data

MDPI

Data Descriptor

SparrKULee: A Speech-Evoked Auditory Response Repository from KU Leuven, Containing the EEG of 85 Participants

Bernd Accou [1,2,*,†], Lies Bollens [1,2,*,†], Marlies Gillis [1], Wendy Verheijen [1], Hugo Van hamme [2] and Tom Francart [1,*]

[1] Experimental Oto-Rhino-Laryngology (ExpORL), Department Neurosciences, KU Leuven, B-3001 Leuven, Belgium; marlies.gillis@kuleuven.be (M.G.)
[2] Processing Speech and Images (PSI), Department of Electrical Engineering (ESAT), KU Leuven, B-3001 Leuven, Belgium; hugo.vanhamme@esat.kuleuven.be
* Correspondence: bernd.accou@kuleuven.be (B.A.); lies.bollens@kuleuven.be (L.B.); tom.francart@kuleuven.be (T.F.)
† These authors contributed equally to this work.

Abstract: Researchers investigating the neural mechanisms underlying speech perception often employ electroencephalography (EEG) to record brain activity while participants listen to spoken language. The high temporal resolution of EEG enables the study of neural responses to fast and dynamic speech signals. Previous studies have successfully extracted speech characteristics from EEG data and, conversely, predicted EEG activity from speech features. Machine learning techniques are generally employed to construct encoding and decoding models, which necessitate a substantial quantity of data. We present SparrKULee, a Speech-evoked Auditory Repository of EEG data, measured at KU Leuven, comprising 64-channel EEG recordings from 85 young individuals with normal hearing, each of whom listened to 90–150 min of natural speech. This dataset is more extensive than any currently available dataset in terms of both the number of participants and the quantity of data per participant. It is suitable for training larger machine learning models. We evaluate the dataset using linear and state-of-the-art non-linear models in a speech encoding/decoding and match/mismatch paradigm, providing benchmark scores for future research.

Keywords: auditory EEG; dataset; speech

check for updates

Citation: Accou, B.; Bollens, L.; Gillis, M.; Verheijen, W.; Van hamme, H.; Francart, T. SparrKULee: A Speech-Evoked Auditory Response Repository from KU Leuven, Containing the EEG of 85 Participants. *Data* **2024**, *9*, 94. https://doi.org/10.3390/data9080094

Academic Editor: Giuseppe Ciaburro

Received: 10 June 2024
Revised: 5 July 2024
Accepted: 16 July 2024
Published: 26 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Summary

In order to study the neural processing of speech, recent studies have presented natural running speech to participants while recording an electroencephalogram (EEG). Currently, regression is used to either decode features of the speech stimulus from the EEG (also known as a backward model) [1–5], to predict the EEG from the speech stimulus [1,6] (forward model), or to transform both EEG and speech stimulus to a shared space [7,8] (hybrid model). Deep neural networks have recently been proposed for auditory decoding and have obtained promising results [4,5,9–12].

All previously mentioned methods require EEG recordings of the participants with strict time alignment to the speech stimulus. This time alignment is necessary due to the time-locked neural tracking of the speech stimulus at a millisecond scale (e.g., auditory brainstem responses (ABRs)), which can last up to 600 ms [13]. As these data are personal and expensive to collect, there is a need for more public datasets that researchers can use to benchmark and train their models.

Table 1 presents an overview of currently available public datasets of EEG recordings of people listening to natural speech. These studies have generated 87.7 h of EEG data from 133 participants listening to clean speech and speech-in-noise in their native language. However, this quantity of data is relatively small compared to datasets in other domains, such as automatic speech recognition, and needs to be increased for training models due

to the low signal-to-noise ratio of auditory EEG. Additionally, combining the data from these studies for model training is challenging due to differences in the authors' signal acquisition equipment, measurement protocols, and preprocessing methods. SparrKULee provides a large homogeneous dataset to train and evaluate models effectively, alleviating the need for combining multiple heterogeneous datasets. Its utilization has been facilitated through two organized challenges [14–16], which have catalyzed the dataset's adoption, leading to a range of advancements and novel model architectures [17–24].

Table 1. Overview of currently publicly available single-speaker datasets.

Dataset	Ref	Speech Material	Language	Participants	Average Time per Participant (min)	Total Time (min)
Broderick	[25]	Clean speech	English	19	60	1140
		Time-reversed speech		10	60	600
		Speech-in-noise		21	30	630
DTU Fuglsang	[26]	Clean speech	Danish	18	8.3	150
Etard	[27]	Clean speech	English	18	10	180
		Speech-in-noise		18	30	540
		Foreign language speech	Dutch	12	40	480
Weissbart	[28]	Clean speech	English	13	40	520
Brennan	[29]	Clean speech	English	49	12.4	610
Vanheuseden	[30]	Clean speech	English	17	24	410
SparrKULee		Clean speech	Dutch	85	110	9320
		Speech-in-noise		26	28.5	740

2. Methods

For our dataset (SparrKULee), we conducted an EEG experiment in which 85 participants were recruited and presented with speech stimuli for a duration ranging between 90 and 150 min, divided into 6 to 10 recordings (i.e., an uninterrupted period in which a participant listens to a stimulus), totaling 168 h of EEG data. To validate the obtained dataset, we employed state-of-the-art linear [2,8,31] and deep learning models [12], in participant-specific and participant-independent training scenarios (note that the cited studies used different, smaller datasets). These models can serve as benchmarks for comparison in future research. Our dataset is publicly available on the RDR KU Leuven website https://doi.org/10.48804/K3VSND (accessed on 1 June 2024).

We defined a trial as an uninterrupted recording lasting around 15 min. We defined a session as the complete set of trials and pre-screening activities that a participant underwent from the moment they entered the room until the moment they left. Stimulus, in our study, referred to the speech audio files that we presented to the participants during the experiment, which were designed to elicit specific responses from their brains. Figure 1 provides a high-level overview of the different parts of a session.

2.1. Participants

Between October 2018 and September 2022, data were collected from 85 participants (74 females/11 males, 21.4 $\pm$ 1.9 years (sd)). Inclusion criteria for this study were young (18–30 years), normal-hearing adults (all hearing thresholds $\leq$ 30 dB HL, for 125–8000 Hz), with Dutch/Flemish as their native language. Before commencing the EEG experiments, participants read and signed an informed consent form approved by the Medical Ethics Committee UZ KU Leuven/Research (KU Leuven, Belgium) with reference S57102. All participants in this dataset explicitly consented to share their pseudonymized data in a publicly accessible dataset. This dataset is a subset of our larger proprietary dataset containing data from participants who did not give consent to share their data. Additionally,

the participants completed a questionnaire requesting general demographic information (age, sex, education level, handedness [32]) and diagnoses of hearing loss and neurological pathologies. Participants indicating any neurological or hearing-related diagnosis were excluded from the study. Last, the medical history and the presence of learning disabilities were questioned as research has shown that serious concussions, the medication used to treat, for example, insomnia [33] and learning disabilities such as dyslexia can affect brain responses [34,35]. Therefore, that information was used to screen out participants with possibly divergent brain responses.

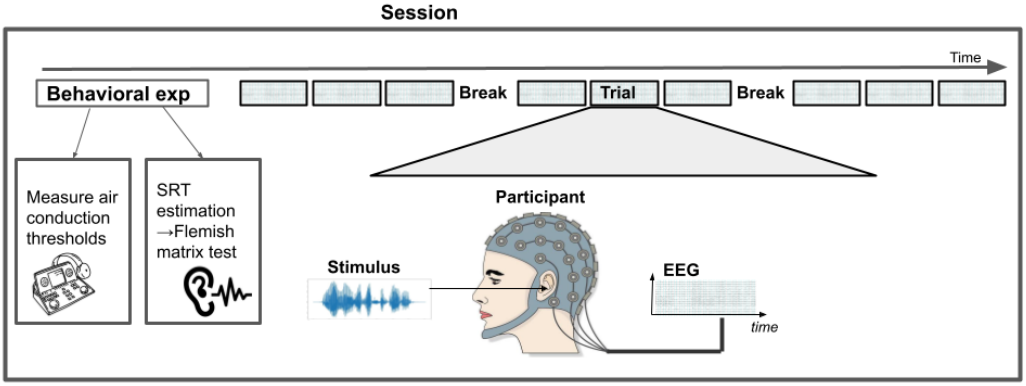

Figure 1. Overview of a session. First, the participant underwent behavioral experiments: air conduction thresholds were measured using the Hughson-Westlake method and the Flemish matrix test estimated the speech reception threshold (SRT). Following the Flemish matrix test, the EEG part of the study started, consisting of multiple trials of EEG recording. A trial was defined as an uninterrupted EEG measurement when a stimulus was playing. In this study, trials were approximately 15 min in length. After three trials, the participants were offered the option to take a short break.

