Proceedings of International Joint Conference on Neural Networks, San Jose, California, USA, July 31 – August 5, 2011

# Classification of EEG During Imagined Mental Tasks by Forecasting with Elman Recurrent Neural Networks

Elliott M. Forney and Charles W. Anderson

*Abstract*— **The ability to classify EEG recorded while a subject performs varying imagined mental tasks may lay the foundation for building usable Brain-Computer Interfaces as well as improve the performance of EEG analysis software used in clinical settings. Although a number of research groups have produced EEG classifiers, these methods have not yet reached a level of performance that is acceptable for use in many practical applications. We assert that current approaches are limited by their ability to capture the temporal and spatial patterns contained within EEG. In order to address these problems, we propose a new generative technique for EEG classification that uses Elman Recurrent Neural Networks. EEG recorded while a subject performs one of several imagined mental tasks is first modeled by training a network to forecast the signal a single step ahead in time. We show that these models are able to forecast EEG well with an RMSE as low as 0.110. A separate model is then trained over EEG belonging to each class. Classification of previously unseen data is performed by applying each model and assigning the class label associated with the network that produced the lowest forecasting error. This approach is tested on EEG collected from two able-bodied subjects and one subject with a high-level spinal cord injury. Classification rates as high as 93.3% are achieved for a two-task problem with decisions made every second yielding a bitrate of 38.7 bits per minute.**

## I. Introduction

The ability to effectively classify electroencephalogram (EEG) recorded while a subject voluntarily alters their mental state may lead to an exciting new form of human-computer interface where a device can be controlled entirely through changes in mental activity. These Brain-Computer Interfaces (BCI) are of immediate interest to those who have lost a level of motor control due to conditions such as high-level spinal cord injury, ALS or stroke. Beyond assistive technology, one might imagine applications in virtual reality, gaming, monitoring of emotional states and every-day interaction between humans and machines. Additionally, better algorithms for the classification of EEG recorded during voluntary changes in mental state would certainly lead to improved automated analysis of EEG in clinical settings. For example, similar approaches may be used to detect abnormal sleep patterns or predict oncoming epileptic seizures.

As exciting as these prospects may be, reliable EEG classification is extremely difficult to achieve. The electrical activity generated by the brain is measured on the microvolt level and, due to the dissipation of the signal as it passes through the skull, meninges and scalp, only superficial activity can be measured. Additionally, EEG tends to be highly contaminated with interference from internal sources, such as ocular movement, sinus rhythms and muscle activity, as well as external sources, such as alternating electrical currents and computer peripherals. Combined, this leads to a very low signal to noise ratio and a large number of artifacts. Perhaps the most challenging aspect of EEG classification is the sheer complexity of the signal. EEG contains information that is embedded in both time and space that is representative of large pools of neurons that each exhibit complex behavior in themselves.

We choose to focus on discriminating between EEG produced while subject performs each of several different imagined mental tasks. If a classifier is able to accurately identify when a subject is performing each mental task, the subject can control a device by switching their mental state from task to task in a controlled manner. We feel that this approach is general, offers many degrees of freedom to BCI users, allows for mutual learning and retains the potential to be extended to other forms of EEG analysis. Although the ability to classify EEG recorded during imagined mental tasks has been demonstrated by several research groups in recent years [1], [2], [3], [4], current classification accuracies are not yet high enough for use in many practical applications. We assert that one of the major barriers encountered by current approaches is a limited ability to capture both spatial and temporal information. In [1] and [2], for example, Millán et al. use binned Power Spectral Densities (PSD) to generate features. Although this approach yields impressive results, PSD's represent only the estimated power across a range of frequencies and cannot readily express differences in phase across multiple sensors or the exact order in which short-term signal events occurred. In [3] and [4], Anderson et al. use Time-Delay Embedding (TDE) in order to capture temporal as well as spatial patterns. Although TDE avoids many of the limitations encountered with PSD's, it is fundamentally limited by the size of the embedding dimension and typically leads to a high dimensional input to the classifier. Another limitation of current approaches may be the prominent use of purely linear classification methods [5]. Despite the fact that a level of success has been met with linear methods, we believe that non-linear methods may offer superior performance when combined with an appropriate feature representation and strict regularization. Given the inherently complex and dynamic nature of the human brain, it seems likely that the most effective EEG classifiers must be capable of identifying

Elliott Forney and Charles Anderson are with the Department of Computer Science, Colorado State University, Fort Collins, Coloardo, USA (email: {idfah, anderson}@cs.colostate.edu).

