a one-dimensional convolutional layer in TensorFlow with a kernel size of 32, corresponding to a 500 ms integration window. The linear decoder is trained on all subjects without individual adaptation.

Piao et al. (2023)
The top-performing model for the regression task by Piao et al.[48] utilizes a feedforward transformer architecture with pre-layer normalization. This model also incorporates an auxiliary global conditioner to integrate subject-specific information, leading to a 13% improvement in correlation values when tested on the within-subject set. However, this conditioner cannot be applied to the held-out subjects' test set.

Experiments

For evaluation, we use the two official tasks of the 2023 ICASSP Auditory EEG decoding challenge[37] and test on the official test set[37]. We use the output embeddings of the same model for both tasks to demonstrate the generalizability of the generated embeddings.

All the models are initialized with random weights, which can introduce slight variations in the final models. To mitigate this, we train multiple (16) models and report the average accuracy across them, where the average is calculated per subject.

When reporting the final mean score, it is important to note that the mean operator is weighted, giving more significance to test set 1 (held-out stories from seen subjects). The total score weights 2/3 of the score to test set 1, comprising 70 test subjects, and 1/3 of the score to test set 2, comprising 14 unseen test subjects. This score complies with the ICASSP challenge rules and is only useful for comparisons within that specific context. However, for comparison with other models, we opted to report these metrics. When assessing the generalizability of the model, metrics computed on test set 2, which involves held-out subjects, should be emphasized as the most realistic evaluation scenario in new scenarios.

Match-mismatch
Task 1 is a classification task based on the concept of matched and mismatched (EEG, speech) pairs[74]. A matched (EEG, speech) pair means that the EEG segment was recorded while the speech segment was presented to the subject. A mismatched (EEG, speech) pair indicates that the EEG segment is not a response to the speech segment. There are many possibilities to define a match-mismatch task based on the concept of matched/mismatched segments[26,27,75].

For this challenge, the setup described by Accou et al.[26] is used. A schematic of the chosen match-mismatch task is illustrated in the left plot of Fig. 2. The model is provided with three inputs of length 3 seconds: (1) a segment of EEG, (2) the time-aligned speech stimulus (match), and (3) an unaligned stimulus (mismatch). The task of the model in this paradigm is to determine which of the two input stimulus segments correspond to the EEG segment. More specifically, we define the mismatched stimulus as temporally close to the matched one by randomly taking the segment starting either one second after the end or 4 seconds before the start of the matched segment. A mismatched segment close to the matched segment makes the training process more challenging for the model (see[20], section "Ablation Study").

The test set for this task consists of the first half or the second half of each EEG recording from the held-out stories set (test set 1, 70 subjects) and half of the recordings from the held-out subjects set (test set 2, 15 subjects). The other half is used for the regression task. The test set contains pairs of (EEG, stimulus 1, and stimulus 2) with a length of 3 seconds, each with a unique identifier and a subject identifier.

For each pair of (EEG, stimulus 1 and stimulus 2), we calculate the embeddings E_i, S_1 and S_2. Then, we calculate the similarity by taking a scaled dot product between (E_i, S_1) and (E_i, S_2). The maximum similarity is then taken as the matched segment. When using an ensemble model, we average the scaled dot products across models before taking the maximum similarity as the matched segment. Then, the mean classification accuracy per subject is calculated, and the mean accuracy is calculated for both test set 1 and test set 2.

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

The test set for the regression task contains either the first or the second half of each EEG recording from the held-out stories set (test set 1, 70 subjects) and half of the recordings from the held-out subject set (test set 2, 15 subjects). All stimuli are held-out stimuli, i.e., they do not appear in the training set. The stimuli are split into

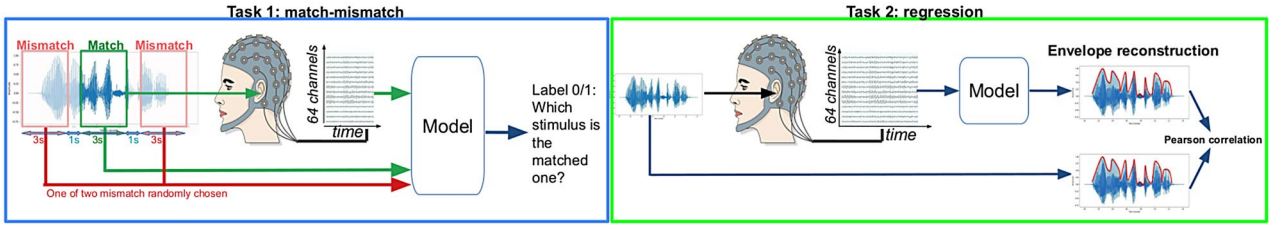

Fig. 2. Schematic Overview of the two tasks which are used for evaluation, as originally defined in[37]. Left: Task 1 (match-mismatch). The model gets three inputs: an EEG segment, the matched (in time) speech segment, and a mismatched segment. The task is to determine which of the two segments is matched. Right: Task 2 (regression). The task is to decode the speech envelope from the EEG brain response. Reconstructed and original envelopes will be compared by computing the Pearson correlation.

segments of 60 seconds. For each segment of 60 seconds, a reconstructed envelope is expected, which is then compared to the original envelope using Pearson correlation. Then, the mean correlation value per subject and the mean correlation value over all subjects are calculated for test sets 1 (seen subjects) and 2 (held-out subjects).

In order to calculate the envelope, we split the EEG segment of 60 seconds up into smaller segments of 3 seconds, with an overlap of 50 %. We then calculate the latent EEG embeddings for all segments of the training set for all subjects. Starting from these embeddings, we train a simple linear model, which predicts the envelope. The linear model is implemented using a one-dimensional convolutional layer with a kernel size of 32 (corresponds to a 500 ms integration window).

This linear model can also be used on the raw EEG data, which is the proposed baseline in the regression task. We recombine the reconstructed envelope segments using an overlap-add algorithm with a Hann window. The ensemble model's output is calculated by predicting the envelope separately for each of the 16 models and then averaging the envelope outputs, yielding one averaged output, for which the correlation is then computed.

Attention visualization

In an attempt to add explainability to the model, we visualize some of the attention weights from the different layers . The explainability of neural networks is a major concern. As shown by Niu et al.[76], several methods have been proposed to visualize the attention weights, and several implementations have been proposed[77,78]. We use attention visualization maps, a graphical representation to depict the attention weights within transformer models. These maps can be used as an attempt to illustrate how the different attention layers assign varying levels of importance to different parts of the input data, thereby highlighting the relationships and dependencies that the model has learned.

For our visualization, we take the story audiobook_2, as this is a story to which 70 of the subjects have listened and which is present in the test set. For each of the 3-second EEG segments of all subjects and this audiobook, we calculate the attention weights as $Softmax(\frac{QK^T}{\sqrt{k}})$. This yields an output of 8 heads per layer, each of dimension

192×192 (3 seconds $\times$ sampling rate of 64 *Hz*), for each of the 10 layers of our model.

