Biomed. Technik
47 (2002), 03-08

Estimating the Mutual Information of an EEG-based Brain-Computer Interface

Quantifikation der Informationsrate einer EEG-basierenden Hirn-Computer-Schnittstelle

A. Schlögl[1]
C. Neuper[2]
G. Pfurtscheller[1,2]

[1]*Department of Medical Informatics, Institute of Biomedical Engineering, University of Technology, Graz*
[2]*Ludwig Boltzmann Institute for Medical Informatics and Neuroinformatics, Graz, Austria*

Key words: Brain-computer interface – single-trial EEG analysis – event-related desynchronization (ERD) – adaptive autoregressive model – time-varying spectrum – non-stationary signal processing – communication theory – mutual information – entropy

An EEG-based Brain-Computer Interface (BCI) could be used as an additional communication channel between human thoughts and the environment. The efficacy of such a BCI depends mainly on the transmitted information rate. Shannon's communication theory was used to quantify the information rate of BCI data. For this purpose, experimental EEG data from four BCI experiments was analyzed off-line. Subjects imaginated left and right hand movements during EEG recording from the sensorimotor area. Adaptive autoregressive (AAR) parameters were used as features of single trial EEG and classified with linear discriminant analysis. The intra-trial variation as well as the inter-trial variability, the signal-to-noise ratio, the entropy of information, and the information rate were estimated. The entropy difference was used as a measure of the separability of two classes of EEG patterns.

Schlüsselwörter: Hirn-Computer Schnittstelle – ereignisbezogene Desynchronisierung – adaptives autoregressives Modell – zeitveränderliches Spektrum – nichtstationäre Signalverarbeitung – Informationstheorie – Entropie

Eine EEG-basierende Gehirn-Computer-Schnittstelle (BCI) könnte als zusätzlicher Kommunikationskanal zwischen dem menschlichen Denken und der Umwelt verwendet werden. Die Effektivität einer solchen Kommunikation hängt entscheidend von der übertragbaren Informationsrate ab. Erstmals wurde die Informationsmenge eines BCI mit Shannons Informationstheorie quantifiziert. Es wurden experimentelle BCI-Daten off line analysiert. Versuchspersonen stellten sich linke und rechte Handbewegungen vor, während das EEG über dem sensomotorischen Areal abgeleitet wurde. Aus dem EEG wurden adaptive-autoregressive Parameter berechnet und anschließend mit dem Gewichtsvektor einer linearen Diskriminante zusammengefaßt. Die Variabilität, sowohl zwischen als auch innerhalb von Einzelversuchen, das Signal-Rausch-Verhältnis, die Entropie der Information und die Informationsrate vom BCI wurden bestimmt. Der Entropieunterschied wurde als Maß für die Unterscheidbarkeit von zwei Klassen von EEG-Mustern verwendet.

1 Introduction

An EEG-based Brain Computer Interface (BCI) analyzes and classifies EEG patterns and transfers those patterns into control signals. A BCI is an additional communication channel between purely mental activity (thoughts or imaginations) and the surrounding physical world. Because no muscle activation is involved, the BCI can be used as an additional communication channel, which might be useful for handicapped people [1–3]. The question is: How much information can be transmitted by a BCI?

The information rate is also important in the learning process of the user. Usually, the various BCI approaches [1–11] rely on feedback. The amount of information provided by the feedback signal determines the subject's ability to learn (see Figure 1). A system without a reliable classification output can not be used for learning. Hence, one reason for unsuccessful BCI attempts might be the lack of information provided by the feedback. For this reason, quantifying the information in the BCI output is important. In this work, the information based on the principle of Shannon's communication theory will be quantified.

It is known that the amount of information of (clinical) EEG recordings (Figure 1: A) is about 8–11 bits [12] multiplied by the sampling rate and the number of channels. This information rate is not useful, but the relevant information needs to be extracted. Actually, the information rate, obtained from the output of an EEG-based BCI, is of interest. This paper introduces a measurement for the Signal-to-Noise Ratio (SNR) and

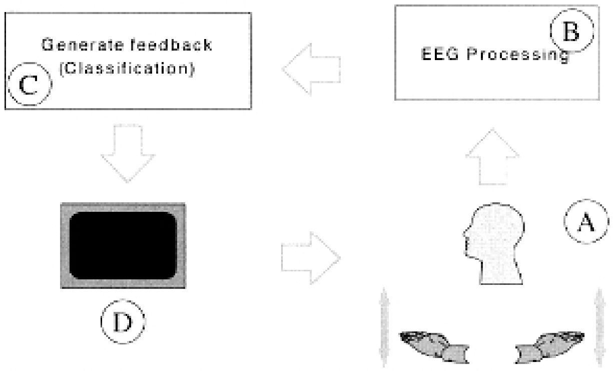


Figure 1. Scheme of a BCI with feedback. The EEG from the subject's scalp is recorded (A); next it has to be processed online (B); the extracted features are combined and classified (C); the output can be used to control a device, e. g. a cursor on a computer screen (D); simultaneously, the output provides feedback to the subject.

the mutual information (MI) [13, 14] between a BCI input and output. The SNR describes the ratio of the variation of the output due to the task, and the variations that are not task-related (unspecific changes, noise). The MI is the average amount of information that an observation (output) provides about a signal (input) [14]. In case of the BCI, it is of interest how much information about the input (cue on the computer screen) can be gained by observing the output.

