

Towards natural human computer interaction in BCI

Ian Daly¹ (Student) and Slawomir J Nasuto¹ and Kevin Warwick¹

Abstract. BCI systems require correct classification of signals interpreted from the brain for useful operation. To this end this paper investigates a method proposed in [1] to correctly classify a series of images presented to a group of subjects in [2]. We show that it is possible to use the proposed methods to correctly recognise the original stimuli presented to a subject from analysis of their EEG. Additionally we use a verification set to show that the trained classification method can be applied to a different set of data.

We go on to investigate the issue of invariance in EEG signals. That is, the brain representation of similar stimuli is recognisable across different subjects.

Finally we consider the usefulness of the methods investigated towards an improved BCI system and discuss how it could potentially lead to great improvements in the ease of use for the end user by offering an alternative, more intuitive control based mode of operation.

1 INTRODUCTION

Brain computer interfaces (BCI's) are able to provide alternative methods of communication. This can allow individuals with severe motor disabilities additional channels for communication with their family and friends and control of their environment [3].

The majority of BCI systems work via the reading and interpretation of cortically evoked electro-potentials across the scalp via an Electro-encephalogram (EEG) or Magneto-encephalography (MEG) system. The subsequent classification and interpretation of this data is an area of much current research as it offers the key to improved performance of BCI systems and improved quality of life for individuals who require such systems.

Traditionally BCI research has focused on signals that are quickly identifiable by simple and fast classification methodologies. This is to get the systems working reliably for end users quickly. However this approach hasn't always resulted in intuitive systems. For example much research has gone into classifying motor movement signals [3] and quickly classifiable events generated by indirect means such as the P300 [2]. While these control methodologies work reliably and can be performed quickly on relatively cheap technology they are not an ideal and intuitive means for humans and computers to interface. It's not necessarily intuitive for the end user to have to attempt to move their foot to spell a word for example.

This is a serious drawback from the human computer interaction perspective. Unintuitive control signal perception tests which are unrelated to the intended outcome make the

interaction cumbersome and may result in subject's fatigue, loss of concentration and an increase in error rates. Such tasks may effectively act to distract the subject from their goal.

To this end this paper details an investigation into the methods proposed originally in a series of papers [1], [4], [5] and [6]. These methods attempt to recognise the original stimuli presented to a subject from EEG recordings made while the subject was exposed to a range of different stimuli.

This potentially opens up the possibility for creating BCI systems whose mode of operation is more direct and task-relevant, and hence more natural and intuitive for the end users. This is could provide a less tiring system which can be operated in a more focused manner and hence with improved levels of overall performance, concentration and motivation for the end user who can now concentrate solely on their desired goal. Subsequently human computer interaction within BCI could be able to more directly meet the needs of the end users.

The authors of the papers investigated here report high recognition rates achieved via their methods. However no apparent attempt is made to verify these results against a third data set. The recognition methods used can be thought of as a type of classification method. Prototype waveforms are matched using a trained and optimised filtering technique to trials from a training set via a Euclidean distance calculation. However unlike standard classification methods no attempt is made to verify the classification by applying the classification results to a verification data set. It is well known in classifier research that typically high classification rates achieved on the training data do not guarantee especially good results on new data sets.

Therefore we attempt to properly evaluate and extend these methods to show whether they can be used to successfully classify the original stimuli presented to subjects during an experiment intended to evoke P300 events for use in a BCI system. Furthermore we attempt to apply the results of our classification to a verification data set.

The data we use comes courtesy of the work of Ulrich Hoffman et. al. [2] and was originally used to train and verify classification methods for identifying P300 events. The P300 events were evoked by the subjects' exposure to a series of six different images. Their intention was to use the most successful classification methods for identifying P300's in a BCI system to assist with communications and control for disabled individuals.

Here we attempt to apply the methods outlined in [1], [4], [5] and [6] to identify which image the subjects from [2] were looking at in each trial.