Since we are interested in which parts of the EEG segments are attended to in order to create a meaningful latent representation to perform auditory downstream tasks afterward, we calculate the attention weights per segment separately and then averaged over all the subjects that listened to the same stimulus segment.

Finetuning

When training a linear model, the choice is typically made to train subject-independent or subject-specific models. Subject-specific models are often chosen for linear models. In contrast, for larger neural networks, this approach might lead to a lack of data, which leads to overfitted networks that do not generalize well. A solution for this can be to take a previously trained model and then continue training (finetune) the model, using only data from a certain subject. The advantage, in this case, is that the networks have often already found general patterns and only have to adapt their weights slightly for optimal performance for one subject. For the 70 subjects of test set 1, we finetune the model separately for each of these subjects and then evaluate the model in the same way. In order to prevent overfitting, we lower the learning rate to $1e-5$ for 50 epochs, using early stopping with patience 10 as an extra metric. The inclusion of seen subjects in test set 1 represents a limitation of the finetuning process: the most realistic case would be finetuning the model with a completely new subject and testing on that same subject. However, in order to keep the results comparable to those of the general model and with the benchmarks as set by the ICASSP auditory EEG challenge, we opted to do the finetuning on the seen subjects of test set 1.

How many subjects are needed?

In order to achieve good performance on held-out subjects, the subject-independent model needs to learn to extract general features from the EEG. These patterns can be directly related to the neural tracking of the stimulus and the discarding of irrelevant subject-specific information. In order to learn these general characteristics, the number of subjects in the training set must be high enough to aid generalization and prevent overfitting on the seen subjects. In this section, we calculate the effect of the number of subjects in the training data on performance. We train the model on 1-194 subjects, at each time randomly adding another subject. We report the performance of the 14 held-out subjects, test set 2 of the challenge.

Ablation study

To inspect the individual contributions of various model components to overall performance, we conducted an ablation study. This involved systematically modifying specific aspects of the model architecture and evaluating the resulting impact on accuracy. The following model variants were compared:

1. **No CLIP:** The contrastive learning mechanism was omitted from the training process. In this scenario, the speech encoder and EEG encoder are kept. However, instead of a contrastive loss, there is a binary classification problem, which can be summarized as *does this segment of EEG correspond to this segment of speech?*
2. **No Transformer:** The transformer component was entirely removed from the EEG processing encoder. We keep only the convolutional block, which is repeated once. This was done to determine whether it is beneficial and necessary to have a large EEG encoder.
3. **Transformer 1 Layer:** There is only 1 repetition of the transformer instead of 10. In this case, the EEG encoder is slightly bigger than in the 'No Transformer' case mentioned above, yet still much smaller than the final EEG encoder.

4. **Small Speech encoder:** The speech encoder uses a simpler encoder, using a convolutional block with a temporal kernel size of 16 and 8 output filters. This is done to determine if it is necessary to have a large speech encoder.
5. **Mel:** Mel spectrograms replaced the wav2vec output within the speech encoder.
6. **Wav2vec 19:** We use the final architecture with the wav2vec layer 19 as speech representation, yet there is no ensemble modeling used, and only one model is trained.
7. **Ensemble Wav2vec 19:** We use the final architecture with the wav2vec layer 19 as speech representation, averaging the output of 16 trained instances with randomly assigned weights before making a final classification.

Results

Match-mismatch

The results of our proposed model in the match-mismatch paradigm can be seen in Fig. 3. For test sets 1 and 2, we reach accuracies of respectively 85.72% and 84.37%. Using an ensemble model, which averages the output of 15 trained models, we reach accuracies of respectively 87.32% and 86.40%, resulting in a final score for task 1 of 87.01%. The winners of the ICASSP 2023 auditory EEG challenge reached scores of respectively 82.71% and 80.98%, for a final score of 82.13%. For both test set 1 and test set 2, we significantly outperform the winners of the challenge. Significance values were computed using one-sided Wilcoxon signed-rank tests with Bonferroni-Holm corrections, and a p-value of 0.001.

Regression

The results of our model for the regression task can be seen in Fig. 4. Our final model scores a regression score of 0.1992 for the seen subjects and a value of 0.1301 for the unseen subjects, leading to a mean total score of 0.1762. When using the linear baseline, i.e., the same complexity of the model, on the raw EEG, we obtain a mean correlation value of 0.105 for the seen subjects and 0.096 for the unseen subjects, leading to a final mean score of 0.102. The winners of the ICASSP auditory EEG challenge reach final correlation scores of 0.1895 for the seen subjects and 0.0976 for the unseen subjects, yielding a final score of 0.1589. When comparing these results, our model significantly outperforms the top model on both tasks. Significance values were computed using one-sided Wilcoxon signed-rank tests with Bonferroni-Holm corrections and a p-value of 0.001.

Attention visualization

Figure 5 shows some sample attention maps for several heads for layers 5 and head 5 for several layers, averaging over 70 subjects that listened to the same stimulus segments. In order to have an idea of what the different layers focus on most, we also plot the average attention values, averaged over different subjects, different EEG segments and different heads. Figure 6 shows the averaged attention maps. There is one single attention map per layer.

As can be seen upon visual inspection, for each token, on average, the tokens that are closely time-aligned to it are being paid the most attention, and after about 40 tokens, there is no general effect visible. When trying to reconstruct the stimulus from the EEG responses, traditional linear decoders typically look at the brain region

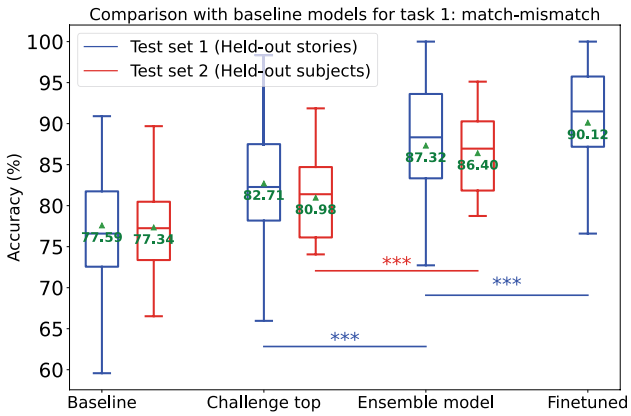

Fig. 3. Comparison of our network with baseline models for task 1: match-mismatch. The baseline network is the dilated convolutional network, which is proposed as a baseline for task 1[72]. The challenge top network is the top-performing architecture, proposed by Thornton et al.[73]. Each point in the boxplot corresponds to the average accuracy for one subject, averaged across all segments from the official test set of the 2023 Auditory EEG decoding challenge. Whiskers extend to the most extreme data points within 1.5 times the interquartile range from the quartiles. The green triangles and corresponding numerical values denote the mean accuracy across all subjects. The proposed ensemble model significantly outperforms the winners of the Auditory EEG challenge ($***: p < 0.0001$) for both the Held-out stories and the Held-out subjects test set. Statistical significance was assessed using single-sided Wilcoxon tests with Holm-Bonferroni corrections for multiple comparisons. Reported p-values are the corrected values. The finetuned model is the ensemble model, which has been finetuned on individual subjects. Since no data is available for the subjects of test set 2 in the training set, finetuning has only been conducted for the subjects of test set 1.