2.2. Experimental Procedure

In this section the experimental procedure for the behavioral screening and subsequent EEG measurement are explained.

2.2.1. Behavioral

First, we measured the air conduction thresholds using the Hughson-Westlake method [36] for frequencies from 125 to 8000 Hz (see Figure 2). Participants with hearing thresholds > 30 dB HL were excluded.

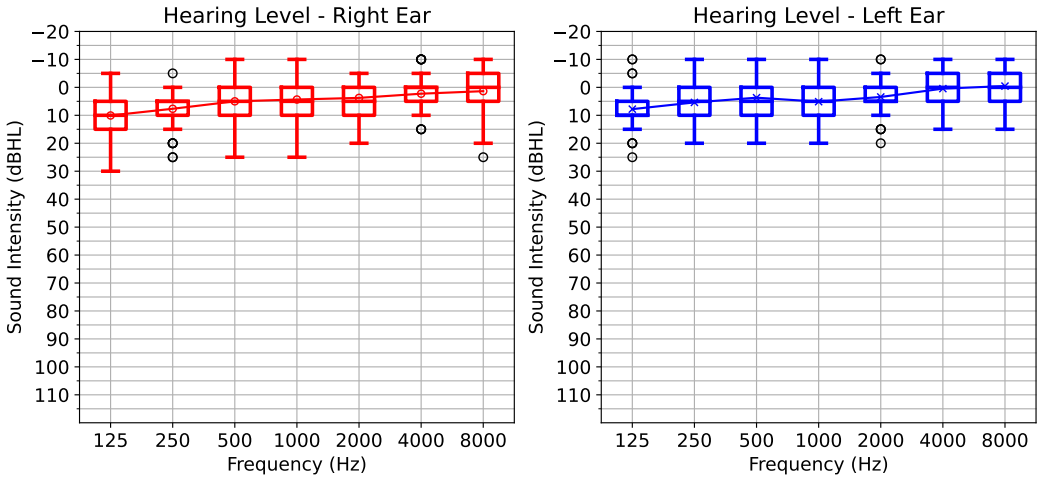

Figure 2. Air conduction thresholds (in dB hearing level (HL)) of the participants.

Secondly, we used the Flemish matrix test [37] to determine each participant's speech reception threshold (SRT, the signal-to-noise ratio (SNR) at which 50% speech understand-

ing is achieved). The test consisted of 3 lists (2 for training, 1 for evaluation) of 20 sentences following the adaptive procedure of Brand et al. [38]. Each sentence had a fixed syntactic structure of 5 words: name, verb, numeral, color, and object [e.g., "Lucas telt vijf gele sokken" ("Lucas counts five yellow socks")]. After each sentence, participants were asked to indicate the heard sentence using a 5×11 matrix containing ten possibilities for each word and a blank option. The order of the three lists was randomized across participants. The last SNR value was used as an estimate of the SRT. The lists were presented to the participants using electromagnetically shielded Ethymotic ER-3A insert phones, binaurally at 62 dBA for each ear. Luts et al. [37] presented the list to the participants monoaurally to the best ear and obtained an average SRT of -8.7 dB SNR when using the results of the third list of the adaptive procedure. During the first repetitions, they reported a significant training effect, which disappeared starting from the third repetition. In our setup, binaural stimulation was chosen to be close to our EEG data acquisition setup. Figure 3 shows the histogram of the obtained SRT over participants in our study. Participants scored an average value of -8.9 dB $\pm$ 0.6 (sd), similar to results obtained by Luts et al. [37].

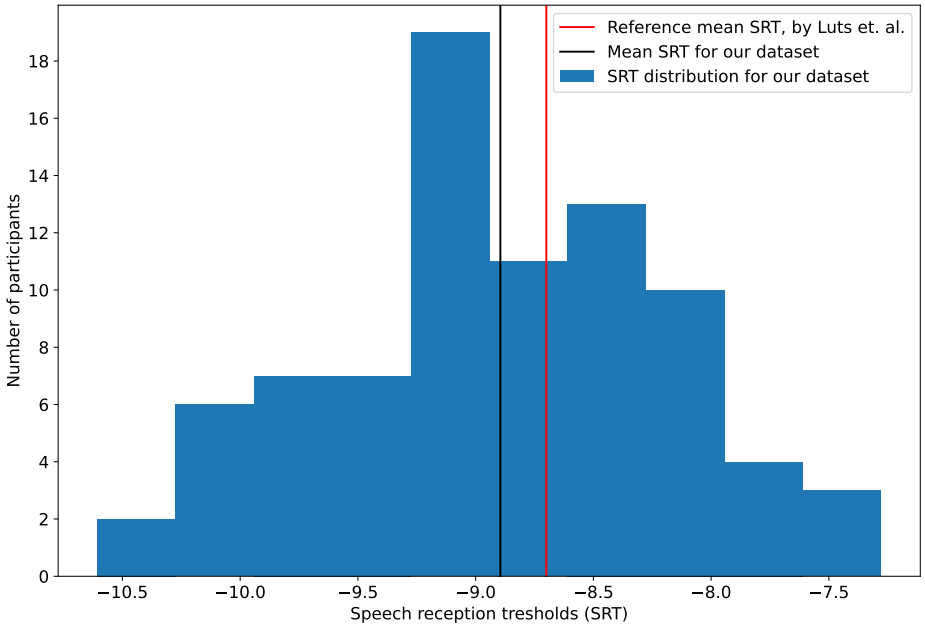

Figure 3. Histogram of the speech reception threshold (SRT), as determined by the MATRIX test [37].

2.2.2. EEG

All participants listened to 6, 7, 8, or 10 trials, each of approximately 15 min. The order of all the trials was randomized per participant. After each trial, a question about the stimulus content was asked to determine attention to and comprehension of all audio stimuli. As the questions were not calibrated, they merely motivated the participant to pay attention to the stimulus, and no further analysis of the answers is provided. After three trials, the participants were asked if they wanted to have a short break. Table 2 shows an overview of the experiment and timing.

We used different categories of stimuli:

- **Reference audiobook** to which all participants listened, made for children and narrated by a male speaker. The length of the audiobook was around 15 min.
- **Audiobooks** made for children or adults. To keep the trial length around 15 min, some audiobooks were split into different partswhen the length exceeded 15 min.
- **Audiobooks with noise** made for children, to which speech-weighted noise was added, as explained below, to obtain an SNR of 5 dB.

- **Podcasts** from the series "Universiteit van Vlaanderen" (University of Flanders) [39]. Each episode of this podcast answers a scientific question, lasts around 15 min, and is narrated by a single speaker.
- **Podcasts with video** from the series "Universiteit van Vlaanderen" (University of Flanders) [39], while video material of the speaker was shown. The video material can be found on the website of Universiteit van Vlaanderen for each podcast separately. The video contains parts where the face of the speaker is visible.

The dynamic range of the podcasts and podcasts with video was compressed by the producers of the stimuli, while that of the audiobooks was not.

The dataset collection consisted of two main session types, ses-shortstories01 and ses-varyingstories, differing in the presented stimuli. Each participant undertook one session. An overview of the experiment and timing can be found in Table 2, while Figure 4 summarizes which stimuli were used for each participant in each session.

Table 2. Overview of the experimental procedure.

