Optimizing Input Layers Improves CNN Generalization and Transfer Learning for Imagined Speech Decoding from EEG

Ciaran Cooney[1], Rafaella Folli[2], and Damien Coyle[1], *Member, IEEE*
[1]*Intelligent Systems Research Centre, Ulster University, Derry, UK*
[2]*Institute for Research in Social Sciences, Ulster University, Jordanstown, UK*
Email: cooney-c@ulster.ac.uk

***Abstract*— A brain-computer interface (BCI) that employs imagined speech as the mode of determining user intent requires strong generalizability for a feasible system to be realized. Research in this field has typically applied data to training algorithms on a within-subject basis. However, even within-subject training and test data are not always of the same feature space and distribution. Such scenarios can contribute to poor BCI performance, and real-world applications for imagined speech-based BCIs cannot assume homogeneity in user data. Transfer Learning (TL) is a common approach used to improve generalizability in machine learning models through transfer of knowledge from a source domain to a target task. In this study, two distinct TL methodologies are employed to classify EEG data corresponding to imagined speech production of vowels, using a deep convolutional neural network (CNN). Both TL approaches involved conditional training of the CNN on all subjects, excluding the target subject. A subset of the target subject data was then used to fine-tune either the input or output layers of the CNN. Results were compared with a standard benchmark using a within-subject approach. Both TL methods significantly outperformed the baseline and fine-tuning of the input layers resulted in the highest overall accuracy (35.68%; chance: 20%).**

I. Introduction

A direct-speech brain-computer interface (DS-BCI) has the potential to facilitate language-based communication between a user and an interlocutor [1]. In harnessing imagined speech as the communicative modality, such a system would require a user to internally pronounce phonemes, words and sentences, without any movement or audible output. Neural recordings (e.g. electroencephalogram (EEG)) corresponding to the production of imagined speech would then be decoded using signal processing and classification algorithms [2]. Approaches to decoding imagined speech have typically employed traditional BCI feature extraction and classification methods. Feature extractors have included common spatial patterns [3] and Riemannian manifold features [4], while algorithms such as support vector machine [2], linear discriminant analysis [5] and Naïve Bayes [5] have been used to classify imagined speech. To-date, no combination of feature extractor and classifier has proven itself as the best approach. Thus, the application of deep learning (DL) methodologies within this context is a logical advance. Moreover, studies investigating the classification of imagined speech EEG have, to-date, implemented training and testing using a within-subject pipeline [2], [4]. Therefore, a system in which information is shared or adapted across subjects must be investigated for its potential generalizability. For these reasons, we investigated a DL framework, namely a convolutional neural network (CNN), with two transfer learning (TL) methods for imagined speech classification.

DL has been successful in fields such as computer vision, with CNNs in particular exhibiting strong performance gains over traditional machine learning classifiers [6]. The success of CNNs in these fields has led to their adoption in the BCI/EEG domain. Steady-state visually-evoked potentials [7] and motor imagery [8] studies have utilized CNNs for classification. More recently, CNNs have been used to classify imagined speech production [9], [10], although the number of studies is still relatively few. Other applications of CNNs in relation to EEG data include automated screening of depression [11] and prediction of drivers' cognitive performance [12]. Given a recent increase in research using CNNs for EEG analysis, several CNN architectures have been designed specifically for decoding EEG. *EEGNet* is a CNN designed to minimize the number of parameters required while retaining strong performance across EEG paradigms [13]. The *Braindecode* repository [14] offers both shallow and deep CNNs, with both designed to mimic the feature extraction method used by the filterbank common spatial patterns (FBCSP) algorithm [15].