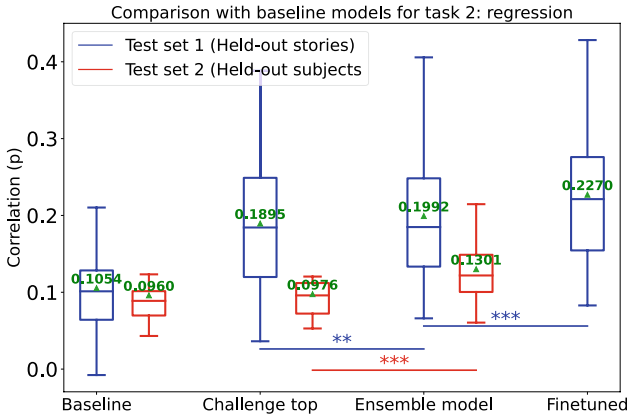

Fig. 4. Comparison of our network with baseline models for task 2: regression. The baseline network is a simple linear model, which is proposed as a baseline for task 2. The challenge top network is the top-performing architecture proposed by Piao et al.[48]. Each point in the boxplot corresponds to the average correlation for one subject, averaged across all segments from the official test set of the 2023 Auditory EEG decoding challenge. The green points denote the mean correlation across all subjects. The proposed ensemble model significantly outperforms the winners of the Auditory EEG challenge for both the Held-out stories ($p < 0.001$) and the Held-out subjects ($p < 0.0001$) test set. Significance has been tested using single-side Wilcoxon-signed rank tests. The finetuned model is the ensemble model, which has been finetuned on individual subjects. Since there is no data available for the subjects of test set 2 in the training set, finetuning has only been conducted for the subjects of test set 1.

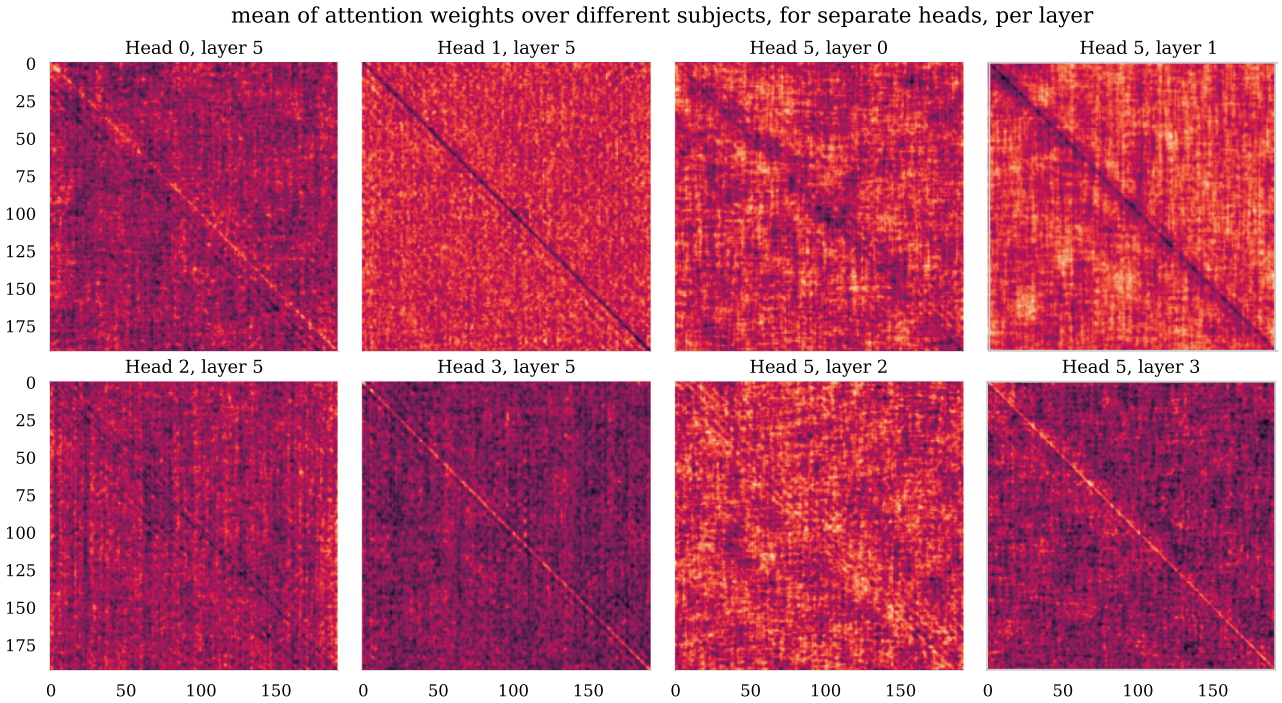

Fig. 5. Mean attention weights, visualized for one segment, averaged over different subjects. We have averaged the outputs of the different heads. The attention of each of the 192 tokens to each of the other 192 tokens is calculated and visualized.

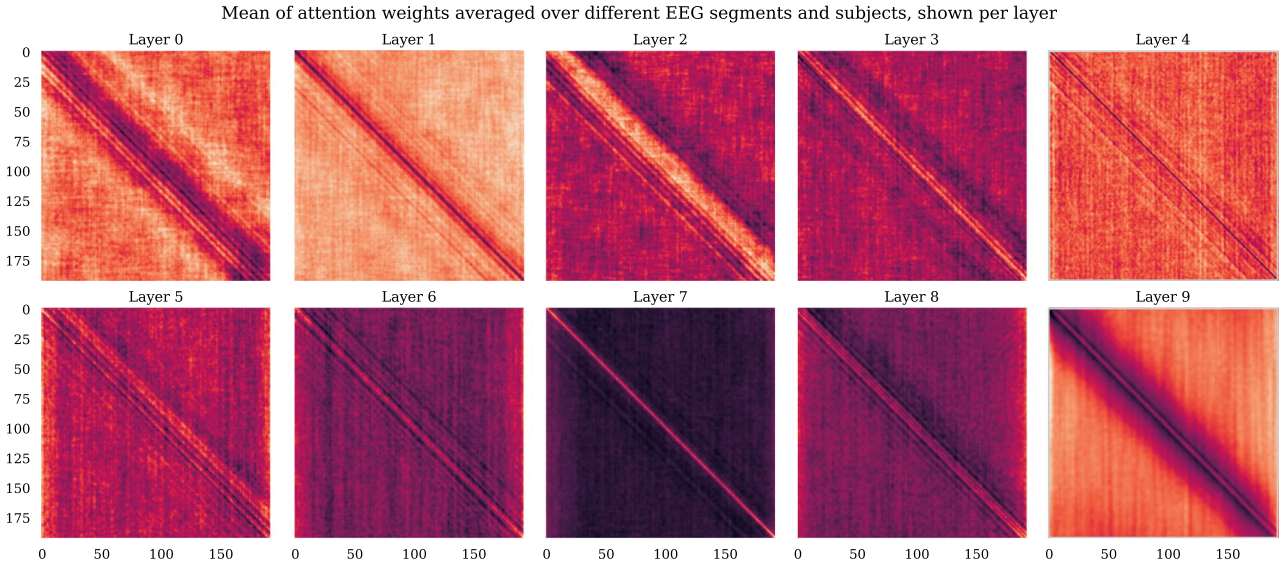

Fig. 6. Mean attention weights, visualized over different segments, over different subjects. We have averaged the outputs of the different heads. The attention of each of the 192 tokens to each of the other 192 tokens is calculated and visualized.