Experimental Procedure	Required Time (min)	Cumulative Time (min)
Fill in informed consent	5	5
Fill in questionnaire	5	10
Pure tone audiometry	15	25
Speech audiometry (matrix test)	25	50
Fit EEG equipment	15	65
Listen to 3 stimuli	50	115
First break	5	120
Listen to 3 stimuli	50	170
Second break	5	175
Krios scan of EEG electrode positions	10	185
Listen to 3 stimuli	50	245

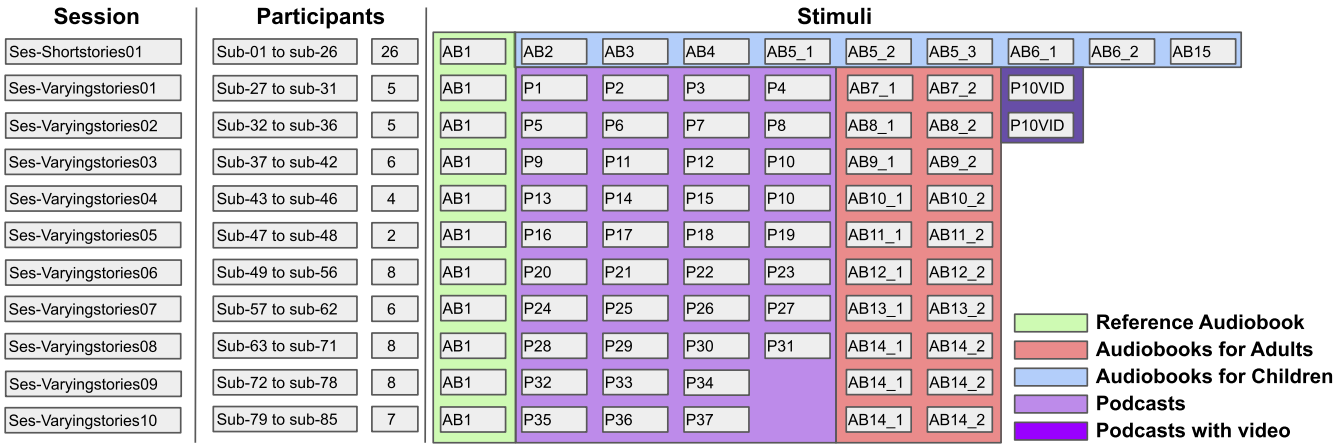

Figure 4. Overview of all the stimuli that were presented, per participant. AB = audiobook, P = podcast. Audiobooks and podcast are numbered. The subscripts _1/2/3 indicate different parts of the same audiobook, each around 15 min in length.

2.2.3. Ses-Shortstories01

For this session type, data from 26 participants are available. It includes ten different parts of audiobooks for children. Two audiobooks, audiobook_1 and audiobook_4, were narrated by male speakers, the other by female speakers. audiobook_3, audiobook_5, and audiobook_6 were narrated by the same speaker. Two out of ten trials were randomly chosen for each participant and presented in speech-weighted noise (SNR = 5 dB). Additionally, 3 subjects listened to a different version of audiobook_1. For that experiment,

the audiobook was cut in 2 halves (audiobook_1_1, audiobook_1_2, respectively), and a pitch-shifted version was created and used for each half (audiobook_1_1_shifted, audiobook_1_2_shifted, respectively). More information about the pitch shifting and additional experiments can be found in the work of Algoet et al. [40]. Finally, there was one control condition in which the first 5 min of audiobook_1 was presented to a subject who had no insert phones inserted (audiobook_1_artefact).

2.2.4. Ses-Varyingstories

For the ses-varyingstories type, data from 59 participants are available. Ses-varyingstories had a fixed reference audiobook_1 (which was presented to all subjects), an audiobook of around 30 min split into two parts, and three to five different podcasts per participant, chosen to keep an even distribution of the sex of the speaker. The stimuli were changed every 2 to 8 participants.

2.3. Data Acquisition

2.3.1. EEG

All recording sessions were conducted at the research group ExpORL of KU Leuven, in a triple-walled, soundproof booth equipped with a Faraday cage to reduce external electromagnetic interference. Participants were instructed to listen to the speech while seated and minimize muscle movements. They were seated in a comfortable chair in the middle of the booth.

We recorded EEG using a BioSemi ActiveTwo system with 64 active Ag-AgCl electrodes and two additional electrodes for the common electrode (CMS) and current return path (DRL). In addition, two mastoid electrodes and the BioSemi head caps were used, containing electrode holders placed according to the 10-20 electrode system.

To ensure proper electrode placement for each participant, we first measured their head size (from nasion to inion to nasion) and selected an appropriate cap. Mastoid locations were scrubbed with Nuprep and cleaned with alcohol gel. The mastoid electrodes were then attached using stickers and held with tape.

The electrode cap was placed on the participant's head from back to front, with ears placed through gaps in the cap. The closing tape at the bottom was secured, and a visual assessment was performed to ensure proper fit. The cap was adjusted so that the distances between the nasion and the electrode Cz and the inion and the electrode Cz were equal, as well as the distance between the left and right ears and the Cz electrode. Electrode gel was applied to the cap holes, and the electrodes were placed gently. The battery, electrode cables, and mastoid electrodes were attached to the EEG amplifier. The participant was then instructed to sit still while EEG was recorded. The subjects were told to keep their eyes open during the measurement. If necessary, additional gel was applied to poorly behaving electrodes, and the electrode offset was checked to ensure proper connection. All offsets were ideally between +20 and −20 mV.

The EEG recordings were digitized at a sampling rate of 8192 Hz and stored on a hard disk using the BioSemi ActiView software, version 7.07 Before digitization, the activeTwo system applies a 5th-order cascaded integrator-comb (CIC) digital filter with a cut-off frequency of 1600 Hz

2.3.2. Digitizer

We acquired a 3D scan of the configuration of the EEG caps for all participants, using a Polaris Krios scanner (NDI; Waterloo ON, Canada), which scans all the electrodes, using a probe to mark three reference points: at the nasion and the height of the tragus at both sides. The Polaris Krios scanner is based on optical measurement technology and uses light reflected by markers to determine the position coordinates.

2.4. Stimulus Preparation

All stimuli were stored at a sampling rate of 48 kHz. For each stimulus file, a trigger file was generated. These triggers were sent from the stimulation equipment (Fireface UC, RME-audio; Haimhausen, Germany) to the BioSemi system. Triggers were generated every second in the form of a block wave. At every second and the beginning and end of the recording, a block pulse with a width of 1 ms was inserted. Based on the stimulus, speech-shaped noise was created at the same root-mean-square value (RMS) as the stimulus. The noise was created by taking white noise and changing the spectrum of the white noise to the spectrum of the speech and then matching the RMS value of the original stimulus file.

Afterward, using one noise file for each RMS value, the stimuli were calibrated with a type 2260 sound-level pressure meter, a type 4189 0.5 in. microphone, and a 2 cm^3 coupler (Bruel & Kjaer, Copenhagen, Denmark).

The auditory stimuli were presented using a laptop connected to a Fireface UC (RME-audio; Haimhausen, Germany) using the APEX software platform, version 4.0.1 [41] and electromagnetically shielded insert phones (ER-3A, Etymotic; Fort Worth, TX, USA), binaurally at 62 dBA for each ear.

2.5. Preprocessed Data

Besides the raw EEG recordings, we also provide EEG and speech stimuli with commonly used preprocessing steps applied. All steps were conducted in Python 3.6, and the code for preprocessing is available on our GitHub repository (https://github.com/exporl/auditory-eeg-dataset (accessed on 1 June 2024)).

2.5.1. EEG

First, EEG data were 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, the EEG was downsampled from 8192 Hz to 1024 Hz and eye-blink artifact removal was applied to the EEG, using a multichannel Wiener filter [42]. Afterward, the EEG was re-referenced to a common average, and finally, the EEG was downsampled to 64 Hz. The bandwidth of the resulting signal was therefore 0.5–32 Hz, as no low-pass filtering was performed (excluding the anti-aliasing filtering as performed by `scipy.signal.resample_poly`).

2.5.2. Speech Stimuli

The initial sampling frequency of the stimuli was 48 kHz. We provide a script to calculate the envelope using a gammatone filterbank [43] with 28 subbands. Each subband envelope was calculated by taking the absolute value of each sample, raised to the power of 0.6. A single envelope was obtained by averaging all these subbands [44]. Then, the envelope was downsampled to 64 Hz.

2.6. Validation

In order to demonstrate the validity of the data, we conducted several experiments on the preprocessed version of the proposed dataset, as explained in more detail in this section. All our results are reproducible using the code on our GitHub: https://github.com/exporl/auditory-eeg-dataset (accessed on 1 June 2024).

For all our experiments, we split each trial into a training, validation, and test set, containing, respectively, 80%, 10%, and 10% of each trial for each participant. The training, validation, and test sets did not overlap, so the test set remained unseen for all the models.

Before usage, we normalized each trial by computing the mean and standard deviation for each of the 64 EEG channels and the envelope stimulus on the training set. We then normalized the training, validation, and test sets by subtracting from each trial the mean and dividing by the standard deviation computed on the training set.

2.6.1. Linear Forward/Backward Modeling

To show the validity of the data, we trained participant-specific linear forward and backward models [1,2] (i.e., models that predict EEG from the stimulus envelope and the stimulus envelope from the EEG, respectively). The backward model was used to detect neural tracking in each recording, i.e., that the speech envelope could effectively be decoded for each participant/story compared to a null distribution of random predictions. The forward model was used to visualize the EEG channels for which the stimulus-related activity could be best predicted.