Additionally [5] and [6] indicate a level of invariance across sessions and subjects. That is, trained classifiers from one subject or session are shown to be able to correctly recognise trials from a different session or subject. Verifying this is important to BCI work as it indicates whether a trained classifier can be used again to classify trials at different times or from different subjects.

¹ Department of Cybernetics, School of Systems Engineering, The University of Reading, RG6 6AY, UK

2 METHODS

Data analysis outlined in [1], [4], [5] and [6] share common fundamental steps with a few problem specific modifications. We based the approach on that outlined in [5] as this paper describes results obtained from visual stimuli, hence the stimuli type was the closest to the one used in our studies.

Table 1. Steps performed by our method.

1. Separate the trials corresponding to each stimulus for an individual session into Prototype, Training and Verification sets.
2. Normalise each of the trials.
3. Subtract out the pre-stimulus baseline.
4. Mirror and smooth the data using a Gaussian function (exploratory).
5. Fourier transform the trials.
6. Optimally filter the trials with a 4th order Butterworth filter.
7. Estimate the inverse of the Fourier transform.

2.1 Recording

The data used was recorded from 8 subjects observing a series of 6 randomly presented images as detailed in [2]. The subjects were asked to count the occurrences of a particular image with the intention of evoking a P300 event which could then be classified.

Both uni-polar and bi-polar montages were used for the analysis. The average of the two mastoid channels was used as the reference for the uni-polar electrodes.

The following uni-polar electrodes were used.

FP1, AF3, F7, F3, FC1, FC5, T3, C3, CP1, CP5, T5, P3, PZ, PO3, O1, OZ, O2, PO4, P4, T6, CP6, CP2, C4, T4, FC6, FC2, F4, F8, AF4, FP2, FZ, CZ.

The following bi-polar electrodes were used.

CZ-C4, CZ-P4, CZ-PZ, CZ-P3, CZ-C3, F4-FP2, F4-F8, C4-F8, C4-T8, C4-P8, P4-P8, P3-P7, C3-P7, C3-T7.

2.2 Data separation

Our data set for a single session is made up of four runs, for each of which a different stimulus was chosen to be the target intended to evoke a P300 event. For the purpose of our study we treat these four runs as a single data set.

From this data set we identify individual trials. The inter-stimulus interval used in the experiment was 400ms; there was also a 400ms pre-stimulus length of recording made for each session. Thus each trial is identified as a 400ms window of observations occurring sometime after the first 400ms of recording and identified explicitly by the meta-data provided with the recordings from [2].

We subsequently separate these trials into 6 sets, one set for each stimulus. We then further separate these 6 sets into 3 subsets, labelled as Prototype, Training and Verification sets with equal (or as close to as the size of our set allows) numbers of trials in each.

The original experiments within a session were performed in different runs over a number of days. In order to account for the potential time dependence of the stimulus waveforms we constructed the stimulus prototypes from waveforms extracted from each run.

The training and verification sets were constructed in an analogous way.

This expands on the work in [1], [4], [5] and [6]. In that work only classifier training was performed and reported with no verification data set used.

2.3 Pre-processing

For each individual trial in each one of these subsets we perform the following pre-processing steps.

Firstly we take the pre-stimulus baseline, that is the first 400ms of recording prior to stimulus onset at the beginning of our session and normalise to the range -1 to 1. We then normalise the individual trials into the same range and subtract the baseline from each normalised trial.

In order to increase the signal to noise ratio the training and verification waves were averaged over five trials giving eight averaged waveforms in each training and verification subset for each stimulus.

We then took every trial in our Prototype subsets and average them all together to form a single Prototype wave form.

As an additional step here we investigated a technique introduced in [4] to use a mirroring technique in conjunction with a Gaussian function to smooth the trials for the subsequent use of Fast Fourier transforms.

2.4 Analysis

The training is performed by calculating a Euclidean distance between a prototype corresponding to a given stimulus and the input wave form at a given pass-band frequency and electrode location. The exhaustive search of pass-band frequencies and electrode locations is used to find the best parameters for the classifier.