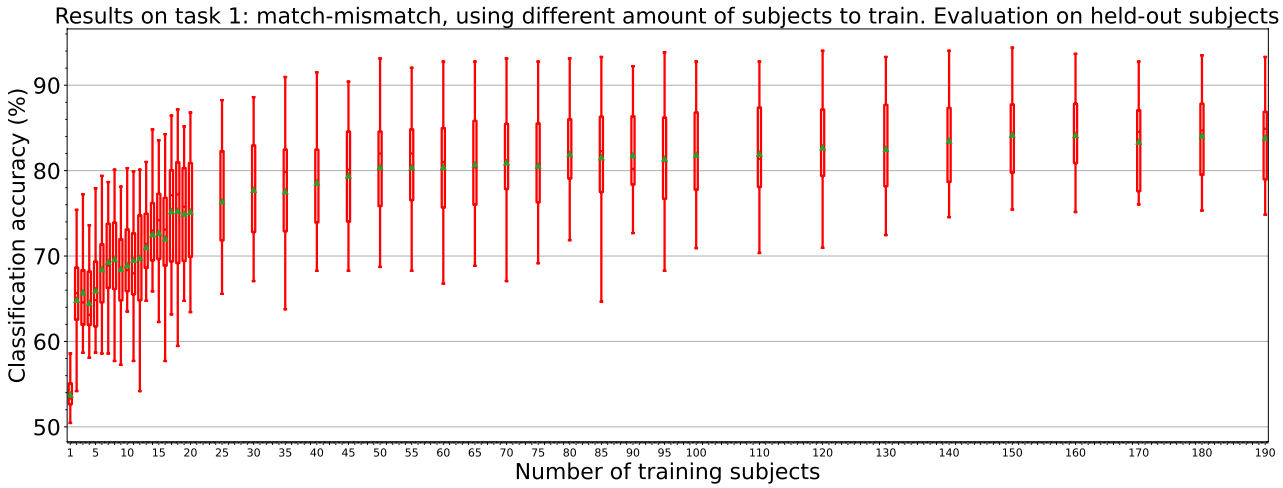

Fig. 7. The network is trained on 1 to 190 subjects and evaluated on the 14 heldout subjects of test set 2. Each point in the boxplot corresponds to the average accuracy of one of these 14 subjects

up until 600 *ms* after the onset of a stimulus[79]. At our sampling rate of 64 *Hz*, this corresponds to ≈ 38 samples, which is roughly the time region in which the attention maps show significant values.

Finetuning

The results of finetuning can be seen in figure 3 and 4. For task 1, the median accuracy increases from 87.32% to 90.12%, and 48 out of 70 subjects perform better after finetuning. The results from the finetuned model are significantly better than those of the general model.

Similarly, for task 2, the median correlation increases from 0.1992 to 0.2270, and 65 out of the 70 subjects perform better after finetuning. The results from the finetuned model are significantly better than those of the general model. Significance values were computed using one-sided Wilcoxon signed-rank tests with Bonferroni-Holm corrections and a p-value of 0.001.

How many subjects are needed?

The results of training on varying number of subjects can be seen in Fig. 7. The mean accuracy and correlation increase with the number of subjects that were seen in training. When training on one subject, a mean accuracy of 53.78% is obtained. To achieve an accuracy higher than 70%, training has to be performed on at least 13 subjects. To achieve an accuracy of at least 80%, training has to be performed on at least 50 subjects. After 50