TL refers to use of knowledge learned in one domain to improve generalization in another [16]. It is a general term for the adaptation of knowledge from a *source* domain (here, multiple subjects' EEG) to enable its use in learning applied to a *target* domain (here, a single subject's EEG) [17]. The target domain often has relatively few, or no, class labels. Depending on the problem under consideration, there are several TL methods and several implementation strategies which can be used [17]. The most common TL approach is *inductive TL,* with its sub-categories of multi-task learning [18] and self-taught learning [19]. Multi-task learning refers to TL settings in which both source and target domain labels are available. The TL methodologies considered in is study are multi-task. As well as having been used across multiple fields where it has improved classification accuracy through use of large datasets as the source domain and relatively small datasets as the target [20], TL has more recently been applied to EEG and BCI. A leave-one-subject-out strategy has been applied to the training and testing of a CNN on EEG data corresponding to error-detection [21]. Lin et al. [22] leveraged existing EEG-based emotion data to train a model for a new subject using a conditional TL system. *Transductive TL,* where target domain labels are unavailable, has been used for to classify epileptic seizures from EEG [23], and TL has also been applied to the classification of imagined speech EEG previously [24]. However, results obtained from with TL did not show significant performance gains.

Here, we investigated whether deep TL approaches could have utility in the extremely difficult imagined speech EEG classification problem. The dataset used consists of EEG recordings corresponding to imagined speech production of the five vowels. Two multi-task TL methodologies were implemented in which a 6-layer CNN was trained on source domain data before being fine-tuned on the target subject data. In TL method 1 (TL1), the input layers of the CNN were fine-tuned. In method 2 (TL2), the final convolution layers were fine-tuned. A within-subject approach to training and testing the CNN was used for baseline comparison. Results obtained indicated that both the TL methods achieved accuracies significantly better than those of the within-subject method. Of the TL methods, fine-tuning of the input layers was more successful than fine-tuning of the final convolutional layers.

II. Methods

A. Data

Data 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. [25]. Overt and imagined speech tasks involving Spanish words and vowels were performed by 15 subjects while EEG signals were recorded. Only signals

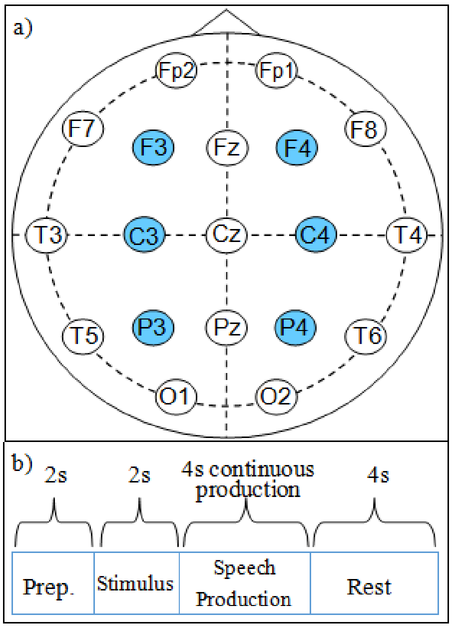

Figure 1 a) 6-electrode montage. b) Experimental protocol.

corresponding to imagined vowel production were analysed for this study. Thus, the data consisted of trials in which participants imagined speaking the five vowels "/a/", "/e/", "/i/", "/o/" and "/u/". The experimental protocol for the imagined vowels task included a 2 second pre-trial stimulus period during which prompts were presented both visually and audibly. Participants then had 4 seconds to continually produce the prompt using imagined speech (Figure 1b). EEG signals were recorded using an 18-channel Grass® analog amplifier model 8-18-36 and a Data Translation® analog-to-digital converter board model DT9816, sampled at 1024 Hz. Electrodes were positioned according to the 10-20 international system over F3, F4, C3, C4, P3 and P4 (Figure 1a).

B. Preprocessing

The original dataset was filtered between 2 Hz and 40 Hz using a finite impulse response bandpass filter [24], so no further filtering was applied. Data were down-sampled to 128 Hz and artefact detection and removal were implemented using Independent Component Analysis with Hessian approximation preconditioning [25]. Classifiers are sensitive to non-stationarities, often leading to poor performance when applied to EEG data from different distributions [26]. As the TL methods require using data with different distributions, Scikit-learn's *robust_scaler* [27] function was used to standardize the data by centering to the median and component-wise scale according to the interquartile range.