Model Training

The models were trained based on the recommendations of Crosse et al. (2021) [31]. The backward model weights were obtained similarly by Equation (1):

$$w_b = (R^T R + \lambda I)^{-1} R^T s \quad (1)$$

where R is a matrix consisting of time-lagged versions of the EEG, s is the stimulus envelope, and λ is the ridge regression parameter. In a similar fashion, Equation (2) was used to obtain the forward model weights:

$$w_f = (S^T S + \lambda I)^{-1} S^T r \quad (2)$$

where S is a matrix consisting of time-lagged versions of the stimulus envelope, r is a matrix containing the EEG response, and λ is the ridge regression parameter.

Both models utilized time lags from -100 ms to 400 ms. Following the recommendations of Crosse et al. (2021) [31], a leave-one-out cross-validation was performed on the recordings in the training set to determine the optimal ridge regression parameter (λ) from a list of values (10^x for $x = [-6, -4, -2, 0, 2, 4, 6]$). Correlation scores were averaged across folds and channels, after which the λ was chosen, corresponding to the highest correlation value.

To evaluate the performance of both models, the Pearson correlation between the predicted and true data was calculated on the test set. In order to detect neural tracking, we followed the procedure of Crosse et al. (2021) [31]. For each recording in the test set, the predictions were (circularly) shifted in time by a random amount $N = 100$ times. By correlating these shifted predictions to the actual signal, a null distribution was constructed for each participant. The 95th percentile of this null distribution was compared to the mean of the obtained scores on the test sets.

The analysis of EEG neural responses is typically performed in specific filter bands. For auditory EEG, the research typically focuses on the delta band (0.5–4 Hz) and the Theta band (4–8 Hz) [2,45,46]. We investigated the effect of filtering the EEG and envelope in different bands: delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–14 Hz), Beta (14–30 Hz), and Broadband (0.5–32 Hz). A 1st-order Butterworth filter was chosen for each of the proposed filtering bands.

The model training and evaluation were performed in Python using Numpy [47] and Scipy.

Analysis

Using the linear backward model, we were able to detect neural tracking for all participants. In 11 of the 666 recordings, we were not able to detect neural tracking in any frequency band with the linear decoder. These recordings are listed in Table 3. The results per frequency band are shown in Figure 5. As previously shown by Vanthornhout et al. [2], the optimal performance was reached when filtering in the delta band (0.5–4 Hz). While correlations are hard to compare between studies because they are heavily influenced by the measurement paradigm, subject selection, preprocessing, and modeling choices, the correlations we found for the delta band were roughly in line with previous studies (median correlation between 0.1 and 0.2 [1,2]).

Table 3. Recordings where no significant tracking was found with the linear backward model.

Subject	Stimulus
sub-002	audiobook_1_artefact
sub-011	audiobook_6_1
sub-051	audiobook_12_1
sub-051	audiobook_12_2
sub-051	podcast_23
sub-054	audiobook_12_2
sub-056	podcast_22
sub-060	podcast_24
sub-064	audiobook_14_2
sub-064	podcast_30
sub-076	audiobook_14_1

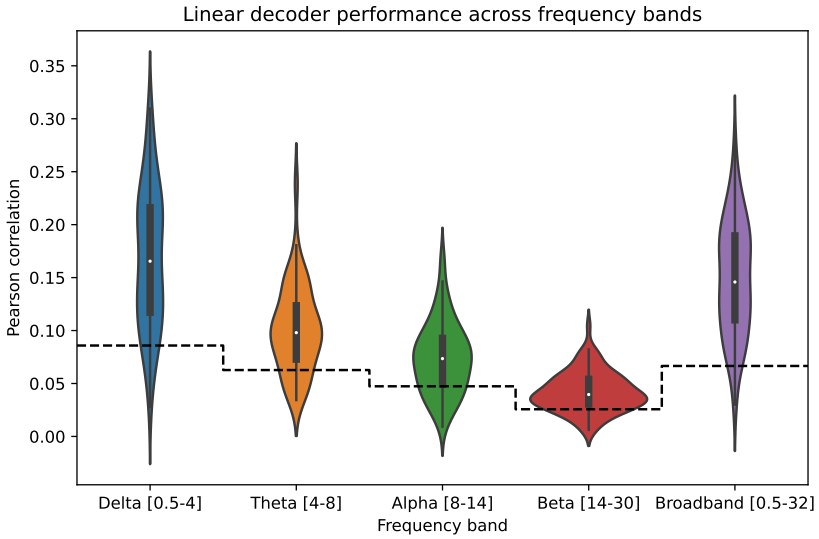

Figure 5. Results of the linear backward model for different frequency bands. Each point in the boxplot is the correlation between the predicted speech envelope and stimulus envelope for one participant, averaged over recordings. Separate models were trained for each participant and frequency band (delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–14 Hz), Beta (14–30 Hz), and Broadband (0.5–32 Hz)). Highest correlations were obtained in the delta band and decreased when going to higher frequency bands. The dashed line represents the significance level ($\alpha = 0.05$).

We compared the linear backward model performance across all stimuli and stimulus types (audiobooks vs. podcast, excluding the audiobook_1 shifted and artifact versions) in the delta band. The results are visualized in Figures 6 and 7, respectively. Note that there was a large variability in decoding scores within and between stimuli. Additionally, a significant difference was found between the audiobook and podcast stimuli (0.184 vs. 0.133 median Pearson correlation, Mann-Whitney U test: $p < 10^{-9}$). A similar observation was made for all the proposed architectures of the 2023 Auditory EEG decoding challenge [14], where the authors compared the performance of the models on stimulus type (audiobook vs. podcast) and observed that models performed significantly better on audiobooks than podcasts. A possible explanation for this might be that the podcasts are recorded using compression techniques, which generate an evenly loud signal. It might also be that the models fail to generalize to different compression techniques rather than to different unseen subjects.

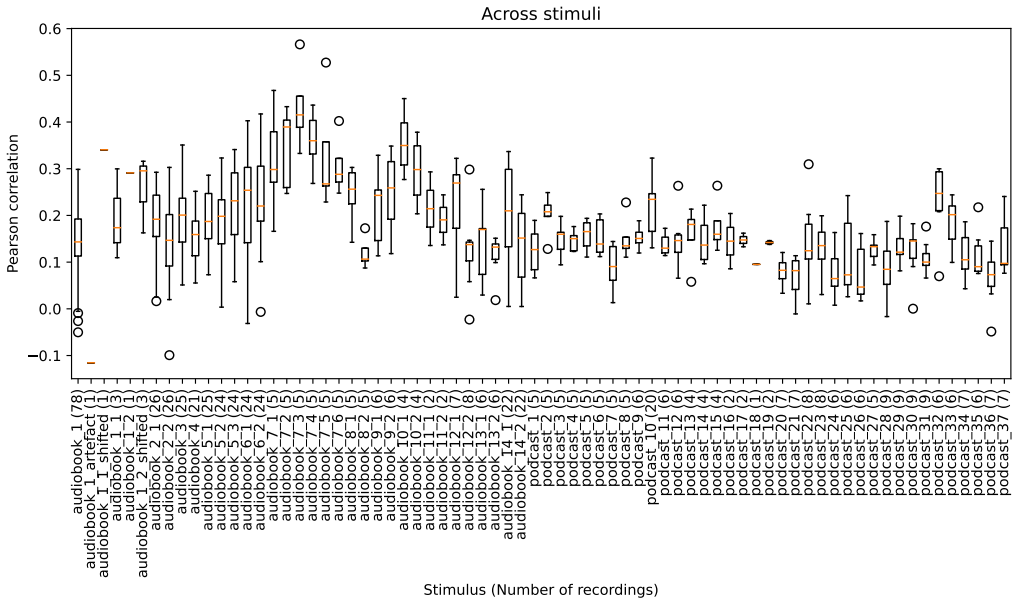

Figure 6. Results of the linear backward model for the different stimuli in the dataset. One model was trained per participant. Each point in the boxplot is the correlation between the predicted speech envelope and stimulus envelope for one recording. Data were filtered in the delta band (0.5–4 Hz). There is high variability across participants and stimuli.

