Extracting the Auditory Attention in a Dual-Speaker Scenario From EEG Using a Joint CNN-LSTM Model

Ivine Kuruvila[1], Jan Muncke[1], Eghart Fischer[2] and Ulrich Hoppe[1]*

[1] *Department of Audiology, ENT-Clinic, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany,* [2] *WS Audiology, Erlangen, Germany*

Human brain performs remarkably well in segregating a particular speaker from interfering ones in a multispeaker scenario. We can quantitatively evaluate the segregation capability by modeling a relationship between the speech signals present in an auditory scene, and the listener's cortical signals measured using electroencephalography (EEG). This has opened up avenues to integrate neuro-feedback into hearing aids where the device can infer user's attention and enhance the attended speaker. Commonly used algorithms to infer the auditory attention are based on linear systems theory where cues such as speech envelopes are mapped on to the EEG signals. Here, we present a joint convolutional neural network (CNN)—long short-term memory (LSTM) model to infer the auditory attention. Our joint CNN-LSTM model takes the EEG signals and the spectrogram of the multiple speakers as inputs and classifies the attention to one of the speakers. We evaluated the reliability of our network using three different datasets comprising of 61 subjects, where each subject undertook a dual-speaker experiment. The three datasets analyzed corresponded to speech stimuli presented in three different languages namely German, Danish, and Dutch. Using the proposed joint CNN-LSTM model, we obtained a median decoding accuracy of 77.2% at a trial duration of 3 s. Furthermore, we evaluated the amount of sparsity that the model can tolerate by means of magnitude pruning and found a tolerance of up to 50% sparsity without substantial loss of decoding accuracy.

Keywords: EEG, cocktail party effect, auditory attention, long short term memory networks, hearing aids, speech enhancement, speech separation, convolutional neural network

1. INTRODUCTION

Holding a conversation in presence of multiple noise sources and interfering speakers is a task that people with normal hearing carry out exceptionally well. The inherent ability to focus the auditory attention on a particular speech signal in a complex mixture is known as the cocktail party effect (Cherry, 1953). However, an automatic machine based solution to the cocktail party problem is yet to be discovered despite the intense research for more than half a century. Such a solution is highly desirable for a plethora of applications such as human-machine interface (e.g., Amazon Alexa), automatic captioning of audio/video recordings (e.g., YouTube, Netflix), advanced hearing aids etc.

In the domain of hearing aids, people with hearing loss suffer from deteriorated speech intelligibility when listening to a particular speaker in a multispeaker scenario. Hearing aids

currently available in the market often do not provide sufficient amenity in such scenarios due to their inability to distinguish between the attended speaker and the ignored ones. Hence, additional information about the locus of attention is highly desirable. In visual domain, selective attention is explained in terms of visual object formation where an observer focuses on a certain object in a complex visual scene (Feldman, 2003). This was extended to auditory domain where it was suggested that phenomena such as cocktail party effect could be better understood using auditory object formation (Shinn-Cunningham, 2008). In other words, brain forms objects based on the multiple speakers present in an auditory scene and selects those objects belonging to a particular speaker during attentive listening (top-down or late selection). However, flexible locus of attention theory was concurrently proposed where the late selection is hypothesized to occur at low cognitive load and early selection is hypothesized to occur at high cognitive load (Vogel et al., 2005). This has inspired investigation into whether cortical signals could provide additional information that helps to discriminate between the attended speaker and interfering speakers. In a dual-speaker experiment, it was observed that the cortical signals measured using implanted electrodes track the salient features of the attended speaker stronger than the ignored speaker (Mesgarani and Chang, 2012). Similar results were obtained using magnetoencephalography and electroencephalography (EEG) (Ding and Simon, 2012; O'Sullivan et al., 2014). In recent years, EEG analyses have become the commonly used methodology in attention research which is lately known as auditory attention decoding (AAD).

Both low level acoustic features (speech envelope or speech spectrogram) and high level features (phonemes or phonetics) have been used to investigate the speech tracking in cortical signals (Aiken and Picton, 2008; Lalor and Foxe, 2010; Di Liberto et al., 2015; Broderick et al., 2019). State-of-the-art AAD algorithms are based on linear systems theory where acoustic features are linearly mapped on to the EEG signals. This mapping can be either in the forward direction (Lalor and Foxe, 2010; Fiedler et al., 2017; Kuruvila et al., 2020) or in the backward direction (O'Sullivan et al., 2014; Mirkovic et al., 2015; Biesmans et al., 2017). These algorithms have been successful in providing insights into the underlying neuroscientific processes through which brain suppresses the ignored speaker in a dual-speaker scenario. Using speech envelope as the input acoustic feature, linear algorithms could generate system response functions that characterize the auditory pathway in the forward direction. These system response functions are referred to as temporal response function (TRF) (Lalor and Foxe, 2010). Analysis of the shape of TRFs has revealed that the human brain encodes the attended speaker different to that of the ignored speaker. Specifically, TRFs corresponding to the attended speaker have salient peaks around 100 and 200 ms which are weaker in TRFs corresponding to the ignored speaker (Fiedler et al., 2019; Kuruvila et al., 2021). Similar attention modulation effects were observed when the acoustic input was modified to using speech spectrogram or higher level features such as phonetics (Di Liberto et al., 2015). Likewise using backward models, the input stimulus can be reconstructed from EEG signals (stimulus reconstruction

method) and a listener's attention could be inferred by comparing the reconstructed stimulus to the input stimuli (O'Sullivan et al., 2014). These findings give the possibility of integrating AAD algorithms into hearing aids which in combination with robust speech separation algorithms could greatly enhance the amenity provided to the users.

It has been well-established that the human auditory system is inherently non-linear (Zwicker and Fastl, 2013) and AAD analysis based on linear systems theory addresses the issue of non-linearity to a certain extend in the preprocessing stage. For example, during speech envelope extraction. Another limitation of linear methods is the longer time delay required to classify attention (Fuglsang et al., 2017; Geirnaert et al., 2019), although there were attempts to overcome this limitation (Miran et al., 2018; Kuruvila et al., 2021). In the last few years, deep neural networks have become popular especially in the field of computer vision and natural language processing. Since neural networks have the ability to model non-linearity, they have been used to estimate the dynamic state of brain from EEG signals (Craik et al., 2019). Similarly in AAD paradigm, convolutional neural network (CNN) based models were proposed where the stimulus reconstruction algorithm was implemented using the CNN model to infer attention (Ciccarelli et al., 2019; de Taillez et al., 2020). A direct classification of attention which bypasses the regression task of stimulus reconstruction, instead classifies whether the attention is to speaker 1 or speaker 2 directly was proposed in Ciccarelli et al. (2019) and Vandecappelle et al. (2021). In a non-competing speaker experiment, classifying attention as successful vs unsuccessful or match vs mismatch was further addressed in Monesi et al. (2020) and Tian and Ma (2020).

All aforementioned neural network models either did not use speech features or made use of only speech envelope as the input feature. As neural networks are data driven models, additional data/information about the speech stimuli may improve the performance of the network. In speech separation algorithms based on neural networks, spectrogram is used as the input feature to separate multiple speakers from a speech mixture (Wang and Chen, 2018). Inspired by the joint audio-visual speech separation model (Ephrat et al., 2018), we present a novel neural network framework that make use the speech spectrogram of multiple speakers and the EEG signals as inputs to classify the auditory attention.

The rest of the paper is organized as follows. In section 2, details of the datasets that were used to train and validate the neural network are provided. In section 3, the neural network architecture is explained in detail. The results are presented in sections 4, 5 provides a discussion on the results.