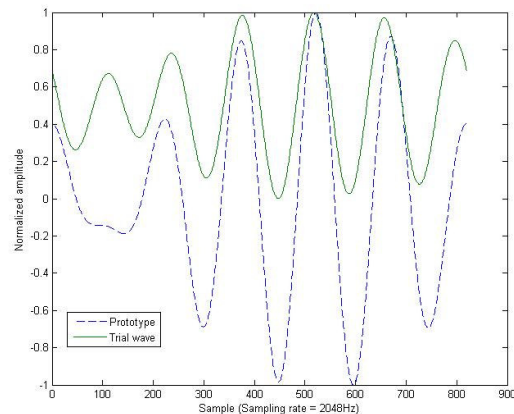


Fig 1. Typical prototype and test waveforms for stimulus 3 at a low pass value of 12Hz and width of 2Hz after a Fourier transform, filtering with an optimised filter and an inverse Fourier transform.

To illustrate further we take a given prototype for a given stimuli and a wave form from each of the training subsets for each stimuli. Thus for example if we were trying to classify stimuli number one (a picture of a television), we would take the prototype waveform for stimuli one and one waveform from the training subset for each stimuli, so one waveform for stimuli one, one waveform for stimuli two etc.

We then performed a fast Fourier transform on both the chosen prototype and each of the chosen training waveforms. We pick the optimal band-pass parameters for a fourth order Butterworth filter and filter both the prototype and each of our training waveforms with these same parameters. We then perform an inverse fast Fourier transform to put all our waveforms back into the time domain.

A typical prototype waveform is shown next to a typical test waveform in Figure 1 and the steps used in this method are summarised in Table 1.

2.5 Mirroring and Gaussian smoothing

Our analysis method involves filtering in the frequency domain to obtain the best match between our Prototypes and the Training trials for the same stimuli. Therefore our methods include both a Fourier transform and an inverse Fourier transform. To achieve faster computation times at this stage we here investigate a method proposed in [4] and used in [4], [5] and [6] but not in [1]. We therefore investigate the effects on the classification results both with and without including this step in our pre-processing.

The technique is as follows. For each waveform (a Prototype, Training or Verification waveform) we took the complete sequence of observations in the waveform and placed them at the centre of a sequence 2,048 observations thus scaling the data to a size of 2^n to allow faster processing of our Fourier transforms in the analysis section.

We then mirrored the first half of our shorter centred sequence onto the first half of the longer sequence before the start of our shorter sequence and the end half of our shorter sequence onto the end of the longer sequence after the end of our shorter sequence. We then positioned a Gaussian function in the centre of our entire longer sequence and set its standard deviation equal to half the length of our shorter sequence.

Finally for every observation in either the beginning or end of our longer sequence outside of our shorter sequence we multiplied the value of the observation by the ratio of the value of our Gaussian function at this point to the value of the Gaussian function at either the beginning or end of our shorter sequence respectively.

This is illustrated in Figure 2.

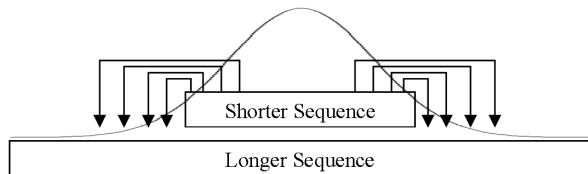


Fig 2. Mirroring and Gaussian smoothing process schematic.

2.6 Recognition surfaces

As we don't know the optimal band-pass parameters we perform the above analysis for each waveform from our group of eight waveforms in the training subsets. When we sum the results together for our groups of eight waveforms it's possible to produce for each electrode a recognition rate surface (correct matches on the training set against the low pass and width of our band-pass filter).

By finding the peak in this surface we can identify the optimal classifier parameters for a given electrode. If these optimal parameters provide good performance when used to classify trials from the verification set then we can say a good rate of verification has been achieved.

