Proceedings of the
8th Graz Brain-Computer Interface Conference 2019 DOI: 10.3217/978-3-85125-682-6-62

CLASSIFICATION OF IMAGINED SPOKEN WORD-PAIRS USING CONVOLUTIONAL NEURAL NETWORKS

C. Cooney[1], A. Korik[1], R. Folli[2], D. Coyle[1]

[1] Intelligent Systems Research Centre, Ulster University, Derry, UK
[2] Institute for Research in Social Sciences, Ulster University, Jordanstown, UK

E-mail: cooney-c@ulster.ac.uk

ABSTRACT: Imagined speech is gaining traction as a communicative paradigm for brain-computer-interfaces (BCI), as a growing body of research indicates the potential for decoding speech processes directly from the brain. The development of this type of direct-speech BCI has primarily considered feature extraction and machine learning approaches typical to BCI decoding. Here, we consider the potential of deep learning as a possible alternative to traditional BCI methodologies in relation to imagined speech EEG decoding. Two different convolutional neural networks (CNN) were trained on multiple imagined speech word-pairs, and their performance compared to a baseline linear discriminant analysis (LDA) classifier trained using filterbank common spatial patterns (FBCSP) features. Classifiers were trained using nested cross-validation to enable hyper-parameter optimization. Results obtained showed that the CNNs outperformed the FBCSP with average accuracies of 62.37% and 60.88% vs. 57.80% ($p<0.005$).

INTRODUCTION

A direct-speech brain-computer interface (DS-BCI) is one in which a user’s imagined speech is harnessed as the mode of communication between themselves and a computer, or interlocutor [1]. Imagined speech is the internal pronunciation of words or sentences, which does not result in any audible output [2]. Imagined speech has received relatively little attention from BCI researchers in comparison with more common paradigms such as motor imagery (MI) or steady-state visually-evoked potentials (SSVEP) (see [3] for a review). However, a DS-BCI does offer the possibility of a more naturalistic form of communication and must therefore be considered an important field of study within the BCI community.

Both invasive and non-invasive approaches to data acquisition have been applied to the recording of imagined speech, primarily through electrocorticography (ECoG) [4] and electroencephalography (EEG) [5]. In this study, we focus specifically on the decoding of imagined speech from EEG recordings.

Approaches to imagined speech decoding have typically employed traditional BCI feature extraction and classification algorithms. Among the features used to represent imagined speech from EEG are autoregressive coefficients [6], common-spatial patterns [7] and spectrotemporal features [8]. More recently, Mel Frequency Cepstral Coefficients (MFCC) [9], [10] and Riemannian manifold features [5] have enabled imagined speech classification.

Several traditional machine learning approaches have been applied to the task of decoding imagined speech from EEG. These include support vector machines (SVM) [9], [11], Linear Discriminant Analysis (LDA) [6], [12], Naïve Bayes [12], k-Nearest Neighbors [10] and Random Forests [13]. Of these, the SVM has been the most-often utilized classification method, resulting in accuracies of 71.3% [7] and 69.3% [14] in binary tasks. However, to-date no combined feature extraction and classification method has proven itself to be the dominant approach. For this reason, research into a deep learning approach to imagined speech classification is a logical undertaking. Deep learning has been enormously successful across fields such as computer vision [15] and automatic speech recognition [16]. More recently, it has been successfully applied to BCI in relation to MI [17] and SSVEP [18] but its application to imagined speech has been relatively sporadic [19]. Of the deep learning approaches available, convolutional neural networks (CNN) have been the most heavily-utilized in relation to BCI/EEG. Among many others, the applicability of CNNs has been demonstrated for automated screening of depression [20], and prediction of drivers’ cognitive performance [21]. For a review into deep learning analysis of EEG, see [22].