This work was partially supported by the National Science Foundation through Grant Number 0208958.

978-1-4244-9637-2/11/$26.00 ©2011 IEEE

highly non-linear and temporal patterns.

In this paper we explore a generative method for addressing these concerns using Elman Recurrent Neural Networks (ERNN), which are powerful finite state machine approximators capable of learning both temporal and spatial patterns. In Section II we begin by describing the collection procedures and preprocessing methods used to assemble three EEG datasets that are used in subsequent offline experiments. In Section III we proceed by describing a method for modeling EEG by training ERNN's to forecast the signal a single step ahead in time. We then demonstrate that ERNN's are capable of modeling EEG better than several types of naive forecasters and that the resulting models can capture complex temporal dynamics. In Section IV we formulate a classification method where a separate ERNN is used to model EEG recorded during each mental task. Classification of previously unseen data can then be performed by applying each ERNN and selecting the class label associated with the model that produced the lowest forecasting error, a technique we refer to as Classification via Forecasting (CVF). We then provide offline experimental results showing that this approach performs well, achieving up to 93.3% correct test classification at one-second intervals for two tasks, or 38.7 bits per minute (bpm). Finally, we offer some discussion regarding the performance of our classifier and potential improvements to this technique in Section V.

It should be noted that while the paradigm for signal classification proposed here is similar to techniques proposed by Coyle, Gupta, and Oeda, [6], [7], [8], our network architecture, training methods and implementation are quite unique. Furthermore, Coyle et al. are, to our knowledge, the only other group that has applied a similar approach to the classification of EEG during imagined mental tasks and their technique differs from ours in several key ways. A brief comparison of our methods and results to those found in [6] is provided in Section V.

## II. EEG Collection and Preprocessing

In the following experiments we examine three datasets, each recorded from different subjects. Each subject was presented with a visual cue on an LCD screen requesting them to perform one of four imagined mental tasks for the duration of the cue. A five-second break was alloted between each sequence during which the subject was instructed to relax. Presentation of the visual cues and data collection were performed using custom software [9].

Subject-A and Subject-B are both able-bodied, right-handed, male volunteers in their mid-twenties. Data collection from Subject-A and Subject-B was performed in the CSU EEG Pattern Analysis Laboratory and the imagined mental tasks were: clenching of right hand, shaking of left leg, visualization of a tumbling cube and counting backward from 100 by 3's. Ten five-second sequences were recorded during each imagined mental task for a cumulative 50 seconds of EEG per subject per task.

Subject-C is a right-handed, male subject in his mid-twenties with quadriplegia due to a complete spinal lesion at vertebrae C4. This dataset was recorded in the subject's home in order to capture real-world environmental conditions. For Subject-C the imagined mental tasks were: counting backward from 100 by 3's, clenching of right hand, visualization of a tumbling cube and silently singing a favorite song. Five ten-second sequences were recorded during each imagined mental task, again yielding 50 seconds of EEG per subject per task.

In order to simplify our initial analysis, we have chosen to look only at two of the four tasks from each dataset. The "imagined right hand movement" and "count backward from 100 by 3's" tasks were chosen because they are common to all three datasets and because past experience has demonstrated that these tasks often perform well. Each dataset was divided into two partitions with the first 60% designated for use in training and cross-validation and the last 40% reserved only for use in testing our final classifiers.

All three datasets were recorded using the relatively inexpensive Neuropulse Mindset-24 amplifier with a sampling rate of 256Hz. An Electrocap using the 19 channel 10-20 system with common earlobe references was used for electrode placement. The Mindset-24 contains a hardware Sallen-Key filter with a passband of 1.5-34Hz with a 48dB per octave roll-off, so our sampling rate is roughly 4 times the Nyquist rate. A Maximum Noise Fraction (MNF) filter was used in software to remove ocular artifacts [10] [11]. A separate MNF filter was generated for each subject using a five second recording outside of both the training and test partitions by removing the single slowest component. The corresponding MNF filter was then applied to each training and test sequence. Additionally, channel F8 was discarded from all three datasets due to a bad connection during data collection from Subject-C. Finally, each dataset was standardized to have zero mean and unit standard deviation across each channel using only the means and standard deviations from the relevant training partition.