C. Convolutional Neural Network

The deep CNN architecture adopted for this study is based on one designed specifically for EEG decoding applications [14]. The structure of the CNN is depicted in Figure 2, including the number and size of filters included in each layer. Considering 1a and 1b as a single unit, the feature extraction section of the network consists of six convolution layers, two more than the architecture presented previously [14]. The input of the CNN (1a and 1b) consists of two convolutional layers, the first to perform temporal convolution and the second for spatial filtering. This construction has been conceived of as a feature extraction stage analogous to that of FBCSP [15], and is designed to decode band power features from EEG [14]. This initial combination of layers is followed by batch-normalization and a leaky rectified linear units (ReLU) activation for non-linearity. No pooling layer is implemented at this stage as we have extended the depth of the CNN to contain two additional convolution layers. Each of the remaining five layers consist of the same basic structure, differing only in the size and number of filters used (Figure 2). This structure is as follows: *dropout, convolution, batch-normalization, non-linear activation* and *mean pooling*. Here, dropout has been set to 0.1 and leaky ReLU implemented as the activation function for each layer. Figure 2 identifies the layers which were fine-tuned as part of the TL process. This is discussed further in Section E. The feature map obtained from the feature extraction stage is passed to the final block for classification. This is a dense softmax layer which produces posterior probabilities and a prediction for one of the classes.

D. Within-Subject Classifier

The baseline method was constructed to classify imagined

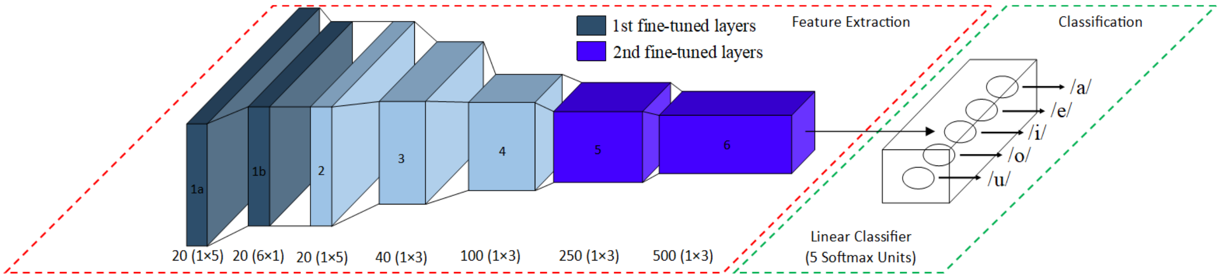

Figure 2 CNN architecture used with layers used for the two fine-tuning approaches highlighted.

speech EEG using a within-subject approach to training and testing. Within-subject classification is the most common approach employed for BCI applications [2], [4]. A 5-fold cross-validation scheme was applied to split the data into training, validation and test sets. For each of the 5 folds, the CNN was trained and validated, and the model with the best classification performance on the validation set selected for making predictions on the test set. Classification accuracy is reported as the mean obtained from the 5 folds.

E. Transfer Learning Methodology

Two TL approaches were tested using the CNN and imagined speech EEG data. The TL task being investigated is a multitask problem, where source data i.e., data from all *other* subjects, is used to transfer knowledge to the target domain - the model being trained for the target subject. Many multitask learning applications, typically with regards to images [28], share knowledge at lower (input) layers and apply fine-tuning to the later layers. This is due to many categories of images containing similar structures such as edges and geometric shapes which facilitate shared lower-level learning, and higher-level task-dependent learning in later layers. Alternatively, an automatic speech recognition (ASR) system for example, may require shared information at the output of a classifier and domain-specific training at the input. An ASR is required to produce accurate representations of sentences at the output but earlier layers of a DL classifier may need to learn representations of speech from a wide range of accents, alternative pronunciations or vocalizations, etc. [16]. The two methods tested here were aimed at determining whether TL in general is applicable to imagined speech EEG and which strategy, if either, is more suited to the task.