2. MATERIALS AND METHODS

2.1. Examined EEG Datasets

We evaluated the performance of our neural network model using three different EEG datasets. The first dataset was collected at our lab and it will be referred to as FAU_Dataset (Kuruvila et al., 2021). The second and third datasets are publicly available and they will be referred to as DTU_Dataset (Fuglsang et al., 2018) and KUL_Dataset (Das et al., 2019), respectively.

2.1.1. FAU_Dataset

This dataset comprised of EEG collected from 27 subjects who were all native German speakers. A cocktail party effect was simulated by presenting two speech stimuli simultaneously using loudspeakers and the subject was asked to attend selectively to one of the two stimuli. Speech stimuli were taken from the slowly spoken news section of the German news website www.dw.de and were read by two male speakers. The experiment consisted of six different presentations with each presentation being approximately five minutes long making it a total of 30 min. EEG was collected using 21 AgCl electrodes placed over the scalp according to the 10–20 EEG format. The reference electrode was placed at the right mastoid, the ground electrode was placed at the left earlobe and the EEG signals were sampled at 2,500 Hz. More details of the experiment could be found in Kuruvila et al. (2021).

2.1.2. DTU_Dataset

This is a publicly available dataset that was part of the work presented in Fuglsang et al. (2017). The dataset consisted of 18 subjects who selectively attended to one of the two simultaneous speakers. Speech stimuli were excerpts taken from Danish audiobooks that were narrated by a male and a female speaker. The experiment consisted of 60 segments with each segment being 50 s long making it a total of 50 min. EEG were recorded using 64 electrodes and were sampled at 512 Hz. The reference electrode was chosen either as the left mastoid or as the right mastoid after visual inspection. Further details can be found in Fuglsang et al. (2017, 2018).

2.1.3. KUL_Dataset

The final dataset that was analyzed is another publicly available dataset where 16 subjects undertook selective attention experiment. Speech stimuli consisted of four Dutch stories narrated by male speakers. Each story was 12 min long which was further divided into two 6 min presentations. EEG was recorded using 64 electrodes and were sampled at 8,196 Hz. The reference electrode was chosen either as TP7 or as TP8 electrode after visually inspecting the quality of the EEG signal measured at these locations. The experiment consisted of three different conditions namely HRTF, dichotic and repeated stimuli. In this work, we analyzed only the dichotic condition which was 24 min long. Additional details about the experiment and the dataset can be found in Das et al. (2016, 2019).

Details of the datasets are summarized again in **Table 1**. A total of 34.9 h of EEG data were examined in this work. However, the speech stimuli used were identical across subjects per dataset and they totalled 104 min of dual-speaker data. In all the three datasets that were analyzed, the two speakers read out different stimuli. Moreover, the stimuli were presented only once to the subject in order to avoid any learning effect. For each subject, the training and the test data were split as 75–25% and we ensured that no part of the EEG or the speech used in the test data was part of the training data. The test data were further divided equally into two halves and one half was used as a validation set during the training procedure.

2.2. Data Analysis

As EEG signals analyzed were collected at different sampling frequencies, they were all low pass filtered at a cut off frequency of 32 Hz and downsampled to 64 Hz sampling rate. Additionally, signals measured at only 10 electrode locations were considered for analysis and they were F7, F3, F4, F8, T7, C3, Cz, C4, T8, Pz. We analyzed four different trial durations in this study namely 2, 3, 4, and 5 s. For 2 s trials, an overlap of 1 s was applied. Thus, there were 118,922 trials in total for analysis. In order to maintain the total number of trials constant, 2 s of overlap was used in case of 3 s trial, 3 s of overlap was used in case of 4 s trial and 4 s overlap was used in case of 5 s trial. EEG signals in each trial were further high pass filtered with a cut off frequency of 1 Hz and the filtered signals were normalized to have zero mean and unit variance at each electrode location.

Speech stimuli were initially low pass filtered with a cut off frequency of 8 kHz and were downsampled to a sampling rate of 16 kHz. Subsequently, they were segmented into trials with a duration of 2, 3, 4, and 5 s at an overlap of 1, 2, 3, and 4 s, respectively. The speech spectrogram for each trial was obtained by taking the absolute value of the short-time Fourier transform (STFT) coefficients. The STFT was computed using a Hann window of 32 ms duration with a 12 ms overlap. Most of the analysis in our work was performed using 3 s trial and other trial durations were used only for comparison purposes. A summary of the dimensions of EEG signals and speech spectrogram after preprocessing for different trial durations is provided in **Table 2**.

3. NETWORK ARCHITECTURE

A top level view of the proposed neural network architecture is shown in **Figure 1**. It consists of three subnetworks namely EEG_CNN, Audio_CNN, and AE_Concat.

TABLE 1 | Details of the EEG datasets analyzed.

Name	Number of subjects	Duration per subject (minutes)	Total duration (hours)	Experiment type	Language
FAU_Dataset	27	30	13.5	Male + Male	German
DTU_Dataset	18	50	15	Male + Female	Danish
KUL_Dataset	16	24	6.4	Male + Male	Dutch

TABLE 2 | Trial duration vs. dimension of the input.

Trial duration (sec)	EEG data (time × num_electrodes)	Speech data (time × freq)
2	128 × 10	101 × 257
3	192 × 10	151 × 257
4	256 × 10	201 × 257
5	320 × 10	251 × 257

