Detection of Imagery Vowel Speech Using Deep Learning

Jigar Patel and Syed Abudhagir Umar

1 Introduction

Vocal speech and physical gestures are typical ways for humans to communicate with each other. However, these normal means of communications are not possible for some individuals due to some special conditions such as advanced stage of amyotrophic lateral sclerosis (ALS) or locked-in syndrome (CIS). Brain–computer interface (BCI) system can translate brain activities into computer commands, and hence, allow an individual to communicate using their brain activities. Recent developments in the area of BCI have resulted in multiple systems that can be helpful to LIS of advanced ALS patients [1–4].

To capture the activity of the brain, electroencephalography (EEG) one of the most popular techniques. EEG records brain activity in the form of electrical signals (which are produced as a result of underlying brain activities) by placing electrodes on the scalp surface. The reasons behind EEG being the most popular technique to be used in the BCI system are its cost-effectiveness and simplicity in the acquisition of data. EEG can monitor minute changes in the brain activity due to high temporal resolution.

The EEG signals vary based on the type of brain activities. These activities can be differentiated into several types such as event-related synchronization/event-related desynchronization (ERD/ERS) [5], steady-state visual evoked potential (SSVEP) [6], mental task [7], and P300 potentials [8]. These types are called paradigms in BCI systems. The BCI systems based on SSVEP and P300 paradigm present visual stimuli to the subject and upon focusing on them, and the brain will produce specific signals

J. Patel (✉) · S. A. Umar
Department of Electronics and Communication Engineering, B V Raju Institute of Technology, Narsapur, Medak District, Telangana, India
e-mail: jigar.patel@bvrit.ac.in

S. A. Umar
e-mail: syedabudhagir.u@bvrit.ac.in

R. C. Bansal et al. (eds.), *Advances in Energy Technology*, Lecture Notes in Electrical Engineering 766, https://doi.org/10.1007/978-981-16-1476-7_23

which can be distinguished. These approaches are mainly used in developing speller systems [9–11]. The ERD/ERS and mental task-based BCI systems require subject to the image or perform some mental activities to produce brain signals which can be detected. The major advantage of this type of paradigm is that any external stimuli are not required and the user can perform a mental activity at his/her pace. These paradigms are mainly used in controlling prosthetics, wheelchairs, etc. [12, 13].

Apart from these four paradigms, there is another paradigm that is based on imagined speech. This paradigm has received comparatively less attention than other paradigms. Earlier studies report that the signals are produced in the motor cortex area of the brain while imagining vowels [14, 15]. These signals can be called speech-related potentials (SAP) and can be used to develop speech prostheses.

An earlier study implemented an imaginary vowel speech paradigm classify imagery vowels 'u' and 'a' using spatial filters and support vector machines [16]. They reported classification accuracy from 68 to 78%. Similar studies are based on the same dataset reported accuracies of 70–80% [17, 18]. A similar approach was used to classify imagined pronunciations of Chinese characters 'zuo' and 'yi'. They reported average classification accuracies from 74 to 95% across eight subjects [19]. The more recent studies implement detections of imaginary speech of five vowels and words. To detect these imagined activities, features based on the time domain, Mel-frequency cepstral coefficients along with the classifiers such as support vector machine [20–22]. A home automation system was also implemented based on the hybrid approach of imagined speech and event-related potentials [23].

Recently, many researchers focus on developing deep learning methods for classification-based tasks. The most popular architectures for deep learning-based models are convolution neural network (CNN) and long short-term memory (LSTM). Deep learning-based approaches using CNN were also implemented to detect imagery speech with five vowels and six words [24, 25].

In this study, three different approaches based on deep learning architectures using CNN and LSTM are proposed to classify between imagined speech containing two vowels 'a' and 'u' along with no activity as a control task.