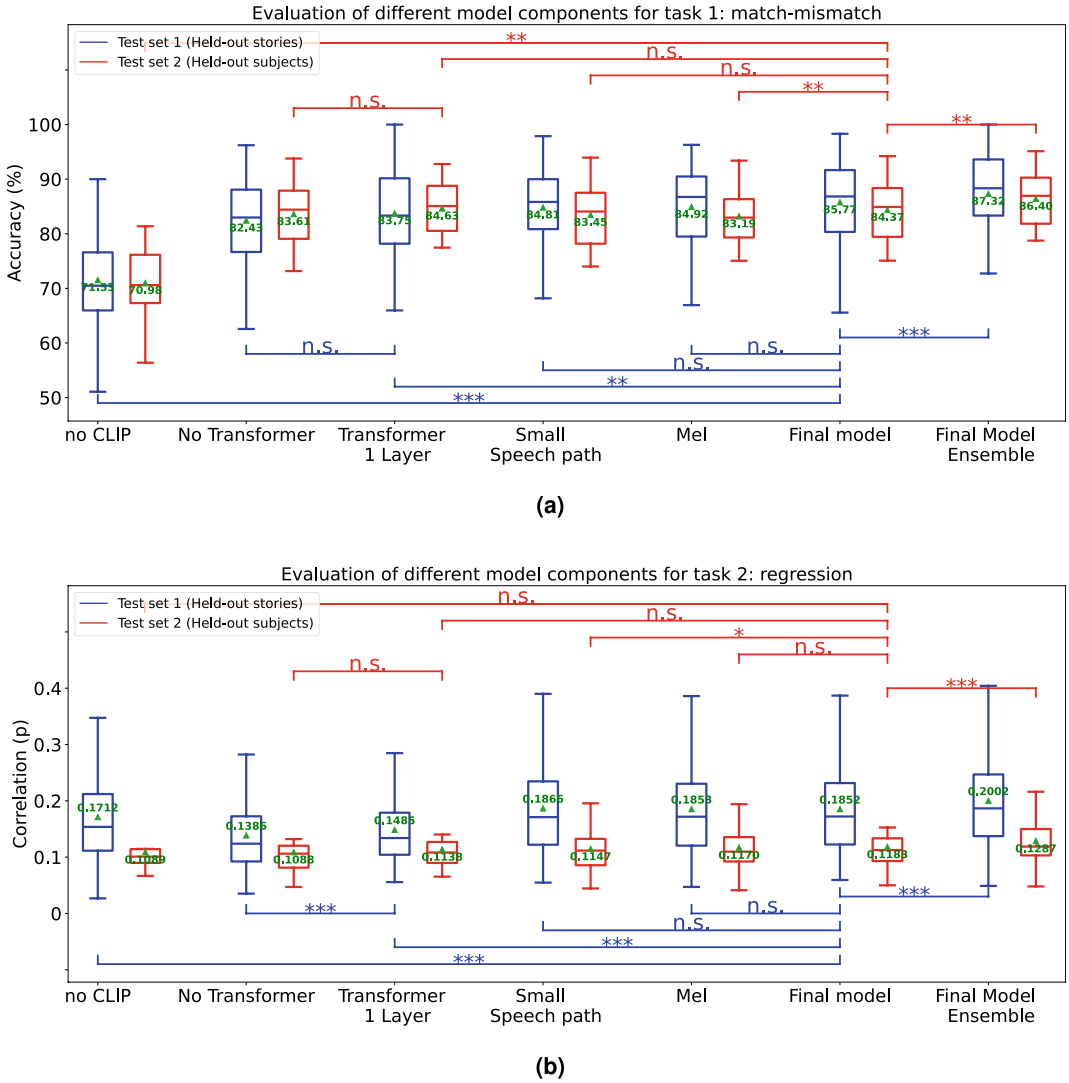

Fig. 8. Ablation study of the network. **(a)** evaluated on task 1: match-mismatch. Each point in the boxplot corresponds to the average accuracy of one subject. The green triangles and corresponding numerical value denote the mean accuracy across all subjects. **(b)** evaluated on task 2: regression. Each point in the boxplot corresponds to the average correlation of one subject. The green points denote the mean correlation across all subjects

subjects, the accuracy further increases as we add subjects, but there are diminishing returns. Similar results have been found in other recent papers[15], which propose a model for the regression task.

Ablation Study

The quantitative outcomes of this ablation study are presented in Fig. 8(a) and (b). Statistical significance was assessed using two-sided Wilcoxon tests with Holm-Bonferroni corrections for multiple comparisons. Reported p-values are the corrected values.

Leaving out the CLIP loss mechanism does result in significantly lower accuracy for task 1, with the classification accuracy dropping down to around 70% (p-values < 0.001). Interestingly, for the regression task, the correlation for the held-out subjects is significantly lower when omitting the CLIP loss than when using the final model ($p < 0.001$). However, the drop in correlation is not nearly as high as for the classification task. Additionally, for the held-out subjects, there is no significant difference when using or omitting the CLIP loss.

Simplifying the EEG encoder results in a significantly lower accuracy. There is a significant difference between using only a convolutional block in the EEG encoder (**No Transformer**) and using a small EEG transformer (**Transformer 1 Layer**) versus when using the full model (**wav2vec_19**)). This difference appears to be most pronounced in the seen subjects, both for task one and task two. There is a small increase in accuracy when adding 1 layer of a transformer block for the within-subjects (p-value < 0.05 for task 1, p-value < 0.001 for task 2). When adding extra repetitions, there is a more significant increase for test set 1 (task 1 ($p < 0.01$) and task 2 ($p < 0.001$)). For the held-out subjects, there is a small numerical increase. However, there is no significant difference for any of the additions.

Using the full speech encoder, as opposed to the **small speech encoder**, leads to a small, yet not statistically significant, increase in accuracy for task 1 for both test sets. When using the wav2vec, layer 19 speech representations, as opposed to the **mel** spectrogram, there is a slight increase in accuracy for task 1 for both the within-subjects (not significant) and held-out subjects ($p<0.01$). Notably, none of these alterations, all of which pertain to the speech encoder only, influence the correlations that are obtained in the regression tasks.

When using the **ensemble** model, which for task 1 averages the logits before taking a final decision, the accuracy for both tests gets significantly better (p-values < 0.001).

For task 2, the ensemble model averages the final envelopes before calculating the correlation. The same observation can be seen here, with significant increases for both test sets, p-values < 0.001). In summary, significant improvements were found using CLIP loss, a transformer architecture and averaging the logits before taking a final decision. Replacing the **mel** spectrogram with the output of layer 19 of wav2vec yielded a significant accuracy increase for held-out subjects only.

Discussion

Our study presents a novel approach for auditory EEG decoding using contrastive learning. It demonstrates its efficacy in enhancing model performance across two distinct tasks: match-mismatch classification and stimulus envelope regression. By learning latent representations that capture the relationship between auditory stimuli and EEG responses, our model significantly outperforms state-of-the-art methods on both tasks, as evidenced by the results on the ICASSP 2023 Auditory EEG Decoding Challenge dataset. Using ensemble learning, for the match-mismatch task, we obtain an accuracy of 87.32% for the held-out stories test set and 86.40% for the held-out subjects test set. For the regression task, we obtain a mean correlation of 0.1992 for the held-out stories test set and 0.1301 for the held-out subjects test set. The highest performance was achieved using an ensemble model, in which each of the 16 models was trained using different initialization weights. This tactic was first introduced in the context of auditory EEG decoding by Thornton et al.[73] for the auditory EEG decoding challenge of 2023 and has been used by 4 out of the five teams for the second auditory EEG decoding challenge. This is in line with what could be expected, as different initializations of the model weights might lead to different patterns that have been found. We obtain competitive performance on two downstream tasks, without explicitly training the model for these tasks. We hypothesize that the contrastive learning mechanism forces the model to learn to encode information in the latent representation that maximally discriminates between different speech features, and therefore, useful information for auditory decoding is retained. Although not tested, we hypothesize that our model would also perform well on other auditory decoding-related downstream tasks without necessary retraining of the model.

When comparing these results to the results of the ICASSP 2023 auditory decoding challenge, we significantly outperform all other proposed architectures for the match-mismatch task. Thornton et al[43] obtained the highest performance, with an accuracy of 82.71% for the held-out stories test set and 80.98% for the held-out subjects test set. In their solution, they used high-frequency envelope modulations as a speech representation but used a relatively simple architecture. In contrast, our model uses more complex speech features in combination with a significantly more complex training architecture. For the regression task, Piao et al.[48] gained the top score, with correlation values of 0.1895 and 0.0976 for, respectively, the held-out stories and held-out subjects. Our model significantly outperforms this model for both tasks. However, for the held-out stories test set, the result is not nearly as big as for the held-out subjects. This might indicate that letting the model create latent representations first and letting it focus on making these representations as discriminative as possible enables the model to learn a more general representation of speech, which helps when creating the representations for held-out subjects. The authors propose a feedforward transformer architecture with an auxiliary global conditioner that integrates subject information in the model. When comparing a single instance of our model, which obtains correlation values of 0.1852 and 0.1188 for respectively the held-out stories and held-out subjects, (as opposed to an ensemble model) against their solution, the difference for the held-out stories disappears, and only the difference in correlation for the held-out subjects remains.

The ablation study sheds light on the role of each model component. First, the CLIP mechanism is essential for attaining high accuracy on both tasks. When omitting the contrastive loss, accuracy for the match-mismatch task drops to about 70%. Conversely, the correlation for the regression drops, but not nearly as much. Without the CLIP mechanism, the training task can be summarized as *does this segment of EEG correspond to this segment of speech?*. Thus, in training, the model sees both the EEG and the mel spectrogram and has to make a decision based on this one instance, which forces the model to find resemblances between EEG and speech. Therefore, it is not unsurprising that the no CLIP model still manages to reconstruct the envelope well. Looking at the other components, it becomes obvious that the strength of this model, when reconstructing an envelope, lies in the transformer architecture. The match-mismatch task, where the task is to classify between two segments of speech, can be seen as a simplified version of a contrastive loss, with just one EEG segment and just two speech segments. Therefore, it is not unsurprising that omitting the contrastive loss in its entirety leads to such a big drop in accuracy.