Methodology common to both TL approaches is depicted in Figure 3a. The first step in this process is the selection of source subjects for use in training. Conditional strategies have previously been used for selecting source data for EEG-based TL [22]. Here, source data is selected on the basis of a positive correlation with the target subject's data, obtained using the Pearson correlation coefficient. Each subject's complete dataset is used here to facilitate a correlation assessment on the entire distribution of the data between subjects. Source subjects whose data shows a negative correlation with the target subject's data are ruled out. Selected source data is combined and the order randomized before it is used to train and validate the CNN. The target data is not used at all during this stage. Source data is only used in training the CNNs, not during model testing.

Method 1: TL1 employed fine-tuning of the input layers to the pre-trained network i.e., layers 1a and 1b in Figure 2. Concretely, the method depicted in Figure 3b, required freezing all weights in the trained network, excluding those in the input temporal and spatial convolution layers only, during backpropagation. The same 5-fold cross-validation scheme was applied to network fine-tuning on target subject data. For each iteration of the cross-validation method, 4 folds were used to train and one split into validation and test sets.

Method 2: TL2 employs a similar approach to the first (Figure3b). Here, the layers for fine-tuning the pre-trained network are the final two feature extractors i.e., layers 5 and 6 in Figure 2. All other weights in the network are frozen during fine-tuning with 5-fold cross-validation, enabling fine-tuning of only the final two layers.

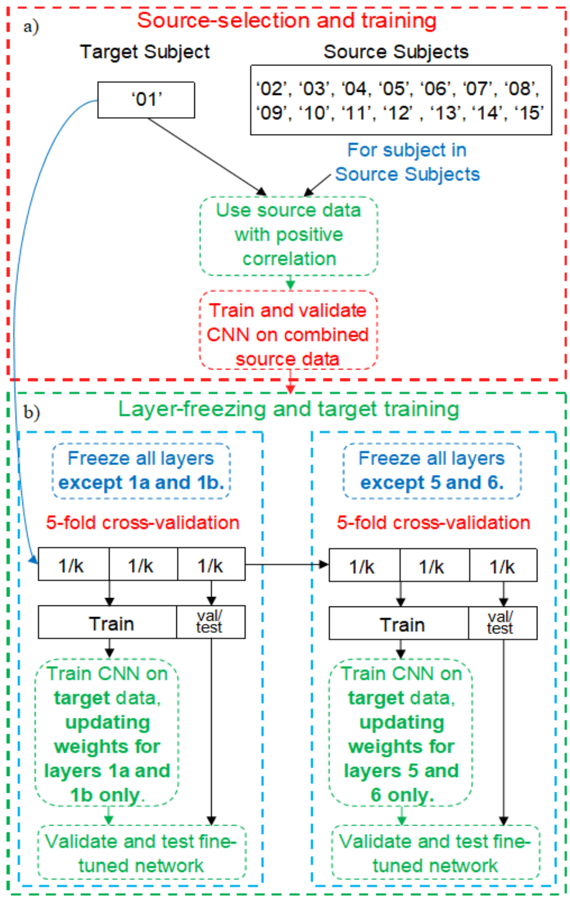

Figure 3 Transfer-learning methodology. a) Selection of source subjects for transfer the CNN. b) Two approaches to fine-tuning.

F. Training

The CNN training protocol was controlled between the two TL methods, with only small differences with the baseline method. All three classifiers were trained using the ADAM optimizer and the cross entropy loss function. Learning rate and batch size was different between the non-TL and TL methods due to the difference in the volume of training data. The TL approaches, with substantially more training data, had learning rates of 0.0001 and batch sizes of 128. The non-TL approach had a learning rate of 0.001 and a batch size of 32. A two-fold stopping criteria was used to manage the number of epochs. First, a maximum number of epochs to reach before stopping was set to 100. Second, a maximum number of epochs to continue training while there is no progress on the validation loss, was set to 50. This *patience* gives the CNNs sufficient training time while protecting against overfitting.

