Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Possibility of reinforcement learning using event-related potential toward an adaptive BCI

Kazuhiro Nomoto, Tadashi Tsubone, Yasuhiro Wada
Department of Electrical Engineering
Nagaoka University of Technology
Nagaoka, Japan

***Abstract*—We applied event-related potential (ERP) to reinforcement signals that are equivalent to reward and punishment signals. We conducted an experiment using an electroencephalogram (EEG) in which volunteers identified the success or failure of an inverted pendulum task. We confirmed that there were differences in the EEG signal depending on whether the task was successful or not and that ERP might be used as a punishment of reinforcement learning. We used a support vector machine (SVM) for recognizing the ERP. We selected the feature vector in SVM that was composed of averages of each 35 msec for each of three channels (F3,Fz,F4) on the frontal area, for a total of 700 msec. Our experimental results suggest that reinforcement learning using ERP can be performed accurately. Finally, we suggest the possibility of developing an adaptive brain-computer interface (BCI) by ERP.**

***Index Terms*—BCI, ERP, Reinforcement learning.**

I. INTRODUCTION

Recently, brain-computer Interfaces (BCIs) that can control various equipment such as robots by brain signal activity have been widely researched. BCIs are tools for direct communication between a human brain and a machine [1][2]. Also, this technology will help severely paralyzed people such as amyotrophic lateral sclerosis (ALS) patients to communicate and interact with the outside world. However, we believe that a control algorithm adapted to each BCI user is needed to accurate perform various tasks using brain signals.

The reinforcement learning algorithm can be used as a technique for designing a control algorithm in various environments. If the reinforcement signal, that is, reward and punishment signals, can be estimated from the brain signals, a BCI adapted to each user can be developed by learning from individual brain signals. Mesencephalon dopamine neurons are thought to be the part of the brain associated with reward, but they are near the central areas of the brain, such as the substantia nigra pars compacta, and they are difficult to measure with an electroencephalograph (EEG). We investigated event-related potential (ERP), which is related to recognition or identification of external phenomena and that generally occurs along the midline of the scalp as an alternative to determining reward.

Yamagishi et al. [6] examined a possible reinforcement signal of P300, which is a type of ERP and has a comparatively large amplitude and an electric potential peak change at around 300 msec, from stimulation. They suggested that a detected P300 signal could be used as a signal for reward in reinforcement learning [3] to realize an adaptive BCI. We discuss here the possibility of developing an adaptive BCI system using reinforcement learning in the framework of a control problem, which is different from the previous work done by Yamagishi et al. [6]. In this paper, we address an inverted pendulum problem as a control problem. Reinforcement learning is performed using the success or failure information of the inverted pendulum control estimated from EEG signals. We examine the possibility of applying the correct learning process. That is, in the previous reinforcement learning algorithm, failure is defined as being pendulum angular, that is, the control becomes a failure when the angle is large. In this paper, we attempt to advance the learning using EEG signals for reward or punishment. We discuss here just the possibility of reinforcement learning using EEG and we describe a feasibility study. An actual learning system will be developed in the near future.

II. EXPERIMENT TO ESTIMATE REINFORCEMENT SIGNAL

We investigated whether ERP could be used to detect the success or failure of a task by conducting the following experiment. Four male volunteer subjects aged from 22 - 24 years old took part in the experiment. All volunteers received an explanation of the experiment beforehand and agreed to participate.

A. Experimental setup

An EEG system (Biosemi, ActiveTwo) was used to measure brain activity. The placement of the 16 scalp electrodes was according to the international 10/20 system shown in Figure 1, which shows the arrangement of electrodes from the top view of a head. The standard electrode was the electrode labeled ”ref” attached to the right earlobe. The EEG data from each electrode were sampled at 2,048 Hz, and filtered with a band pass filter of 0.5-7 Hz.