2 Data Acquisition

The EEG data utilized in this study was acquired from the 'speech imagery dataset' from 'https://www.brainliner.jp' and was recorded by Dasalla et al. [16]. The EEG data was collected using BioSemi Active Two system manufactured by BioSemi B. V., Amsterdam, Netherland. The EEG was recorded with a sampling rate of 2048 Hz using Ag–AgCl electrodes. To reduce the size, the data has been down-sampled to 256 Hz of sampling rate. Also, to remove electronic noise and low-frequency baseline shifts, a zero-phase bandpass filter with the range 1–45 Hz was applied to the recorded data. The data was recorded from the locations Fz, C3, C4, and Cz of the international 10–20 system for EEG.

2.1 Subjects

The data was recorded from three subjects, two males and one female (right-handed) with the ages from 26 to 29 years. The subjects did not report any health problems or neurological disorders and were fluent in English.

2.2 Experimental Procedure

The subjects were settled in a comfortable chair and asked to perform one of the three tasks based on a visual stimulus. The three tasks were

1. Vowel 'a': Imagine vocalization of 'a'
2. Vowel 'u': Imagine vocalization of 'u'
3. Rest 're': no action, rest.

The subjects were trained beforehand in addition to the rehearsal with real movements to ensure correct task execution. The visual cue will be displayed on the monitor place at approx. 1 m away from the subject. The experiment consists of multiple trials. A fixation cross is placed in the screen for a duration between two to three seconds to help the subject focus on the upcoming tasks on remove any potentials from the previous task. The trial begins with a beep sound followed by a visual cue that is randomly selected. The cue will be displayed for two seconds on the screen followed by a blank screen for three seconds. The participants were instructed to perform appropriate tasks for the duration the cue is displayed on the screen. Each task is repeated 50 times and each task is considered as a trial. So, each subject performed total 150 tasks or trials. The experiment procedure is depicted in Fig. 1.

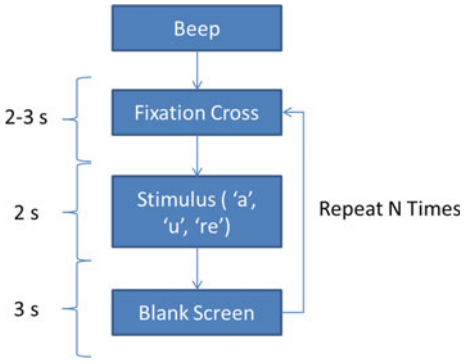


Fig. 1 Experiment task

3 Methodology

There are three deep learning models implemented in this study. These three models are the CNN model, the LSTM model, and the CNN-LSTM model. The data was not pre-processed apart from the bandpass filter with range 1–45 Hz which was applied on the dataset beforehand. The hyper-parameters β_1 and β_2 were set to 0.9 and 0.999 for all three models. The hyper-parameters learning rate and decay rate were tuned to obtain the best performance from the network. All three networks were trained using Adam optimizer. The 70% of data was used for training the models and remaining 30% data was used for testing. The data was randomized before the split to avoid any training bias.

3.1 CNN Model

Convolution neural networks have been used extensively in implementing image processing methods such as object detection and face recognition. In this study, a similar model is implemented to classify imaginary vowel pronunciations.

The model can be divided into two sections: convolution layers and dense or fully connected layers. The convolution layers are mainly utilized for extracting features from the data while the dense layers will be used to map the patterns and classify the data into distinct groups.

There are three convolution layers in the proposed model, and the one-dimension convolutions are used in this model as the input data consists of EEG data series. A batch normalization layer was inserted after each of the convolution layers. The three dense layers of fully connected layers were connected to the last convolution layer after applying to flatten. Figure 2 illustrates the architecture of the CNN model.

The kernel size for the convolution was set at 10 for all the layers. Reducing or increasing this kernel size resulting in the dropping of the classification accuracy. The number of filters used in the convolution layers is 64, 64, and 128 for the respective convolution layers. The dense layers contain 128 and 32 nodes followed by the output layer with 3 node. The model contains dropout layers with a coefficient of 0.3 after each convolution blocks and dense layer blocks to reduce overfitting.