G. Statistics

A repeated measures analysis of variance (ANOVA) was performed on the classification accuracies with the Tukey Honest Significance Difference test performed *post-hoc*. Precision, sensitivity and f-scores were also calculated.

III. Results

Here we report classification accuracies obtained by each of the three methodologies applied to the 5-class imagined vowel decoding task. Mean classification accuracies and standard deviations obtained from each approach are presented in Table 1, and all subject accuracies presented in Figure 4. Both the TL approaches outperform the baseline here, with fine-tuning on the initial convolution layers (TL1) resulting in the highest mean accuracy of 35.68% (stdev.3.01%; chance: 20%). For all but one subject (1), the TL methods returned higher scores than the non-TL method, with TL1 achieving the best performance for 10 of the 15 subjects (Figure 4). This includes the best single-subject performance of 39.09% (subject 5). The repeated measues ANOVA indicated that overall differences in performance were significant ($F(2, 44) = 13.462$, $p<0.001$). The Tukey *post-hoc* tests indicated that the null hypothesis (that the classifiers should perform at the same level) should be rejected between each of the TL methods and the non-TL approach,

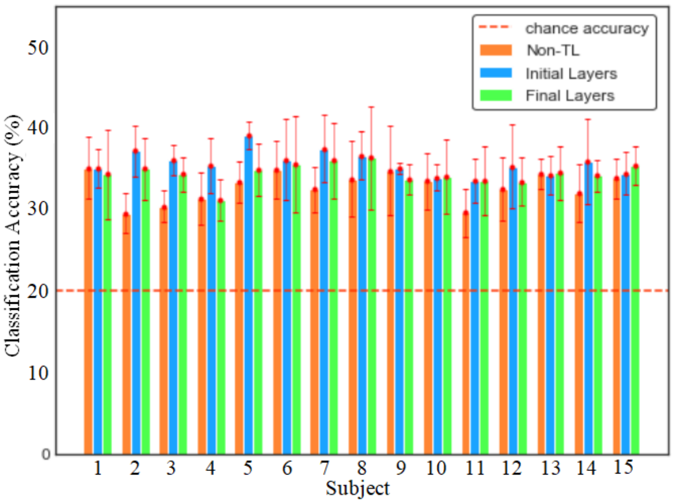

Figure 4 Classification accuracies obtained from each method.

TABLE I. Classification Accuracies for each approach

	Non-TL (%)	*TL1 (%)*	*TL2 (%)*
Mean Accuracy	32.75	35.68	34.41
Standard Deviation	3.23	3.01	3.68

thus indicating that the TL-methods' superioriority was statistically-significant. The Tukey tests did not inidicate significant difference between the two TL methods.

The confusion matrices in Figure 5 show that the three methods produced similar prediction distributions for each of the five classes. Across the three approaches, all classes were predicted with greater than 30% accuracy but the darker shading of the diagonal for the matrices asociated with TL (c and e) are indicative of their greater performance. Of the individual classes, the vowel /u/ received greater prediction accuracy across the classifiers despite not being classified best by either TL method. Receiver operating characteristic (ROC) curves and values for area under the curve (AUC) are presented in Figure 5 (b,d and f). In a highly-accuracte model the ROC curve maps closely to the upper-left of the plot. Here, the ROC curves clearly indicate that none of the the approaches evaluated result in a model that could reasonably be considered feasible for real-world applications. However, TL1 is better than the other methods. The mean AUC is greatest for TL1 (0.6; Figure 5d), indicating that this approach is the most suitable of the three. This result is corrobarated by the statistics in Table II, in which TL1 scored the highest precision, senstivity and accuracy scores.