The EEG transformer architecture, particularly with multiple repetitions, proves vital for capturing the temporal dynamics of the EEG responses to auditory stimuli. Conversely, modifications to the speech encoder, such as using mel spectrograms or simplified convolutional layers, yield only marginal improvements, suggesting that the EEG representation is the primary driver of performance. When decoding a speech envelope from the EEG embeddings, the speech representation is not directly used. Hence, the EEG encoder should be the most important one. Previous work has found that using richer speech representations in combination with neural data does enhance the final accuracies. There is a trend of increased accuracy with wav2vec features, but the difference is small and not significant. Further research is necessary to determine whether this is due to a limited speech encoder, which might not be able to exploit the full capacity of these features, or whether there is a ceiling

in the accuracies that are being reached. Furthermore, when one would like to use the model in a real-life setting, it may be more convenient to use the mel spectrogram as opposed to the wav2vec features due to their much lower dimensionality (28 vs. 1024).

Our investigation into the impact of training set size reveals a clear trend: increasing the number of subjects in the training set leads to improved performance on held-out subjects. The greatest impact can be seen when adding the first subjects. After 50 subjects, an accuracy of 80% has been reached and adding more subjects (up to 194) yields a higher accuracy score. However, there are obviously diminishing returns, as the final score does not exceed 84%. It remains an open question whether this indicates that a ceiling in the obtainable accuracy has been reached due to the quality of the measured data, the inherent attention of the subjects, or whether there is a limitation in the current networks, and whether bigger, more complex models might be able to extract more information from the EEG.

We finetuned the models for all the subjects from test set 1. For task 1 (classification), the median accuracy improved from 87.32% to 90.12%, with a majority (48 out of 70) of subjects showing better performance after finetuning. For task 2 (regression), the median correlation increased from 0.1992 to 0.2270, with an even larger majority (65 out of 70) of subjects benefiting from the finetuning process. These results highlight the effectiveness of finetuning as a strategy for improving model performance on subject-specific tasks, particularly when dealing with larger neural networks.

In general, neural networks are notoriously hard to interpret, especially large ones. This might hinder the applicability in clinical settings, where understanding the reasoning behind a model's prediction is crucial. Future research could focus on more accurate interpretability and explainability metrics, which could better visualize the learned representations and identify the relevant EEG features. Our current model has approximately 3.5 million parameters, and the ensemble model increases this to around 56 million. While these numbers are not large by contemporary standards, practical applications, such as neuro-steered hearing aids, may demand smaller, faster, and more energy-efficient models. This trade-off between performance and computational feasibility must be carefully considered for specific use cases.

Another critical limitation is the lack of clarity regarding which features of the speech stimulus are represented or utilized in the latent space. While the attention maps provide some insights into the EEG data used for creating certain representations, there is a limited understanding of the specific features of the wav2vec model. These features, often derived from intermediate layers (in our case, layer 19), have been shown to encode lexical, acoustic, and other information[63]. Previous studies have demonstrated the advantages of using more complex speech features over mel spectrograms or envelopes[56,80,81], underscoring their utility in linking EEG to speech.

Feature engineering has historically played a significant role in auditory EEG decoding. Despite advances in large models capable of internal feature extraction, many researchers still achieve competitive results using heavily engineered speech features[43,44].

The performance of models in auditory EEG decoding remains difficult to standardize due to the diversity of datasets and training paradigms across research groups. Our evaluation relied on the sparrKULee dataset[30] and the official test set of the ICASSP 2023 auditory EEG decoding challenge[37], enabling comparisons with other architectures. However, these datasets are relatively homogeneous, consisting of data collected in controlled, soundproof environments using high-quality audio stimuli.

The inherent noisiness of EEG signals and the challenges of training deep learning models necessitate the use of such controlled datasets in initial studies. Expanding to more diverse datasets, including data from broader demographic groups, varying age ranges, and individuals with hearing impairments, is crucial for enhancing the model's generalizability. Furthermore, real-life scenarios often involve multi-speaker environments and background noise, conditions not captured in our current setup. Future work should focus on incorporating these factors into datasets to create robust models suitable for practical applications.

In conclusion, this study presents a novel contrastive learning framework for auditory EEG decoding, achieving state-of-the-art performance in match-mismatch classification (87% accuracy) and stimulus envelope regression (0.176 mean Pearson correlation). By learning robust latent representations, the model generalizes effectively across tasks and subjects, with ensemble methods proving particularly impactful. While finetuning further improves performance, challenges such as the interpretability of neural networks and computational demands of ensemble models remain. Expanding evaluations to diverse, real-world datasets and refining architectures for efficiency and scalability are essential for practical applications. This work marks a significant advancement in auditory EEG decoding, with promising potential for both research and clinical innovations.

Data availability

The full dataset analyzed during the current study is available from the corresponding author on reasonable request and in compliance with the participant's consent. A subset of the total dataset, comprising 105 participants, is available and has previously been published[30,37,40].

Received: 1 September 2024; Accepted: 25 July 2025
Published online: 06 August 2025

References

1. Crosse, M. J., Di Liberto, G. M., Bednar, A. & Lalor, E. C. The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience* **10**, 1–14. https://doi.org/10.3389/fnhum.2016.00604 (2016).
2. Vanthornhout, J., Wouters, J., Simon, J. Z. & Francart, T. Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. *Journal of the Association for Research in Otolaryngology* **19**, 181–191. https://doi.org/10.1007/s10162-018-0654-z (2018).