3.2 LSTM Model

The LSTM model is seen as more effective compared to feed-forward or CNN models in terms of sequence prediction. An LSTM model is usually implemented to process and classify time-series or sequence data. The proposed model contains two LSTM blocks followed by output block. Each LSTM block is followed by the batch normalization block. The diagram in Fig. 3 depicts the model architecture.

Fig. 2 CNN model

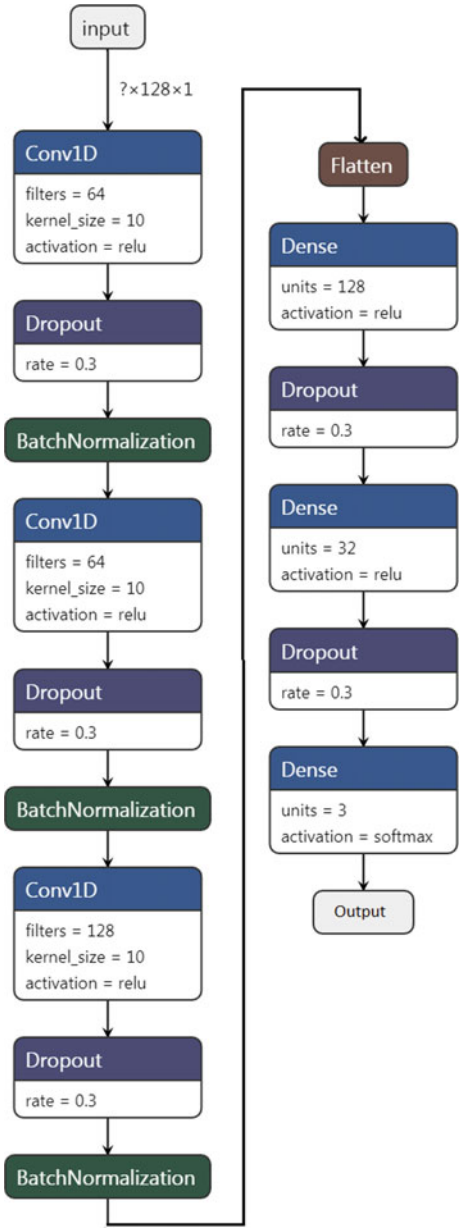
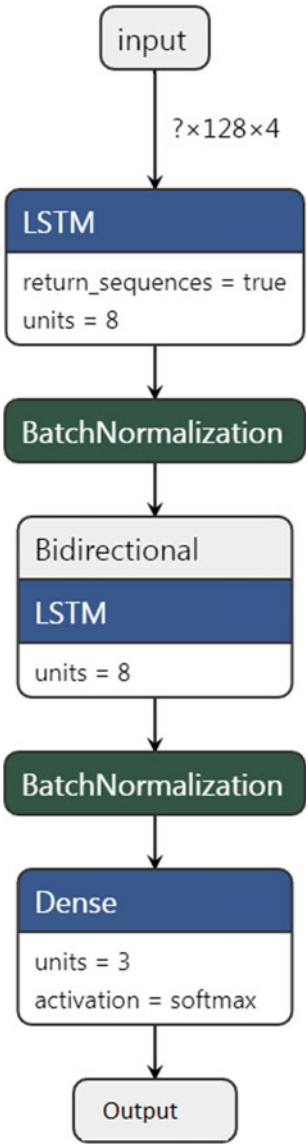


Fig. 3 LSTM model



3.3 CNN + LSTM Model

The proposed model contains a single CNN block followed by batch normalization, a bidirectional LSTM block, and an output layer. The 'softmax' activation method was used at the output layer. The convolution layer consists of 32 filters with a kernel size of 16. The LSTM layer is consists of 16 units. Both CNN and LSTM blocks are

followed by batch normalization blocks. Figure 4 illustrates the architecture of the proposed model.

