ISSN 2063-5346 *Section A-Research paper*

Deep Learning model for Decoding Audio Source Localization through Speech and Neural Sensor Data for Hearing-aid and BCI Applications

Anudeep Peddi[1],
[1]Department of Computer Science and Engineering (Data Science),
RVRJCCE, Guntur, A.P., India.
peddianudeep88@gmail.com
Dr. T. Venkata Ramana[2]
[2]Department of EECE, GIT,
GITAM deemed to be University, Visakhapatnam, Andhra Pradesh, India.
vteppala@gmail.com
DOI: 10.53555/ecb/2023.12.5.3022023.29/05/2023

ABSTRACT
When there is a speech stream with numerous speakers and background noise, the human auditory system can focus on a single speaker of interest and disregard the others. Studies on this scenario have demonstrated that cortical activity follows the speech envelope where the attended speech envelope was stronger than the unattended speech. By connecting speech signals and neural activity, it has been shown that electroencephalography (EEG) signals can be used to determine which speaker a listener is listening to. The human brain is inherently non-linear where deep learning techniques may use to analyze neural signals (EEG data) to decode the changing state of the brain. The non-linear methods proposed in the recent studies have not used any speech streams or have used only envelope of the speech stream as input features. The neural networks performance can be increased by providing information about the speech stream. To extract auditory attention, we introduce a hybrid convolutional neural network and a long short-term memory model (CNN-LSTM). The CNN-LSTM model is employed as an input to decipher attention in a two-speaker situation using cortical recordings and the spectrogram of multiple speech streams. The decoding accuracy of the model in short trial durations are measured. The model CNN-LSTM proposed for audio source localization can be used for developing neuro-steered hearing devices and brain computer interface (BCI) applications.

Keywords:

Brain Computer Interface, Audio source localization, Speech processing, Neural sensor data processing, Convolutional neural network, Bidirectional long-short term memory.

1. INTRODUCTION
Humans have extremely complicated auditory systems and the research in this field is gaining traction in recent times due to the advance in medical equipment. There is a big motivation to understand the response of human brain to audio stimuli as it can lead to progress in the fields of neuroscience, robotics, and brain-computer interfaces. But, till date, not a lot of work has been done in this field due to the lack of access to measuring devices and the absence of standardized datasets.
Cognition is the capability to process information through stimuli that we get from the environment around us. There are different types of cognitive processes and attention is the process that allows us to concentrate on certain activities or stimuli. Attention is used in most of the daily tasks that are to be performed and it controls and regulates the other cognitive processes like perception, thought, language and learning. The focus of this paper is on

auditory attention, more specifically, selective auditory attention. It involves the auditory cortex of the human brain, and it is signified as the action which enables people to pay attention to certain sounds or speech stimuli.

The "cocktail party problem" was coined in the 1950s to describe the situation of a party when several sources of sound are heard at once [1], [2]. Figure 1 shows the pictorial representation for the same. It claimed that people were adept at restricting all other noises and concentrating their attention on a particular source of sound that they were interested in listening to. Familiar characteristics of the speaker like their tone of voice and distance from the speaker helped in filtering out the other sounds. To completely understand the underlying processes in the brain going on, this area has been gaining more recent recognition from researchers.

Every human ear can concentrate on one particular sound even though there exist multiple sounds in the environment or surroundings. It happens for all the living beings in the environment. Basically, every human voice which are present in noisy environment overlap with the frequency and time, which leads to acoustic interference and can impair the clarity of speech. It has been submitted that one's sensory memory subconsciously removes the entire unwanted event that evokes a specific functional reaction in an organ and identifies the pieces of information which is required and transfers it to the human brain. This is the effect in which most of the people able to listen to one particular voice rather in a group of noises.

Figure 1. Cocktail Party Effect

This is a similar phenomenon which occurs when one suddenly detects word which has high importance rather than the unwanted event that evokes a specific functional reaction in an organ. Along with humans the problems occur in animals also, most of the animals such as insect choruses, frog, songbird chorus, colonial and flocking birds also face the similar problem. Animals which communicate in groups, the problem of receiving the signal is similar to the problems faced by the human cocktail party problem. Where humans face the problems of masking and interference but when it comes to animals decrease in the ability to recognize, differentiate among various signal variants, and increase in signal detection thresholds leads to the problem. Most of the work by the researchers are displaying that the cocktail party problem is great to overcome.

Conversing in multiple noise environments and the presence of annoying speakers is a specialty. The Cocktail Party Effect refers to the ability of people who have normal hearing to direct auditory attention to an audio signal in a complex composition. [3], [4]. However, no automated solution has yet been found for cocktail party issues, even after more than half a century of intensive research. This type of solution is in great demand for various users and applications, for example, a man-machine interface such as Amazon Alexa, audio and video automated subtitles recording (YouTube, Netflix, etc.), modern hearing aids, etc., [5].

People with aural problem suffers from diminished speech accessibility while listening to a particular speaker in a setting with several speakers [6]. In such scenarios, public hearing aids in the market are often inadequate because they cannot distinguish between attending and ignored speakers. Therefore, more details regarding the focus point are most desired. The development of visual objects in the eye serves as an illustration of selective attention [7]. In constructing visual objects, the observer focuses on objects in a critical visual scene. This theory has been extended to the auditory domain, implying occurrences like the Cocktail Party Effect. Figure 2 depicts the idea of how estimated source signal can be identified in a cocktail party effect. It can be deduced from the auditory object's development. In other words, when listening closely, the human brain builds things based on different speakers existing in the hearing situation and chooses those items associated with a specific speaker. But at the same time, a theory of flexible attention trajectories was proposed. This assumes that slow selection occurs with low cerebral load and early selection occurs with high cerebral load. This stimulated the investigation of whether brain signals could provide extra information, which helps distinguish between current and disturbing speakers. In experiments with two speakers, it was observed that the brain signal measured at the embedded electrodes traced more pronounced attributes of the current speaker than the neglected speaker. Both EEG and MEG produced comparable outcomes. Auditory Attention Decoding (AAD), sometimes referred to as EEG analysis, has emerged as the most popular technique for studying attention in recent years [8], [9].