3. Iotzov, I. & Parra, L. C. EEG can predict speech intelligibility. *Journal of Neural Engineering* **16**, 036008. https://doi.org/10.1088/1741-2552/ab07fe (2019).
4. Sriraam, N. EEG based automated detection of auditory loss: A pilot study. *Expert Systems with Applications* **39**, 723–731. https://doi.org/10.1016/j.eswa.2011.07.064 (2012).
5. M P, P., Subramaniam, K., Yaccob, S., Hamid, A. & Hema, C. EEG Based Detection of Conductive and Sensorineural Hearing Loss using Artificial Neural Networks. *Journal of Next Generation Information Technology* **4**, 204–212, https://doi.org/10.4156/jnit.vol4.issue3.24 (2013).
6. Petersen, E. B., Wöstmann, M., Obleser, J. & Lunner, T. Neural tracking of attended versus ignored speech is differentially affected by hearing loss. *Journal of neurophysiology* **117**, 18–27 (2017).
7. Decruy, L., Vanthornhout, J. & Francart, T. Hearing impairment is associated with enhanced neural tracking of the speech envelope. *Hearing Research* **393**, 107961 (2020).
8. Fuglsang, S. A., Märcher-Rørsted, J., Dau, T. & Hjortkjær, J. Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention. *Journal of Neuroscience* **40**, 2562–2572 (2020).
9. Cardin, V. Effects of aging and adult-onset hearing loss on cortical auditory regions. *Frontiers in Neuroscience* **10**, 199 (2016).
10. Peelle, J. E., Gross, J. & Davis, M. H. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral cortex* **23**, 1378–1387 (2013).
11. Ahissar, E. et al. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences* **98**, 13367–13372 (2001).
12. Yi, H. G., Leonard, M. K. & Chang, E. F. The encoding of speech sounds in the superior temporal gyrus. *Neuron* **102**, 1096–1110 (2019).
13. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P. & De Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences* **119**, e2201968119 (2022).
14. Reddy Katthi, J. & Ganapathy, S. Deep correlation analysis for audio-eeg decoding. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **29**, 2742–2753, https://doi.org/10.1109/TNSRE.2021.3129790 (2021).
15. Accou, B., Vanthornhout, J., Van hamme, H. & Francart, T. Decoding of the speech envelope from EEG using the VLAAI deep neural network, https://doi.org/10.1101/2022.09.28.509945 (2022). Pages: 2022.09.28.509945 Section: New Results.
16. Krishna, G., Tran, C., Han, Y., Carnahan, M. & Tewfik, A. H. Speech synthesis using eeg. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1235–1238, https://doi.org/10.1109/ICASSP40776.2020.9053340 (2020).
17. Sakthi, M., Tewfik, A. & Chandrasekaran, B. Native language and stimuli signal prediction from eeg. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3902–3906, https://doi.org/10.1109/ICASSP.2019.8682563 (2019).
18. Thornton, M., Mandic, D. & Reichenbach, T. Robust decoding of the speech envelope from eeg recordings through deep neural networks. *Journal of Neural Engineering* **19**, 046007 (2022).
19. de Taillez, T., Kollmeier, B. & Meyer, B. Machine learning for decoding listeners' attention from EEG evoked by continuous speech. *European Journal of Neuroscience* **51**, https://doi.org/10.1111/ejn.13790 (2017).
20. Puffay, C. et al. Relating EEG to continuous speech using deep neural networks: a review. *Journal of Neural Engineering* https://doi.org/10.1088/1741-2552/ace73f (2023).
21. Hosseini, M., Celotti, L. & Plourde, E. Speaker-independent brain enhanced speech denoising. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1310–1314, https://doi.org/10.1109/ICASSP39728.2021.9414969 (2021).
22. Katthi, J. R. & Ganapathy, S. Deep multiway canonical correlation analysis for multi-subject eeg normalization. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1245–1249 (2021).
23. Bollens, L., Francart, T. & Hamme, H. V. 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)*, 1256–1260, https://doi.org/10.1109/ICASSP43922.2022.9747297 (2022).
24. Krishna, G., Tran, C., Carnahan, M., Han, Y. & Tewfik, A. H. Generating eeg features from acoustic features. In *2020 28th European Signal Processing Conference (EUSIPCO)*, 1100–1104, https://doi.org/10.23919/Eusipco47968.2020.9287498 (2021).
25. de Cheveigné, A. et al. Decoding the auditory brain with canonical component analysis. *NeuroImage* **172**, 206–216. https://doi.org/10.1016/j.neuroimage.2018.01.033 (2018).
26. Accou, B., Jalilpour-Monesi, M., Montoya-Martínez, J., hamme, H. V. & Francart, T. Modeling the relationship between acoustic stimulus and eeg with a dilated convolutional neural network. *2020 28th European Signal Processing Conference (EUSIPCO)* 1175–1179 (2021).
27. Monesi, M. J., Accou, B., Montoya-Martinez, J., Francart, T. & Van hamme, H. An LSTM based architecture to relate speech stimulus to EEG. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 941–945, https://doi.org/10.1109/ICASSP40776.2020.9054000 (2020). ISSN: 2379-190X.
28. Jalilpour Monesi, M., Accou, B., Van hamme, H. & Francart, T. Extracting different levels of speech information from EEG using an lstm-based model. *Proceedings Interspeech 2021* 526–530 (2021).
29. Puffay, C., Van Canneyt, J., Vanthornhout, J., Van hamme, H. & Francart, T. Relating the fundamental frequency of speech with EEG using a dilated convolutional network. *Proceedings Interspeech 2022* (2022-09-18).
30. Accou, B. et al. Sparrkulee: A speech-evoked auditory response repository from ku leuven, containing the eeg of 85 participants. *Data* **9**, https://doi.org/10.3390/data9080094 (2024).
31. Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J. & Lalor, E. C. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology* **28**, 803-809.e3. https://doi.org/10.1016/j.cub.2018.01.080 (2018).
32. Fuglsang, S. A., Wong, D. D. & Hjortkjær, J. *EEG and audio dataset for auditory attention decoding* https://doi.org/10.5281/zenodo.1199011 (2018).
33. Etard, O. & Reichenbach, T. *EEG Dataset for "Decoding of selective attention to continuous speech from the human auditory brainstem response" and "Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise"*. https://doi.org/10.5281/zenodo.7086209 (2022).
34. Weissbart, H., Kandylaki, K. & Reichenbach, T. *EEG Dataset for "Cortical Tracking of Surprisal during Continuous Speech Comprehension"* https://doi.org/10.5281/zenodo.7086168 (2022).
35. Brennan, J. R. EEG Datasets for Naturalistic Listening to "Alice in Wonderland" (2018).
36. Vanheusden, F. J. et al. *Dataset for: Hearing aids do not alter cortical entrainment to speech at audible levels in mild-to-moderately hearing-impaired subjects* https://doi.org/10.5258/SOTON/D1140 (2019).
37. Monesi, M. J. et al. Auditory eeg decoding challenge for icassp 2023. *IEEE Open Journal of Signal Processing* (2024).
38. Bollens, L. et al. Icassp 2023 auditory eeg decoding challenge. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–2 (IEEE, 2023).
39. Bollens, L. et al. Icassp 2024 auditory eeg decoding challenge. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 127–128, https://doi.org/10.1109/ICASSPW62465.2024.10626137 (2024).
40. Bollens, L. et al. Auditory eeg decoding challenge for icassp 2024. *TechrXiv preprint* https://doi.org/10.36227/techrxiv.172504102.27360831 (2024).