## III. Modeling EEG

Recurrent Artificial Neural Networks (RNN) are an interesting and potentially powerful tool for modeling EEG. RNN's consist of a number of simple computational units with weighted interconnections, including delayed feedback connections. These feedback connections give RNN's an intrinsic state and the ability to learn tasks that require memory. Combined with non-linear activation functions, RNN's are capable of learning complex spatiotemporal patterns.

For the experiments conducted here, we have chosen to use Elman Recurrent Neural Networks (ERNN). Elman networks consist of two fully connected layers: A hidden layer with recurrent connections followed by a strictly feedforward visible layer, as shown in Figure 1. ERNN's have a history of application to novel time series problems and it has been theoretically demonstrated that ERNN's are universal approximators of finite state machines [12], [13]. That is, an ERNN can approximate any finite state machine with arbitrary precision given enough hidden units and the proper weight values. Training of our networks is performed using a batch gradient descent. In order to estimate the error gradient, the network is unrolled a number of steps back through time so that it resembles a multilayer feedforward network in a process known as Back-Propagation Through Time (BPTT) [14], [15]. We then iteratively update the network's weight values using Scaled Conjugate Gradients (SCG), a fast gradient descent algorithm pioneered by Møller that estimates second order gradient information [16]. Although there is relatively little found in current literature regarding the use of SCG to train RNN's, it is demonstrated favorably in [17].

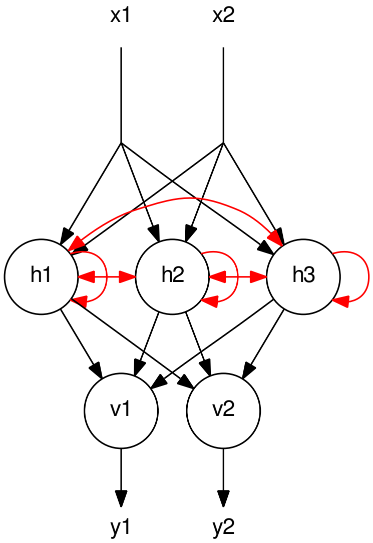


Fig. 1. An Elman Recurrent Neural Network with two inputs, three hidden units and two outputs. Inputs are denoted $\mathbf{x}$, hidden units are $\mathbf{h}$, visible units are $\mathbf{v}$ and outputs are $\mathbf{y}$. Note that the hidden layer has full recurrent connections.

The activation function used in the hidden layer of our Elman Networks is a variant of the hyperbolic tangent suggested by LeCun et al. in [18] while the activation in the visible layer is strictly linear. The initial weight values for all units are drawn from a uniform random distribution, again using a strategy outlined in [18]. Our networks are unrolled 38 steps back in time, an empirically determined value after which little performance improvement is seen. All recurrent inputs are initially set to zero. The network is then allowed an initial transient period, $\rho = 38$, in order for the recurrent dynamics to acclimate to the signal. Our implementation is custom written in the R programming language [19].

We begin to explore the ability of ERNN's to model EEG by training a network to forecast the $N$-dimensional signal a single step ahead in time, where $N = 18$ is the number of channels. In other words, the network is configured so that

$$\mathbf{y}(t+1) = \mathbf{ernn}(\mathbf{x}(t)) \tag{1}$$

where $t$ is the current timestep, $\mathbf{y}$ is the $N$x1 dimensional forecast value of the signal, $\mathbf{x}$ is the $N$x1 dimensional current value of the signal and **ernn** is the application of our Elman network. During training, we minimize the mean squared error (MSE)

$$\frac{1}{(T-\rho)\cdot N}\sum_{t=\rho}^{T}\sum_{n=1}^{N}(y_n(t) - x_n(t))^2 \tag{2}$$

where $T$ is the length of the EEG sequence and $\rho = 38$ is the initial transient period during which the ERNN is allowed to build initial dynamics.

In order to provide a benchmark for our forecasting performance, we also define two naive error measures. The naive repeated, or Naive-R, error is defined as

$$\frac{1}{(T-\rho)\cdot N}\sum_{t=\rho}^{T}\sum_{n=1}^{N}(x_n(t-1) - x_n(t))^2 \tag{3}$$

and is equivalent to forecasting by repeating the previous input. The naive interpolated, or Naive-I, error is defined as

$$\frac{1}{(T-\rho)\cdot N}\sum_{t=\rho}^{T}\sum_{n=1}^{N}(l_n(t) - x_n(t))^2 \tag{4}$$

where

$$\mathbf{l}(t) = 2\left[\mathbf{x}(t-1) - \mathbf{x}(t-2)\right] + \mathbf{x}(t-2) \tag{5}$$

and is equivalent to forecasting by performing a linear interpolation of the previous two inputs.