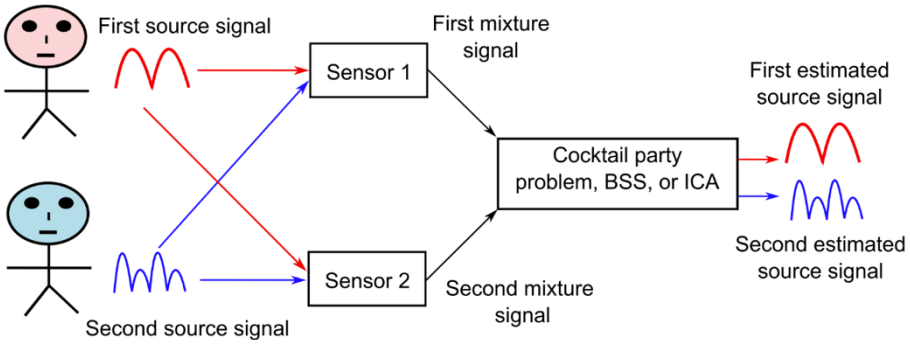

Figure 2. Estimated source signal identification in a cocktail party effect [10]

In a competitive scenario with two speakers, neural activity was shown using EEG or EMG by continuously recording the dynamic changes in the arriving speech envelope at auditory processing. Supervised audio envelopes are usually more pronounced than unsupervised audio envelopes—this neural tracking of stimuli issued to regulate auditory attention [11]. The most common technique is stimulus reconstruction, which makes use of brain activity to interpret and recreate the stimulus envelope after stimulation. The original stimulus envelope and the reconstructed envelope are then correlated, and the envelope with the greatest correlation is used to represent the current speaker [12]. The other methods of decoding attention consist of a forward modelling approach, which is anticipating EEG from auditory stimuli, Canonical Correlation Analysis (CCA) -based method, and Bayesian state-space modelling.

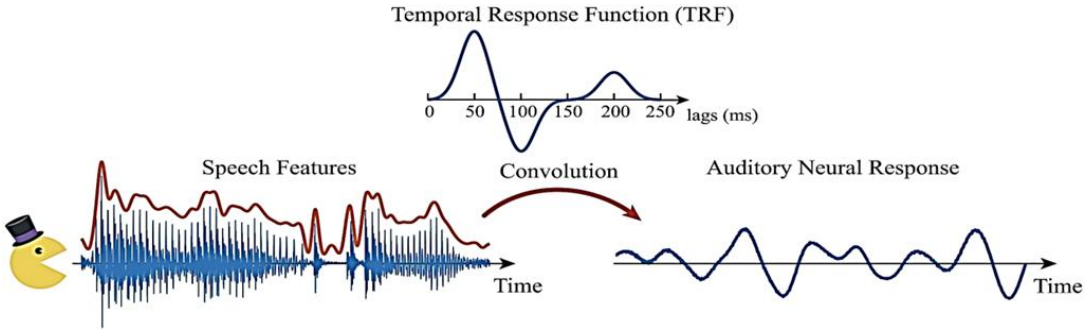

Figure 3. Reproduction of neural signal from speech signal using TRF [13]

Both low-level acoustic characteristics (voice envelope) and high-level characteristics (Phonemes and phonetics) were required to research speech tracing of brain signals. The acoustic characteristics are linearly connected to the linear system in the AAD algorithm, which is based on linear systems theory. This assignment may be completed either forwards or backward direction. These algorithms provide a good understanding of the fundamental neuroscientific mechanism by which, in multi-speaker situations, the brain suppresses ignored speakers. By the speech envelope as an input feature, the linear approach may create a system response function that characterizes the forward auditory route. This system response function is called the temporal response function (TRF). TRF is a linear stimulus-response model that provides a linear relationship between the provided input signal, which is speech, and the output signals, i.e., cortical response. Figure 3 gives a pictorial understanding on how reproduction of neural signal from speech signal using TRF.

TRF is used to predict the cortical response from the speech envelope, which is termed a forward model. Similarly, the equations can be altered so that the speech is predicted from the available cortical response, which is called the backward model. The backward model involves fewer complex calculations to find the TRF coefficients and is relatively easier to implement when compared to the forward model. Forward models are also called generative or encoding models as they define how the system generates or encodes information. Figure 4 is a conceptual sketch of envelop extraction and comparison between speech and EEG signal.

$$Q = PW \qquad (1)$$

where Q – model prediction of time dimension 't' vector
P – model input matrix with the channel dimension 'c' and dimension 't'
W – linear TRF model parameter

$$P = \begin{bmatrix} p_{11} & p_{12}\ldots\ldots & p_{1p} \\ p_{21} & p_{22}\ldots\ldots & p_{2p} \\ \vdots & \vdots & \vdots \\ p_{n1} & p_{n2} & p_{np} \end{bmatrix}, W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix}, Q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix}$$

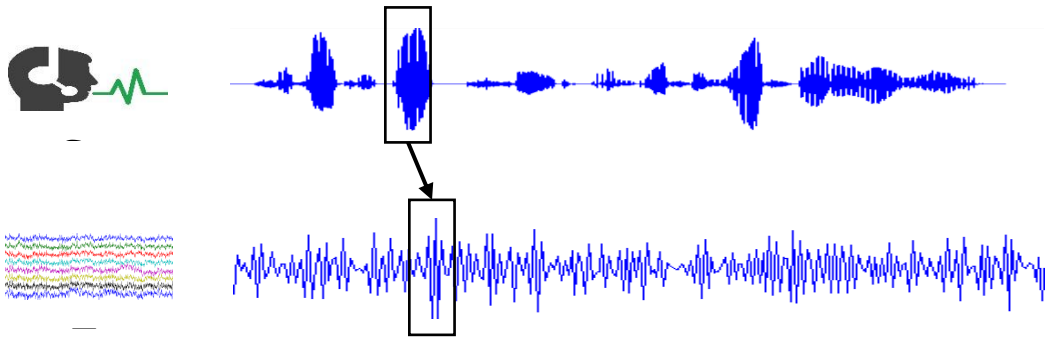

Figure 4. A conceptual sketch of envelop extraction and comparison between speech and EEG signal

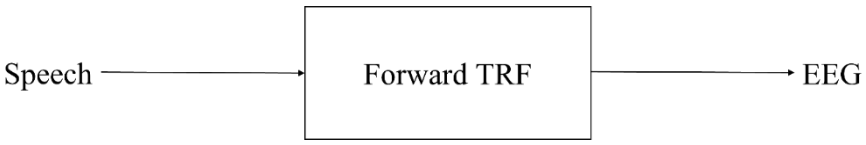

Figure 5. Forward TRF model block diagram

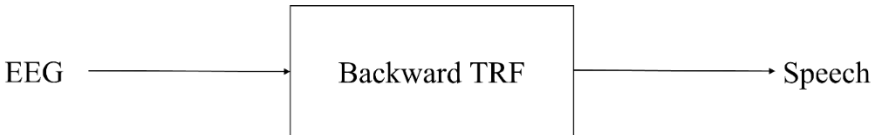

Figure 6. Backward TRF model block diagram

Figure 5 and 6 shows the block diagram representation of the forward model of TRF and the backward model of TRF. Figure 7 shows the forward model of TRF where EEG is predicted from the stimulus. The two speech streams are given to the pre-processing units, where the raw speech streams are converted to the required format. Pre-processing unit takes care of missing and noisy data in the input, if any. It converts the raw data into how the forward model of TRF needs the input data. After the pre-processed signals are passed, we may get the predicted EEG as output.

On the other hand, original cortical recordings of the experiment are applied to the pre-processing unit to process the noisy parts of the data. The output from the pre-processing of cortical recordings includes 66 EEG signals. The predicted EEG out of the TRF and pre-processed cortical signals are applied to the correlation blocks separately. For each speech stream, correlation coefficients are produced out of the correlation block. These correlation coefficients are the features to decide which speech stream listener has attended to.