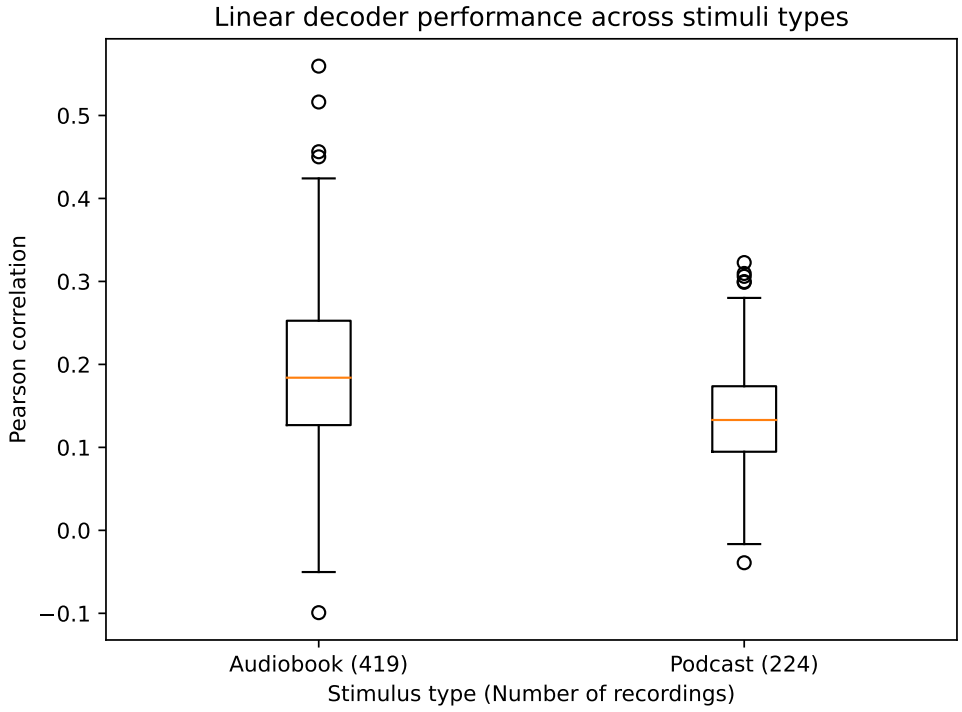

Figure 7. Results of the linear backward model for the different stimuli in the dataset. One model was trained per participant. Each point in the boxplot is the correlation between the predicted speech envelope and stimulus envelope for one participant, averaged across recordings. Significantly higher correlations were obtained for the audiobooks (0.184 vs. 0.133 median Pearson correlation, Mann-Whitney U test: $p < 10^{-9}$).

Additionally, the backward model performance for the podcast with and without video is compared for the delta band in Figure 8. In that case, a decoder was trained separately for each recording on the training set with cross-validation and evaluated on the

test set. No significant difference was found between the podcast presented with or without video. This is in contrast to previous studies that saw better neural tracking when the face of the speaker was presented [48,49]. A possible explanation for not finding an effect is that the face of the speaker was not continually present during the whole recording, removing the added benefit for at least some portion of the recording. Alternatively, the sample size contained within SparrKULee might be too small to find an effect across participants.

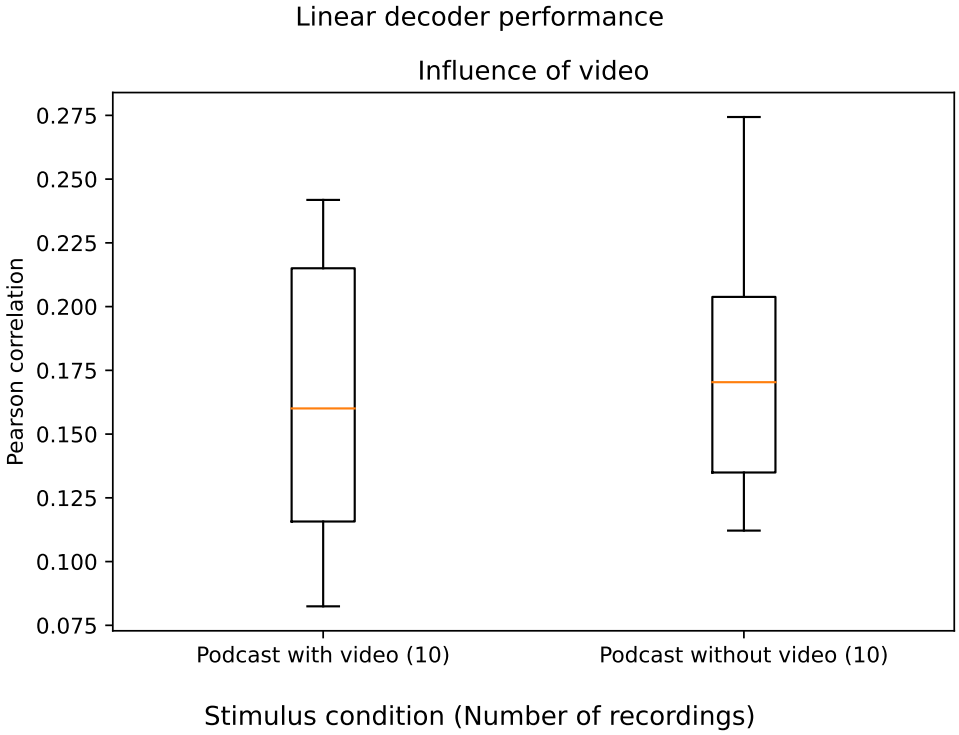

Figure 8. Results of the linear backward model for the (same) podcast with and without video. One model was trained per recording. Each point in the boxplot is the correlation between the predicted speech envelope and stimulus envelope for one recording. No significant difference was found (Mann-Whitney U test: $p = 0.73$).

For the forward model, we show topomaps averaged across participants for each frequency band and stimulus type in Figure 9. As with the backward model, we observed the highest correlations between predicted and actual EEG signals in the delta band. The highest correlations were obtained for the channels in the temporal and occipital regions.

2.6.2. Non-Linear Models: Match-Mismatch Paradigm

For the non-linear models, we used the match-mismatch paradigm [7,12]. In this paradigm, the models were given three inputs: a segment of the EEG recording, the time-matched stimulus envelope segment, and a mismatched (imposter) stimulus envelope segment. As specified by Monesi et al. [10], the imposter was taken 1 s after the matched stimulus envelope segment. If extracting an imposter (at the end of each set) was impossible, the segment was discarded from the dataset. We extracted overlapping windows with 80% overlap. We included an analysis using a dilated convolutional model [12] to show typical match-mismatch performance across different input segment lengths.

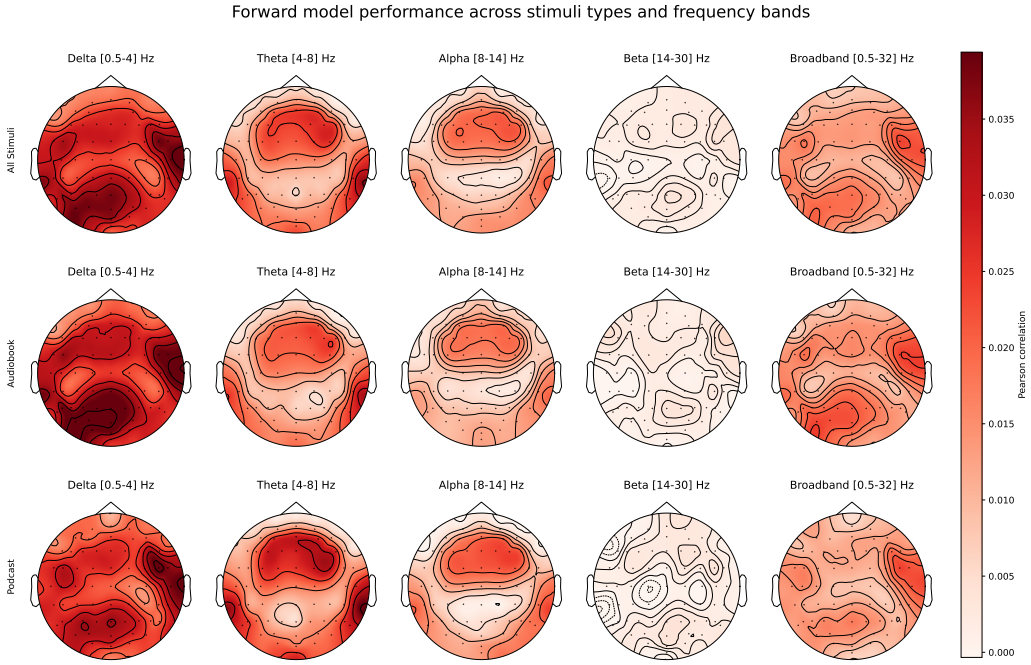

Figure 9. Results of the forward linear model for different stimulus types and frequency bands. For each channel, the correlation between actual and predicted EEG is shown and averaged across participants. One model was trained per participant. The highest correlations were obtained in the delta band for the channels in the temporal and occipital regions.