B. Experimental task

Figure 2 shows a trial sequence displayed on a monitor, and Figure 3 shows the time course of one trial. First, Figure 2 shows an inverted pendulum system and a ball in the bottom of the figure. When a session starts, the pole is controlled by moving the cart to the left or the right to maintain the inverted pendulum. The ball is hit by the pole with a sound when the inverted pendulum is a failure. A successful trial is defined as when the inverted pendulum is maintained for five seconds. When the pendulum falls down in less than five seconds, the

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

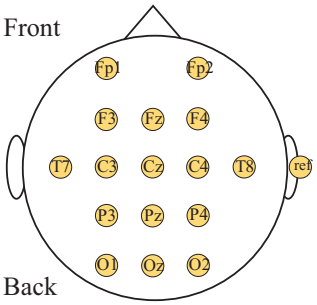

Fig. 1. Arrangement of electrodes

control is defined as a failure. The inverted pendulum was controlled by prepared rules; that is, the inverted pendulum succeeds or fails with a probability. The cart control was done using prepared pendulum state files, which contain cart positions and pendulum angles every 20 msec from start to success or failure. In the failure case, the pendulum motion is simulated by a physical model using velocity and acceleration just after failure. A total of 250 state files, that is, 50 files for success and 200 files for failure, were prepared. A file was randomly selected with a probability. Fig. 4 shows trajectories of pendulum angle for success (green) and failure (red and blue). The failure probability was 20%; that is, one session contained ten failure trials. Probabilities for trials of three seconds to four seconds and for trials of four seconds to five seconds were generated with equal probabilities.

A base time for a successful trial was defined as being when the pendulum stopped, and a base time for failure was defined as being when the pendulum hit the ball with a sound. Subjects were instructed to silently count the number of failure trials just after the sound of hitting the ball in all sessions. In addition, the subjects were directed to think about nothing during the trial. After 2000 msec, the pendulum moved to the initial state with a beep sound, and then subjects were allowed to blink. The next session started after another 1000 msec. The above sequence was one trial; one session contained fifty trials, and one subject participated in eight sessions.

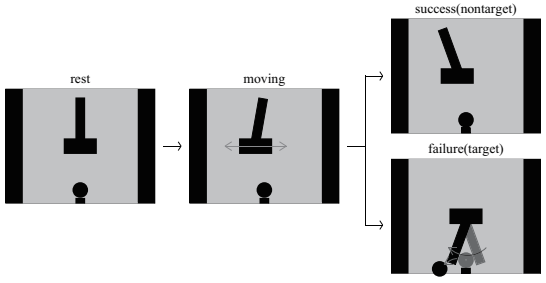

Fig. 2. Sequence on monitor

III. Data Analysis

Trials in which the volunteers were instructed to count were called ”target trials,” and ones in which they were not

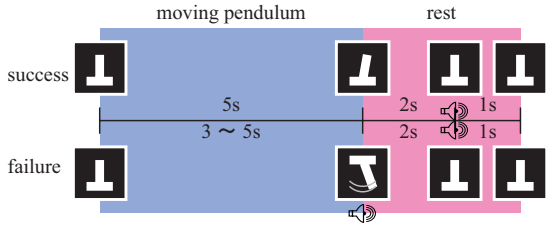

Fig. 3. Time course of a trial

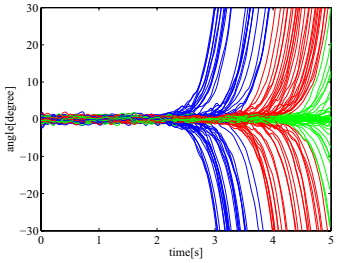

Fig. 4. Pole angular trajectories

instructed to count were ”non-target trials.”

A. Preprocessing

First, the EEG data were resampled at 256 Hz. High-frequency components were removed by using a low-pass filter (third-order Butterworth filter) with a cut-off frequency of 7 Hz because ERPs consist of low-frequency components, and α wave components should be removed. Then, the data between 1000 msec and -3000 msec were clipped where time 0 was defined as the time of the trigger signal described above. Finally, the mean between -100 msec and 0 msec was deducted from the data from every trial as a baseline adjustment. Trials that included a blinking or ocular movement were excluded. Table I shows the number of trials excluding the trials with artifacts.