The backward model of TRF is shown in Figure 8, where the envelope of the audio is approximated from the EEG. After pre-processing of cortical signals, the result is applied to TRF, which produces the predicted audio. From each speech stream, a correlation coefficient is produced from each correlation block. The decision block decides the attended speech stream by comparing the correlation coefficients, whichever the more significant is the attended.

TRF shape analysis on the human brain has encoded monitored speakers differently than ignored speakers. In particular, the TRF adjacent to the current speaker has illustrated peak points around 100ms and200ms and weak in TRF corresponding to ignored speakers. A nearer attention modulation effect is noticed when the acoustic input was changed to the following uses: higher-order functions such as speech spectrogram or phonetics. Similar to the inverse model, the EEG signal can be used to reconstruct the input stimuli (stimuli reconstruction technique), listener attention is derived by comparing with reconstructed

stimuli with the input. Based on these findings, the AAD algorithm supports hearing aids in union with a sturdy voice dissociation algorithm, conveniently available to users.

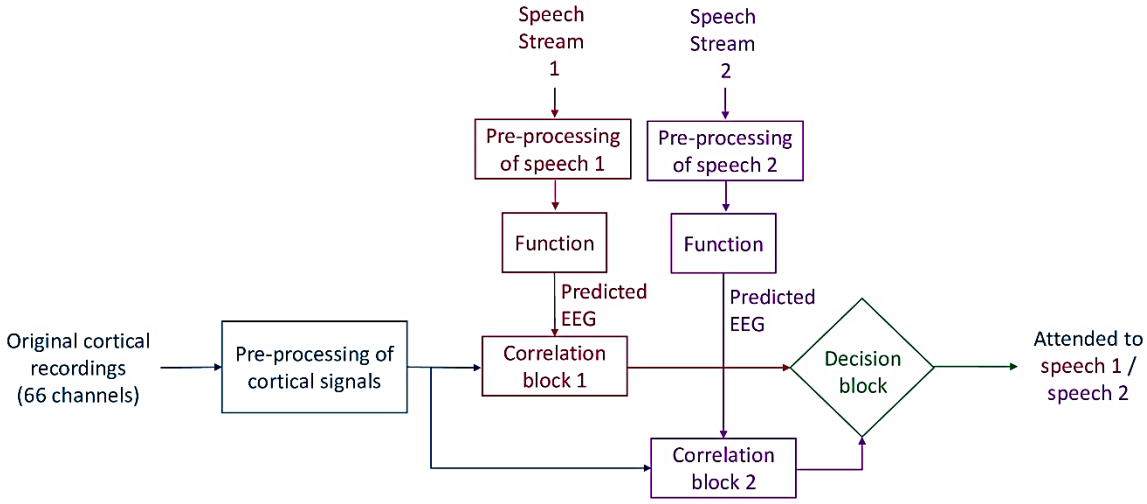

Figure 7. Forward model block diagram

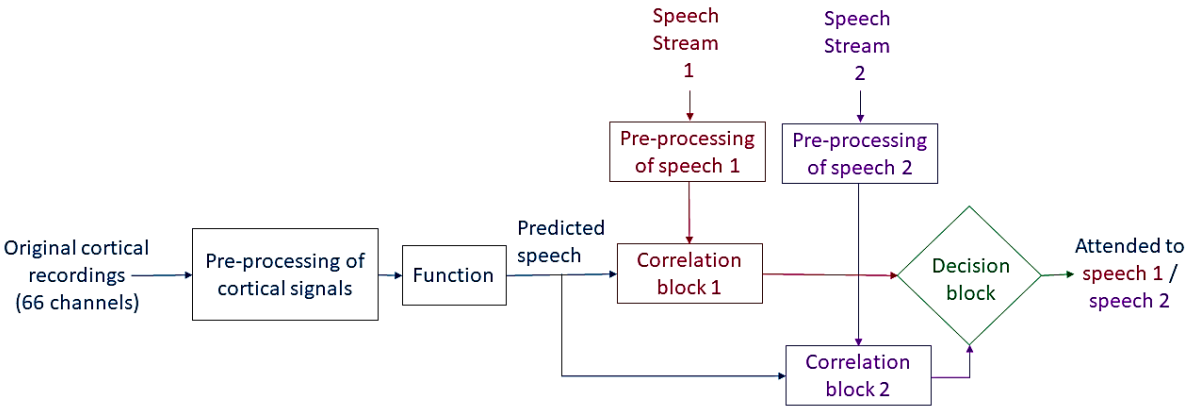

Figure 8. Backward model block diagram

The nonlinear nature of the human auditory system has been well demonstrated, and AAD evaluation has traditionally relied solely on linear structures. Within the pre-processing step, the concept lectures the nonlinearity problem to a certain extent. Another disadvantage of linear techniques in the duration of speech envelope extraction is the more time needed to categorize attention, despite attempts to overcome this limitation. Neural networks' popularity has increased in recent years, particularly in automatic vision and natural language processing. Due to its capability to model nonlinearity, Electroencephalogram Data has been used to replicate the dynamic condition of the brain. using neural networks.

Similarly, to comprehend the AAD, a convolutional neural network (CNN)-based model has been developed. The stimulus reconstruction set of rules uses the CNN version to derive focus. A direct type of attention bypasses the stimulus reconstruction regression process and instead categorizes whether the focus is on the first or second speaker. Categorizing interest as a hit vs. flop or fit vs. unfit was also addressed in a non-competing speaker experiment.

In addition to deciphering which audio it is also possible to determine the spatial location of the particular attention by comparing the envelope to the present speaker. That means it doesn't decode which speakers are in attendance but wherein the room. The advantage of this approach to neurally controlled artificial hearing is that it does not require access to pure

speech stimuli. This was investigated based on differences in EEG entropy characteristics, but the performance was less for practical benefits (less than 70% in a 60-second window). Recent studies indicate that path of auditory attention is neuron-encoded, and it is viable to decipher the location or orbit of the accompanying sound from the EEG. Some studies by MEG propose that The place of choosing auditory attention may be found by tracing the alpha power band.

In addition to adopting a stimulus reconstruction strategy akin to that of linear approaches, the attended speaker can also be identified by nonlinear systems derived from (deep) neural networks using only the audio and EEG. (Aka direct classification). However, given the short dataset sizes that are commonly used in AAD research, these nonlinear approaches are more susceptible to overfitting. There are various tactics and network topologies of the offered nonlinear methods in order to understand the distinctions between the many neural network based AAD systems now in use.

The neural network prototypes shown above either don't employ audio functions or the audio envelope curve as an input characteristic. Because neural networks are data-driven approaches, adding more data or speech stimuli might help them perform better. The input function to segregate several speakers from one audio composition using a voice separation algorithm in neural networks is spectrogram. We introduce a new neural network architecture, which uses a few voice spectrograms of speakers and EEG data as inputs for classifying auditory attention, which a standard audio-visual speech separation model inspires.