Model Training

The dilated convolutional network consisted of four steps. First, the EEG channels were combined, from 64 to 8, using a 1D convolutional layer with a kernel size of 1 and a filter size of 8. Second, there were N dilated convolutional layers with a kernel size of K and 16 filters. These N convolutional layers were applied to both EEG and envelope stimulus segments. After each convolutional layer, a rectified linear unit (ReLU) was applied. Both stimulus envelope segments shared the weights for the convolutional layers. After these non-linear transformations, the EEG was compared to both stimulus envelopes, using cosine similarity. Finally, the similarity scores were fed to a single neuron, with a sigmoid non-linearity, to create a prediction of the matching stimulus segment.

The model was implemented in Tensorflow and used the Adam optimizer, with a learning rate of 0.001 and binary-cross entropy as the loss function. Models were trained for a maximum of 50 epochs, using early stopping based on the validation loss, with a patience factor of 5. We trained the models with an input segment length of 5 s and in a participant-independent way, i.e., all participant data were given simultaneously to the model. We report results for input testing lengths of 1, 2, 3, 5, and 10 s. Since the trained dilation model did not have fixed input lengths, we used the same model with different input lengths.

Analysis

The results of this analysis can be seen in Figure 10. The accuracy of the model increased with longer window lengths. We see the same trend as in [12]. In order to test the generalizability of the model, we also tested the model with an arbitrarily chosen mismatch segment, as opposed to the fixed 1 s. There was no significant difference between these two testing conditions, which is in line with the experiment conducted in [50].

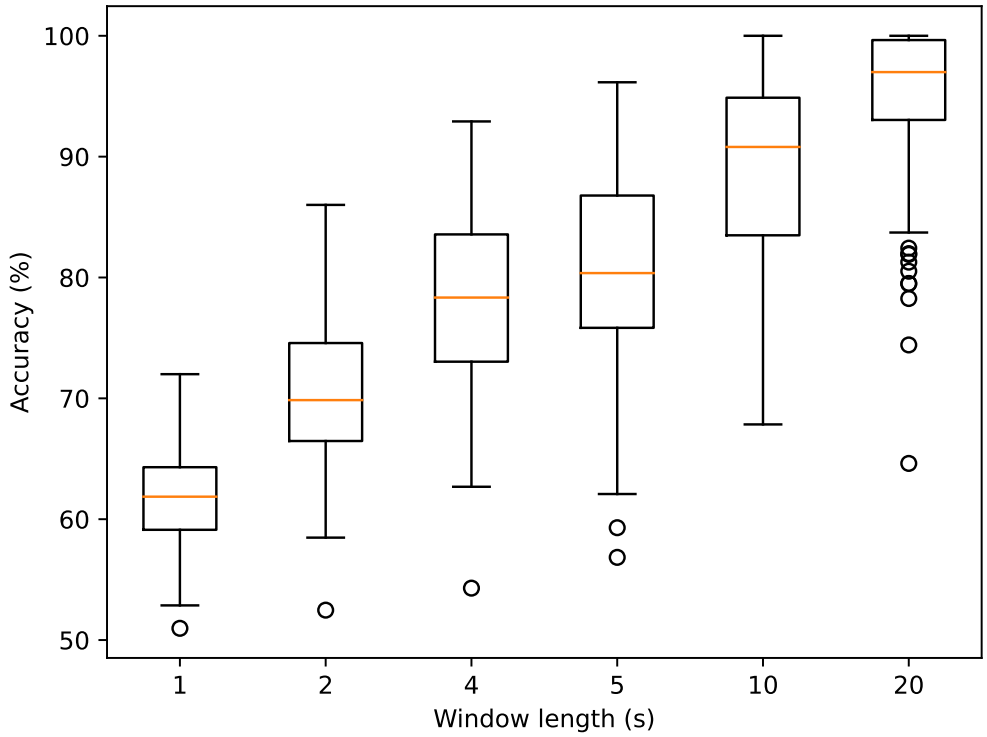

Figure 10. Results of the non-linear dilation model, in the match-mismatch paradigm. Each point in the boxplot is the match-mismatch accuracy for one participant, averaged across recordings. The imposter envelope segment starts one second after the end of the true segment. One model was trained across all participants.

3. Data Description

All data were organized according to EEG-BIDS [51], an extension to the Brain Imaging Directory Structure (BIDS) [52] for EEG data. EEG-BIDS allows the storage of EEG data with relevant extra information files, e.g., about the experiment, the stimuli, and the triggers, enabling quick usage of the data and linking the auditory stimuli to the raw EEG files. A schematic overview of our repository is shown in Figure 11. The dataset consists of 3 parts: (1) raw data, in a folder per participant, (2) the auditory stimuli, in zipped Numpy (.npz) [47] format, and (3) the preprocessed data records, as described above, in the derivatives folder.

3.1. Raw Data

The raw data were structured as one folder per participant. For each participant (1 to 85), a folder `sub-xxx` is available in the root folder. In this folder, there is a folder indicating the session, which can be either `ses-shortstories01` or `ses-varyingstoriesxx` ($xx = 01 \ldots 09$).

Each session folder contains a subfolder `beh`, containing the results of the behavioral matrix SRT estimation. These files were named according to the participant, the session, the task, which is always listeningActive, and the behavioral experiment run, which goes from one to three.

The data of the EEG experiment were stored in a subfolder in the session folder, named `eeg`. The EEG experiment data were named according to the participant, the session, the task, and the run. When the participant listened to a stimulus, the task was `listeningActive`. When the participant listened to silence, which happened at the start and end of the experiment, the task was `restingState`. The run suffix chronologically numbers the different trials starting at 01. Each trial has four corresponding files, differing only in their ending, after the run suffix: (1) raw gzipped file of EEG data in BioSemi Data Format (BDF), sampled at 8192 Hz, ending in `eeg.bdf.gz`; (2) a descriptive apr file

`eeg.apr`, containing extra information about the experiment, such as the answers to the questions that were asked; (3) a stimulation file to link EEG to the corresponding stimulus `stimulation.tsv`; and (4) `events.tsv`, which describes which stimuli were presented to the participants at which time.