3.1. EEG_CNN

The EEG subnetwork comprised of four different convolutional layers as shown in **Table 3**. The kernel size of the first layer was chosen as 24 and it corresponded to a latency of 375 ms in the time domain. A longer kernel was chosen because previous studies have shown that the TRFs corresponding to attended and unattended speakers differ around 100 and 200 ms (Fiedler et al., 2019; Kuruvila et al., 2021). Therefore, a latency of 375 ms could help us to extract features that modulate the attention to different speakers in a dual-speaker environment. All other layers were initialized with kernels of shorter duration as shown in **Table 3**. All convolutions were performed using a stride of 1 × 1 and after the convolutions, max pooling was used to reduce the dimensionality. To prevent overfitting on the training data and improve generalization, dropout (Srivastava et al., 2014), and batch normalization (BN) (Ioffe and Szegedy, 2015) were applied. Subsequently, the output was passed through a non-linear activation function which was chosen as rectified linear unit (ReLU). The dimension of the input to EEG_CNN varied according to the length of the trial (**Table 2**) but the dimension of the output was fixed at 48 × 32. The max pooling parameter was slightly modified for different trial durations to obtain the fixed output dimension. The first dimension (48) corresponded to the temporal axis and the second dimension (32) corresponded to the number of convolution kernels. The dimension of the output that mapped the EEG signals measured at different electrodes was reduced to one by the successive application of max pooling along the electrode axis.
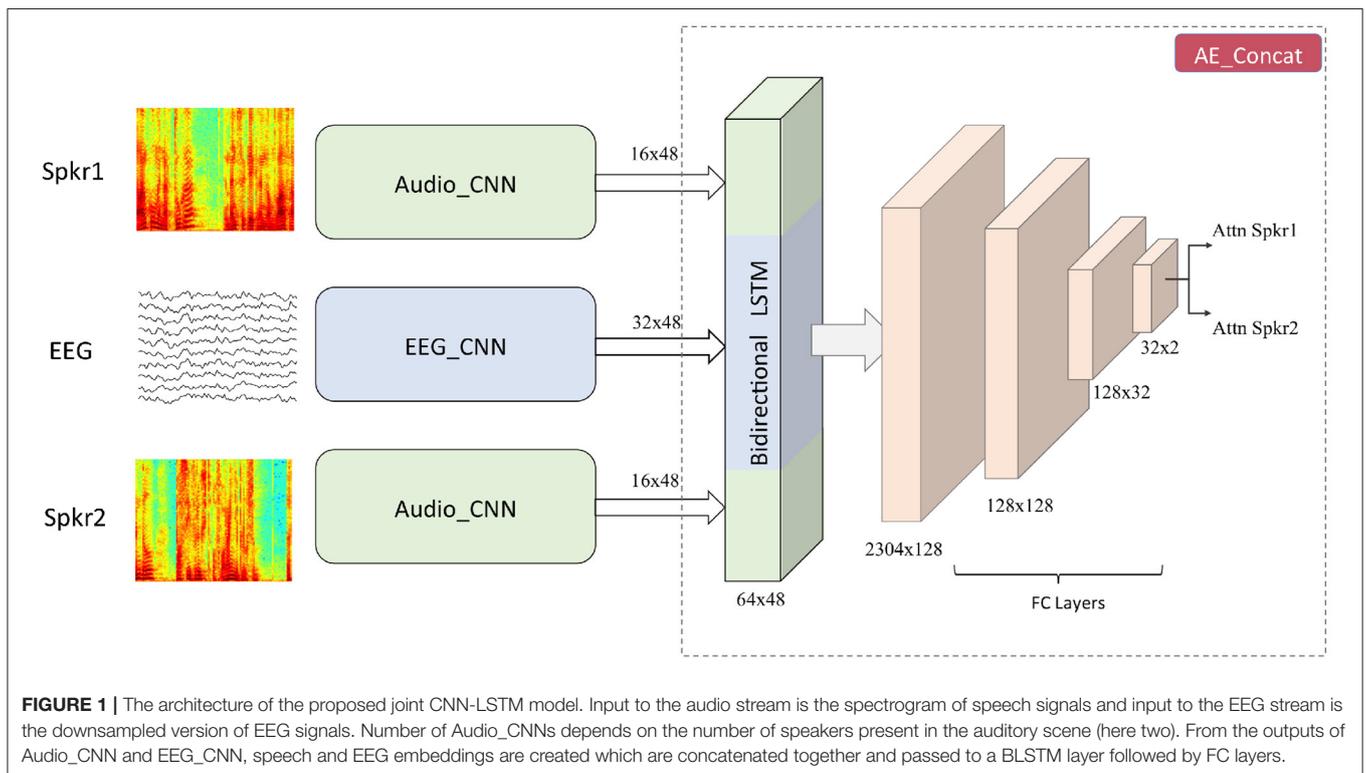
3.2. Audio_CNN

The audio subnetwork that processed the speech spectrogram consisted of five convolution layers whose parameters are shown in **Table 4**. All standard procedures such as max pooling, batch normalization, dropout, and ReLU activation were applied to the convolution output. Similar to the EEG_CNN, dimension of the input to the Audio_CNN varied according to the trial duration (**Table 2**) but the dimension of the output feature map was always

TABLE 3 | CNN parameters of the EEG subnetwork.

	Number of kernels	Kernel size	Dilation	Padding	Maxpool
Layer 1	32	24 × 1	1.1	12.0	2.1
Layer 2	32	7 × 1	2.1	6.0	1.2
Layer 3	32	7 × 5	1.1	3.2	2.5
Layer 4	32	7 × 1	1.1	3.0	1.1

TABLE 4 | CNN parameters of the Audio subnetwork.

	Number of kernels	Kernel size	Dilation	Padding	Maxpool
Layer 1	32	1 × 7	1.1	0.3	1.1
Layer 2	32	7 × 1	1.1	0.0	1.4
Layer 3	32	3 × 5	8.8	0.16	1.2
Layer 4	32	3 × 3	16.16	0.16	1.1
Layer 5	1	1 × 1	1.1	0.0	2.2

FIGURE 1 | The architecture of the proposed joint CNN-LSTM model. Input to the audio stream is the spectrogram of speech signals and input to the EEG stream is the downsampled version of EEG signals. Number of Audio_CNNs depends on the number of speakers present in the auditory scene (here two). From the outputs of Audio_CNN and EEG_CNN, speech and EEG embeddings are created which are concatenated together and passed to a BLSTM layer followed by FC layers.

fixed at 48 × 16. As the datasets considered in this study were taken from dual-speaker experiments, the Audio_CNN was run twice resulting in two sets of output.

3.3. AE_Concat

The feature maps obtained from EEG_CNN and Audio_CNN were concatenated along the temporal axis and the dimension of the feature map after concatenation was 48 × 64. In this way, we ensured that half of the feature map was contributed from the EEG data and half of the feature map was contributed from the speech data. This also provides the flexibility to extend to more than two speakers such as the experiment performed in Schäfer et al. (2018). The concatenated feature map was passed through a bidirectional long short-term memory (BLSTM) layer (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997) which was followed by four fully connected (FC) layers. For the first three FC layers, ReLU activation was used and for the last FC layer, softmax activation was applied which helps us to classify the attention to speaker 1 or speaker 2.

The total number of EEG samples and audio samples (trials) available was 118,922 and 75% of the total available samples (89,192) were used to train the network and the rest of the available samples (29,730) were equally split as validation and test data. The network was trained for 80 epochs using a mini batch size of 32 samples and with a learning rate of $5 * 10^{-4}$. The drop out probability was set to 0.25 for the EEG_CNN and the AE_Concat subnetworks but it was increased to 0.4 for the Audio_CNN subnetwork. A larger drop out probability was used for the Audio_CNN because speech stimuli were identical across subjects for a particular dataset. Hence, when trained on data from multiple subjects, the speech data remain identical and the network may remember the speech spectrogram of the training data. The network was optimized using Adam optimizer (Kingma and Ba, 2014) and the loss function used was binary cross entropy. As neural network training can result in random variations from epoch to epoch, the test accuracy was calculated as the median accuracy of the last five epochs (Goyal et al., 2017). The network was trained using an Nvidia Geforce RTX-2060 (6 GB) graphics card and took ~36 h to complete the training. The neural network model was developed in PyTorch and the python code is available at: https://github.com/ivine-GIT/joint_CNN_LSTM_AAD.