We evaluate the forecasting performance of our ERNN's by applying a 6-fold cross-validation over the training partition for the first task recorded from Subject-A. Training is terminated after 250 iterations of SCG because empirical evidence suggests that little improvement is seen with more training epochs and regularization can be controlled by limiting the number of hidden units. In Figure 2 we see the training and validation forecasting errors versus the naive errors as the number of hidden units in our ERNN is varied. It quickly becomes clear that our ERNN's are able to model EEG well, achieving a training root mean squared error (RMSE) as low as 0.110 and a validation RMSE as low as 0.125 while Naive-R and Naive-I achieve an RMSE of 0.392 and 0.310 respectively. Validation RMSE drops below Naive-R with 6 hidden units and below Naive-I with 8.

Interestingly, the training and validation errors begin to noticeably diverge after roughly 15 hidden units even though they both continue to fall, albeit slowly after about 40 hidden units. If the ERNN is overfiting the signal, we would expect to see the validation error rise. Instead, Figure 2 suggests that while aspects of the signal are becoming overfit, the networks continue to learn components of the signal that are common to all of the EEG sequences.

An interesting way to evaluate the temporal dynamics of our ERNN models is to first train the ERNN as described by equations (1) and (2) and then place a feedback loop between the network's outputs and inputs so that

$$\mathbf{y}(t+1) = \mathbf{ernn}(\mathbf{y}(t)) \tag{6}$$

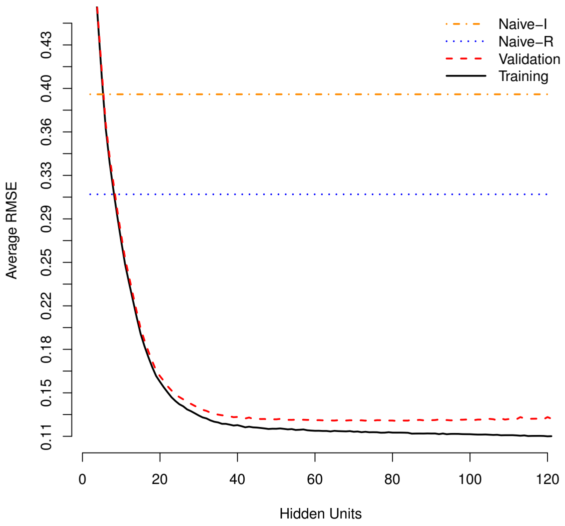


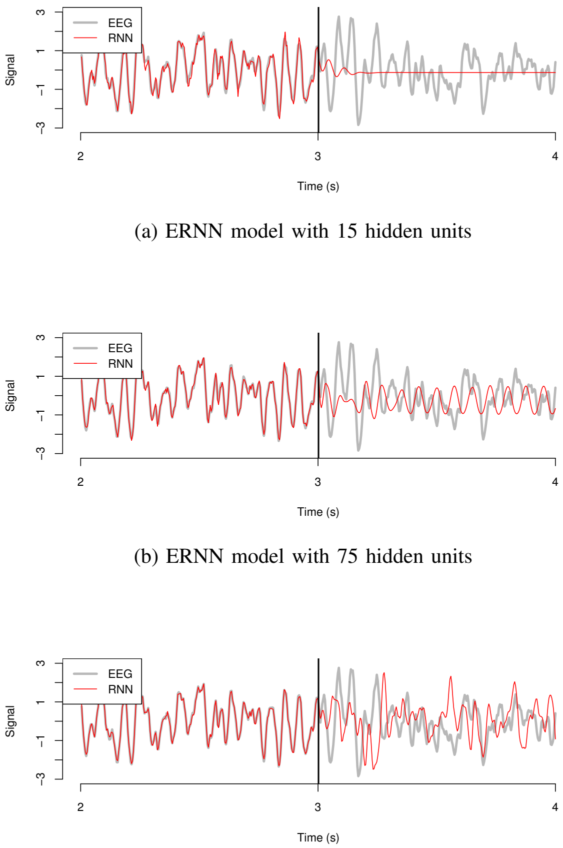
Fig. 2. Average modeling error as the number of hidden units is varied using 6-fold cross-validation for the first task recorded from Subject-A.