41. Borsdorf, M. et al. Multi-head attention and gru for improved match-mismatch classification of speech stimulus and EEG response. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–2, https://doi.org/10.1109/ICASSP49357.2023.10096959 (2023).
42. Kulasingham, J. P. et al. High gamma cortical processing of continuous speech in younger and older listeners. *NeuroImage* **222**, 117291. https://doi.org/10.1016/j.neuroimage.2020.117291 (2020).
43. Thornton, M., Mandic, D. P. & Reichenbach, T. Decoding envelope and frequency-following eeg responses to continuous speech using deep neural networks. *IEEE Open Journal of Signal Processing* (2024).
44. Thornton, M., Auernheimer, J., Jehn, C., Mandic, D. & Reichenbach, T. Detecting gamma-band responses to the speech envelope for the ICASSP 2024 auditory EEG decoding signal processing grand challenge. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 55–56, https://doi.org/10.1109/ICASSPW62465.2024.10626244 (2024).
45. Baevski, A., Zhou, H., Mohamed, A. & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR* **abs/2006.11477** (2020). arXiv2006.11477.
46. Wang, B. et al. Self-supervised speech representation and contextual text embedding for match-mismatch classification with EEG recording. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 111–112, https://doi.org/10.1109/ICASSPW62465.2024.10625969 (2024).
47. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
48. Piao, Z., Kim, M., Yoon, H. & Kang, H.-G. HappyQuokka system for ICASSP 2023 auditory EEG challenge. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–2, https://doi.org/10.1109/ICASSP49357.2023.10095477 (2023).
49. Xu, X. et al. Convconcatnet: A deep convolutional neural network to reconstruct mel spectrogram from the EEG. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 113–114, https://doi.org/10.1109/ICASSPW62465.2024.10626859 (2024).
50. Qiu, Z., Gu, J., Yao, D., Li, J. & Yan, Y. BMMSNet: Bidirectional mapping and multilevel similarity comparison for EEG-speech match-mismatch problem. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 117–118, https://doi.org/10.1109/ICASSPW62465.2024.10626977 (2024).
51. Oord, A. v. d. et al. WaveNet: A Generative Model for Raw Audio (2016). ArXiv:1609.03499 [cs].
52. Van Dyck, B., Yang, L. & Van Hulle, M. M. Decoding auditory EEG responses using an adapted Wavenet. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–2, https://doi.org/10.1109/ICASSP49357.2023.10095420 (2023).
53. Fang, Y., Li, H., Zhang, X., Chen, F. & Gao, G. Cross-attention-guided wavenet for mel spectrogram reconstruction in the ICASSP 2024 auditory EEG challenge. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 7–8, https://doi.org/10.1109/ICASSPW62465.2024.10627006 (2024).
54. Oord, A. v. d., Li, Y. & Vinyals, O. Representation Learning with Contrastive Predictive Coding (2019). ArXiv:1807.03748 [cs, stat].
55. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607 (PMLR, 2020).
56. Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O. & King, J.-R. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence* **5**, 1097–1107 (2023).
57. Banville, H., Chehab, O., Hyvärinen, A., Engemann, D.-A. & Gramfort, A. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering* **18**, 046020 (2021).
58. Radford, A. et al. Learning transferable visual models from natural language supervision (2021). arXiv2103.00020.
59. Somers, B., Francart, T. & Bertrand, A. A generic EEG artifact removal algorithm based on the multi-channel Wiener filter. *Journal of Neural Engineering* **15**, 036007. https://doi.org/10.1088/1741-2552/aaac92 (2018).
60. Grosman, J. Fine-tuned XLSR-53 large model for speech recognition in Dutch. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-dutch (2021).
61. Millet, J. et al. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems* **35**, 33428–33443 (2022).
62. Vaidya, A. R., Jain, S. & Huth, A. G. Self-supervised models of audio effectively explain human cortical responses to speech. arXiv preprint arXiv:2205.14252 (2022).
63. Pasad, A., Chou, J.-C. & Livescu, K. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 914–921 (IEEE, 2021).
64. Bollens, L., Accou, B., Van hamme, H. & Francart, T. A Large Auditory EEG decoding dataset, https://doi.org/10.48804/K3VSND (2023).
65. COREN, S. The lateral preference inventory for measurement of handedness, footedness, eyedness, and earedness - norms for young-adults. *BULLETIN OF THE PSYCHONOMIC SOCIETY* **31**, 1–3, https://doi.org/10.3758/bf03334122 (1993).
66. Francart, T., van Wieringen, A. & Wouters, J. Apex 3: a multi-purpose test platform for auditory psychophysical experiments. *Journal of neuroscience methods* **172**, 283–93. https://doi.org/10.1016/j.jneumeth.2008.04.020 (2008).
67. Song, Y., Zheng, Q., Liu, B. & Gao, X. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **31**, 710–719 (2022).
68. Lawhern, V. J. et al. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering* **15**, 056013. https://doi.org/10.1088/1741-2552/aace8c (2018).
69. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer normalization. arXiv preprint arXiv:1607.06450 (2016).
70. Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016).
71. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
72. Accou, B., Monesi, M. J., Van hamme, H. & Francart, T. Predicting speech intelligibility from EEG using a dilated convolutional network. arXiv:2105.06844 [cs, eess, q-bio] (2021). ArXiv: 2105.06844.
73. Thornton, M., Mandic, D. P. & Reichenbach, T. Decoding envelope and frequency-following eeg responses to continuous speech using deep neural networks. *IEEE Open Journal of Signal Processing* 1–18, https://doi.org/10.1109/OJSP.2024.3378593 (2024).
74. de Cheveigné, A. et al. Decoding the auditory brain with canonical component analysis. *NeuroImage* **172**, 206–216. https://doi.org/10.1016/j.neuroimage.2018.01.033 (2018).
75. de Cheveigné, A., Slaney, M., Fuglsang, S. A. & Hjortkjaer, J. Auditory stimulus-response modeling with a match-mismatch task. *Journal of Neural Engineering* **18**, 046040. https://doi.org/10.1088/1741-2552/abf771 (2021).
76. Niu, Z., Zhong, G. & Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **452**, 48–62 (2021).
77. Yeh, C. et al. Attentionviz: A global view of transformer attention. *IEEE Transactions on Visualization and Computer Graphics* (2023).
78. Vig, J. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 37–42, https://doi.org/10.18653/v1/P19-3007 (Association for Computational Linguistics, Florence, Italy, 2019).
79. Brodbeck, C. & Simon, J. Z. Continuous speech processing. *Current Opinion in Physiology* **18**, 25–31 (2020).
80. Monesi, M. J., Vanthornhout, J., Van hamme, H. & Francart, T. The role of vowel and consonant onsets in neural tracking of natural speech, https://doi.org/10.48550/arXiv.2308.00161 (2023). ArXiv:2308.00161 [eess].

81. Caucheteux, C., Gramfort, A. & King, J.-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour* **7**, 430–441 (2023).

Acknowledgements
Funding was provided by the KU Leuven Special Research Fund C24/18/099 (C2 project to Tom Francart and Hugo Van hamme) and an FWO-SB fellowship to Lies Bollens (1SB1423N). The authors thank Amelie Algoet, Jolien Smeulders, Lore Kerkhofs, Sara Peeters, Merel Dillen, Ilham Gamgami, Amber Verhoeven and Wendy Verheijen for their help with data collection.

Author contributions
L.B. created the network architecture, ran experiments and analyzed the results. B.A., H.V.h and T.F. helped to create the network architecture and thoroughly analyzed the results. All authors have reviewed and contributed to the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-13646-4.

Correspondence and requests for materials should be addressed to B.A. or T.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025