The classifier training can therefore be thought of as a kind of dimensionality reduction. We take a set of parameters for a given subject and session, EEG channel, band-pass parameters etc. and reduce them down to a single point (the peak in the recognition surface). The parameters at this point are said to be trained and we use them to classify data in our verification data. Thus only the peaks in the recognition surfaces are considered for analysis.

By way of illustration, if we have a peak in our recognition rate surface for the training data with a low pass frequency of 5Hz and a band-pass width of 3Hz at electrode C3 we should get a high rate of classification when attempting to match the waveforms in the verification set to the prototype using these parameters. We therefore select the same electrode and filter using these parameters on the verification set before calculating using the Euclidean distance calculations to find the recognition rate. If we get a similarly high rate of recognition with these parameters in the verification data set we can say we have a correctly verified classification method.

3 RESULTS

We present the classification rates achieved from different subjects and sessions and look at the verification of these classification methods by applying the trained classifiers to our verification data sets. We will investigate how classifiers trained on one subject perform classifying data from another subject to test inter-person invariance. Analogously the temporal invariance is investigated by testing how classifiers trained on one session perform on data from another session.

3.1 Training vs. Verification surfaces

We constructed recognition rate surfaces for the training and verification sets. These allowed a visual inspection and comparison between the locations of the optimal parameters for best recognition (the highest peaks in the recognition rate surfaces) in the two different sets.

For a given session, for a given subject, recognition rate surfaces are compared for each channel and for each stimulus. A large correlation was found between the training and verification surfaces for some stimuli and for several electrode locations. The results of these correlations are not presented here due to space constraints.

Table 2 lists the best recognition rates achieved with the training data and the corresponding rate of recognition when these trained parameters are applied to the verification data set.

Table 2. Optimal filter parameters obtained on the training data against the recognition rate achieved when applying these parameters to the verification data. Uni-polar electrodes only.

Sub ject	Training			Recon. Rate (%)	Recon rate (%)
	Optimal filter				
	Lower cut off (Hz)	Upper cut off (Hz)	Best EEG sensor		
S1	11	18	AF3	100	75
S2	39	43	FP1	100	87.5
S3	21	30	F4	100	100
S4	12	14	FC1	87.5	75
S6	17	22	CP1	87.5	62.5
S7	16	20	P8	100	50
S8	8	10	FP1	100	100
S9	9	11	P7	100	100

Table 3. Recognition rate achieved using the following optimal parameters from subject 6 on the training data. EEG channel = PO4, band-pass lower cut off = 9Hz and band-pass upper cut off = 12Hz, against recognition rate achieved by applying these parameters to the verification data.

Prototype source (subject / session)	Trials source (subject / session)	Training	Verification
		Recon. Rate (%)	Recon. Rate (%)
6 / 1	6 / 2	100	87.5
6 / 1	6 / 3	100	100
6 / 1	6 / 4	100	62.5
6 / 1	7 / 1	100	75
6 / 1	8 / 1	100	100

Table 4. Confusion matrices. Target stimuli against the results of each classifier. Results are averaged across subjects, sessions and waveform presentations and are hence out of 256. Uni-polar electrode montage.

Training data								Verification data							
Target stimuli								Target stimuli							
Classifier result	TV	Phone	Lamp	Door	Window	Radio		Classifier result	TV	Phone	Lamp	Door	Window	Radio	
	TV	205	16	0	12	11	12			TV	63	50	4	52	42
Phone	13	206	0	8	15	14		Phone	54	58	5	47	42	50	
Lamp	2	1	251	0	1	1		Lamp	10	10	212	6	12	6	
Door	11	17	1	204	13	10		Door	48	49	6	53	39	61	
Window	6	14	3	9	212	12		Window	50	49	9	44	52	52	
Radio	15	11	1	9	13	207		Radio	46	39	7	49	56	59	

These results are for session four, for stimuli number three for each subject.