Here we evaluate the performance of two CNNs tasked with decoding imagined speech from EEG. The data used consisted of fifteen word-pairs extracted from a dataset containing six Spanish words produced with imagined speech. The performance of the CNNs are rated in comparison with a regularized-LDA (rLDA) trained on FBCSP features. A nested approach to cross-validation is implemented to facilitate parameter-optimization and improve the robustness of results. The results obtained show that the CNNs perform significantly better than the rLDA classifier, and that the performance of the deep CNN was significantly better than that of the shallow CNN.

MATERIALS AND METHODS

The methodology implemented in order to classify imagined speech production from EEG signals is

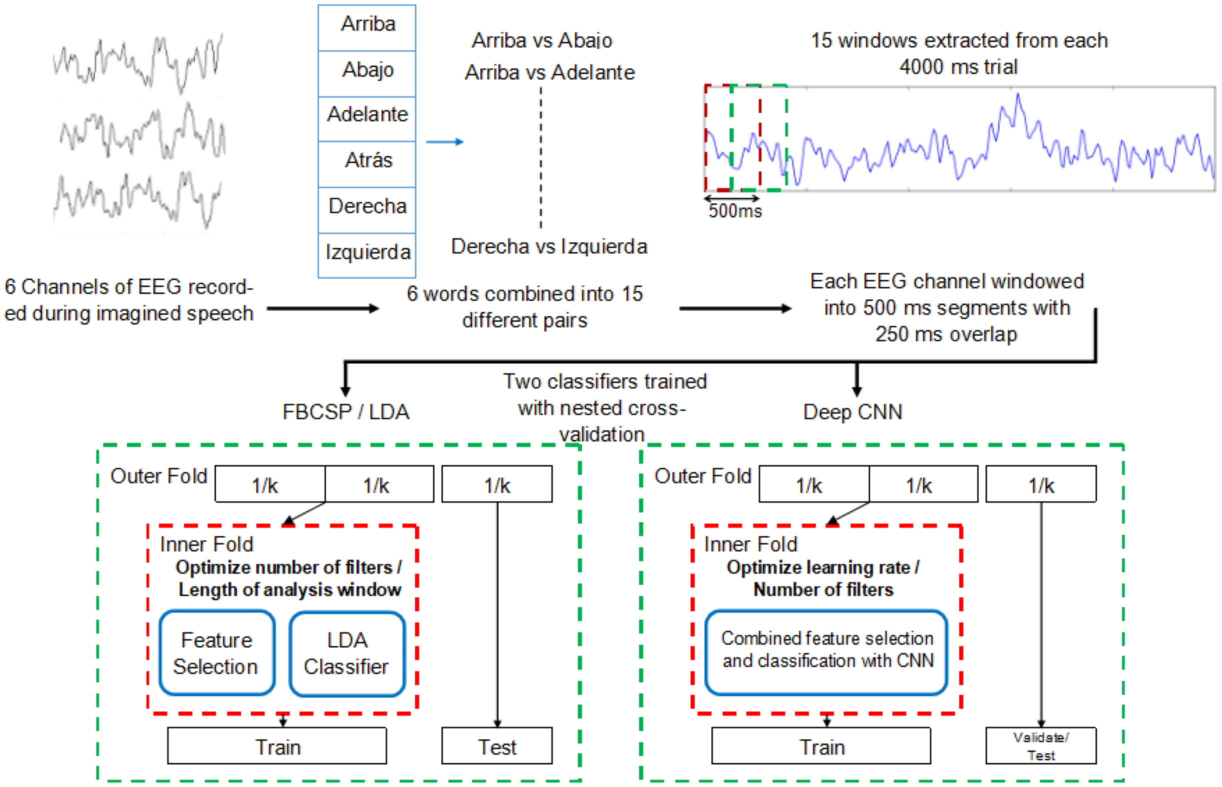
Arriba
Abajo
Adelante
Atrás
Derecha
Izquierda
Arriba vs Abajo
Arriba vs Adelante
Derecha vs Izquierda
15 windows extracted from each 4000 ms trial
500ms
6 Channels of EEG recorded during imagined speech
6 words combined into 15 different pairs
Each EEG channel windowed into 500 ms segments with 250 ms overlap
Two classifiers trained with nested cross-validation
FBCSP / LDA
Deep CNN
Outer Fold
1/k
1/k
1/k
Inner Fold
Optimize number of filters / Length of analysis window
Feature Selection
LDA Classifier
Train
Test
Outer Fold
1/k
1/k
1/k
Inner Fold
Optimize learning rate / Number of filters
Combined feature selection and classification with CNN
Train
Validate/ Test

Figure 1Depiction of the methodology followed for this study

depicted in Figure 1 and described throughout the remainder of this section, beginning with the dataset.