2. EXPERIMENT AND DATA

The efficacy of neural networks was assessed using three different EEG datasets. These are publicly available and accessible. The FAU dataset has the EEG data of 27 testers, all native German speakers, is included. Two speech stimuli were presented simultaneously over a speaker to simulate the cocktail party effect, and participants were instructed to pay particular attention to a single speaker of those stimuli. Audio stimuli were collected from two male speakers reading from the German news site www.dw.de's slowly spoken news section. Six different presentations, each lasting around 5.30 minutes, make up the experiment. EEG was recorded in a 10-20 EEG style with 21 SilverChloride electrodes positioned on the scalp. On the right side, the reference electrode is inserted in the mastoid. The EEG signal was obtained at 2500Hz using the ground electrode on the left earlobe. In DTU dataset the sample comprises 18 testers who chose two simultaneous speakers to listen to. The speech stimulus comes from a male and female speaker reading an excerpt from a Danish audiobook. The experiment is divided into 60 segments, each lasting 50 seconds, for 50 minutes. The EEG was recorded using 64 electrodes, at a frequency sample of 512 Hz. After visual inspection, the reference electrode is taken as either left or right mastoid. The KUL dataset contains 16 testers who took part in a selective attention experiment. A male speaker delivered four Dutch stories as the speech stimulus. Each story spanned 12 minutes and was broken into two 6-minute halves. EEG (electroencephalography) is captured on 64 electrodes and sampled at 8196Hz. After visual examination of the standard of the EEG signal obtained at TP7 or these sites, the reference electrode was employed as a TP8 electrode. Three separate conditions were used in the experiment: HRTFs, binaural separation, and repetitive stimulation. This only looked at the dichotic state persisted for 24 minutes. 34.9 hours' worth of EEG data were evaluated in this procedure. However, the speech stimuli used in each dataset, which includes two speaker data of 104 minutes, are the same for all subjects. Two speakers read the 3 datasets that were analyzed and a variety of stimuli. To prevent the stimulus learning effect, the stimulus was only presented to the participant once. Data from training and testing for each participant were split into 75% and 25% of each, respectively. The test data present

in the training data do not employ any of the speech or EEG components. then half test data is divided, with one half serving as an authentication or validation parameter during the training record of proceedings.

Table. 1 How experiment is conducted for the first eight trails.

Sl. No	Input to the left ear	Input to the right ear	Direction of the ear attended
1	1.1	2.1	Left
2	2.2	1.2	Right
3	3.1	4.1	Left
4	4.2	3.2	Right
5	2.1	1.1	Left
6	1.2	2.2	Right
7	4.1	3.1	Left
8	3.2	4.2	Right

X.Y = Story. Part of the story

The analyzed brain signals (EEG) are recorded at different sampling frequencies, a low pass filter was used to filter each of them. 32 Hz serves as the cut-off frequency, down sampling to a sampling rate of 64Hz. Further measured signals at only ten electrode positions are considered in the analysis; these are F3, F4, F7, F8, T7, C3, C4, Cz, T8, Pz. This study examined 4 test periods: 2, 3, 4, and 5 seconds respectively. A one-second overlap is applied to the 2-second attempt; that's why a total of 118922 studies for analysis. To keep the total trials constant, 2 seconds overlap was used for 3-second trials, and the 3-second overlap is used for 4-second trials; for a 5-second trial, a 4-second overlap was used.
The EEG signal in each experiment was even filtered with high pass The filtered signal is normalized to zero average and unity mean at each electrode location using a 1 Hz cut-off frequency. First, low pass filters were used to down sample audio stimuli at an 8 kHz cut-off frequency. The sampling rate is up to 16kHz. It was then divided into experiments that lasted for 2, 3, 4, and 5 seconds and overlap is of 1, 2, 3, and 4 seconds respectively. Each experiment's audio spectrogram was obtained by using the STFT's absolute value (short-time Fourier transform). A Han window is used to calculate the STFT with a duration of 32ms with 12ms overlap. Most of the detailed analysis was done in 3 seconds trial and other trials; the run time was used only for comparison purposes.

Table 2. Elements and attributes of the data sets considered for the experiment.

Name	Number of subjects	Duration per subject in minutes	Total duration in hours	Experiment type
Data set 1	16	24	6.4	Male & Male
Data set 2	27	30	13.5	Male & Male
Data set 3	18	50	15	Male & Female

3. METHODOLOGY

Artificial Intelligence techniques now-a-days are trying to build potential systems that bridge gap between the humans and machines [14], [15]. Computer vision is one such area where

there is a monumental growth in enabling machines to perceive the world for performing multitude of jobs [15]-[19].

Several convolutional layers with a non-linear activation function make up a convolutional neural network (CNN), which is subsequently followed by a pool layer [20]-[24]. The local data features were extracted using one or more convolutional filters which are applied to the input. The pooling layer comes next, which combines the output after computing mean and other metrics. Like the other types of neural networks, the CNNs are enhanced by lowering the loss function and the enhancing characteristics are calculated using optimization algorithms such as stochastic gradient descent [25]-[30].

Convolutional-Neural-Networks is used for decoding the orientation of auditory attention where the 64XT matrix is the input. Here, the variable 'T' stands in for a sample decision window, while In the dataset, there are 64 total EEG channels, which is represented by the number 64. The initial step of the model is convolutional layer. Five autonomous spatiotemporal filters are moved across the input matrix it is the initial dimension which is same as the total number of channels. Every output is a time series of 1xT. Here '17' is 130ms at 128Hz, and 130ms are the optimal filter widths which is longer or shorter decision window. The length resulted in higher losses in the validation set. After the convolution process, the rectified linear unit's activation function is utilized.

The data is averaged throughout the temporal dimension during the mean pooling procedure by periodically decreasing series to a single digit value. Here After the pooling process, there are two completely connected layers. The 1st layer consists of 5 neurons (1 per time series), followed by the sigmoidal activation function, and 2nd layer consists of 2 output neurons. Here the two output neurons are concatenated to the cross-entropy loss function. EEG subnet includes four different layers of convolution. Size of kernel of the first layer was chosen to be 24, which accommodated a delay of 375ms in the time domain. Previous studies have shown that there is a difference from 100ms to 200ms in TRFs adjacent to attended and ignored speakers. Therefore, In a situation with two speakers, a 375ms delay aids in bringing out elements that control attention to various speakers. The shorter core was used to initialize each of the subsequent layers. Figure 9 shows the visionary sketch of finding subjects attention when there are two competing speakers.

Up until the point where the max pooling was utilized to reduce the dimensionality, all convolutions were carried out in 1x1 stages. To avoid and improve over-adaptation of training data and generalization respectively, Batch normalization and dropout were used. The output is then sent via a rectified linear unit (ReLU), a nonlinear activation function.