3.4. Sparse Neural Network: Magnitude Pruning

Despite neural network learning being a sophisticated algorithm, it is still not widely used in embedded devices due to the high memory and computational power requirements. Sparse neural networks have been recently proposed to overcome these challenges and enable running these models on embedded devices (Han et al., 2015). In sparse networks, majority of the model parameters are zeros and zero-valued multiplications can be ignored thereby reducing the computational requirement. Similarly, only non-zero weights need to be stored on the device and for all the zero-valued weights, only their position needs to be known reducing the memory footprint. Empirical evidences
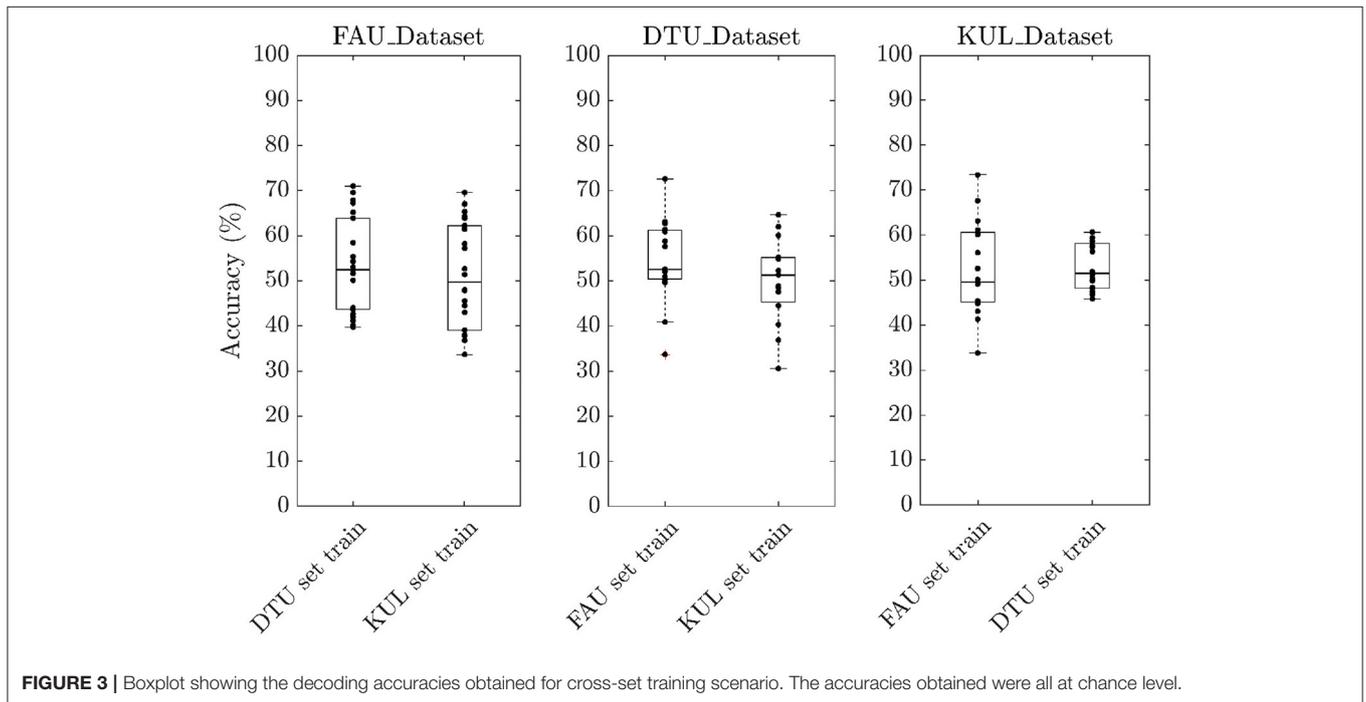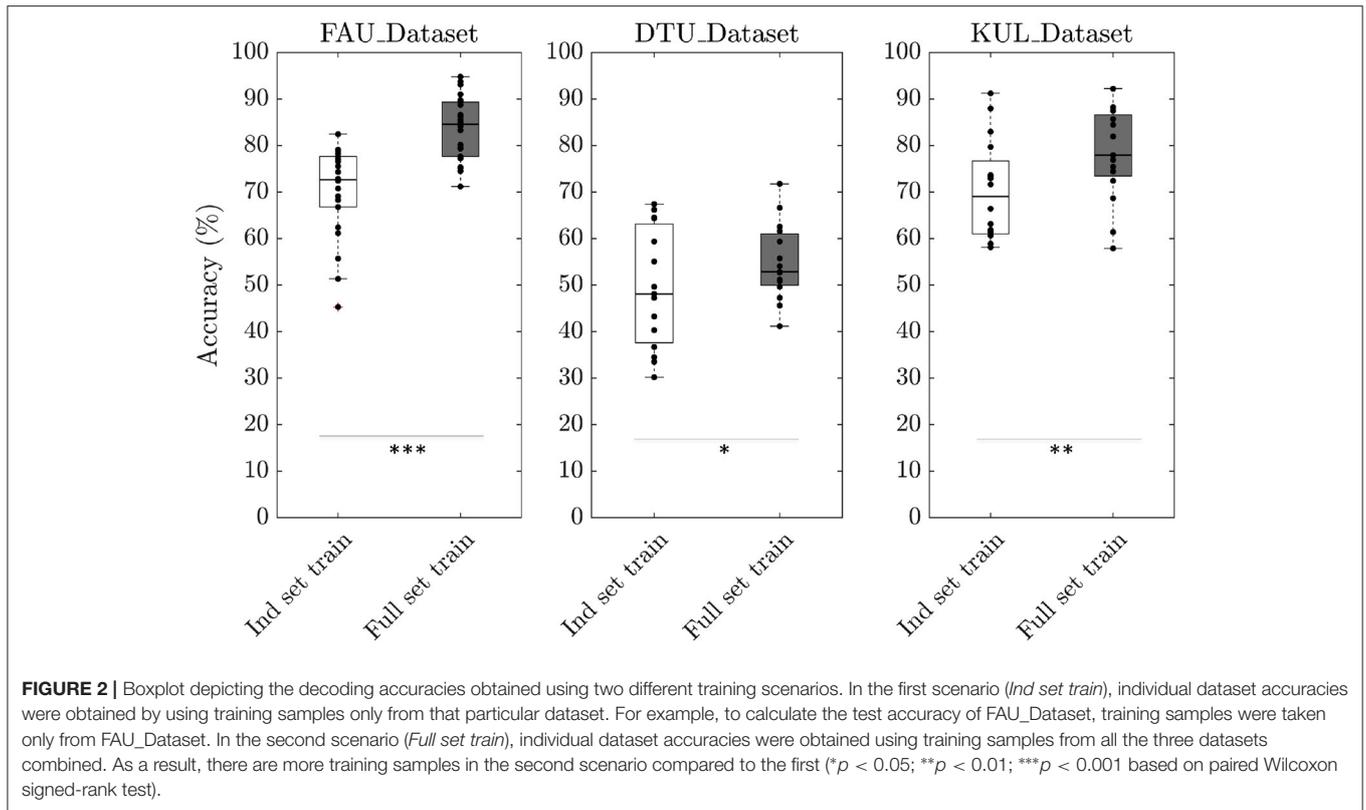
have shown that neural networks tolerate high level of sparsity (Han et al., 2015; Narang et al., 2017; Zhu and Gupta, 2017).

Sparse neural networks are found out by using a procedure known as network pruning. It consists of three steps. First, a large over-parameterized network is trained in order to obtain a high test accuracy as over-parameterization has stronger representation power (Luo et al., 2017). Second, from the trained over-parameterized network, only important weights based on certain criterion are retained and all other weights are assumed to be redundant and reinitialized to zero. Finally, the pruned network is fine-tuned by training it further using only the retained weights so as to improve the performance. Searching for the redundant weights can be based on simple criteria such as magnitude pruning (Han et al., 2015) or based on complex algorithms such as variational dropout (Molchanov et al., 2017) or L0 regularization (Louizos et al., 2017). However, it was shown that introducing sparsity using magnitude pruning could achieve comparable or better performance than complex techniques such as variational dropout or L0 regularization (Gale et al., 2019). Hence, we will present results based on only magnitude pruning in this work.

4. RESULTS

4.1. Attention Decoding Accuracy

To evaluate the performance of our neural network, we trained the model under different scenarios using a trial duration of 3 s. In the first scenario (*Ind set train*), attention decoding accuracies were calculated per individual dataset. In other words, to obtain the test accuracy of subjects belonging to FAU_Dataset, the model was trained using training samples only from FAU_Dataset leaving out DTU_dataset and KUL_Dataset. Similarly, to obtain the test accuracy for DTU_Dataset, the model was trained using training samples only from DTU_Dataset. The same procedure was repeated for KUL_Dataset. The median decoding accuracy was 72.6% for FAU_Dataset, 48.1% for DTU_Dataset, and 69.1% for KUL_Dataset (**Figure 2**). In the second scenario (*Full set train*), accuracies were calculated by combining training samples from all the three datasets together. The median decoding accuracies obtained in this scenario were 84.5, 52.9, and 77.9% for FAU_Dataset, DTU_Dataset, and KUL_Dataset, respectively. The results from the second scenario showed a clear improvement over the first scenario ($p_{-FAU} < 0.001$; $p_{-DTU} < 0.05$; $p_{-KUL} < 0.01$) suggesting that the model generalizes better in the *Full set train*. Furthermore, to evaluate the cross-set training performance, we trained the model using one dataset and tested it on the other two datasets. For example, the training would be performed using FAU_Dataset and testing would be performed on both DTU and KUL datasets. The same procedure was repeated by training using the DTU dataset and the KUL dataset. The decoding accuracies obtained were all at chance level across the three cross-set training scenarios (**Figure 3**). Consequently, all results presented further in this paper are based on *Full set train*. The statistical analyses are based on paired Wilcoxon signed-rank test with sample sizes given in **Table 1**.