Table 4 shows confusion matrices which illustrate the performance of this method. Notice that in the training data a large number of correct recognitions are made. However these high rates of training data recognition only translate into a high rate of classification within the verification set for the lamp stimulus.

To assess the statistical significance of these results we considered the null hypothesis that the results were obtained by random chance. This is modelled by a binomial distribution with $p=0.5$. The results for the correct classification of the lamp from the verification set was 212 correct matches out of 256 and had a statistical significance of $p < 1\%$ from our null hypothesis distribution. Therefore we can say this is a statistically significant result.

By way of contrast we also assessed the probability of 63 correct classifications (as occurs for the TV stimulus in our verification set). The probability of this occurrence from our null hypothesis binomial distribution shows that this result is not statistically significant.

It's important to note that in general the results showed some level of correlation between training and verification for each stimuli. However the only statistically significant peaks were produced by stimuli three and it is these results that are presented here. This is discussed further later.

3.2 Mirroring and Gaussian smoothing

As the original work we are investigating uses the technique of mirroring and Gaussian smoothing in [4], [5] and [6] but not

in [1]. We investigated the effects on the classification and verification processes both with and without using this technique.

When Gaussian smoothing was used the general effect was to lower the size of the peaks in our recognition rate surface. That is when using the Gaussian smoothing technique when training our classifier we were unable to achieve as high a rates of recognition as we could without using it. Therefore while all the analysis was done both with and without the use of the Gaussian smoothing technique we here present only the results obtained from not using it.

3.3 Invariance between subjects and sessions

One of the key claims of the [1], [4], [5] and [6] is that there exists a level of invariance between the representation of a stimulus across multiple sessions or subjects. That is; a prototype for a given stimulus could be used to correctly classify that stimulus in multiple trials across different subjects or over time.

Assessing the validity of this claim is very important for any potential applications of this technique to a BCI system. If the same trained classifier can recognise stimuli from different subjects or across different days then the need to retrain the classifier is reduced to a practical amount for the end users needs.

The similarity of the peaks in the recognition surfaces for the training and verification sets already show a certain level of invariance between different trials within one session. However for our purposes it's necessary for us to show that this invariance extends across sessions and across subjects.

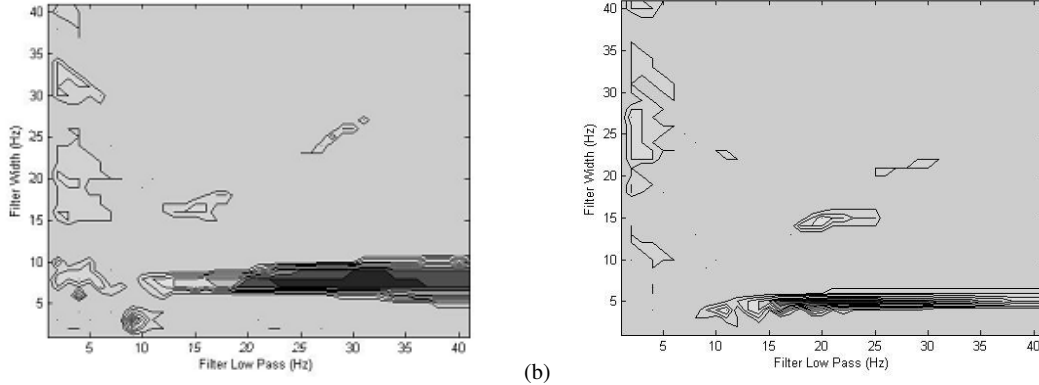


Fig 3. Recognition rate surfaces for unipolar electrode montage, channel FP1 (a) and bipolar electrode montage, channels C4 – T6 (b) The darker the surface the higher the rate of recognition at these filter parameters.

To assess the invariance levels we take a trained prototype from one subject / session and apply it's parameters to verify trials from a different subject / session. For example to show that there exists a level of invariance between the representations of the stimuli of a picture of a lamp within the EEG of subjects one and two we take the trained prototype for this stimuli from subject one and attempt to use it to classify the trials from subject two.