Dataset: The dataset used for this research was recorded at the offices of the Laboratorio de Ingeniería en Rehabilitación e Investigaciones Neuromusculares y Sensoriales (LIRINS) in the Faculty of Engineering at the National University of Entre Ríos (UNER) by Pressel Coretto et al. [23]. EEG signals were recorded while 15 subjects performed overt and imagined speech tasks corresponding to the production of Spanish words and vowels. Only the EEG associated with imagined word production was analysed for this study. Thus, the EEG data used consisted of those trials recorded while participants imagined the production of six Spanish words: "arriba", "abajo", "derecha", "izquierda", "adelante" and "atrás" (corresponding to the English words up, down, left, right, backward and forward). These words were selected as commands a user might make when interacting with a BCI. The experimental protocol for the imagined words tasks required participants to imagine speaking one of the prompted words at three audibly-cued time-points during the 4-second trial-period. Prior to the trial-period, stimuli were presented in both visual and auditory form, showing each subject the word for 2 seconds before being removed during the trial. EEG signals were recorded using a six-channel system, sampled at 1024 Hz. Electrodes were positioned according to the 10-20 international system over F3, F4, C3, C4, P3 and P4 (Figure 2).

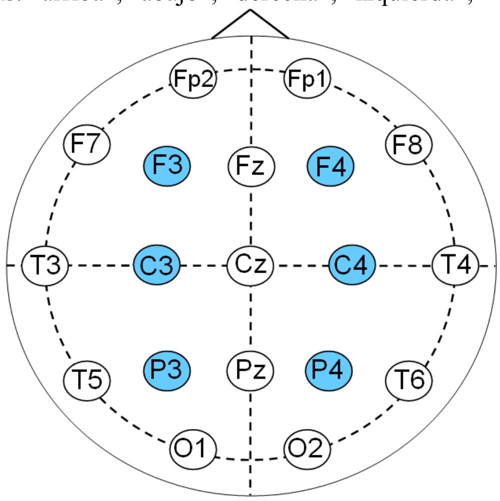

Figure 2 The EEG montage used to acquire data.

Preprocessing: The original dataset was filtered between 2 Hz and 40 Hz using a finite impulse response bandpass filter [23], so no further filtering was applied for this study. Artefact detection and removal were implemented using Independent Component Analysis with Hessian approximation preconditioning [24].

Data splitting: In order to facilitate the analysis of multiple binary classifiers, all possible pairs of words were extracted from the dataset (Figure 1B). This resulted in 15 different pairs of imagined words for binary classification (e.g. *arriba* vs. *abajo*). The number of trials per class varied across subjects, with a maximum of 51. However, all pairs were balanced prior to training. Due to the high computational load associated with the nested cross-validation scheme employed for this study (see below), a 500 ms segment was extracted from each 4000 ms trial (Figure 1C) to act as the classification window. The selection of this window was based on the description of the experimental protocol described in

[23], in which three audible clicks directed participants when to imagine speaking. Therefore, a window was extracted to encompass the first of these periods of imagined speech production, about the 1-second mark of the overall trial. Concretely, this was the 500 ms segment between 750 ms and 1250 ms of the overall trial window.