TABLE I
NUMBER OF TRIALS AFTER ARTIFACT REJECTION

Subject	Correct(Nontarget)	Incorrect(Traget)
S.T	295	68
O.M	310	69
N.T	270	49
A.S	307	72

IV. Results

A. Measurement results

Figure 5 shows the average profile of each channel from four sessions of successful trials and failure trials for each subject. A positive electrical change, which peaked from 200 msec to 400 msec, was confirmed around electrode positions Cz and Fz located at the top, and towards the front of the head. No such feature was found in the successful trials. This

feature, that is, a positive electrical peak around 300 msec, resembles a measured EEG signal in an oddball paradigm. It is well known that P300 can be observed at the back of the head; however, the electrical change that was observed in the experiment was stronger at the top of the head than at the back of the head. Moreover, the negative electrical change could be measured just before the positive electrical change for subject O.M. Therefore, signals measured in the experiment are slightly different from P300.

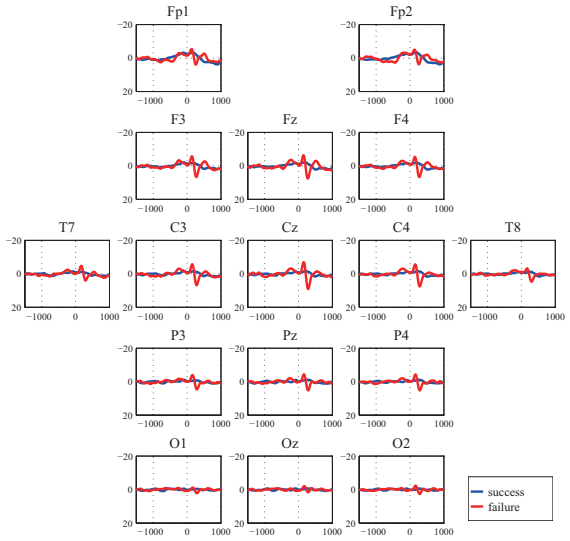

Fig. 5. Average profile of each channel (volunteer O.M.)

V. Possibility of Reinforcement Learning by Using ERP

Here, we consider whether reinforcement learning is possible using EEG signals, that is, whether signals that distinguished target and non-target trials can be used as a reward or punishment in reinforcement learning.

A. Support vector machine

An SVM is a two-class classification machine used in supervised learning and for designing identification functions from training data and for classifying test data by that identification function. The data from sessions 1 - 4 and those from 5 - 8 were used as training data x_i and test data x, respectively. An identification function was defined from Equation (1).

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b\right), \tag{1}$$

where $\mathrm{sgn}(\cdot)$ outputs 1 if the value inside the parenthesis is positive and outputs -1 if the value is negative. Here, $y_i \in \{1, -1\}$ and n show a class label of the training data and the number of training data, respectively. $K(\cdot)$ is a kernel function; a Gaussian kernel function (2) was used in the paper.

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|}{2\sigma^2}\right), \tag{2}$$

where σ expresses a parameter for adjusting the shape of the function, and b from Equation (1) is calculated from Equation (3).

$$b = y_s - \sum_{i=1}^{n} y_i \alpha_i K(x_i, x_s). \tag{3}$$

Here, x_s and y_s represent the note support vector and its class label, respectively, and α_i is the optimal solution of the quadratic optimization problem under the condition of $0 \leq \alpha_i \leq C$. In this paper, two parameter values of σ and C, which gave a better discrimination rate, were found by using a search method.

B. Feature vector

A feature vector is selected to classify a single trial EEG using SVM. It is well known that when a feature vector has a large number of dimensions, it conduces a lower discrimination rate because of overfitting. Therefore, it is important to reduce the number of feature vector dimensions in order to achieve a lower calculation cost and increase the discrimination rate. Generalization capability can be expected with a feature vector that has a proper number of dimensions. In the paper, we selected proper features for a discrimination model of SVM by searching spatial information (EEG channels) and temporal information (time domain and time step).

(1) Channel selection

Figure 6 shows the SVM discrimination rate results using a single EEG channel for the same conditions in time domain and time step. We confirmed that the discrimination rate at the top of the head was higher than that at the back of the head; in particular, the point around Fz (F3, Fz, F4) at the top of the head gave the highest discrimination rate.