The training, validation and test loss computed by the training algorithm using cross entropy is presented in Figure 6. Loss here has been averaged across the five folds of the cross-validation scheme. The Non-TL method has a different scale to the TL approaches here due to it having a much greater initial loss. Early-stopping after epoch 65 meant that less training time was required by the within-subject non-TL approach. This is an effect of the training, validation and testing data coming from the same source. Given the random initialization of weights with this method, it is unsurprising that loss decreased rapidly for the duration of the first 20 epochs before settling down to smaller incremental improvements. The loss plots corresponding to the two TL methodologies depict their differences in training. For most of the training time, the training loss of TL2 was much lower than either the validation loss or the test loss. This is indicative of a model that is overfitting and results in more training before convergence (100 epochs). This is in contrast with TL1, in approach which the validation and test losses are much closer to the training loss throughout, indicating that this generalizes better. Training and validation loss are almost identical at the point at which training

TABLE II. Precision, Sensitivity and F1-Scores.

	Precision (%)	*Sensitivity (%)*	*F-score (%)*
Non-TL	33.00	32.91	33.17
TL1	35.64	35.63	36.65
TL2	34.38	34.34	34.45

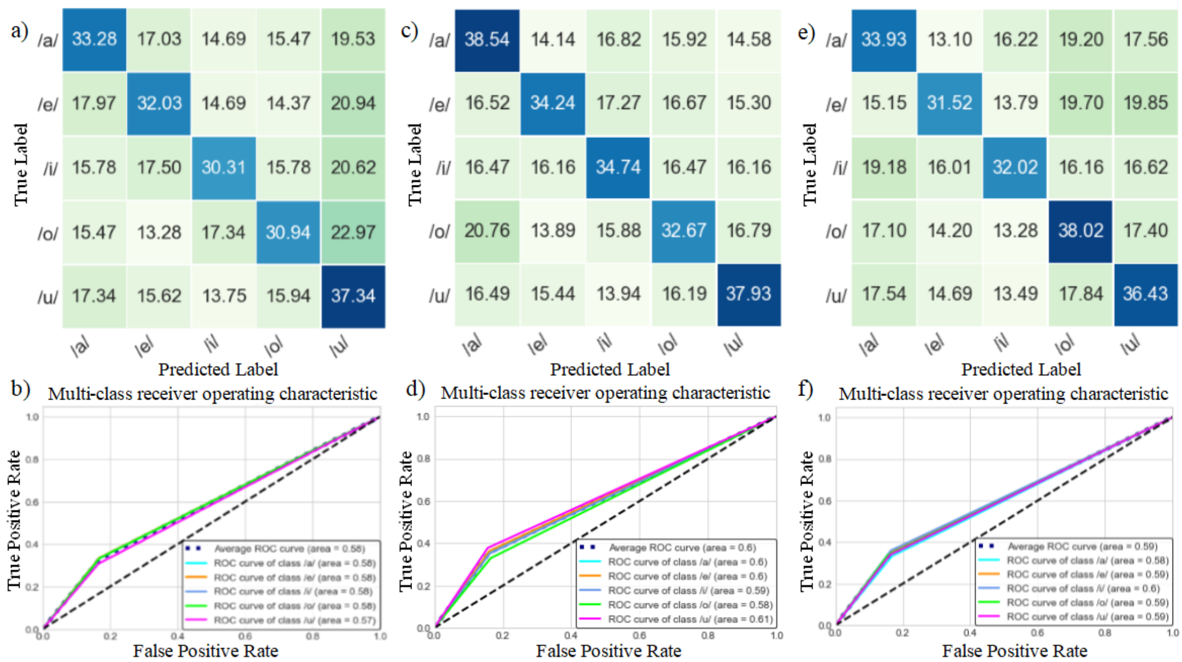

Figure 5 Confusion Matrices for each of the three methods. a) Non-TL, b) Initial layers fine-tuning, c) Final layers fine-tuning.

finishes due to early-stopping (94 epochs), indicating that the model has not been overfit and explaining some of the model's superior performance.