Classification methods: Three distinct methods of classification were applied to the imagined words EEG data. The first of these methods, was the use of FBCSP features to train a rLDA (Figure 1D). The rLDA classifier is a regularized version of the LDA algorithm [25], which reduces the dispersion of the eigenvalues of the sample covariance matrix when a diverging dimension *p* is large. It has been employed elsewhere as a classification method for EEG signals [26], and is used here to provide a baseline reference for the performance of the CNNs. Unlike CNNs, the rLDA requires separate feature engineering and classification, and thus the type of features must be selected prior to training. FBCSP is a widely-used feature extraction algorithm across multiple BCI paradigms [27]. Linear combinations of the EEG channels are computed to enhance discrimination of band power features between classes. FBCSP has been proven successful in MI tasks, including as the winner in several EEG decoding competitions (e.g. [28]). Its proven results as a decoding algorithm in BCI, and the fact that there is no clear benchmark specifically for imagined speech, has led to selection of FBCSP as a reasonable baseline in this study.

The second classifier tested was a deep CNN designed by Schirrmeister et al. [29] specifically for EEG decoding applications (Figure 1E). The network architecture is based on similar CNNs used in computer vision [30] and is constructed to extract a wide range of features from the EEG signals. Figure 3 depicts the composition of the deep CNN. The input block of the CNN consists of two convolutional layers, one to perform convolution over time and one for spatial filtering. This first block also contains batch normalization, a non-linear activation function and a mean-pooling layer. Following this are three identical convolution blocks, each containing dropout, convolution, batch normalization, non-linear activation and mean-pooling. Finally, the output consists of a dense softmax layer for classification. The second CNN has been constructed with a shallow architecture and designed to decode band power features from EEG [29]. The shallow CNN is constructed of the same series of layers featured in the input block to the deep CNN (Figure 3), followed by dropout and a softmax classification layer. Here, we set dropout to 0.1 and selected the leaky rectified linear units activation function to add non-linearity into the two networks. Both CNNs used the ADAM optimizer [31] and the cross entropy loss function, and were allowed to train for 60 epochs with a *patience* of 30. A batch size of 64 was used. The CNNs were implemented in PyTorch [32], using the Braindecode repository [29].

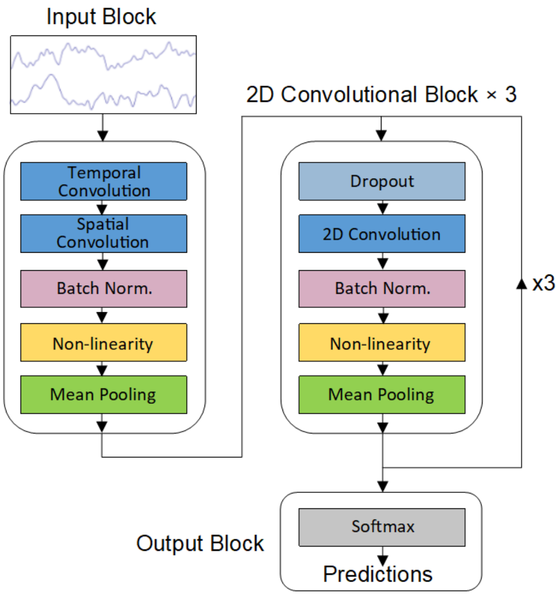

Figure 3 Deep CNN architecture designed by [22].

Nested cross-validation: A nested approach to cross-validation has been applied to training and testing of both the rLDA and the CNNs, with only small differences implemented when required by the respective classifiers (Figure 1D and E). Although not typically employed in deep learning contexts, nested-cross validation is utilized here to improve the robustness of results and to facilitate hyper-parameter optimization. Based on principles described in [33], the data are first split into 4-folds, one of which is retained in the outer-fold. An inner fold is then instantiated using the remaining 3 folds. The combined inner-fold is then re-split into 4 folds, with each fold iteratively acting as the test-set. The inner-fold facilitates training and testing of the two classifiers using each possible combination of hyper-parameters. The hyper-parameter combination with the best average accuracy across all inner-folds is then used to train the classifier on the entire inner-fold data. The outer-fold is then used as the test-set, or in the case of the CNNs, both validation- and test-sets. The classification accuracy is reported as the average accuracy across all 4 outer folds.