FIGURE 2 | Boxplot depicting the decoding accuracies obtained using two different training scenarios. In the first scenario (*Ind set train*), individual dataset accuracies were obtained by using training samples only from that particular dataset. For example, to calculate the test accuracy of FAU_Dataset, training samples were taken only from FAU_Dataset. In the second scenario (*Full set train*), individual dataset accuracies were obtained using training samples from all the three datasets combined. As a result, there are more training samples in the second scenario compared to the first (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$ based on paired Wilcoxon signed-rank test).

FIGURE 3 | Boxplot showing the decoding accuracies obtained for cross-set training scenario. The accuracies obtained were all at chance level.

4.2. Decoding Accuracy vs. Trial Duration

To analyse the effect of trial duration on the attention decoding accuracy, the model was trained using trials of length 2, 3, 4, and 5 s. For every trial, only 1 s of new data were added and the remaining data were populated by overlapping to the previous trial using a sliding window. Specifically, for 2 s trial, 1 s of
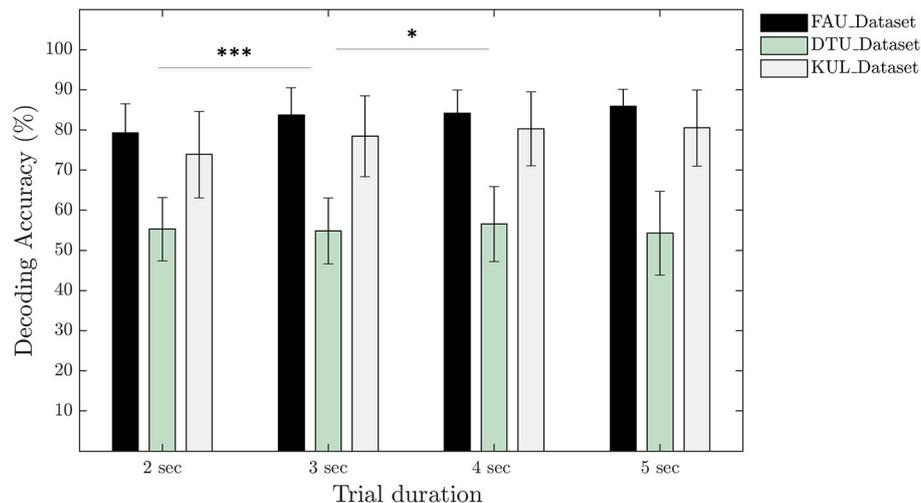
FIGURE 4 | Comparison of the decoding accuracies calculated for different trial durations per dataset. Statistical analysis based on paired Wilcoxon signed-rank test and pooled over all subjects together from the three datasets (*$p < 0.05$; ***$p < 0.001$).

overlap was used and for 3 s trial, 2 s of overlap was used, and so on. In this way, total number of training samples remained constant for different trial durations considered in our analysis. The mean decoding accuracy across all subjects and all datasets in case of 2 s trial duration was $70.9 \pm 13.2\%$. The mean accuracy improved to $73.9 \pm 14.8\%$ when the trial duration was increased to 3 s ($p < 0.001$, $r = 0.60$). Using a trial duration of 4 s, the mean accuracy obtained was $75.2 \pm 14.3\%$ which is a slight improvement over 3 s trials ($p < 0.05$, $r = 0.31$). For 5 s trials, our neural network model resulted in a mean accuracy of $75.5 \pm 15.7\%$ that was statistically identical to the accuracy obtained using 4 s trials ($p > 0.05$, $r = 0.10$). **Figure 4** depicts the accuracy calculated for individual datasets.

4.3. Ablation Analysis

In order to gain further insights into the architecture and understand the contribution of different parts of our neural network, we performed ablation analysis using a trial duration of 3 s. To this end, we modified the neural network architecture by removing specific block such as the BLSTM layer or the FC layers one at a time and retrained the modified network. Similarly, to understand the importance of the audio input feature, decoding accuracies were calculated by zeroing out the EEG input and to understand the importance of the EEG input feature, decoding accuracies were calculated by zeroing out the audio input. As shown in **Figure 5**, the median decoding accuracy by zeroing out the EEG input was 48.6% whereas zeroing out the audio input resulted in an accuracy of 53.6% resulting in no significant difference ($p > 0.05$). When the network was retrained by removing the BLSTM layer only, the median decoding accuracy obtained was 68.3% and on removing the FC layers only, median decoding accuracy was 74.7%. Hence, the BLSTM layer contributes more toward the network learning than the FC layer ($p < 0.001$). To compare, the
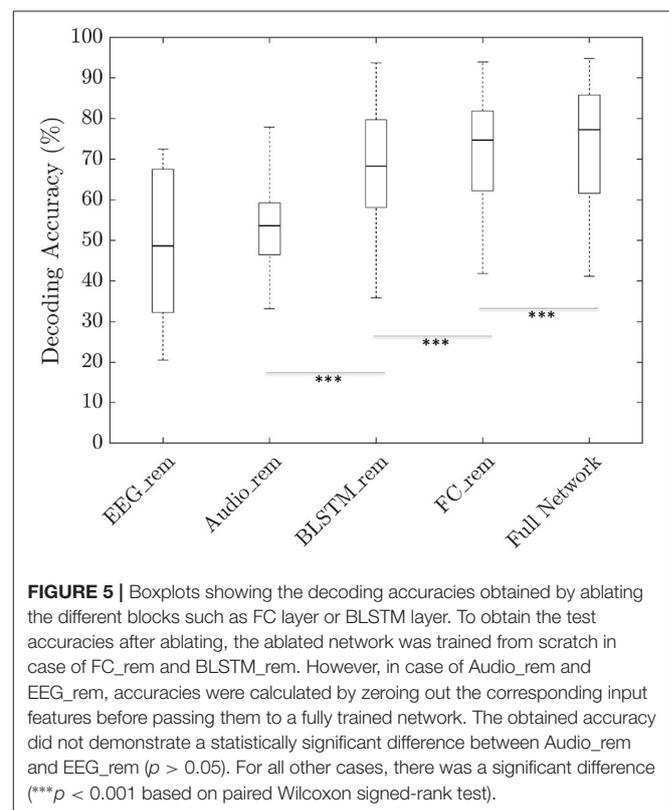
FIGURE 5 | Boxplots showing the decoding accuracies obtained by ablating the different blocks such as FC layer or BLSTM layer. To obtain the test accuracies after ablating, the ablated network was trained from scratch in case of FC_rem and BLSTM_rem. However, in case of Audio_rem and EEG_rem, accuracies were calculated by zeroing out the corresponding input features before passing them to a fully trained network. The obtained accuracy did not demonstrate a statistically significant difference between Audio_rem and EEG_rem ($p > 0.05$). For all other cases, there was a significant difference (***$p < 0.001$ based on paired Wilcoxon signed-rank test).

median decoding accuracy calculated using the full the network was 77.2%.