The trial length determines the input dimensions for EEG_CNN, and output dimensions are set to 48x32. Maximum pooling parameters has changed slightly over various test periods to maintain a fixed performance dimension. The temporal axis is one dimensional (48), whereas the number of convolutional kernels is two dimensional (32). Output dimensions representing EEG signals noted in various electrodes are lowered to unity by applying Max Pooling continuously down the electrode axis.

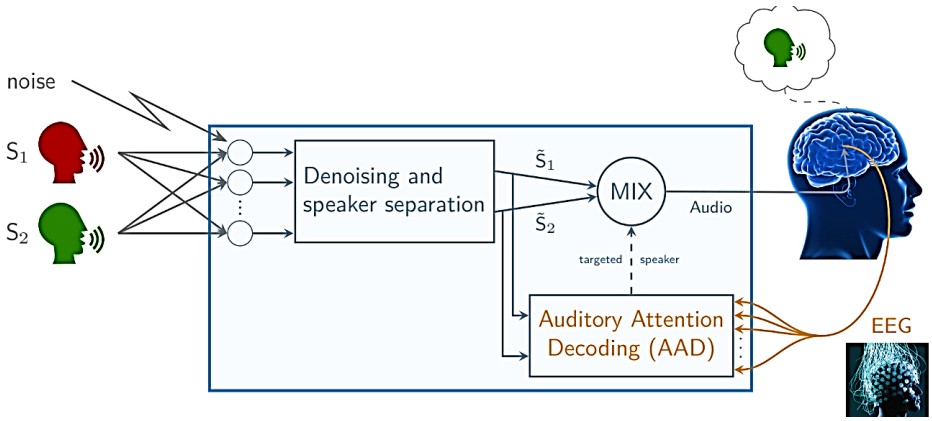

Figure 9. A visionary sketch of finding subjects attention when there are two competing speakers.

The audio sub network that analyses the audio spectrogram has five layers of convolution, and the convolutional output has been subjected to all conventional approaches such as activation of ReLU, batch normalisation, max pooling, and dropout. The duration of the experiment had an impact on both the input dimensions and the audio CNN of EEG-CNN, while the dimensions of output feature map were set as 48x16 always. Because the dataset used in this work came from a two-speaker experiment, the audio CNN was run two tomes, yielding two sets of output.

The characteristic maps produced by EEG-CNN and Audio CNN was connected down the time axis. The size of the attribute map after chaining was 48x64. Then half of it is provided by EEG data, and other half is by audio data. This even issues the feasibility to extend to more than one speaker. Connected feature maps have been slide a bidirectional long-short term memory (BLSTM) layer followed by 4 fully connected (FC) layers. ReLU activated for the initial 3 FC layers and soft max activation was applied to the end FC layer to help categorize the attention of speaker one or speaker two. There are 1,18,922 total EEG and audio trials available, 75% of the total available samples, which are 89192, were required to train the network and the remaining samples are 29730, were equally split between test data and validation data. The network was trained in a mini batch of 80 epochs with a 32-sample size, the learning rate is $5*10^{-4}$. The dropout probability was adjusted to 0.25 for EEG-CNN and AEConcat subnets but incremented to 0.4 on audio CNN subnets.

Since the voice stimuli were the same everywhere, the audio CNN was more likely to drop out. Therefore, when training with data from multiple subjects, audio data maintain identity, the network can store the speech spectrum of training data. The network was optimized using the Adam Optimizer with binary cross entropy as the loss function. Neural network training causes random variation from epoch to epoch, so the accuracy of the test is measured as the average accuracy of the last 5 epochs. The network is trained using Nvidia GeForce RTX2060 (6 GB) graphics card, which lasted about 36 hours training and the figure 10 shows the block diagram of deep learning-based source localization technique from speech and EEG mixture.

Neural-network study is an advanced and complex algorithm, which is not yet widely utilized in embedded systems gadgets because of great memory and computational power requirements. By applying these models in embedded system devices, sparse neural networks have been briefly shown to solve these issues. Most model parameters in sparse networks are zero and zero-valued multiplication, which reduces computational effort. Similarly, it is not equal to zero weights must be preserved on the device, and only their position should be specified for all zero-valued weights. It is well-known to reduce memory requirements. Factual evidence shows that neural networks allow high sparsity degree.

Sparse neural networks are made utilizing a process called network pruning. It includes three steps mainly. Firstly, vast networks with over parameters are trained to achieve a high level of test accuracy as excessive parameterization has more powerful display capabilities. Second, only important weights based on criteria are kept in an over parameterized trained network, while other weights are considered verbose and are reset to '0'. Lastly, the pruned network is refined by continuing training with only retained weights to upgrade performance. Simple methods, such as magnitude thinning, or more complicated algorithms, such as variation dropouts and L0 regularization, can be used to identify inessential weights. However, it has been shown that the introduction of sparsity can be achieved using magnitude pruning and performance will equal to or better than sophisticated techniques such as variability dropouts and L0 arrangement.