Next, we checked multiple channel conditions to find out whether the results were the same as the single channel conditions. Figure 7 illustrates the seven conditions that we checked. Condition 3 consisting of F3, Fz, and F4, which gave the highest discrimination rates for single-channel conditions, had the best rate. Condition 2, which contained F3, Fz, F4, and three other electrodes, and condition 7, which also contained Fz, were even better. These results show that F3, Fz, and F4 should be selected to get the highest discrimination rate.

(2) Time domain condition

We searched the time domain for a feature vector. The time domain was searched from 100 msec to 700 msec every 50 msec, as listed in Table II, and we decided on a time domain condition of around 350 msec. Basically, the discrimination rate increased with a longer time domain, and the maximum time domain of 700 msec gave the best discrimination rate.

(3) Time step condition

Finally, the time step, which is a down-sampling rate of the measured EEG to a feature vector, was searched. The search areas are listed in Table III. The average of the EEG in a time domain was used as the feature vector. The results were almost

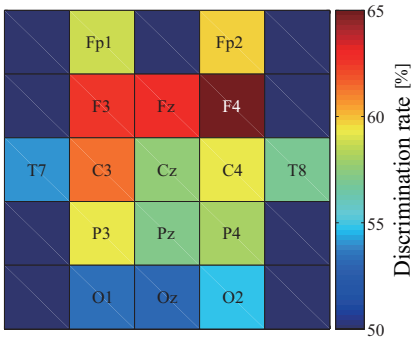

Fig. 6. Discrimination results for single channel

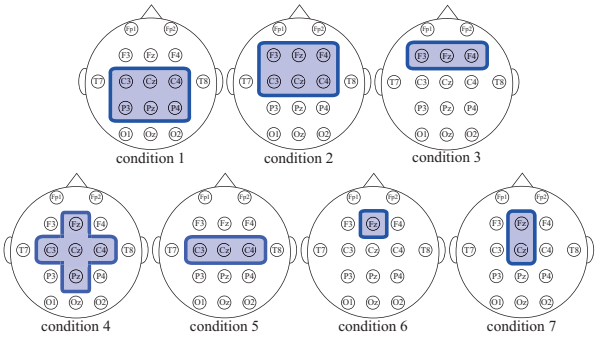

Fig. 7. Channel conditions

the same, although 35 msec gave the highest discrimination rate.

From the above considerations, we selected F3, Fz, and F4 as EEG channels, 700 msec as the time domain, and 35 msec as the time step. Therefore, there were 60 feature vector dimensions, that is, 3 channels x 700 msec / 35 msec.

C. Discrimination results - offline

Figure 8 plots the discrimination results in offline cases using selected feature vectors. We used sessions one to four as training data sets and sessions five to eight as test data sets. There were four subjects, S.T, O.M, N.T, and A.S. The

TABLE II
TIME DOMAIN CONDITIONS

Condition No.	Time Domain	Condition No.	Time Domain
1	0 - 700 msec	11	200 - 550 msec
2	50 - 700 msec	12	150 - 500 msec
3	0 - 650 msec	13	200 - 500 msec
4	50 -650 msec	14	250 - 500 msec
5	100 - 650 msec	15	200 - 450 msec
6	50 - 600 msec	16	250 - 450 msec
7	100 - 600 msec	17	300 - 450 msec
8	150 -600 msec	18	250 - 400 msec
9	100 - 550 msec	19	300 - 400 msec
10	150 - 550 msec	-	-

TABLE III
TIME STEP CONDITIONS

Condition No.	Time Step
1	100 msec
2	75 msec
3	50 msec
4	35 msec
5	20 msec
6	10 msec

discrimination rate for non-target trials was 100% for all subjects shown in the figure. Moreover, the discrimination rates for the target trials for all subjects were greater than about 5%, and the rates for subject 2 and 4 were around 50%, which is relatively higher. These results suggest that reinforcement learning in which a reward is given for the target trial might be possible in the inverted pendulum system for these four subjects.