4.4. Sparse Neural Network Using Magnitude Pruning

To investigate the degree of sparsity that our neural network can tolerate, we pruned the model at 40, 50, 60, 70, and 80% sparsity

using the 3 s trial duration. In order to fine-tune the pruned neural network, there are two options: (1) sequential or (2) one-shot. In sequential fine-tuning, weights of the trained original model are reinitialized to zero in smaller steps per epoch until the required sparsity is attained. In one-shot fine-tuning, weights of the trained original model are reinitialized to zero at one shot in the first epoch and the sparse model is further trained to improve performance. We observed that the sequential fine-tuning is less efficient than one-shot fine-tuning in terms of training time budget. Therefore, all results presented here are based on one-shot fine-tuning. We achieved a median decoding accuracy of 76.9% at a sparsity of 40% which is statistically identical to the original model at 77.2% ($p > 0.05$). When the sparsity was increased to 50%, the median decoding accuracy decreased to 75.7% which was lower than the original model ($p < 0.001$). Increasing the sparsity level further resulted in deterioration of decoding accuracy reaching 63.2% at a sparsity of 80% (**Figure 6**). Total number of learnable parameters in our model was 416,741 and to find the sparse network, we pruned only the weights leaving the bias and BN parameters unchanged.

5. DISCUSSION

People with hearing loss suffer from deteriorated speech intelligibility in noisy acoustic environments such as multispeaker scenarios. Increasing the audibility by means of hearing aids has not shown to provide sufficient improvement to the speech intelligibility. This is because the hearing aids are unable to estimate *apriori* to which speaker the user intends to listen. Hence, hearing aids amplify both the wanted signal (attended speaker) and interfering signals (ignored speakers). Recently, it has been shown that the cortical signals measured

FIGURE 6 | Plots comparing the trade off between decoding accuracies and sparsity level (*$p < 0.05$; ***$p < 0.001$ based on paired Wilcoxon signed-rank test).

using EEG could infer the auditory attention by discriminating between the attended speaker and the ignored speaker in a dual-speaker scenario (O'Sullivan et al., 2014). Linear system analysis has been the commonly used methodology to analyse the EEG signals measured from a listener performing selective attention. However, in recent years, non-linear analyses based on neural networks have become prominent, thanks to the availability of customized hardware accelerators and associated software libraries.

In this work, we developed a joint CNN-LSTM model to infer the auditory attention of a listener in a dual-speaker environment. CNNs take the EEG signal and spectrogram of the multiple speakers as inputs and extract features through successive convolutions. These convolutions generate an intermediate embeddings of the inputs which are then given to a BLSTM layer. As LSTMs fall under the category of recurrent neural networks, they can model the temporal relationship between the EEG embedding and the multiple spectrogram embeddings. Finally, the output of the BLSTM is processed through FC layers to infer the auditory attention. The effectiveness of the proposed neural network was evaluated with the help of three different EEG datasets collected from subjects who undertook dual-speaker experiment.

There are many choices for the acoustic cues of speech signal that could be given as input to the neural network. They are speech onsets (Howard and Poeppel, 2010), speech envelopes (Aiken and Picton, 2008), speech spectrograms (Pasley et al., 2012), or phonemes (Di Liberto et al., 2015). Due to the hierarchical processing of speech, all of the aforementioned cues could be tracked from the cortical signals measured using EEG (Hickok and Poeppel, 2007; Ding and Simon, 2014). Speech envelope is the most commonly used acoustic cues in the linear system analysis of EEG signal. However, we decided to use spectrogram due to its rich representational power of the corresponding speech signal and the ability of neural networks to index these multidimensional inputs efficiently.

5.1. Attention Decoding Accuracy
We analyzed the performance of our neural network in two different training scenarios. In the first scenario, individual dataset accuracy was found out by training the network using samples taken only from that particular dataset. In the second scenario, individual dataset accuracy was found out by training using samples combined from all three datasets together. The accuracies obtained in the second scenario were higher than the first scenario by 10.8% on average, which is in agreement with the premise of neural network learning that larger the amount of training data, the better the generalization. The decoding accuracies obtained for subjects belonging to the DTU_Dataset were markedly lower than the other two datasets similar to the observation made in Geirnaert et al. (2020). While the exact reason for the lower performance is unclear, a major difference of the DTU_Dataset compared to the other two datasets was that the former consisted of attention to male and female speakers whereas the latter consisted of attention to only male speakers. Therefore, training with additional EEG data that consist of attention to female speakers can provide more insights into
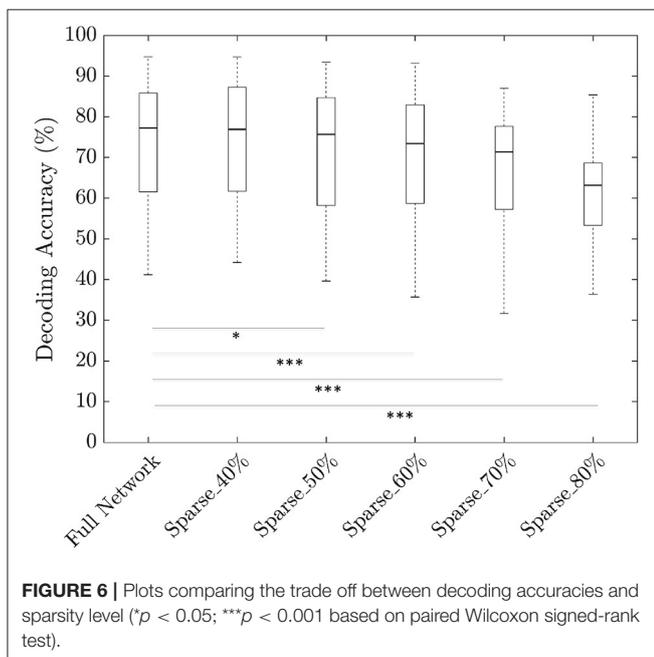
the lower performance. Additionally, we investigated the cross-set performance by training the model using one dataset and testing using the other two datasets. The accuracies obtained were all at chance level as seen in **Figure 3**. This is not against our expectation because if the underlying training set is not representative, neural networks will not generalize. Specifically, features in the training set and the test set are different since they were recorded in different audio settings, languages, and EEG devices. This further affirms the importance of having a large and diverse training set for the neural networks to function efficiently.

5.2. Decoding Accuracy vs. Trial Duration

One of the major challenges that AAD algorithms based on linear system theory faces is the deteriorated decoding performance when the trial duration is reduced. To this end, we calculated the accuracies using our neural network for different trial durations of 2, 3, 4, and 5 s. We observed a clear performance improvement when trial duration was increased from 2 to 3 s whereas for all other trial durations, accuracies did not improve substantially (**Figure 4**). However, increasing the trial duration will result in larger latency needed to infer the auditory attention that can adversely affect applications which require real-time operation. Hence, 3 s trial duration may be an optimal operation point as it is known from a previous study that human brain tracks the sentence phrases and phrases are normally not longer than 3 s (Vander Ghinst et al., 2019). Similarly, our analysis made use of 10 electrodes distributed all over the scalp but future work should investigate the effect of reducing the number of electrodes. This will help in integrating algorithms based on neural networks into devices such as hearing aids. We anticipate that the current network will require modifications with additional hyperparameter tuning in order to accommodate for the reduction in number of electrodes, as the fewer is the number of electrodes, the lower is the amount of data available for training.