4 Results and Discussion

Here, we report the performance of all three proposed models applied to detect imaginary vowel pronunciations. The first model we trained and evaluated is CNN model. CNN was proved to be very successful in classifying images and detecting objects in images. However, in this scenario, the CNN model performed with the average classification accuracy of 51% across all the subjects. The model was trained with the hyper-parameters learning rate = 0.0001, decay rate = 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The models were trained for 20 epochs. The model was found to be overfitting the data even after adding dropout layers. The training and testing classification accuracies and loss for subject 'S1' are depicted in Fig. 5.

The second model which was trained and evaluate is the LSTM model. The LSTM models have been known to be performing well for the time-series of sequence classification tasks. In this EEG signal classifications, the LSTM model performed better than the CNN model with the average classification accuracy of 63%. The hyper-parameters used to train the model were learning rate of 0.001, β_1 and β_2 values same as CNN model, 0.9 and 0.999, respectively, and decay rate of 0.001. The total number of epochs was 50 with batch size being 10. The classification accuracies and loss for train and test data are illustrated in Fig. 6.

The last model to train and evaluate is the combination of CNN and the LSTM model. The model was trained with the same β_1 and β_2 values as per the previous model. The learning rate and decay rates were set to 0.0001 and 0.001, respectively. The batch size was set to 10 and the number of epochs is set to 50. This model proved to be better performing than the previous models with the average classification accuracy of 84% across three subjects (Fig. 7).

The subject-wise classification accuracies are presented in Table 1. The overall performance of CNN + LSTM was best among all three proposed models as discussed earlier. Among the subjects, the data from the subject 'S3' yielded the best classification accuracy for all the models across subjects.

5 Conclusion

In this study, three deep learning models to classify the EEG data belonging to the imagined activities of pronouncing vowels 'a', 'u', and no activity have been proposed. The multiclass classification models are based on deep learning architectures and that based CNN, LSTM, and the combination of both is implemented to classify brain activity. The data used for the analysis was not pre-processed in any