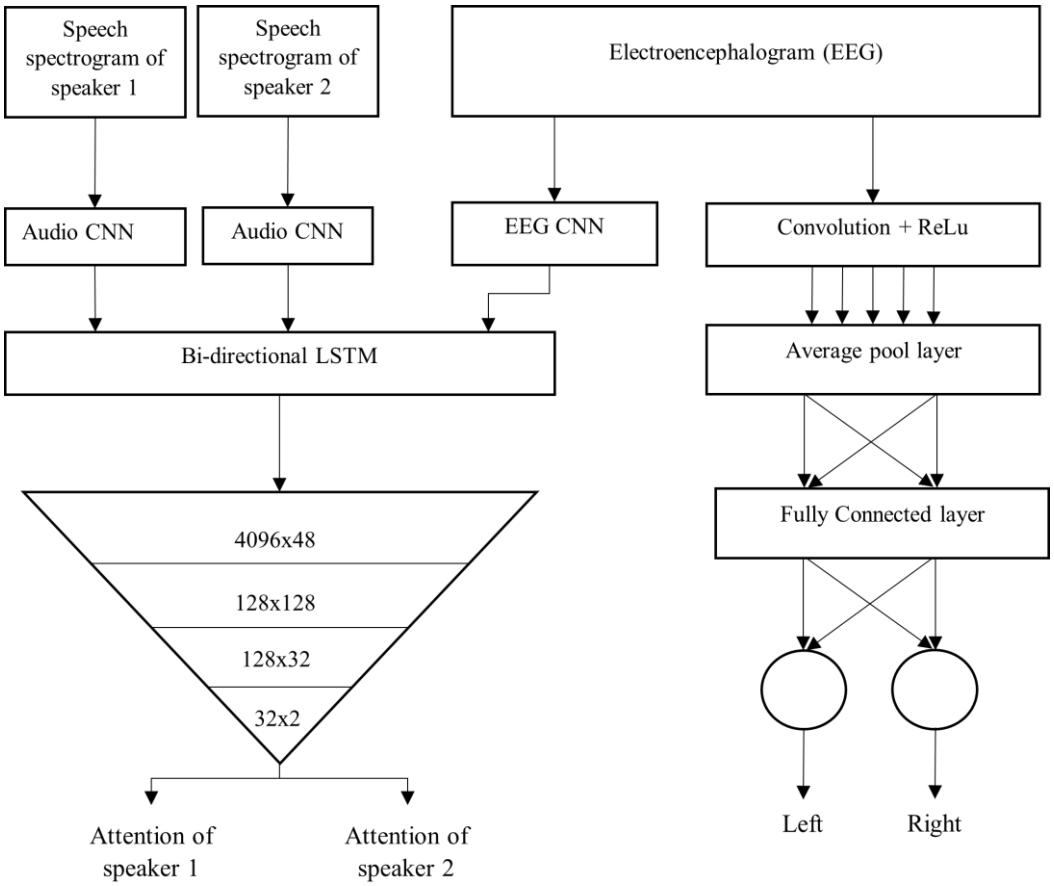

Figure 10. Block diagram of deep learning-based source localization technique from speech and EEG mixture

4. TRAINING AND EVALUATION:

The data from every subject was used to train the model. It creates a subject-specific decoder that can use dxata from other subjects as a data extension strategy to avoid overfitting the data under test to a given amount of training. To keep the model from over-fitting to a single story, we ran it through four cross-validations. In other words, one story was held out and trained over the other three. Overfitting isn't an issue with a basic linear model, but it can be with the proposed CNN. Even if only EEG responds to a specific segment of a tale, It can result in the model picking up certain characteristics unique to the story-specific, which in turn might lead to reliable results. when the approach is introduced in an EEG response to other parts of the same story, the result is too optimistic. The figure 11 shows a visionary sketch of a continuous input reformation decoder.

Similarly, because every speaker has unique story-telling characteristics (such as speech pace and intonation) as well as distinct voice tones, EEG reacts to different things as speakers may vary. Therefore, we only maintained the folds that did not contain the same speaker in both the test set and the training set since the model benefits from having an EEG response to a specific speaker. Only two instances were left in the end.
An approach that combines cross-validation is called Leave-on-story + speaker-out. In further experiments, we scrutinized the dependence of subject of the model. In additional to the story-speaker cross-validation, we also performed topic cross-validation. That is, instead of using 'N' subjects, 'N-1' subjects were used for training and testing, and a subject that was held out was used for testing.

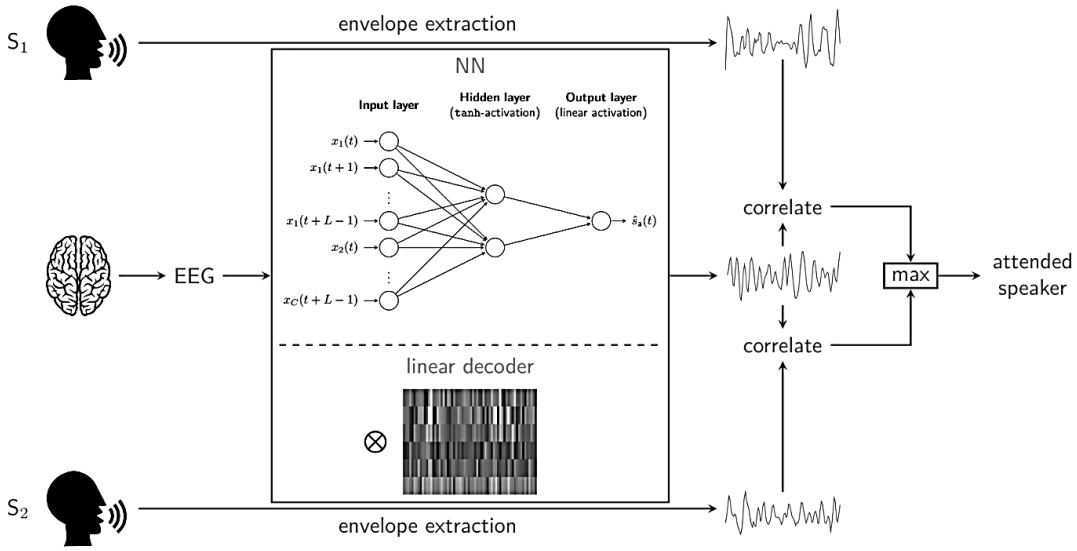

Figure 11. A visionary sketch of a continuous input reformation decoder

One benefit of this methodology is that new subjects won't have to undergo the potentially expensive and time-consuming retraining procedure. This makes it more practical for real-world use. The performance disparities between the two paradigms determine if it is a finer alternative than the subject specific re-training. Subject-specific re-training can be a considerable cost if the variation is large enough. Figure 12 shows the attention decoding performance when one specific frequency band is removed.

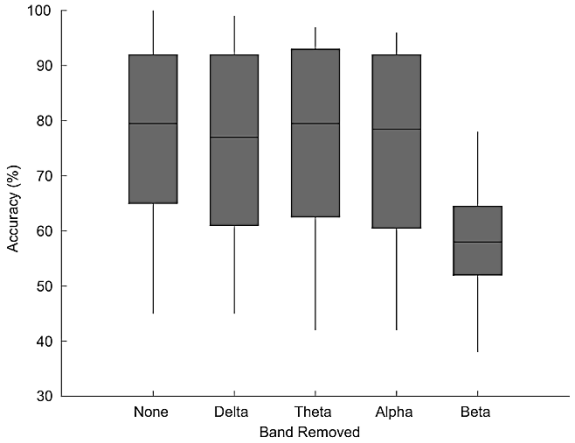

Figure 12. Attention decoding performance when one specific frequency band is removed.

By lowering the cross-entropy betwixt the outputs of network and the appropriate label, we were able to train the network (supervised ear). With momentum of 0.9 and initial learning rate of 0.09, a mini-batch stochastic gradient descent was used to ensure convergence, we used a step collapse(decay) learning plan and lowered the learning rates after epochs 10 and 35 to 0.045 and 0.0225, respectively. Because of the memory limit, the batched size was limited to 20, and there was no substantial improvement with bigger batched sizes.

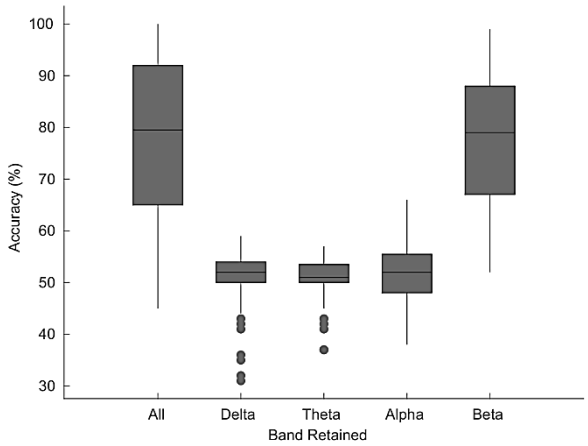

Figure 13. Attention decoding performance when exclusive band is used.