when $t > T$. In this way our model becomes an iterated, autonomous system. In Figure 3 we see such a model as it transitions from the behavior described in equation (1) to that of equation (6). The model was trained over all 18 channels of the first three seconds of EEG recorded from Subject-A. After the three second mark, the model's previous outputs are fed back into its inputs and the result is superimposed over the remaining two seconds of EEG. Although Figure 3 only shows one second before and after this transition on channel P4, each channel typically exhibits similar behavior. With only 15 hidden units the autonomous model quickly dampens to zero. With 75 hidden units the model quickly and indefinitely falls into a periodic state. A relatively large network of 150 units, on the other hand, appears to produce rich and long-lasting dynamics. Although the autonomous model quickly diverges from the true EEG, they do appear to have similar characteristics and a subsequent analysis using continuous wavelet transforms has revealed that these autonomous models have a spectrum that is similar to the true EEG.

Based on the above experiments, it seems clear that our ERNN's are able to model EEG well. They are able to forecast the signal considerably better than the proposed naive solutions and they are able to form autonomous models with rich temporal dynamics by placing a feedback loop between the network's outputs and inputs. It appears, however, that a relatively large number of hidden units may be required in order to achieve long-term dynamics.

## IV. Classification of EEG

Next, we examine a method for EEG classification that utilizes the errors produced by the ERNN models described in Section III. We refer to this method as Classification via Forecasting (CVF). In this approach, a separate ERNN is trained to forecast sample EEG recorded while a given subject performs each imagined mental task. Thus, if we have $K$ imagined mental tasks, then we must train $K$ different ERNN models. In this way, we have an expert at forecasting EEG belonging to each class. Previously unseen EEG can then be classified by applying each ERNN and assigning the class label associated with the model that produced the lowest forecasting error. In order to improve performance and provide classifications at a rate that is sensible for use in BCI, the error measure used for classification is the MSE across all channels over one-second windows.



(a) ERNN model with 15 hidden units

(b) ERNN model with 75 hidden units

(c) ERNN model with 150 hidden units

Fig. 3. To the left of the 3 second mark we see an ERNN forecasting EEG a single step ahead. To the right of the 3 second mark we see the ERNN operating autonomously, driven only by its previous predictions. Both sides are superimposed over the true EEG. 3a) with 15 hidden units the autonomous model quickly dampens to zero. 3b) with 75 hidden units the autonomous model falls into clearly periodic state. 3c) with 150 hidden units the autonomous model has very rich and EEG-like dynamics.

Most of the parameters used to train our ERNN's are carried over from Section III. However, the appropriate number of hidden units for classification remains to be determined because a different level of regularization may be required. In Figure 4 we see our training and validation classification accuracies for Subject-A versus the number of hidden units used in our ERNN models.

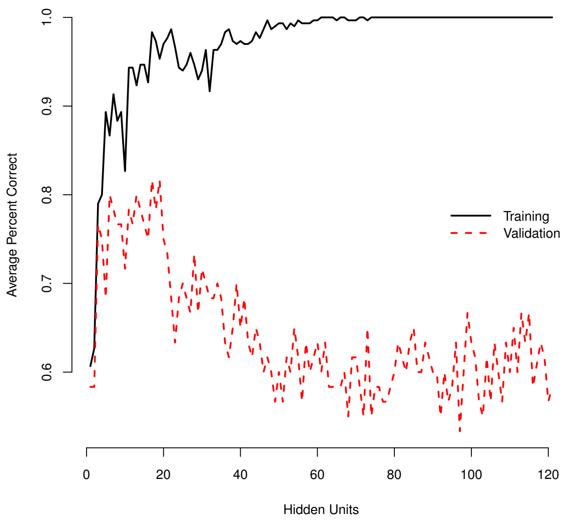


Fig. 4. Average classification accuracies as the number of hidden units is varied using a 6-fold cross validation over the training partition recorded from Subject-A.

Surprisingly, classification accuracy appears to peak with a mere 15-20 hidden units. Despite the evidence from Section III that ERNN's with a large number of hidden units produce lower forecasting errors and are better able to capture long-term temporal dynamics, it appears that our classification problem is quickly overfit. Although this may seem paradoxical, there several are several potential explanations. First, there may be common signal components across EEG sequences recorded during different mental tasks. In other words, as the number of hidden units is increased beyond 20 the ERNN may continue learning common signal components, causing both the training and validation forecasting errors to fall, while simultaneously learning signal noise and non-discriminative components, causing classification accuracy drop. This possibility is supported by the fact that the training and validation errors in Figure 2 begin to diverge even though they both continue to fall.