5.3. Ablation Analysis

Performing ablation analysis gives the possibility to evaluate the contribution of different inputs and modules in a neural network. To our model, when only the speech features were given as input, the median decoding accuracy was 48.6% whereas only EEG features as input resulted in an accuracy of 53.6% (**Figure 5**). However, statistical analysis revealed that there is no significant difference between the two. This is contrary to our anticipation because we expected the model to learn more from the EEG features than from the audio features, as the EEG signal is unique to the subject while the audio stimulus was repeated among subjects per dataset. Nevertheless, in future care must be taken to design the experiment in such a way as to incorporate diverse speech stimuli. Further analysis ablating the BLSTM layer and the FC layers revealed that the BLSTM layer was more important than the FC layers. This is probably due to the ability of the LSTM layer to model the temporal delay between speech cues and the EEG. However, we anticipate that when the training datasets become larger and more dissimilar, FC layers will become more

important due to the improved representation and optimization power of dense networks (Luo et al., 2017).

5.4. Sparse Neural Networks

Although neural networks achieve state-of-the-art performances for a wide range of applications, they have large memory footprint and require extremely high computation power. Over the years, neural networks were able to extend their scope of applications was by scaling up the network size. In 1998, the CNN model (LeNet) that was successful in recognizing handwritten digits consisted of under a million parameters (LeCun et al., 1998), whereas AlexNet that won the ImageNet challenge in 2012 consisted of 60 million parameters (Krizhevsky et al., 2017). Neural networks were further scaled up to the order of 10 billion parameters and efficient methods to train these extremely large networks were presented in Coates et al. (2013). While these large models are very powerful, running them on embedded devices poses huge challenges due to the large memory and computation requirements. Sparse neural networks are a novel architecture search where redundant weights are reinitialized to zero thereby reducing the computation load.

Investigation into the amount of sparsity that our neural network can tolerate revealed a tolerance of upto 50% sparsity without substantial loss of accuracy (**Figure 6**). However, standard benchmarking on sparsity has found that deep networks such as ResNet-50 can tolerate upto 90% sparsity (Gale et al., 2019). One of the potential reasons for the lower level of sparsity in our model is due to its shallow nature. That is, our model is comprised of less than half a million learnable parameters while deep networks such as ResNet-50 is comprised of over 25 million learnable parameters. It is also interesting to note that the accuracy obtained by removing the FC layer in our ablation analysis was 74.6% compared to the full network accuracy of 77.2%. And the ablated network consisted of 105,605 parameters which is approximately only a quarter of the total number of parameters (416,741) of the original network. This shows that by careful design choices, we can reduce the network size considerably compared to an automatic sparse network search using magnitude pruning.

Sparsification of neural network has also been investigated as a neural network architecture search rather than merely as an optimization procedure. In the lottery ticket hypothesis presented in Frankle and Carbin (2018), authors posit that, inside the structure of an over-parameterized network, there exist subnetworks (winning tickets) that when trained in isolation reaches accuracies comparable to the original network. The prerequisite to achieve comparable accuracy is to initialize the sparse network using the original random weight initialization that was used to obtain the sparse architecture. However, it was shown that with careful choice of the learning rate, the stringent requirement on original weight initialization can be relaxed and the sparse network can be trained from scratch for any random initialization (Liu et al., 2018).

One of the assumptions that we have made throughout this paper is the availability of clean speech signal to obtain the spectrogram. In practice, only noisy mixtures are available and speech sources must be separated before the spectrogram can

be calculated. This is an active research field and algorithms are already available based on classical signal processing such as beamforming or based on deep neural networks (Wang and Chen, 2018). Another challenge in neural network learning and in particular, its application in EEG research is the scarcity of labeled data to train the network. This limits the ability of network to generalize well to unseen EEG data. To mitigate the aforementioned limitation, data augmentation techniques are widely used in neural network training. Data augmentation is a procedure to generate synthetic dataset that spans unexplored input signal space but corresponding to the true labels (Wen et al., 2020). In auditory attention paradigm, linear system analyses have shown that the TRF properties differ between attended and ignored speakers (Fiedler et al., 2019; Kuruvila et al., 2021). As a result, synthetic EEG can be generated by performing a linear convolution between TRFs and the corresponding speech signal cues (Miran et al., 2018). The signal-to-noise ratio of the synthesized EEG can be varied by adding appropriate noise to the convolved signal. The most commonly used speech cue is the signal envelope obtained using Hilbert transform. However, more sophisticated envelope extraction methods such as the computational models simulating the auditory system could improve the quality of synthesized EEG signals (Kates, 2013; Verhulst et al., 2018). It must be noted that the data augmentation techniques must only be used to train the network. The validation and the testing procedure must still be performed using real datasets.

6. CONCLUSION

Integrating EEG to track the cortical signals is one of the latest proposals to enhance the quality of service provided by hearing aids to the users. EEG is envisaged to provide neuro-feedback about the user's intention thereby enabling the hearing aid to infer and enhance the attended speech signals. In the present study, we propose a joint CNN-LSTM network to classify the attended speaker in order to infer the auditory attention of a listener. The proposed neural network uses speech spectrograms and EEG signals as inputs to infer the auditory attention. Results obtained by training the network using three different EEG datasets collected from multiple subjects who undertook a dual-speaker experiment showed that our network generalizes well to different scenarios. Investigation into the importance of different constituents of our network architecture revealed that adding an LSTM layer improved the performance of the model considerably. Evaluating sparsity on the proposed joint CNN-LSTM network demonstrates that the network can tolerate upto 50% sparsity without considerable deterioration in performance. These results could pave way to integrate algorithms based on neural networks into hearing aids that have constrained memory and computational power.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics committee, University of Erlangen-Nuremberg (Protocol Number: 314_18B) on 18th September 2018. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

UH and EF conceived the study and contributed to data analysis, supervision, and manuscript writing. JM contributed to the results interpretation and data analysis. IK is the main author and performed most of the data analysis. All authors contributed to the article and approved the submitted version.

REFERENCES

Aiken, S. J., and Picton, T. W. (2008). Human cortical responses to the speech envelope. *Ear Hear*. 29, 139–157. doi: 10.1097/AUD.0b013e31816453dc

Biesmans, W., Das, N., Francart, T., and Bertrand, A. (2017). Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario. *IEEE Trans. Neural Syst. Rehabil. Eng*. 25, 402–412. doi: 10.1109/TNSRE.2016.2571900

Broderick, M. P., Anderson, A. J., and Lalor, E. C. (2019). Semantic context enhances the early auditory encoding of natural speech. *J. Neurosci*. 39, 7564–7575. doi: 10.1523/JNEUROSCI.0584-19.2019

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am*. 25, 975–979. doi: 10.1121/1.1907229

Ciccarelli, G., Nolan, M., Perricone, J., Calamia, P. T., Haro, S., O'Sullivan, J., et al. (2019). Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods. *Sci. Rep*. 9, 1–10. doi: 10.1038/s41598-019-47795-0

Coates, A., Huval, B., Wang, T., Wu, D., Catanzaro, B., and Andrew, N. (2013). "Deep learning with COTS HPC systems," in *International Conference on Machine Learning* (Atlanta, GA), 1337–1345.