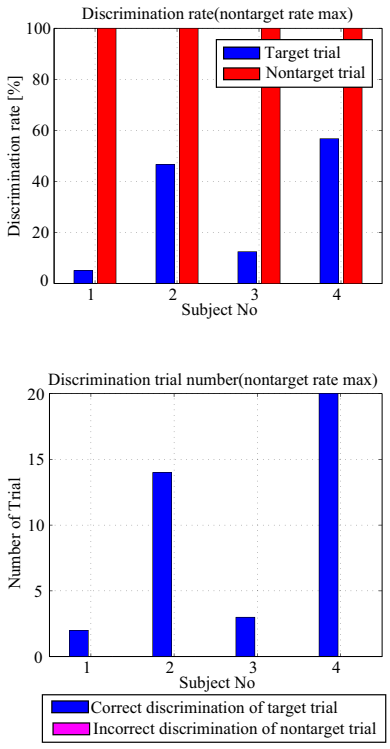

Fig. 8. Discrimination rate of target trials and non-target trials (top). Number of correctly recognized target trials and the number of incorrectly recognized non-target trials (bottom) calculated using maximum discrimination rates for non-target trials.

D. Possibility of reinforcement learning

We investigated the possibility of applying the reinforcement learning model such that a punishment "*a*" is given in the target trial by estimating P300 using an SVM, and a reward "0" is given in the non-target trial by estimating non-ERP. Reward

r is given by Equation (4) for each trial.

$$r = \begin{cases} a & \cdots \text{Target trial (punishment)} \\ 0 & \cdots \text{Non-target trial (reward)} \end{cases}, \tag{4}$$

where a is an arbitrary constant. In the case, a is negative.

The trials in which rewards in target trials consisted of correctly recognized target trials and incorrectly recognized non-target trials in the condition of non-perfect discrimination. The incorrectly recognized non-target trials were factors of misleading learning. In this case, for correct learning, the number of correctly recognized target trials had to exceed the number of incorrectly recognized non-target trials. In addition, the incorrectly recognized target trials, in which reward "0" was given, did not effect learning. Therefore, we can estimate the possibility of reinforcement learning by the number of trials distinguished as target trials.

VI. Realtime Discrimination Results

In this section, the possibility of the reinforcement learning is discussed by using real-time data. The recognition process is the same as described in a previous study [6]. The processes were as follows:

1) Obtain data to recognize from present time to 800 msec before.
2) Perform a baseline correction by average of first 100 msec in 1)
3) Check artifacts in the remaining 700 msec. If artifacts are detected, we estimate that a non-target trial has occurred.
4) Compute feature vectors every 35 msec.
5) Classify the data by SVM.
6) Repeat steps 1) to 5) every 20 msec.

As described in the previous section, we used sessions one to four as training data sets, and sessions five to eight as test data sets. Test data were processed previously by down-sampling and filtering. SVM parameters that were searched in the previous section were used.

A. Discrimination results - real-time

Figure 9 plots the results of real-time discrimination. The figure indicates the results of ten trials for one subject. The blue lines denote measured EEG channel profiles that were used as features. Squares in the lower part of the figure indicate discrimination results. Red squares indicate that a target trial was detected by SVM within the 700-msec period prior to a red square. Green background trials show discrimination results for target trials. Figure 9 shows that target trials are correctly recognized at around 350 msec for three out of four trials. Also, in non-target trials, represented by a white background, target trials are not detected around the same time. In that case, we estimate that only one step-detection of a target trial (one red square) shows target trial occurrence. This means that incorrect target trials increased. Therefore, we estimate that supernumerary consecutive detected steps (consecutive red squares) indicate a target trial. Then, we determined values of parameters for SVM. We selected the following parameters for the maximum real-time discrimination rates.