Additionally, the fact that CVF appears to reach its optimal classification rate with relatively few hidden units might suggest that much of the discriminating information contained in the EEG sequences lies in frequency bands near the upper corner of the passband, i.e. 34Hz. As the memory capacity of the ERNN grows, there may be a trade-off between modeling short-term, high-frequency patterns that discriminate between classes and long-term, low-frequency patterns that contribute largely to the modeling error but are not discriminative. In Section V we will briefly discuss methods for exploring these possibilities and potentially overcoming them.

Next, we apply the same classification algorithm and validation procedure to the remaining two datasets in order to find the optimal number of hidden units and classification accuracy for each subject. For Subject-A, Subject-B and Subject-C, 17, 9 and 15 hidden units are found to optimal respectively. In Table I we see the optimal number of hidden units (NH) as well as the training, validation and test classification accuracies for all three subjects. Training and validation accuracies appear to remain relatively high for all three subjects. Subject-C produces the best test classification rate achieving a test accuracy of 93.3%, while Subject-B performs relatively poorly achieving a test accuracy of only 58.8%. Since the validation error for Subject-B was 86.7%, we conjecture that Subject-B may have been distracted during part of the recording session or that the test partition may contain unusual artifacts.

| | NH | Training | Validation | Test |
|---|---|---|---|---|
| Subject-A | 17 | 98.3% | 81.7% | 78.3% |
| Subject-B | 9 | 99.0% | 86.7% | 58.8% |
| Subject-C | 15 | 97.5% | 86.7% | 93.3% |

TABLE I
PERCENT CORRECT CLASSIFICATION

In order to provide a more universal indicator of our classification performance, we also provide information transfer rates in bits per minute (bpm). This value is calculated according to [20], [21]

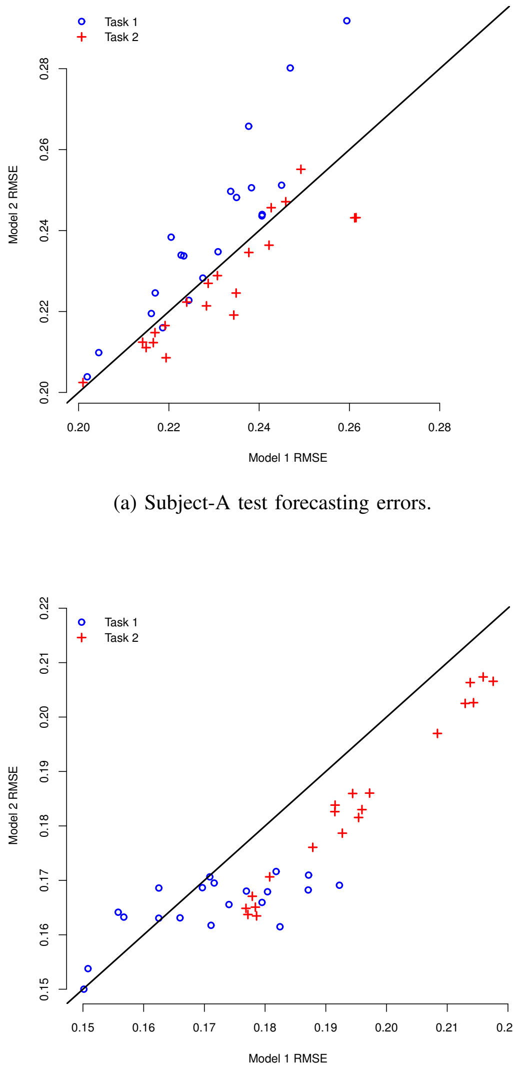
$$V\left(\log_2 K + P\log_2 P + (1-P)\log_2\frac{1-P}{K-1}\right) \quad (7)$$

where $V = 60$ is the classification rate in decisions per minute, $K = 2$ is the number of available classes and $P$ is the classification accuracy as the fraction of correct decisions over total decisions. In Table II we see our classification performance in terms of bitrates. These results suggest that our formulation of CVF is likely to deliver an information transfer rate roughly in the range of 14-38bpm.

| | NH | Training | Validation | Test |
|---|---|---|---|---|
| Subject-A | 17 | 52.5bpm | 18.8bpm | 14.7bpm |
| Subject-B | 9 | 55.2bpm | 26.1bpm | 1.35bpm |
| Subject-C | 15 | 49.9bpm | 26.1bpm | 38.7bpm |