Hyper-parameters: Two hyper-parameters were selected for optimization with each classifier (Table 1). In the case of the rLDA, hyper-parameters required in the computation of FBCSP features were used. The first of these is the number of selected spatial filter pairs (1,3,4,5). The second hyper-parameter used here was the mutual information quantization level, with the values considered being 6, 8, 10 and 12. Hyper-parameters selected for the deep CNN were learning-rate and the number of filters implemented in the final convolutional layer (Table 1). The four learning-rates evaluated were

Table 1 Hyper-parameters optimized using nested cross-validation.

	Hyper-parameter 1	Hyper-parameter 2
FBCSP	# spatial filters: (1,3,4,5)	mutual information: (6,8,10,12)
Deep CNN	learning–rate: (1,0.1,0.01,0.001)	# final layer filters: (100,500,1000,1500)
Shallow CNN	learning-rate: (1,0.1,0.01,0.001)	# spatial filters: (20,40,60,80)

1.0, 0.1, 0.01 and 0.001 and the number of filters used in the final layer of convolution was 100, 500, 1000 and 1500. The same learning-rate range was used for the shallow CNN but the second parameter considered was the number of spatial filters (Table 1).

Statistics: Wilcoxon signed-rank tests were used to determine whether or not differences between the classifiers were statistically significant.

RESULTS

Here we report classification accuracies for each subject in the cohort and for each word-pair used to train the classifiers. Cross-subject classification accuracies are presented in Figure 4. Here, the highest classification accuracy obtained was 65.67%, achieved by subject 13 with the CNN. The shallow CNN showed similar peak performance with 65.28% average accuracy for subject 8. Results obtained by the baseline rLDA and the CNNs are significantly above chance accuracy (50%) for all word-pairs across all subjects. The average classification accuracies for the word-pairs when trained on the rLDA and the two CNNs were 57.80%, 62.37% and 60.88% respectively (Figure 5). The Wilcoxon signed rank tests determined that the greater performance of both the CNNs across word-pairs was significant in comparison with the FBCSP (p<0.005). The greater performance of the deep CNN was also significant in relation to the shallow network (p<0.05). Accuracies for the different word-pairs do not deviate substantially from the mean for any of the combinations (Figure 5). The highest average classification accuracy for a single word-pair was 64.55%, achieved by the *abajo vs derecha* pair, using the deep CNN (Figure 5). The highest single-subject accuracy obtained for a single word-pair was 78.33%, achieved by subject 11, also for the *abajo vs derecha* pair with the deep CNN.

The number of spatial filters used for FBCSP feature extraction was selected by the nested cross-validation as 5 (Table 2), although the difference between selecting 5, 4 or 3 was minimal. A mutual information coefficient of 8 was most often selected for optimization. In the case of the CNNs, the hyper-parameter optimization selected a learning-rate of 0.001 more often than any of the other options. 1000 filters were selected for the final analysis of the convolution layer of the deep network and 20 were selected for spatial convolution in the shallow CNN.

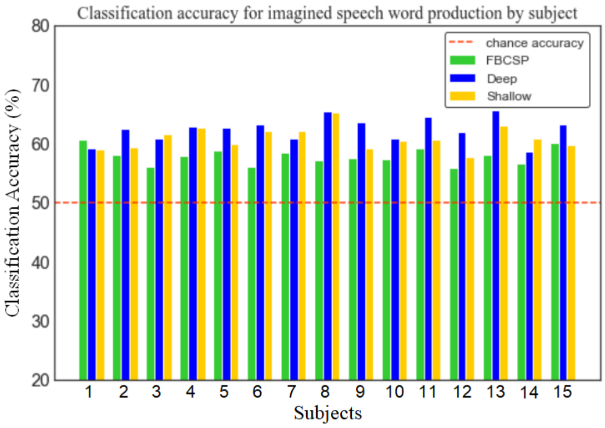

Figure 4 Subject classification accuracies for rLDA and CNNs.

Table 2 Hyper-parameters selected with nested cross-validation.