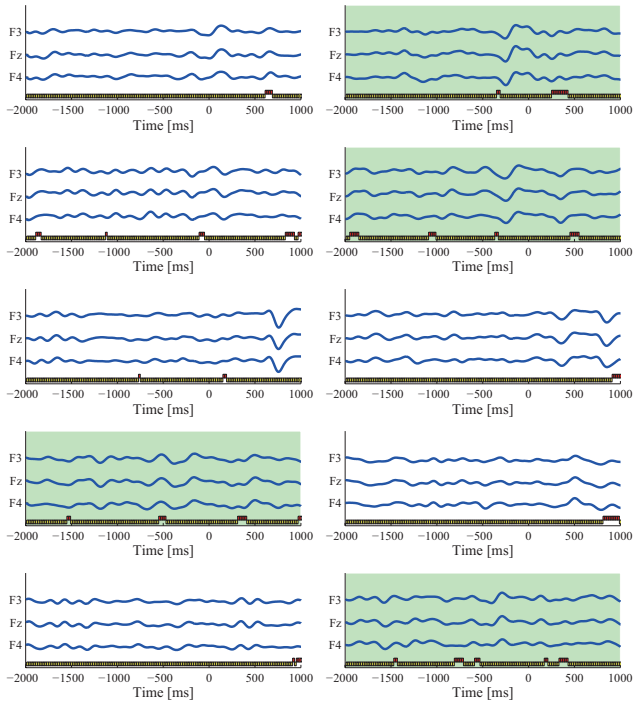

Fig. 9. Discrimination results in real-time

TABLE IV
SVM Parameters for Real-Time Detection

	σ	C
Subject S.T	15.95	0.7
Subject O.M	47.00	0.4
Subject N.T	48.88	4.0
Subject A.S	53.84	0.7

The number of consecutive steps were as follows; 9 for S.T, 5 for O.M, 11 for N.T, and 11 for A.S. Figure 10 shows the number of correctly recognized target trials and the number of incorrectly recognized non-target trials. All four volunteers correctly recognized more target trials than incorrectly recognized non-target trials. Therefore, we can estimate that learning proceeded correctly by using appropriate parameters such as Table IV.

VII. Conclusion

We examined the possibility of whether ERP could be used as a reward or punishment in reinforcement learning in order to design a control model in a BCI system. First, we confirmed that ERP was observed by the recognition of success or failure in an inverted pendulum task.

Next, an SVM was used to recognize ERP. The feature vector used in the SVM was composed of the averages at each 35 msec for each of six channels (F3, Fz, F4) for a total of 700 msec. Finally, the possibility of application to reinforcement learning was investigated. By bringing the discrimination rate

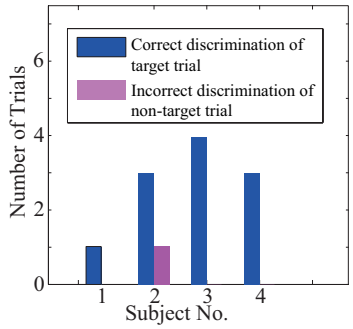

Fig. 10. Number of correctly recognized target trials and number of incorrectly recognized non-target trials calculated using maximum discrimination rates for non-target trial

for non-target trials to almost 100% by decreasing the negative effect of reward of incorrectly recognized non-target trials in reinforcement learning, we showed that all volunteers learned in the correct manner. In other words, in a real-time case such as this study, we can suggest that it is possible to apply reinforcement learning using ERP to actual problems such as an inverted pendulum problem [5].

REFERENCES

[1] R. P. Hasegawa, Y. T. Hasegawa, M. A. Segraves, Single trial-based prediction of a go/no-go decision in monkey superior colliculus, Neural Networks, 19(8), 2006, pp. 1223-1232.

[2] T. M. Vaughan, D. J. McFarland, G. Schalk, W. A. Sarnacki, D. J. Krusienski, E. W. Sellers, J. R. Wolpaw, The Wadsworth BCI Research and Development Program: at Home With BCI, IEEE Trans Neural Syst Rehabil Eng., 14(2), 2006, pp.229-233.

[3] A. G. Barto, R. S. Sutton, C. W. Anderson, Neuronlike adaptive elements that can solve difficult learning control problems, IEEE Transactions on Systems, Man and Cybernetics, 13, 1983, pp. 834-846.

[4] D. J. McFarland, T. Lefkowicz, and J. R. Wolpaw, Design and operation of an EEG-based brain-computer interface with digital signal processing technology, Behavior Research Methods, Instruments, and Computers, 29(3), 1997, pp. 337-345.

[5] T. Tsubone, K. Sugiyama, Y. Wada, Robot task learning based on reinforcement learning in virtual space, Neural Information Processing - Letters and Reviews 11(7), 2007, pp.165-174.

[6] Y.Yamagishi, T. Tsubone, Y.Wada, Possibility of reinforcement learning based on event-related potential. Proc of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2008), 2008, pp.654-657