Craik, A., He, Y., and Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *J. Neural Eng*. 16:031001. doi: 10.1088/1741-2552/ab0ab5

Das, N., Biesmans, W., Bertrand, A., and Francart, T. (2016). The effect of head-related filtering and ear-specific decoding bias on auditory attention detection. *J. Neural Eng*. 13:056014. doi: 10.1088/1741-2560/13/5/056014

Das, N., Francart, T., and Bertrand, A. (2019). Auditory attention detection dataset KULeuven (Version 1.0.0) [Data set]. *Zenodo*. doi: 10.5281/zenodo.3377911

de Taillez, T., Kollmeier, B., and Meyer, B. T. (2020). Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech. *Eur. J. Neurosci.* 51, 1234–1241. doi: 10.1111/ejn.13790

Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.* 25, 2457–2465. doi: 10.1016/j.cub.2015.08.030

Ding, N., and Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.* 107, 78–89. doi: 10.1152/jn.00297.2011

Ding, N., and Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Front. Hum. Neurosci.* 8:311. doi: 10.3389/fnhum.2014.00311

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., et al. (2018). Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*. doi: 10.1145/3197517.3201357

Feldman, J. (2003). What is a visual object? *Trends Cogn. Sci.* 7, 252–256. doi: 10.1016/S1364-6613(03)00111-6

Fiedler, L., Woestmann, M., Graversen, C., Brandmeyer, A., Lunner, T., and Obleser, J. (2017). Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. *J. Neural Eng.* 14:036020. doi: 10.1088/1741-2552/aa66dd

Fiedler, L., Wöstmann, M., Herbst, S. K., and Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *NeuroImage* 186, 33–42. doi: 10.1016/j.neuroimage.2018.10.057

Frankle, J., and Carbin, M. (2018). The lottery ticket hypothesis: finding sparse, TRainable neural networks. *arXiv preprint arXiv:1803.03635*.

Fuglsang, S. A., Dau, T., and Hjortkjær, J. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage* 156, 435–444. doi: 10.1016/j.neuroimage.2017.04.026

Fuglsang, S. A., Wong, D. D. E., and Hjortkjær, J. (2018). EEG and audio dataset for auditory attention decoding (Version 1) [Data set]. *Zenodo*. doi: 10.5281/zenodo.1199011

Gale, T., Elsen, E., and Hooker, S. (2019). The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*.

Geirnaert, S., Francart, T., and Bertrand, A. (2019). An interpretable performance metric for auditory attention decoding algorithms in a context of neuro-steered gain control. *IEEE Trans. Neural Syst. Rehabil. Eng.* 28, 307–317. doi: 10.1101/745695

Geirnaert, S., Vandecappelle, S., Alickovic, E., de Cheveigné, A., Lalor, E., Meyer, B. T., et al. (2020). Neuro-steered hearing devices: decoding auditory attention from the brain. *arXiv preprint arXiv:2008.04569*. doi: 10.1109/MSP.2021.3075932

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., et al. (2017). Accurate, large minibatch SGD: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Han, S., Pool, J., Tran, J., and Dally, W. (2015). Learning both weights and connections for efficient neural network. *Adv. Neural Inform. Process. Syst.* 28, 1135–1143.

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Howard, M. F., and Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J. Neurophysiol.* 104, 2500–2511. doi: 10.1152/jn.00251.2010

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning* (Lille: PMLR), 448–456.

Kates, J. (2013). "An auditory model for intelligibility and quality predictions," in *Proceedings of Meetings on Acoustics ICA2013, Vol. 19* (Montréal, QC: Acoustical Society of America), 050184. doi: 10.1121/1.4799223

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Kuruvila, I., Demir, K. C., Fischer, E., and Hoppe, U. (2021). Inference of the selective auditory attention using sequential LMMSE estimation. *IEEE Trans. Biomed. Eng.* doi: 10.1109/TBME.2021.3075337

Kuruvila, I., Fischer, E., and Hoppe, U. (2020). "An LMMSE-based estimation of temporal response function in auditory attention decoding," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Montréal, QC), 2837–2840. doi: 10.1109/EMBC44109.2020.9175866

Lalor, E. C., and Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *Eur. J. Neurosci.* 31, 189–193. doi: 10.1111/j.1460-9568.2009.07055.x

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Liu, Z., Sun, M., Zhou, T., Huang, G., and Darrell, T. (2018). Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*.

Louizos, C., Welling, M., and Kingma, D. P. (2017). Learning sparse neural networks through L_0 regularization. *arXiv preprint arXiv:1712.01312*.

Luo, J.-H., Wu, J., and Lin, W. (2017). "Thinet: a filter level pruning method for deep neural network compression," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 5058–5066. doi: 10.1109/ICCV.2017.541

Mesgarani, N., and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485, 233–236. doi: 10.1038/nature11020

Miran, S., Akram, S., Sheikhattar, A., Simon, J. Z., Zhang, T., and Babadi, B. (2018). Real-time tracking of selective auditory attention from M/EEG: a Bayesian filtering approach. *Front. Neurosci.* 12:262. doi: 10.3389/fnins.2018.00262

Mirkovic, B., Debener, S., Jaeger, M., and De Vos, M. (2015). Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. *J. Neural Eng.* 12:046007. doi: 10.1088/1741-2560/12/4/046007

Molchanov, D., Ashukha, A., and Vetrov, D. (2017). Variational dropout sparsifies deep neural networks. *arXiv preprint arXiv:1701.05369*.

Monesi, M. J., Accou, B., Montoya-Martinez, J., Francart, T., and Van Hamme, H. (2020). "An LSTM based architecture to relate speech stimulus to EEG," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona), 941–945. doi: 10.1109/ICASSP40776.2020.9054000

Narang, S., Elsen, E., Diamos, G., and Sengupta, S. (2017). Exploring sparsity in recurrent neural networks. *arXiv preprint arXiv:1704.05119*.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* 25, 1697–1706. doi: 10.1093/cercor/bht355

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biol.* 10:e1001251. doi: 10.1371/journal.pbio.1001251

Schäfer, P. J., Corona-Strauss, F. I., Hannemann, R., Hillyard, S. A., and Strauss, D. J. (2018). Testing the limits of the stimulus reconstruction approach: auditory attention decoding in a four-speaker free field environment. *Trends Hear.* 22:2331216518816600. doi: 10.1177/2331216518816600

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681. doi: 10.1109/78.650093

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends Cogn. Sci.* 12, 182–186. doi: 10.1016/j.tics.2008.02.003

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Tian, Y., and Ma, L. (2020). Auditory attention tracking states in a cocktail party environment can be decoded by deep convolutional neural networks. *J. Neural Eng.* doi: 10.1088/1741-2552/ab92b2

Vandecappelle, S., Deckers, L., Das, N., Ansari, A. H., Bertrand, A., and Francart, T. (2021). EEG-based detection of the locus of auditory attention with convolutional neural networks. *bioRxiv [Preprint]*. doi: 10.7554/eLife.56481.sa2

Vander Ghinst, M., Bourguignon, M., Niesen, M., Wens, V., Hassid, S., Choufani, G., et al. (2019). Cortical tracking of speech-in-noise develops from childhood to adulthood. *J. Neurosci.* 39, 2938–2950. doi: 10.1523/JNEUROSCI.1732-18.2019

Verhulst, S., Altoé, A., and Vasilkov, V. (2018). Computational modeling of the human auditory periphery: auditory-nerve responses, evoked potentials and hearing loss. *Hear. Res.* 360, 55–75. doi: 10.1016/j.heares.2017.12.018

Vogel, E. K., Woodman, G. F., and Luck, S. J. (2005). Pushing around the locus of selection: evidence for the flexible-selection hypothesis. *J. Cogn. Neurosci.* 17, 1907–1922. doi: 10.1162/089892905775008599

Wang, D., and Chen, J. (2018). Supervised speech separation based on deep learning: an overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* 26, 1702–1726. doi: 10.1109/TASLP.2018.2842159

Wen, Q., Sun, L., Song, X., Gao, J., Wang, X., and Xu, H. (2020). Time series data augmentation for deep learning: a survey. *arXiv preprint arXiv:2002.12478.*

Zhu, M., and Gupta, S. (2017). To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878.*

Zwicker, E., and Fastl, H. (2013). *Psychoacoustics: Facts and Models, Vol. 22.* Berlin: Springer Science & Business Media.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.