We created a recognition rate surface for the prototype from one subject / session applied to the verification data from a different subject / session and compared it to the recognition rate surface for the training data.

Table 3 shows examples of the optimal parameters from the training being applied to classify data from a different subject / session.

We noticed a strong intersession correlation between the classification and verification surfaces. We also noticed a strong inter-subject correlation in the recognition surface peaks between the classification and verification sets. This suggests a level of invariance in the brains response to stimulus across a short period of time and across different subjects.

3.4 Montages

The analysis was performed using both uni-polar and bi-polar montage systems.

The effects of the use of these different montages are negligible. Generally equally good recognition rates appear to have been achieved from both the uni-polar and the bipolar montages. The only significant difference is in the shape of the recognition surfaces produced. The recognition surfaces produced with the uni-polar montage are in general fatter than the surfaces produce by the bipolar montage.

Two such example surfaces are compared in figure 3 (a) and (b).

4 DISCUSSION

Our results are able to verify the classification methods investigated for individual stimuli during individual sessions.

That is, for a given session with a given subject we have shown that a peak in the recognition rate surface of the training set for certain specific stimuli has a correlation with the peak in the recognition surface of the verification set for that same stimuli.

It's important to note that we obtained very different recognition surfaces for different stimuli. For stimuli number three the recognition surfaces were of a very distinct shape with one or occasionally two clearly defined peaks that were closely matched in the classification and verification surfaces.

While the optimal classification parameters from the training set led to in general high recognition rates on the verification set the recognition rate surfaces for other stimuli were less distinct. This is most clearly indicated by our confusion matrix for our verification set in Table 4. We achieve high rates of correct classification with stimulus 3 but much lower rates with the other stimuli.

This implies that while it is clearly possible to classify original stimuli from EEG recordings as shown by the trials for stimuli three, certain stimuli do produce more distinct and easily classifiable signals than others. Why this should be is unclear at this stage and may be a course of further investigation.

In the original methods, [1], [4], [5] and [6], the recognition rate surfaces for each stimuli were summed together to get a higher more distinct signal peak. The parameters at this single peak could then be used to recognise any stimuli using the appropriate prototype. We were unable to verify this summation technique due to the unclear results produced by some of the stimuli. However if different original stimuli were presented to the subject that each produced clear recognition surfaces it may be possible to use this technique to achieve better results.

The methods used in our data source [2] count the occurrences of a P300 event triggered by the subject counting the occurrences of a particular image. This means that for some of the trials we are using there will be a P300 event within the EEG. We therefore must consider here the possibility of such events causing a distortion on our results.

Due to the averaging of trials described in section 2.2 the effects of a P300 event within an individual trial are thought to be negated. This is further supported by our experimental results which don't show any distortion that could be attributed to a P300 event.

4.1 Invariance

The results achieved from using the prototype from one session to classify trials from a different session or even from a different subject are very encouraging. We can very clearly see a level of invariance here across sessions and subjects. That is, a prototype trained on one session for one subject can be used as the basis for producing verified classification results for a different subject or a different session for that same subject.

We can therefore say that there exists a level of invariance across subjects in the representation of stimuli within the EEG. Because all the sessions for a subject are recorded within a relatively short period of time (for every subject the time between the first session and the last session was less than two weeks), we cannot confidently claim the results to show invariance over time. However as there exists a level of invariance over subjects then a level of invariance over time is thought to be likely.

4.2 Applicability to BCI

The usefulness of these results within a BCI system is an important point of consideration. The possibility of correctly classifying the original stimuli presented to subjects from their EEG alone offers a lot of potential to end users of BCI systems.

There exists considerable research to suggest that the representation of actual events within the brain and the imagination of the same events are closely correlated [3]. For example the brain representation of movement and the imagination of the same movement are very similar. It is this principle that forms the basis of much research in motor control based BCIs [7]. The end user imagines moving some part of their body and this is interpreted as a movement by the classifier. As both able bodied and disabled individuals are able to imagine movement this allows disabled individuals the ability to use a BCI system by imagining motor movement.