	Hyper-parameter 1	Hyper-parameter 2
FBCSP	# of filters = 5	mutual info. = 8
Deep	lr = 0.001	# of filters = 1000
Shallow	lr = 0.001	# of filters = 20

DISCUSSION

The results presented here support the assertion that employing deep learning methodologies to the task of decoding imagined speech from EEG is a reasonable undertaking. For each subject, and for each word-pair, the CNNs outperformed the FBCSP-trained rLDA, achieving accuracies significantly above chance in each case. Despite indicating promise, the results also show that this level of performance is not yet close to what would be required of a functional DS-BCI. However, the greater overall performance of the deep architecture in relation to the shallow CNN does indicate the potential of deep learning for imagined speech decoding. Hyper-parameter optimization through nested cross-validation enabled selection of parameters most appropriate to the current task. Here, we determined that 5 spatial filters and a mutual information coefficient of 8, resulted in greater performance. Of the CNNs, 0.001 was the optimum learning-rate for use with the ADAM optimizer. The number of filters selected for the final convolutional layer was 1000. While this is greater than the number in the original paper [29], it provided the best accuracies here.

A weakness of the present study is the selection of a single 500 ms classification segment from the 4000 ms trial window. Although this approach was followed in the interests of computational efficiency, it is likely that a sliding-window would have improved overall classification performance. Furthermore, the number of trials per class was quite small, ranging from 39 to 51 for a single word. This relatively small volume of data constrains the classifiers' ability to infer classes by recognizing common patterns. Clearly, if deep learning is going to become a useful decoding approach for DS-BCI, larger datasets are required.

Interestingly, the results presented in Figure 5 do not suggest any significant differences in the effects of the linguistic content of the word-pairs. This may be a direct result of the choice of words used for this study. However, we agree with views expressed elsewhere [1], [2] that neurolinguistics research into imagined speech can aid the design of experiments in future research.