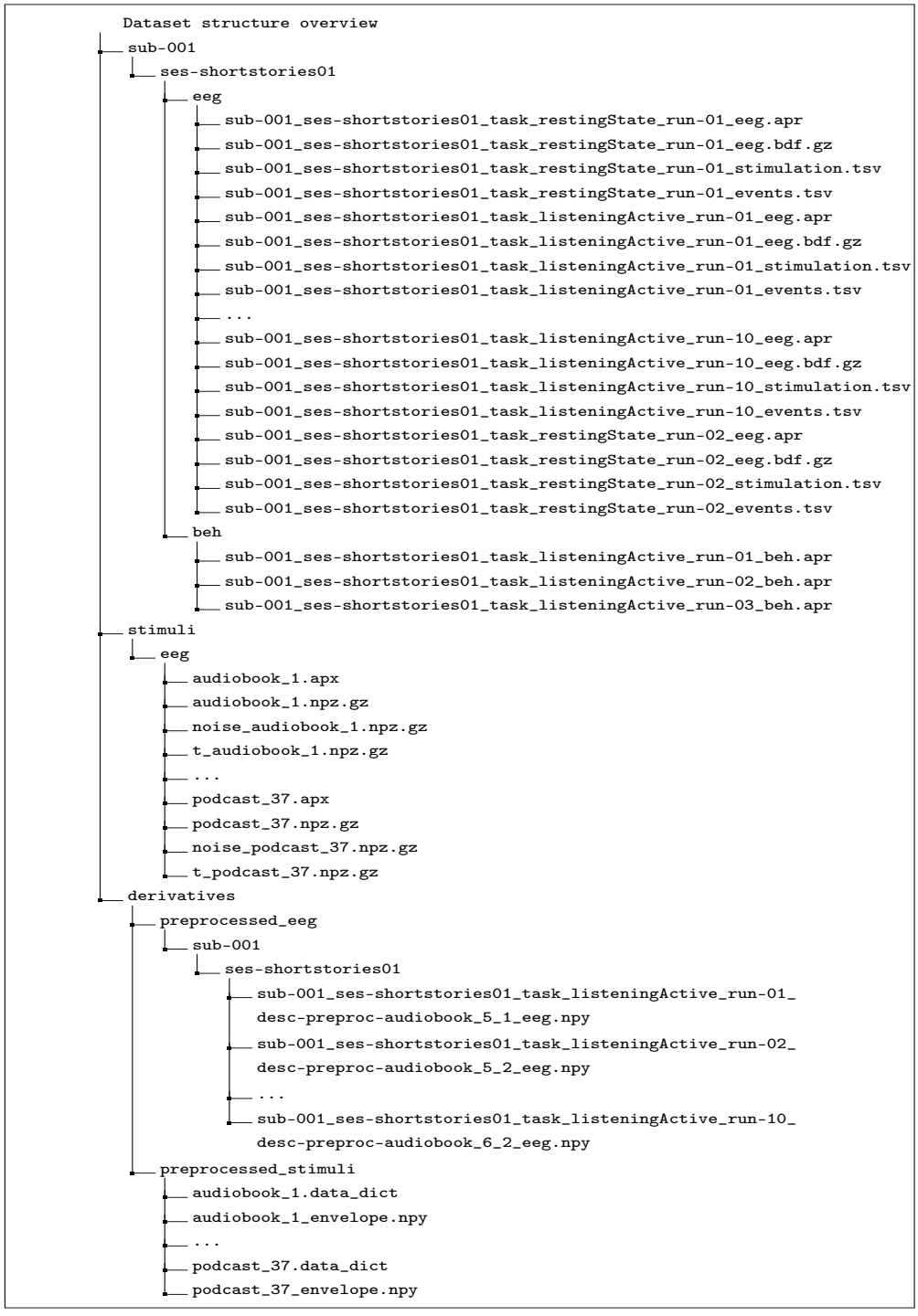

Figure 11. Tree depicting the structure of our dataset. All data are structured according to the EEG-BIDS standard.

3.2. Stimuli

All the stimuli were saved in the folder `stimuli/eeg`. For each stimulus, we provide four corresponding files, stored in the npz format with additional gzipping to reduce storage, which is easily readable in Python: (1) the stimulus, stored at 48 kHz

`stimulusName.npz.gz`, (2) the associated noise file `noise_stimulusName.npz.gz`, (3) the associated trigger file `t_stimulueName.npz.gz`, and (4) the experiment description file `stimulusName.apx`.

The stimuli were named according to their type: either `audiobook_xx` or `podcast_xx`, where xx indexes unique stimuli. Whenever an audiobook was split into multiple consecutive parts, an extra suffix denotes which part of the audiobook is referred to.

3.3. Preprocessed Data

For all data, we also provide a preprocessed, downsampled, and time-aligned version of the data. These data can be found in the `derivatives/preprocessed_eeg` folder. Similar to the raw data, the preprocessed data were structured in a folder per participant, per session, which could be either ses-shortstoriesxx or ses-varyingstoriesxx. The preprocessed files derived their name from the raw EEG file used to create the preprocessed version. To avoid confusion, a suffix `desc-preproc` was added, such that no two files have the same name. After the `desc-preproc` suffix, the name of the stimulus the participant listened to was added to facilitate linking the EEG brain response to the auditory stimulus for downstream tasks.

Author Contributions: B.A., W.V., L.B., H.V.h. and T.F. conceived the experiments; B.A., L.B., M.G. and W.V. conducted/supervised the experiments; B.A. and L.B. analyzed the results. All authors reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

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

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of KU/UZ Leuven (protocol code S57102 and date of approval 8 October 2014).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Stimulus data originate from the Radioboeken project of deBuren (https://soundcloud.com/deburen-eu/ (accessed on 1 June 2024). AB_1, AB_3, AB_{xp1} and AB_{xp2} for $x = 7 \ldots 14$). Other stimulus data were obtained from the Universiteit van Vlaanderen (https://www.universiteitvanvlaanderen.be) (accessed on 20 October 2022). All stimuli in the dataset can only be used/shared for non-commercial purposes. When republishing (adaptations of) the stimuli, explicit permission should be acquired from the original publishing organization(s) (i.e., deBuren or Universiteit van Vlaanderen). EEG data are shared under an Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). Per the request of 11 participants, access to their data is restricted (107 EEG recordings in total). Readers can request access by mailing the corresponding authors at sparrkulee@kuleuven.be, stating what they want to use the data for. Access will be granted to non-commercial users, complying with the CC-BY-NC-4.0 license. The dataset is available on the RDR KU Leuven platform (https://rdr.kuleuven.be/dataset.xhtml?persistentId=doi:10.48804/K3VSND (accessed on 1 June 2024)), or alternatively at https://homes.esat.kuleuven.be/~lbollens/ (accessed on 1 June 2024). All code used for the technical validation can be found online: https://github.com/exporl/auditory-eeg-dataset (accessed on 1 June 2024). We used the mne-python library [53]. For using the data, we recommend using the code on our GitHub repository to get started, which consists of two main parts: (1) code to create the preprocessed EEG and preprocessed stimuli from the raw data and (2) code to perform the experiments as discussed in the technical validation. The README file contains detailed technical instructions.

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

Conflicts of Interest: The authors declare no conflicts of interest.

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. *Front. Hum. Neurosci.* **2016**, *10*, 604. [CrossRef] [PubMed]
2. Vanthornhout, J.; Decruy, L.; Wouters, J.; Simon, J.Z.; Francart, T. Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. *JARO-J. Assoc. Res. Otolaryngol.* **2018**, *19*, 181–191. [CrossRef] [PubMed]
3. Iotzov, I.; Parra, L.C. EEG can predict speech intelligibility. *J. Neural Eng.* **2019**, *16*, 036008. [CrossRef] [PubMed]
4. Thornton, M.; Mandic, D.; Reichenbach, T. Robust decoding of the speech envelope from EEG recordings through deep neural networks. *J. Neural Eng.* **2022**, *19*, 046007. [CrossRef]
5. Accou, B.; Vanthornhout, J.; Hamme, H.V.; Francart, T. Decoding of the speech envelope from EEG using the VLAAI deep neural network. *Sci. Rep.* **2023**, *13*, 812. [CrossRef] [PubMed]
6. Lesenfants, D.; Vanthornhout, J.; Verschueren, E.; Francart, T. Data-driven spatial filtering for improved measurement of cortical tracking of multiple representations of speech. *J. Neural Eng.* **2019**, *16*, 066017. [CrossRef] [PubMed]
7. de Cheveigné, A.; Di Liberto, G.M.; Arzounian, D.; Wong, D.D.; Hjortkjær, J.; Fuglsang, S.; Parra, L.C. Multiway canonical correlation analysis of brain data. *NeuroImage* **2019**, *186*, 728–740. [CrossRef] [PubMed]
8. de Cheveigné, A.; Slaney, M.; Fuglsang, S.A.; Hjortkjaer, J. Auditory stimulus-response modeling with a match-mismatch task. *J. Neural Eng.* **2021**, *18*, 046040. [CrossRef] [PubMed]
9. de Taillez, T.; Kollmeier, B.; Meyer, B. Machine learning for decoding listeners' attention from EEG evoked by continuous speech. *Eur. J. Neurosci.* **2017**, *51*, 1234–1241. [CrossRef] [PubMed]
10. Monesi, M.J.; Accou, B.; Montoya-Martinez, J.; Francart, T.; Van hamme, H. An LSTM Based Architecture to Relate Speech Stimulus to Eeg. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing–Proceedings, Barcelona, Spain, 4–8 May 2020; pp. 941–945. [CrossRef]
11. Jalilpour Monesi, M.; Accou, B.; Francart, T.; Van hamme, H. Extracting different levels of speech information from EEG using an LSTM-based model. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; International Speech Communication Association: Grenoble, France, 2021; pp. 526–530.
12. Accou, B.; Monesi, M.J.; Van hamme, H.; Francart, T. Predicting speech intelligibility from EEG in a non-linear classification paradigm. *J. Neural Eng.* **2021**, *18*, 066008. [CrossRef]
13. Ding, N.; Simon, J.Z. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 11854–11859. [CrossRef] [PubMed]
14. Monesi, M.J.; Bollens, L.; Accou, B.; Vanthornhout, J.; Van Hamme, H.; Francart, T. Auditory EEG decoding challenge for ICASSP 2023. *IEEE Open J. Signal Process.* **2024**, *5*, 652–661. [CrossRef]
15. Bollens, L.; Monesi, M.J.; Accou, B.; Vanthornhout, J.; Van Hamme, H.; Francart, T. ICASSP 2023 Auditory EEG Decoding Challenge. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–2.
16. Auditory EEG Challenge–ICASSP 2024. Available online: https://exporl.github.io/auditory-eeg-challenge-2024/ (accessed on 1 June 2024)
17. Yang, L.; Van Dyck, B.; Van Hulle, M.M. Sea-Wave: Speech envelope reconstruction from auditory EEG with an adapted WaveNet. *IEEE Open J. Signal Process.* **2024**, *5*, 686–699. [CrossRef]
18. Thornton, M.; Mandic, D.P.; Reichenbach, T. Decoding Envelope and Frequency-Following EEG Responses to Continuous Speech Using Deep Neural Networks. *IEEE Open J. Signal Process.* **2024**, *5*, 700–716. [CrossRef]
19. 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. *arXiv* **2024**, arXiv:2401.17380.
20. Thornton, M.; Mandic, D.; Reichenbach, T. Relating EEG Recordings to Speech Using Envelope Tracking and the Speech-FFR. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–2. [CrossRef]
21. Borsdorf, M.; Pahuja, S.; Ivucic, G.; Cai, S.; Li, H.; Schultz, T. Multi-Head Attention and GRU for Improved Match-Mismatch Classification of Speech Stimulus and EEG Response. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–2. [CrossRef]
22. Cui, F.; Guo, L.; He, L.; Liu, J.; Pei, E.; Wang, Y.; Jiang, D. Relate Auditory Speech to EEG by Shallow-Deep Attention-Based Network. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–2. [CrossRef]
23. Piao, Z.; Kim, M.; Yoon, H.; Kang, H.G. HappyQuokka System for ICASSP 2023 Auditory EEG Challenge. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–2. [CrossRef]
24. Van Dyck, B.; Yang, L.; Van Hulle, M.M. Decoding Auditory EEG Responses Using an Adapted Wavenet. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–2. [CrossRef]
25. 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. *Curr. Biol.* **2018**, *28*, 803–809. [CrossRef] [PubMed]