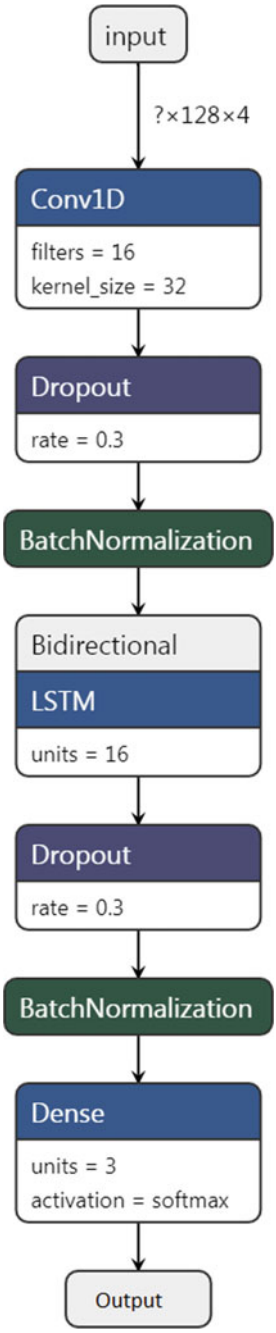


Fig. 4 CNN + LSTM model

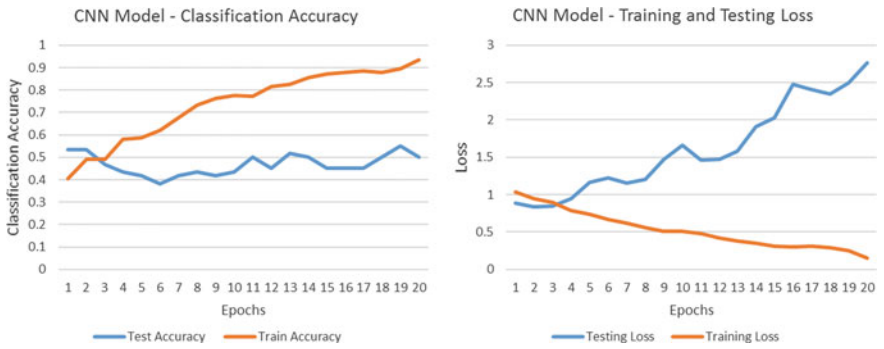


Fig. 5 CNN model—classification accuracy and loss

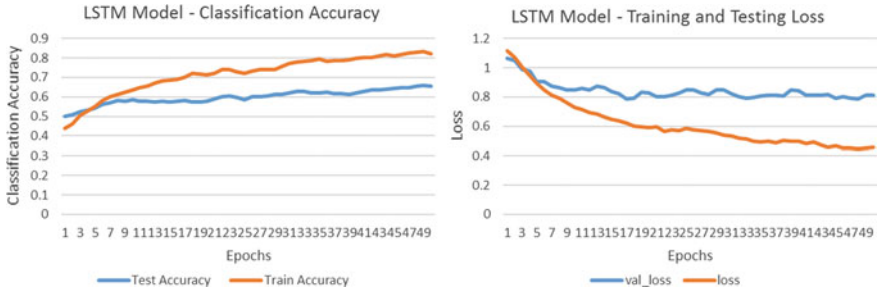


Fig. 6 LSTM model—classification accuracy and loss

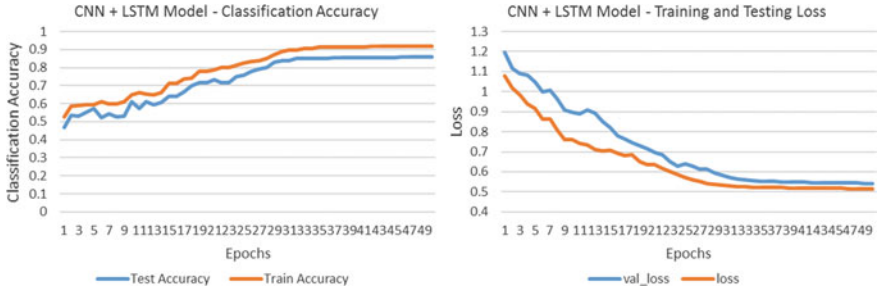


Fig. 7 CNN + LSTM model—classification accuracy and loss

Table 1 Subject-wise classification accuracies

Subject/model	CNN	LSTM	CNN + LSTM
*S*1	50	65	85
*S*2	51	60	82
*S*3	53	64	86
Avg	51	63	84

way other than the bandpass filtering. The models performed with the average classification accuracy of 51% for CNN-based model, 63% for LSTM-based model, and 83% CNN + LSTM model. While the accuracies are comparable to the other studies employing the same dataset [16–18, 26], these studies employ pair-wise classification compared to multiclass classification employed in this study. This study can be extended to classify for more imaginary speech activities.