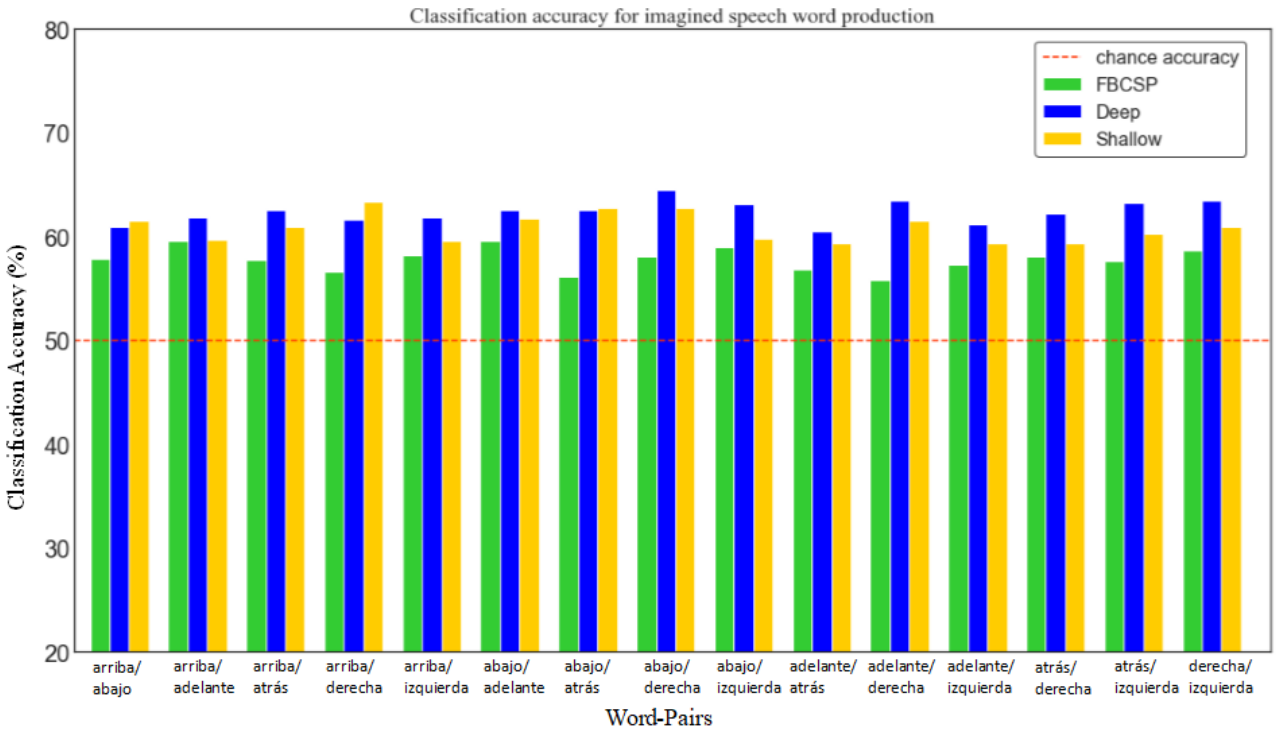

Figure 5 Classification accuracies for imagined speech production, by word-pair

CONCLUSION

In this study, we trained three different types of classifier with the purpose of decoding imagined speech from EEG. A rLDA using FBCSP features, and two CNNs, were trained on a 500 ms classification window extracted from trials where subjects imagined speaking Spanish words. 15 word-pairs were extracted from the dataset to enable multiple binary classifications. Nested cross-validation was employed to facilitate hyper-parameter optimization during training.
Results showed that the CNNs performed significantly better than the rLDA classifier with average accuracies of 62.37% and 60.88% vs. 57.80%. The differences observed between the two CNNs were significant, with the deeper network performing better. Results also indicated that differences in the accuracies obtained between the different word-pairs were not significant. The results suggest that, while further work is required in the field, deep learning is a realistic decoding methodology for imagined speech EEG.

ACKNOWLEDGEMENTS

The authors acknowledge the dataset's authors for the data used in this study. This work is supported by a Northern Ireland Department for the Economy studentship.

REFERENCES

[1] O. Iljina *et al.*, “Neurolinguistic and machine-learning perspectives on direct speech BCIs for restoration of naturalistic communication,” *Brain-Computer Interfaces*, vol. 4, no. 3, pp. 186–199, 2017.

[2] C. Cooney, R. Folli, and D. Coyle, “Neurolinguistics Research Advancing Development of a Direct-Speech Brain-Computer Interface,” *IScience*, vol. 8, pp. 103–125, 2018.

[3] R. A. Ramadan and A. V. Vasilakos, “Brain computer interface: control signals review,” *Neurocomputing*, vol. 223, no. October 2016, pp. 26–44, 2017.

[4] S. Martin *et al.*, “Word pair classification during imagined speech using direct brain recordings,” *Sci. Rep.*, vol. 6, no. 1, 2016.

[5] C. H. Nguyen, G. Karavas, and P. Artemiadis, “Inferring imagined speech using EEG signals: a new approach using Riemannian Manifold features,” *J. Neural Eng.*, 2017.

[6] Y. Song and F. Sepulveda, “Classifying speech related vs. idle state towards onset detection in brain-computer interfaces overt, inhibited overt, and covert speech sound production vs. idle state,” *IEEE 2014 Biomed. Circuits Syst. Conf. BioCAS 2014 - Proc.*, pp. 568–571, 2014.

[7] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, “Single-trial classification of vowel speech imagery using common spatial patterns,” *Neural Networks*, vol. 22, no. 9, pp. 1334–1339, 2009.

[8] S. Zhao and F. Rudzicz, “Classifying phonological categories in imagined and articulated speech,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2015–Augus, pp. 992–996, 2015.

[9] C. Cooney, R. Folli, and D. Coyle, “Mel Frequency Cepstral Coefficients Enhance

Imagined Speech Decoding Accuracy from EEG," *ISSC 2018 Irish Signals Syst. Confernece, 2018*, 2018.