It follows that the brain representation of stimuli and the imagination of the same stimuli by the individual are likely to have a close correlation for a wide range of different types of stimuli. Furthermore if it is, as we show here, possible to recognise original stimuli in the form of images from the EEG then it should be possible to recognise the imagination of the same stimuli from the EEG.

This leads us to the potential of more natural forms of human computer interaction within a BCI system. Instead of controlling a system indirectly via an external device or directly but by none intuitive means that can often actually distract from the task they are attempting to perform, such as moving a toe to spell a word. The possibility is suggested that much more natural and direct control could be achieved by recognising stimuli or imagined stimuli such as command words from the EEG.

This suggests the possibility for a form of human computer interaction within BCI that allows much more direct, less tiring and potentially faster control for the user. As an example consider a speller which works from the users imagining of command words. It's easy to see how this methodology could provide a higher level of natural, task relevant control to the user.

The encouraging results in the invariance experiments furthermore suggest that this method could be applied over time

and over different subjects resulting in more intuitive natural ways for humans and computers to interact that can be applied in the real world.

It is important to note that although there is evidence to suggest that trained classifiers on one subject / session can be applied to a different subject / session there also is a level of variability in the best EEG sensor. That is, the EEG sensor with the highest recognition rate for one subject may not have the highest recognition rate on a different subject. The amount of effect this will have on the applicability of this technique to a BCI system is a subject for further investigation.

Therefore the applicability of this technique in a BCI requires further evidence to quantify. However this study indicates that this is an area of research worth investigating further.

5 CONCLUSIONS

The results obtained during the course of this investigation are promising. It is possible to use the EEG recording made while subjects are exposed to a range of different stimuli to recognise the original stimuli presented to the subject. We show that high classification results can be achieved for certain stimuli by using a prototype (or template) of the stimuli and a trained, optimised filter. We go on to verify that this trained classification method can be used to achieve high rates of recognition in a third data set.

We also show that there exists a level of invariance across different subjects and different sessions for the same subject. Prototype data from one subject or session can be used to achieve high classification results with a different subject or session.

These results indicate a new approach to more intuitive and natural forms of, either goal based or control based, human computer interfacing.

5 ACKNOWLEDGMENTS

The authors would like to thank Dr Ian Bland for his considerable technical assistance and support during the course of this work. They would also like to thank Ulrich Hoffman et. al.[2] for the availability of the data sets used in this work and the referees for their comments which helped improve this paper.

REFERENCES

1. PATRICK SUPPES, Z.-L.L., BING HAN, *Brain wave recognition of words*. PNAS, 1997. **94**(1): p. 14965-14969.
2. Ulrich Hoffmann, J.-M.V., Touradj Ebrahimi, Karin Diserens, *An efficient P300-based brain-computer interface for disabled subjects*. Journal of Neuroscience methods, 2007.
3. Wolpaw, J.R., *Brain-computer interfaces as new brain output pathways*. The Journal of Physiology, 2007. **579**(3): p. 613-619.
4. PATRICK SUPPES, B.H., ZHONG-LIN LU, *Brain-wave recognition of sentences*. PNAS, 1998. **95**(1): p. 15861-15866.
5. Patrick Suppes, B.H., Julie Epelboim, and Zhong-Lin Lu, *Invariance between subjects of brain wave representations of language*. PNAS, 1999. **96**(22): p. 12953-12958.
6. Patrick Suppes, B.H., Julie Epelboim, and Zhong-Lin Lu, *Invariance of brain-wave representations of simple visual images and their names*. PNAS, 1999. **96**(25): p. 14658-14663.
7. PFURTSCHELLER Gert, N.C., *Motor imagery and direct brain-computer communication*. Proceedings of the IEEE, 2001. **89**(7): p. 1123-1134.