IV. Discussion

The results presented here support the hypothesis that TL methodologies can aid the generalizability of classifiers used for decoding imagined speech EEG. In both cases, the TL approaches outperformed the baseline CNN. Although the improvements to classification accuracy were quite small

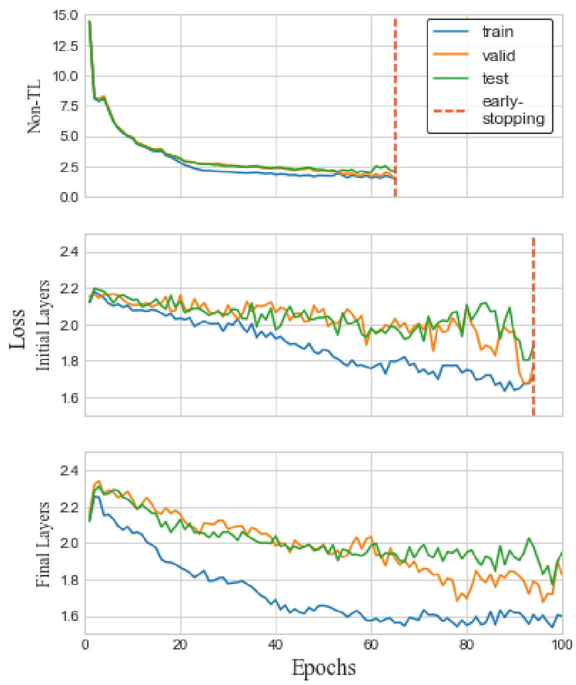

Figure 6 Training, validation and test loss resulting from each method.

(<3%), they were statistically significant and consistent across subjects. Although *post-hoc* tests did not indicate significant difference between the TL methods, it may be the case that fine-tuning input layers is more suited to imagined speech EEG data. Results indicate that the CNN performs better when knowledge is shared among the higher level (output) layers, with domain-specific fine-tuning on the lower level (input) layers. This configuration is analogous to the ASR system discussed in section II-E [16] in which TL performs better when a network attempts to find common output features having been fine-tuned for disparate input representations. However, this observation is based on results obtained from a single dataset and requires validation before it can be confirmed. It is also important that other TL methods are evaluated. Techniques with potential benefits include fine-tuning a randomly-weighted softmax classifier with a pre-trained network [29] and gradual unfreezing of layers aimed at preserving low-level, and adapting, high level features [30].

Research in the field of imagined speech EEG decoding still suffers from a relative scarcity of data. Systems often perform well when training and testing is performed on the same corpus [21], as it has been here. However, it is not practical to expect a model learned on a single corpus to generalize well to new data [31]. Therefore, future work on cross-corpus TL of imagined speech EEG is recommended, as is cross-paradigm TL where different words or sentences produced with imagined speech may be used for knowledge transfer to other imagined speech tasks. Additionally, further study of a conditional system for determining when and how TL methods can be most effectively utilized, as in [22], [32], is an important area of future research.

V. Conclusion

In this study, we investigated the use of TL approaches to classifying imagined speech EEG with CNNs. A dataset consisting of EEG data corresponding to imagined speech production of five vowels was used for testing. Two different TL methods were tested. TL1 involved training the CNN on all source subject data before target subject data was used to fine-tune the input layers of the CNN. TL2 employed the same training strategy, but fine-tuning was implemented on later layers of the CNN. These TL methods were compared to a non-TL approach to training with the same CNN architecture.

The results obtained indicated a statistically significant difference among the performances of the classifiers ($p<0.001$), with mean accuracies of 35.68%, 34.41% and 32.75% respectively. Of the TL strategies, TL1 achieved the higher scores, although this result was not statistically-significant. Overall, the results indicate that there is potential in TL methods to aid decoding of imagined speech EEG. This is an important finding, as increasing cross-subject generalization is essential for any feasible DS-BCI. Equally important is the finding that fine-tuning the input layers of the CNN shows greater performance improvement than fine-tuning the later layers. This indicates that the CNN gains more performance through fine-tuning to the input representations of the imagined speech EEG rather than the output feature representations. Although further work is required to ascertain the applicability of other TL techniques, the results presented here do indicate its potential for enhanced generalizability in a DS-BCI context.