[10] N. Hashim, A. Ali, and W.-N. Mohd-Isa, "Word-Based Classification of Imagined Speech Using EEG," in *Computational Science and Technology*, 2018, pp. 195–204.

[11] T. Kim, J. Lee, H. Choi, H. Lee, I. Y. Kim, and D. P. Jang, "Meaning based covert speech classification for brain-computer interface based on electroencephalography," *Int. IEEE/EMBS Conf. Neural Eng. NER*, pp. 53–56, 2013.

[12] X. Chi, J. B. Hagedorn, D. Schoonover, and M. D. Zmura, "EEG-Based Discrimination of Imagined Speech Phonemes," *Int. J. Bioelectromagn.*, vol. 13, no. 4, pp. 201–206, 2011.

[13] E. F. González-Castañeda, A. A. Torres-García, C. A. Reyes-García, and L. Villaseñor-Pineda, "Sonification and textification: Proposing methods for classifying unspoken words from EEG signals," *Biomed. Signal Process. Control*, vol. 37, pp. 82–91, 2017.

[14] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "Online EEG Classification of Covert Speech for Brain–Computer Interfacing," *Int. J. Neural Syst.*, vol. 27, no. 8, p. 1750033, 2017.

[15] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," pp. 1–13, 2016.

[16] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. 6, pp. 6645–6649, 2013.

[17] B. E. Olivas-padilla and M. I. Chacon-murguia, "Classification of multiple motor imagery using deep convolutional neural networks and spatial filters," *Appl. Soft Comput. J.*, vol. 75, pp. 461–472, 2019.

[18] N. Waytowich *et al.*, "Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials," *J. Neural Eng.*, vol. 15, no. 066031, 2018.

[19] A. Rezazadeh Sereshkeh, R. Trott, A. Bricout, and T. Chau, "EEG Classification of Covert Speech Using Regularized Neural Networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 12, pp. 2292–2300, 2017.

[20] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, "Automated EEG-based screening of depression using deep convolutional neural network," *Comput. Methods Programs Biomed.*, vol. 161, pp. 103–113, 2018.

[21] M. Hajinoroozi, Z. Mao, T. P. Jung, C. T. Lin, and Y. Huang, "EEG-based prediction of driver's cognitive performance by deep convolutional neural network," *Signal Process. Image Commun.*, vol. 47, pp. 549–555, 2016.

[22] Y. Roy and A. Gramfort, "Deep learning - based electroencephalography analysis : a systematic review," *arXiv Prepr. arXiv1901.05498v2*, 2019.

[23] G. A. Pressel Coretto, I. E. Gareis, and H. L. Rufiner, "Open access database of EEG signals recorded during imagined speech," p. 1016002, 2017.

[24] P. Ablin, J. Cardoso, and A. Gramfort, "Faster independent component analysis by preconditioning with Hessian approximations arXiv : 1706 . 08171v3 [stat . ML] 8 Sep 2017," pp. 1–23, 2017.

[25] J. H. Friedman, "Regularized Discriminant Analysis," *J. Am. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, 1988.

[26] E. Neto, F. Biessmann, H. Aurlien, H. Nordby, and T. Eichele, "Regularized Linear Discriminant Analysis of EEG Features in Dementia Patients," vol. 8, no. November, pp. 1–10, 2016.

[27] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter Bank Common Spatial Pattern (FBCSP) in brain-computer interface," *Proc. Int. Jt. Conf. Neural Networks*, pp. 2390–2397, 2008.

[28] M. Tangermann *et al.*, "Review of the BCI competition IV," vol. 6, no. July, pp. 1–31, 2012.

[29] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, 2017.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[31] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Comput. Res. Repos.*, 2017.

[32] A. Paszke, S. Gross, S. Chintala, and G. Chanan, "Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration." 2017.

[33] A. Korik, R. Sosnik, N. Siddique, and D. Coyle, "Decoding imagined 3D hand movement trajectories from EEG: Evidence to support the use of mu, beta, and low gamma oscillations," *Front. Neurosci.*, vol. 12, no. MAR, pp. 1–16, 2018.