Figure 13. Attention decoding performance when exclusive band is used. Early experiments have examined that the optimal decoder was normal, the best decoder was found between epoch 70 and epoch 95 of the training set, which lasted 100 epochs. Regularization includes decay of load with a value of $5x10^{-4}$. After training, a decoder with iterations is chosen that have minimal validation losses. Note that adding data from other subjects could also be considered as a regularized method which further lowers the risk of over fitting. A grid search on a reasonable set of values was used to determine all the above hyper-parameters. A validation set was used to assess performance during these grid searches. Decoding accuracy is properly categorized as a percentage test set decision window in this task, averaged over the two folds mentioned above.

5. RESULTS

There are significant differences observed in event-related potentials (ERP) at various time instances in frontal and parietal channels. The contour pattern of the difference of ERP was almost similar with a slight difference in the positivity lateralization in 1900 ms and 2200 ms. The syllable played at spatialized locations is affected at the above moments.

We can decode source localization or attention of the subject with a better average accuracy. In a binary classification, the average accuracy in all combination of trails is between 74.95% to 76.83%. The decoding accuracy is correlated with each participant's behavioral performance. There is a strong correlation detected between behavior accuracy and decoding accuracy. The following figure 14 shows the histogram for decoding accuracy between number of subjects. The following figure 14 shows the histogram for decoding accuracy between number of subjects.

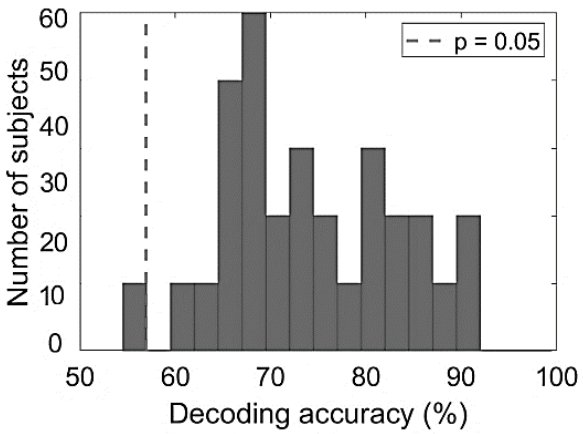

Figure 14. Histogram for decoding accuracy between number of subjects

From the literature, it is evident that 60 to 5s is the approximate window range to test the decision whether the subject is attended to right or left. For the decision windows 1 and 10s though there is a large variability in inter-subject the median decoding accuracy is higher. From 57.1 to 81.2%, the median accuracy increased. when CNN model is applied at 1s decision window. Similarly, highest median accuracy of 86.1% is achieved by the CNN model at a decision window size of 10s. The CNN model outperforms the linear model at both the window sizes with a huge margin. The CNN model performed unsatisfactorily for the stories with a below 50% accuracy where the accuracy for the stories 3 and 4 stands above 80% which is higher. The longer window lengths contain more information than shorter window lengths and hence is the reason for substandard performance. The following figure 15 shows the scatter plots for behavioral performance and decoding accuracy.

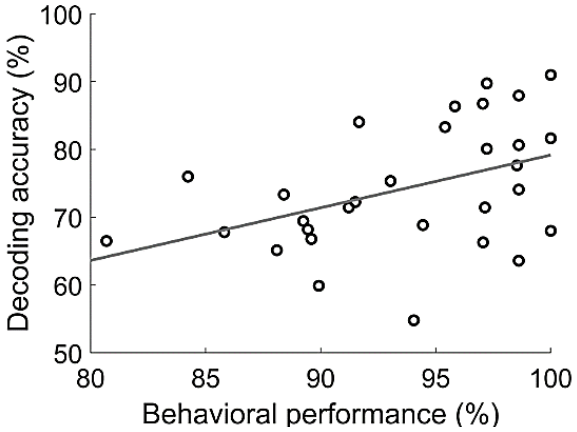

Figure 15. Scatter plots for behavioral performance and decoding accuracy

The frequency ranges of cortical recordings are defined as delta (1 to 4Hz); theta (4 to 8Hz); alpha (8 to 14Hz); and beta (14 to 32Hz). The beta band has more information that network can model. In the convolution layer the filter weights are finding which channel is important. The results show that frontal, temporal and occipital regions have good activations. While working with neural networks, it is difficult to understand which parameter played important role in achieving better accuracy. The approach followed during evaluation of the system is leave one story + speaker out. The findings show that the adopted strategy and the conventional approach both produce same decoding accuracy.

When the input applied to the system are only speech features the median decoding accuracy is observed to be low. While on the other hand when the model is exercised with EEG features the better median decoding accuracy is observed. However, the statistical analysis's findings are at odds with our assumptions, showing that applying the model using speech characteristics as opposed to EEG features does not significantly alter the median decoding accuracy. A bidirectional long short-term memory (BLSTM) is effective in handling the recorded delay between EEG and the speech. In real time scenario the speech available is mixture of huge noise which is difficult to sperate and plot the spectrogram. But in the experiment considered we have data recorded in a clean environment which is easy to process and plot a spectrogram. For training the neural network, to make up for the lack of EEG labelled data needed to train the network, data augmentation techniques are frequently employed. There is a need to generate synthetic EEG to apply linear convolution to the corresponding speech cues and the TRFs.

6. CONCLUSION AND FUTURE SCOPE

Many current algorithms need supervised training and are improved over time. To adjust to the EEG's time-varying data without necessitating onerous a training session in advance for each user individually, Adaptive AAD algorithms that don't require training are essential. Even if that field has made some progress, The results of this investigation show that a practical solution is still a long way off. To create closed-loop systems for hearing devices with neuro-steered, these online adaptive AAD algorithms are essential. This let the user engage with the speech-enhancement system and AAD algorithm. Interaction between the hearing aid's computer algorithms and the individual who wears it may produce neurofeedback effects that considerably enhance the hearing aid's functionality. Finally, real-world testing of these AAD algorithms is required, considering a variety of realistic listening conditions and possible hearing device users. To create the individual parts of a neuro-steered hearing device, In addition to a miniaturized EEG sensor system, a reliable and low-latency speaker separation method, and an intelligent gain control system are required for ADD algorithm. Future generations of hearing devices might work and be accepted by users much more effectively if neuro-steered hearing devices are used as a neuro rehabilitative assistive technology. However, there are still numerous obstacles to overcome.