26. Fuglsang, S.A.; Wong, D.D.; Hjortkjær, J. EEG and Audio Dataset for Auditory Attention Decoding. 2018. Available online: https://zenodo.org/records/1199011 (accessed on 1 June 2024).
27. 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'. 2022. Available online: https://zenodo.org/records/7086209 (accessed on 1 June 2024).
28. Weissbart, H.; Kandylaki, K.; Reichenbach, T. EEG Dataset for 'Cortical Tracking of Surprisal during Continuous Speech Comprehension'. 2022. Available online: https://zenodo.org/records/7086168 (accessed on 1 June 2024).
29. Brennan, J.R. EEG Datasets for Naturalistic Listening to "Alice in Wonderland". 2018. Available online: https://deepblue.lib.umich.edu/data/concern/data_sets/bn999738r (accessed on 1 June 2024)
30. Vanheusden, F.J.; Kegler, M.; Ireland, K.; Georgia, C.; Simpson, D.; Reichenbach, T.; Bell, S. Dataset for: Hearing Aids Do Not Alter Cortical Entrainment to Speech at Audible Levels in Mild-to-Moderately Hearing-Impaired Subjects. 2019. Available online: https://eprints.soton.ac.uk/438737/ (accessed on 1 June 2024).
31. Crosse, M.; Zuk, N.; Di Liberto, G.; Nidiffer, A.; Molholm, S.; Lalor, E. Linear Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological Considerations for Applied Research. *Front. Neurosci.* **2021**, *15*, 705621. [CrossRef] [PubMed]
32. Coren, S. The lateral preference inventory for measurement of handedness, footedness, eyedness, and earedness—Norms for young-adults. *Bull. Psychon. Soc.* **1993**, *31*, 1–3. [CrossRef]
33. van Lier, H.; Drinkenburg, W.H.; van Eeten, Y.J.; Coenen, A.M. Effects of diazepam and zolpidem on EEG beta frequencies are behavior-specific in rats. *Neuropharmacology* **2004**, *47*, 163–174. [CrossRef]
34. De Vos, A.; Vanvooren, S.; Vanderauwera, J.; Ghesquière, P.; Wouters, J. Atypical neural synchronization to speech envelope modulations in dyslexia. *Brain Lang.* **2017**, *164*, 106–117. [CrossRef]
35. Power, A.J.; Colling, L.J.; Mead, N.; Barnes, L.; Goswami, U. Neural encoding of the speech envelope by children with developmental dyslexia. *Brain Lang.* **2016**, *160*, 1–10. [CrossRef] [PubMed]
36. Hughson, W.; Westlake, H. Manual for program outline for rehabilitation of aural casualties both military and civilian. *Trans. Am. Acad. Ophthalmol. Otolaryngol.* **1944**, *48*, 1–15.
37. Luts, H.; Jansen, S.; Dreschler, W.; Wouters, J. Development and Normative Data for the Flemish/Dutch Matrix Test. 2014. Available online: https://lirias.kuleuven.be/retrieve/293640 (accessed on 1 June 2024)
38. Brand, T.; Kollmeier, B. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *J. Acoust. Soc. Am.* **2002**, *111*, 2801–2810. [CrossRef] [PubMed]
39. Universiteit van Vlaanderen. Available online: https://www.universiteitvanvlaanderen.be/podcast (accessed on 20 October 2022).
40. Algoet, A. Invloed van het Geslacht van de Spreker en Luisteraar en Persoonlijke Appreciatie van het Verhaal op de Neurale Tracking van de Spraakomhullende. 2020. Available online: https://repository.teneo.libis.be/delivery/DeliveryManagerServlet?dps_pid=IE14186261& (accessed on 1 June 2024)
41. Francart, T.; Van Wieringen, A.; Wouters, J. APEX 3: A multi-purpose test platform for auditory psychophysical experiments. *J. Neurosci. Methods* **2008**, *172*, 283–293. [CrossRef] [PubMed]
42. Somers, B.; Francart, T.; Bertrand, A. A generic EEG artifact removal algorithm based on the multi-channel Wiener filter. *J. Neural Eng.* **2018**, *15*, 036007. [CrossRef] [PubMed]
43. Søndergaard, P.; Majdak, P. The Auditory Modeling Toolbox. In *The Technology of Binaural Listening*; Blauert, J., Ed.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 33–56.
44. Biesmans, W.; Das, N.; Francart, T.; Bertrand, A. Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2017**, *25*, 402–412. [CrossRef] [PubMed]
45. Ding, N.; Simon, J.Z. Cortical entrainment to continuous speech: Functional roles and interpretations. *Front. Hum. Neurosci.* **2014**, *8*, 311. [CrossRef] [PubMed]
46. Sharon, R.A.; Narayanan, S.; Sur, M.; Murthy, H.A. An Empirical Study of Speech Processing in the Brain by Analyzing the Temporal Syllable Structure in Speech-input Induced EEG. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019. [CrossRef]
47. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef] [PubMed]
48. Crosse, M.J.; Butler, J.S.; Lalor, E.C. Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *J. Neurosci.* **2015**, *35*, 14195–14204. [CrossRef]
49. Park, H.; Kayser, C.; Thut, G.; Gross, J. Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife* **2016**, *5*, e14521. [CrossRef]
50. Puffay, C.; Accou, B.; Bollens, L.; Monesi, M.J.; Vanthornhout, J.; Van hamme, H.; Francart, T. Relating EEG to continuous speech using deep neural networks: A review. *J. Neural Eng.* **2023**, *20*, 041003. [CrossRef] [PubMed]
51. Pernet, C.R.; Appelhoff, S.; Gorgolewski, K.J.; Flandin, G.; Phillips, C.; Delorme, A.; Oostenveld, R. EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Sci. Data* **2019**, *6*, 103. [CrossRef] [PubMed]

52. Gorgolewski, K.J.; Auer, T.; Calhoun, V.D.; Craddock, R.C.; Das, S.; Duff, E.P.; Flandin, G.; Ghosh, S.S.; Glatard, T.; Halchenko, Y.O.; et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* **2016**, *3*, 160044 . [CrossRef] [PubMed]
53. Gramfort, A.; Luessi, M.; Larson, E.; Engemann, D.A.; Strohmeier, D.; Brodbeck, C.; Goj, R.; Jas, M.; Brooks, T.; Parkkonen, L.; et al. MEG and EEG Data Analysis with MNE-Python. *Front. Neurosci.* **2013**, *7*, 267. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.