References

1. Wolpaw JR (2010) Brain–computer interface research comes of age: traditional assumptions meet emerging realities. J Motor Behav 42(6):351–353
2. Allison B (2007) The I of BCIs: next generation interfaces for brain–computer interface systems that adapt to individual users. In: International conference on human-computer interaction. Springer, Berlin, Heidelberg, pp 558–568
3. Birbaumer N, Cohen LG (2007) Brain–computer interfaces: communication and restoration of movement in paralysis. J Physiol 579(3):621–636
4. Hwang HJ, Kim S, Choi S, Im CH (2013) EEG-based brain-computer interfaces: a thorough literature survey. Int J Hum-Comput Interact 29(12):814–826
5. Pfurtscheller G, Brunner C, Schlögl A, Da Silva FL (2006) Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks. Neuroimage 31(1):153–159
6. Lee PL, Yeh CL, Cheng JYS, Yang CY, Lan GY (2011) An SSVEP-based BCI using high duty-cycle visual flicker. IEEE Trans Biomed Eng 58(12):3350–3359
7. Faradji F, Ward RK, Birch GE (2009) Plausibility assessment of a 2-state self-paced mental task-based BCI using the no-control performance analysis. J Neurosci Methods 180(2):330–339
8. Salvaris M, Cinel C, Citi L, Poli R (2011) Novel protocols for P300-based brain–computer interfaces. IEEE Trans Neural Syst Rehabil Eng 20(1):8–17
9. Kaper M, Meinicke P, Grossekathoefer U, Lingner T, Ritter H (2004) BCI competition 2003-data set IIb: support vector machines for the P300 speller paradigm. IEEE Trans Biomed Eng 51(6):1073–1076
10. Nijboer F, Sellers EW, Mellinger J, Jordan MA, Matuz T, Furdea A, Halder S, Mochty U, Krusienski DJ, Vaughan TM, Wolpaw JR (2008) A P300-based brain–computer interface for people with amyotrophic lateral sclerosis. Clin Neurophysiol 119(8):1909–1916
11. Volosyak I, Moor A, Gräser A (2011) A dictionary-driven SSVEP speller with a modified graphical user interface. In: International work-conference on artificial neural networks. Springer, Berlin, Heidelberg, pp 353–361
12. Pfurtscheller G, Neuper C (2001) Motor imagery and direct brain-computer communication. Proc IEEE 89(7):1123–1134
13. Carlson T, Leeb R, Chavarriaga R, Millán JDR (2012) The birth of the brain-controlled wheelchair. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. IEEE. pp 5444–5445
14. Callan DE, Callan AM, Honda K, Masaki S (2000) Single-sweep EEG analysis of neural processes underlying perception and production of vowels. Cogn Brain Res 10(1–2):173–176
15. Fujimaki N, Takeuchi F, Kobayashi T, Kuriki S, Hasuo S (1994) Event-related potentials in silent speech. Brain Topogr 6(4):259–267
16. DaSalla CS, Kambara H, Sato M, Koike Y (2009) Single-trial classification of vowel speech imagery using common spatial patterns. Neural Netw 22(9):1334–1339
17. Idrees BM, Farooq O (2016) EEG based vowel classification during speech imagery. In: 2016 3rd international conference on computing for sustainable global development (INDIACom). IEEE, pp 1130–1134

18. Patel, J., Pasha, I.A. and Krishna, D.H.: Classification of imagery vowel speech using EEG and cross correlation. International Journal of Pure and Applied Mathematics, 118(24) (2018).
19. Wang L, Zhang X, Zhong X, Zhang Y (2013) Analysis and classification of speech imagery EEG for BCI. Biomed Signal Process Control 8(6):901–908
20. Min B, Kim J, Park HJ, Lee B (2016) Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram. BioMed Res Int
21. Riaz A, Akhtar S, Iftikhar S, Khan AA, Salman A (2014) Inter comparison of classification techniques for vowel speech imagery using EEG sensors. In: The 2014 2nd international conference on systems and informatics (ICSAI 2014). IEEE, pp 712–717
22. Watanabe H, Tanaka H, Sakti S, Nakamura S (2019) Synchronization between overt speech envelope and EEG oscillations during imagined speech. Neurosci Res
23. Kim HJ, Lee MH, Lee M (2020) A BCI based Smart Home System Combined with Event-related Potentials and Speech Imagery Task. In: 2020 8th international winter conference on brain-computer interface (BCI). IEEE, pp 1–6
24. Cooney C, Raffaella F, Coyle D (2019) Optimizing input layers improves CNN generalization and transfer learning for imagined speech decoding from EEG. In: IEEE international conference on systems, man, and cybernetics, 2019: Industry 4.0
25. Tamm MO, Muhammad Y, Muhammad N (2020) Classification of vowels from imagined speech with convolutional neural networks. Computers 9(2):46
26. Yoshimura N, Satsuma A, DaSalla CS, Hanakawa T, Sato MA, Koike Y (2011) Usability of EEG cortical currents in classification of vowel speech imagery. In: 2011 international conference on virtual rehabilitation. IEEE, pp 1–2