TABLE II
CLASSIFICATION BITRATES

Next, we briefly examine the effectiveness of our strategy for assigning class labels. Although our winner-takes-all strategy has proven successful, it may not be optimal to choose class labels based simply on the lowest forecasting error. For example, it may be possible for all models to produce a lower forecasting error for one specific task, even if the errors are separable. In other words, it is possible for both models to be biased toward one class. In Figure 5 see we our modeling errors plotted against each other for each task in the test sets for Subject-A and Subject-B. When viewed this way, the discriminant function is along the diagonal. The diagonal does indeed separate the two tasks quite well for Subject-A. For Subject-B, however, it appears that there is a bias toward model two and, therefore, class two. Clearly, a better discriminant function could be found for the test set of Subject-B, although it is not clear how well this function might generalize without further experimentation.



(a) Subject-A test forecasting errors.

(b) Subject-B test forecasting errors.

Fig. 5. Forecasting errors for model 1 versus model 2 for each class. Since we assign class labels based on the lowest forecasting error, the diagonal can be viewed as the discriminant function. 5a) For Subject-A, the diagonal discriminates between the classes extremely well. 5b) For Subject-B, the erros are biased toward model 2 resulting in poor classification.

Furthermore, averaging our forecasting errors over one-second windows may not be ideal. While this process does increase the separability between classes by accumulating information and reducing noise in our modeling error, it is also likely that some information is lost. Alternately, our forecasting errors could be viewed as features and the entire window could be fed into another classifier. This does, however, introduce another layer of complexity and, potentially, additional regularization parameters.

## V. Discussion and Future Work

We have outlined a technique where RNN's are used to model EEG by forecasting the signal a single step ahead in time. A classifier can then be built by training a separate RNN over EEG recorded during each of the mental states of interest. Previously unseen data is subsequently identified using the CVF approach. That is, each model is applied to the test EEG and the class label associated with the network that produced the lowest mean forecasting error over a brief window is selected.

In the above experiments we have demonstrated that our ERNN's are able to closely model EEG as well as capture some of the temporal information that they contain. We have also shown that relatively high classification rates can be achieved using our implementation of ERNN's and CVF. These classification rates and the parameters used to train our ERNN's appear to be fairly robust across subjects and recording environments even when a relatively inexpensive EEG acquisition system is used. In fact, our highest test classification rate was achieved with Subject-C. As a person with quadriplegia using the system in their home environment, Subject-C certainly belongs to a target demographic for this technology.

Although the approach described here is proposed to address concerns with PSD and TDE feature representations, a direct comparison with each of these techniques has yet to be performed. Furthermore, comparison with current literature regarding the classification of EEG during imagined mental tasks is difficult due to the use of varying decision and rejection strategies and mutual learning through user feedback, making bitrates and classification accuracies a poor performance measure. Nevertheless, an examination of [2] and [4] reveals that these state-of-the-art approaches achieve anywhere from 52% correct on three tasks every 0.5 seconds to 78% correct on five tasks every 1.8 seconds before any user feedback. Entering these values into (7) yields bitrates of 12bpm and 37bpm respectively. This suggests that the experiments performed here are competitive with current PSD and TDE based approaches.

Coyle et al. have also experimented with a method for classifying EEG during imagined motor movement using forecasting errors produced by neural networks [6]. However, their approach differs from ours in several key ways. First of all, feedforward networks were used in conjunction with TDE, as opposed to a recurrent architecture. While this provides an important point of reference, the use of TDE can be problematic for the same reasons mentioned in Section I. Second, the window of forecasting errors is viewed as a vector of high dimensional features that is fed to a Linear Discriminant Analysis classifier in a fashion similar to that described in the latter part of Section IV. In contrast, our approach simply averages the error over a window and assigns class labels in a winner-takes-all approach. Finally, it should be mentioned that the bitrates achieved in [6] do not exceed 9.96bpm, significantly lower than the rate achieved here.