REFERENCES

1. E. C. Cherry, “Some Experiments on the Recognition of Speech, with One and with Two Ears,” J Acoust Soc Am, vol. 25, no. 5, p. 975, Jun. 2005, doi: 10.1121/1.1907229.
2. E. C. Cherry, “Some Experiments on the Recognition of Speech, with One and with Two Ears,” Journal of the Acoustical Society of America, vol. 25, no. 5, 1953, doi: 10.1121/1.1907229.
3. C. EC, “Some experiments on the recognition of speech, with one and with two ears,” J Acoust Soc Am, 1953.
4. P. G. Parande and T. G. Thomas, “A study of the cocktail party problem,” in 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Nov. 2017, pp. 1–5. doi: 10.1109/ICECTA.2017.8251979.
5. L. Fiedler, J. Obleser, T. Lunner, and C. Graversen, “Ear-EEG allows extraction of neural responses in challenging listening scenarios - A future technology for hearing

aids?" Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, vol. 2016-October, pp. 5697–5700, Oct. 2016, doi: 10.1109/EMBC.2016.7592020.

6. E. Holmes, P. T. Kitterick, and A. Q. Summerfield, "Peripheral hearing loss reduces the ability of children to direct selective attention during multi-talker listening," Hear Res, vol. 350, pp. 160–172, Jul. 2017, doi: 10.1016/J.HEARES.2017.05.005.
7. J. Feldman, "What is a visual object?," Trends Cogn Sci, vol. 7, no. 6, pp. 252–256, Jun. 2003, doi: 10.1016/S1364-6613(03)00111-6.
8. Y. Lu, M. Wang, Q. Zhang, and Y. Han, "Identification of Auditory Object-Specific Attention from Single-Trial Electroencephalogram Signals via Entropy Measures and Machine Learning," Entropy 2018, Vol. 20, Page 386, vol. 20, no. 5, p. 386, May 2018, doi: 10.3390/E20050386.
9. Z. Zhang, G. Zhang, J. Dang, S. Wu, D. Zhou, and L. Wang, "EEG-based short-time auditory attention detection using multi-task deep learning," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2020-October, pp. 2517–2521, 2020, doi: 10.21437/INTERSPEECH.2020-2013.
10. Tharwat, A. (2021), "Independent component analysis: An introduction", Applied Computing and Informatics, Vol. 17 No. 2, pp. 222-249. https://doi.org/10.1016/j.aci.2018.08.006.
11. S. J. Aiken and T. W. Picton, "Human cortical responses to the speech envelope," Ear Hear, vol. 29, no. 2, pp. 139–157, Apr. 2008, doi: 10.1097/AUD.0B013E31816453DC.
12. T. Wolbers, P. Zahorik, and N. A. Giudice, "Decoding the direction of auditory motion in blind humans," Neuroimage, vol. 56, no. 2, pp. 681–687, May 2011, doi: 10.1016/J.NEUROIMAGE.2010.04.266.
13. S. Miran, A. Presacco, J. Z. Simon, M. C. Fu, S. I. Marcus, and B. Babadi, "Dynamic estimation of auditory temporal response functions via state-space models with Gaussian mixture process noise," PLoS Comput Biol, vol. 16, no. 8, p. e1008172, Aug. 2020, doi: 10.1371/JOURNAL.PCBI.1008172.
14. Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature 2015 521:7553, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
15. C. Li, Y. Qi, X. Ding, J. Zhao, T. Sang, and M. Lee, "A Deep Learning Method Approach for Sleep Stage Classification with EEG Spectrogram," Int J Environ Res Public Health, vol. 19, no. 10, May 2022, doi: 10.3390/IJERPH19106322.

16. H. Ansari, P. J. Cherian, A. Caicedo, G. Naulaers, M. de Vos, and S. van Huffel, "Neonatal Seizure Detection Using Deep Convolutional Neural Networks," Int J Neural Syst, vol. 29, no. 4, May 2019, doi: 10.1142/S0129065718500119.
17. S. Anwar, K. Hwang, and W. Sung, "Structured Pruning of Deep Convolutional Neural Networks," ACM Journal on Emerging Technologies in Computing Systems (JETC), vol. 13, no. 3, Feb. 2017, doi: 10.1145/3005348.
18. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," J Neural Eng, vol. 16, no. 3, p. 031001, Apr. 2019, doi: 10.1088/1741-2552/AB0AB5.
19. S. Belsare, M. Kale, P. Ghayal, A. Gogate and S. Itkar, "Performance Comparison of Different EEG Analysis Techniques Based on Deep Learning Approaches," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2021, pp. 490-493, doi: 10.1109/ESCI50559.2021.9396856.
20. H. Polat and M. S. Özerdem, "Automatic Detection of Cursor Movements from the EEG Signals via Deep Learning Approach," 2020 5th International Conference on Computer Science and Engineering (UBMK), Diyarbakir, Turkey, 2020, pp. 327-332, doi: 10.1109/UBMK50275.2020.9219507.
21. R. Mathumitha and A. Maryposonia, "A Secure Authentication System based on Emotion Analysis of EEG Signals using Deep Learning Technique," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, pp. 1232-1237, doi: 10.1109/ICICCS53718.2022.9788367.
22. X. Jia et al., "Multi-Channel EEG Based Emotion Recognition Using Temporal Convolutional Network and Broad Learning System," 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 2020, pp. 2452-2457, doi: 10.1109/SMC42975.2020.9283159.
23. D. Acharya, S. Goel, H. Bhardwaj, A. Sakalle and A. Bhardwaj, "A Long Short Term Memory Deep Learning Network for the Classification of Negative Emotions Using EEG Signals," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207280.
24. M. Wu, B. Nan and C. Li, "EEG Classification Based On Deep Learning," 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2022, pp. 14-21, doi: 10.1109/ITAIC54216.2022.9836779.
25. D. Truong, M. Milham, S. Makeig and A. Delorme, "Deep Convolutional Neural Network Applied to Electroencephalography: Raw Data vs Spectral Features," 2021 43rd

Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021, pp. 1039-1042, doi: 10.1109/EMBC46164.2021.9630708.

26. Jorge J. Palacios-Venegas, "Deep Learning Assisted Biofeedback", Advances in Non-Invasive Biomedical Signal Sensing and Processing with Machine Learning, pp.289, 2023.
27. Y. Sun and U. Kintak, "CNN-Based Scheme on EEG Hand-Motion Recognition Without Signal Preprocessing," 2022 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Toyama, Japan, 2022, pp. 72-76, doi: 10.1109/ICWAPR56446.2022.9947127.
28. T. Agarwal, S. Raturi, T. Vybhav and M. Singh, "Classification of EEG signal using LSTMs under Audiovisual Stimuli," 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 1229-1232, doi: 10.1109/ICCSP48568.2020.9182092.
29. N. M. N. Leite, E. T. Pereira, E. C. Gurjão and L. R. Veloso, "Deep Convolutional Autoencoder for EEG Noise Filtering," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 2018, pp. 2605-2612, doi: 10.1109/BIBM.2018.8621080.
30. X. Zhang, L. Yao, Q. Z. Sheng, S. S. Kanhere, T. Gu and D. Zhang, "Converting Your Thoughts to Texts: Enabling Brain Typing via Deep Feature Learning of EEG Signals," 2018 IEEE International Conference on Pervasive Computing and Communications (PerCom), Athens, Greece, 2018, pp. 1-10, doi: 10.1109/PERCOM.2018.8444575.