ACKNOWLEDGMENT

The authors acknowledge the dataset's authors for the data used in this study [25]. This work is supported by a Northern Ireland Department for the Economy studentship.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] C. Cooney, R. Folli, and D. Coyle, "Classification of imagined spoken word-pairs using convolutional neural networks," in *Proceedings 8th International Brain-Computer Interface Conference*, 2019.

[10] P. Saha, M. Abdul-mageed, and S. Fels, "SPEAK YOUR MIND! Towards Imagined Speech Recognition With Hierarchical Deep Learning," *arXiv Prepr. arXiv1801.05746V1*, pp. 1–5, 2019.

[11] 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.

[12] 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.

[13] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet : A Compact Convolutional Network for EEG - based Brain - Computer Interfaces," pp. 1–19, 2017.

[14] 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.

[15] 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.

[16] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[17] S. J. Pan and Q. Y. Fellow, "A Survey on Transfer Learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[18] T. Evgeniou and B. De Constance, "Regularized Multi – Task Learning," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 109–117.

[19] R. Raina and A. Y. Ng, "Self-taught Learning : Transfer Learning from Unlabeled Data," in *Proceedings of the 24th International Conference on Machine Learning*, 2007, pp. 759–766.

[20] Y. Sawada and K. Kozuka, "Transfer Learning Method using Multi-Prediction Deep Boltzmann Machines for a small scale dataset," *2015 14th IAPR Int. Conf. Mach. Vis. Appl.*, pp. 110–113, 2015.

[21] M. Völker, R. T. Schirrmeister, L. D. J. Fiederer, W. Burgard, and T. Ball, "Deep Transfer Learning for Error Decoding from Non-Invasive EEG," in *2018 6th International Conference on Brain-Computer Interface (BCI)*, 2018, pp. 1–6.

[22] Y. Lin and T. Jung, "Improving EEG-Based Emotion Classification Using Conditional Transfer Learning," *Front. Hum. Neurosci.*, vol. 11, no. June, pp. 1–11, 2017.

[23] Y. Jiang *et al.*, "Seizure Classification From EEG Signals Using Transfer Learning , Semi-Supervised Learning and TSK Fuzzy System," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 12, pp. 2270–2284, 2017.

[24] J. S. García-salinas, L. Villase, C. A. Reyes-garcía, and A. A. Torres-garcía, "Transfer learning in imagined speech EEG-based BCIs," *Biomed. Signal Process. Control*, vol. 50, pp. 151–157, 2019.

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

[26] F. Lotte, "Defining and quantifying users ' mental imagery-based BCI skills : a first step," *J. Neural Eng.*, vol. 15, no. 046030, p. 23pp, 2018.

[27] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[28] A. K. Reyes, J. C. Caicedo, J. E. Camargo, and U. A. Nari, "Fine-tuning Deep Convolutional Networks for Plant Recognition," in *CLEF (WORKING NOTES)*, 2015, p. 1391.

[29] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual Training of Deep Neural Networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7319–7323.

[30] J. Howard, "Universal Language Model Fine-tuning for Text Classification," *arXiv Prepr. arXiv1801.06146*, 2018.

[31] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer Learning for Improving Speech Emotion Classification Accuracy," *arXiv Prepr. arXiv1801.06353*, 2018.

[32] C. Wei, Y. Lin, Y. Wang, C. Lin, and T. Jung, "A subject-transfer framework for obviating inter- and intra-subject variability in EEG-based drowsiness detection," *Neuroimage*, vol. 174, no. June 2017, pp. 407–419, 2018.