A number of questions about the approach described here remain to be answered in future work. First of all, the exact cause of the contradiction between many hidden units for low modeling error and relatively few hidden units for optimal classification should be explored further. If this problem is in fact caused by components of EEG that are common across each mental task, then perhaps dimensionality reduction and source separation techniques such PCA, ICA or MNF could remove some of these common elements. Additionally, if there is indeed a trade-off between modeling short-term, high-frequency and long-term, low-frequency temporal patterns in EEG, it may be beneficial for RNN's to model the signal at multiple time-scales. This could be achieved by training the models to simultaneously forecast the signal multiple steps ahead in time or by training a single RNN to model multiple incarnations of the same EEG filtered with different passbands and decimated to different sampling rates. Finally, it is important that CVF using ERNN's is directly compared with other approaches and that it is tested in an online setting in order to gauge its true performance and overall user experience.

## References

[1] J. Millán, P. Ferrez, F. Galán, E. Lew, and R. Chavarriaga, "Non-invasive brain-machine interaction," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 20, no. 10, pp. 1 – 13, 2007.

[2] J. Millán, J. Mourino, M. Franze, F. Cincotti, M. Varsta, J. Heikkonen, and F. Babiloni, "A local neural classifier for the recognition of eeg patterns associated to mental tasks," *Neural Networks, IEEE Transactions on*, vol. 13, no. 3, pp. 678 – 686, 2002.

[3] C. Anderson, E. Forney, D. Hains, and A. Natarajan, "Reliable identification of mental tasks using time-embedded eeg and sequential evidence accumulation," *Journal of Neural Engineering*, vol. 8, pp. 25 – 23, 2011.

[4] C. W. Anderson and J. A. Bratman, "Translating thoughts into actions by finding patterns in brainwaves," *Proceedings of the Fourteenth Yale Workshop on Adaptive and Learning Systems*, pp. 1 – 6, 2008.

[5] K. Muller, C. Anderson, and G. Birch, "Linear and nonlinear methods for brain-computer interfaces," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 11, no. 2, pp. 165 – 169, 2003.

[6] D. Coyle, G. Prasad, and T. McGinnity, "A time-series prediction approach for feature extraction in a brain-computer interface," *Neural Systems and Rehabilitation Engineering, IEEE Transactions oni*, vol. 13, no. 4, pp. 461 – 467, 2005.

[7] L. Gupta, M. McAvoy, and J. Phegley, "Classification of temporal sequences via prediction using the simple recurrent neural network," *Pattern Recognition*, vol. 33, no. 10, pp. 1759 – 1770, 2000.

[8] I. K. Shinichi Oeda and T. Ichimura, "Time series data classification using recurrent neural network with ensemble learning," *Lecture Notes in Computer Science*, vol. 4253, pp. 742 – 748, 2006.

[9] "CEBL: CSU EEG Brain-Computer Interface Lab," 2010. [Online]. Available: http://www.cs.colostate.edu/eeg/eegSystem.html

[10] J. Knight, "Signal fraction analysis and artifact removal in eeg," *Masters Thesis, Department of Computer Science, Colorado State University, Fort Collins, CO.*, 2003.

[11] C. Anderson, J. Knight, T. O'Connor, M. Kirby, and A. Sokolov, "Geometric subspace methods and time-delay embedding for eeg artifact removal and classification," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 14, no. 2, pp. 142 – 146, 2006.

[12] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179 – 211, 1990.

[13] S. C. Kremer, "On the computational power of elman-style recurrent networks," *IEEE Transactions on Neural Networks*, vol. 6, pp. 1000 – 1004, 1995.

[14] R. J. Williams and J. Peng, "An efficient gradient-based algorithm for on-line training of recurrent network trajectories," *Neural Computation*, vol. 2, pp. 490 – 501, 1990.

[15] S. Haykin, *Neural networks and learning machines*. New Jersey, USA: Prentice Hall, 2009.

[16] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *NEURAL NETWORKS*, vol. 6, no. 4, pp. 525 – 533, 1993.

[17] C. Gruber and B. Sick, "Fast and efficient second-order training of the dynamic neural network paradigm," in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, vol. 4. IEEE, 2003, pp. 2482 – 2487.

[18] Y. LeCun, L. Bottou, G. Orr, and K. Müller, "Efficient backprop," *Neural networks: Tricks of the trade*, pp. 546 – 546, 1998.

[19] R. D. C. Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/

[20] J. Pierce, *An introduction to information theory: symbols, signals & noise*. Dover Pubns, 1980.

[21] J. Wolpaw, H. Ramoser, D. McFarland, and G. Pfurtscheller, "Eeg-based communication: improved accuracy by response verification," *Rehabilitation Engineering, IEEE Transactions on*, vol. 6, no. 3, pp. 326 – 333, 1998.