2 Data and methods

Data from a previous BCI study was used for the present analysis. The experiment is described in detail in [15–17]. The task was cued by visual stimuli on a computer screen. From 3.0–4.25 s an arrow pointing either to the left or to the right was displayed. Depending on the direction, the subject was instructed to imagine a movement of the left or right hand. From 6.0–7.0 s feedback was given; the feedback was calculated from the band power of the most reactive frequency bands. These frequency components were found by using the Distinction Sensitive-LVQ algorithm [6]. The EEG data was recorded using two bipolar channels over left and right central areas. The two EEG channels were derived from two electrodes each placed 2.5 cm anterior and 2.5 cm posterior to the electrode positions C3 and C4, respectively. The EEG signals were amplified and bandpass filtered between 0.5 Hz and 35 Hz by a Nihon Khoden amplifier, and then sampled and digitized at 128 Hz using 12 bits. The results of selected sessions of 4 subjects (subject f3, session 10; f5, session 6; f7, session 6; g3, session 7) with 80–79, 78–78, 80–80 and 76–75 (L)eft-(R)ight trials, respectively, are presented in [17].

An AAR model was chosen for feature extraction, because it provides estimates with a time-resolution as high as the sampling rate and the AAR estimation algorithm can also be used on-line [18]. Furthermore, it is not necessary to select subject-specific frequency bands. The AAR parameters were estimated with the recursive least squares (RLS) algorithm as decribed in [15, 16, 17, 19]. The update coefficient $UC = 0.007$ and a model order of $p = 10$ was chosen (Figure 1: Part B).

The AAR parameters were estimated for every sample time point. From each trial the AAR parameters were taken at a specific time point t from both EEG channels C3 and C4 [16, 17]. The $2*10$ (two channels, model order 10) AAR parameters span a 20-dimensional feature space. Using linear discriminant analysis [20] a weight vector w_t was found that describes the maximal discriminating hyperplane between the two classes L and R (left and right cue). The time course of the error rate (with and without cross-validation) was calculated. The time point with the lowest error rate (i.e. largest separability) gives the optimal classification time point Tc.

2.1 *Time-varying signed distance (TSD) & Feature-of-Interest*

Next, a weight vector $\mathbf{w}_{Tc}$ was obtained by applying LDA to AAR parameters at time $t = Tc$. This weight vector $\mathbf{w}_{Tc}$ is applied to the AAR parameters of both EEG channels in the following way:

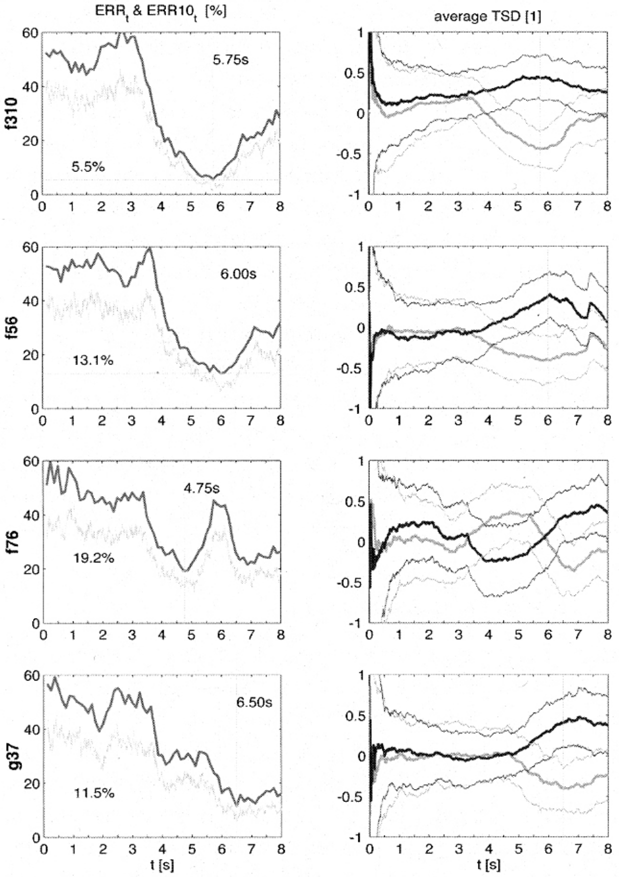
$$D_t = [a_{1,t}^{C3}, \ldots, a_{p,t}^{C3}, a_{1,t}^{C4}, \ldots, a_{p,t}^{C4}, -1] \cdot \mathbf{w}_{Tc} \qquad (1)$$

Note that the term -1 takes into account the offset or threshold incorporated in $\mathbf{w}_{Tc}$. We call D_t the time-varying signed distance function (TSD), because D_t varies in time with the AAR parameters; the sign of D_t describes whether the classification is left or right and D_t expresses the distance to the separating hyperplane described as $\mathbf{w}_{Tc}$.

The advantage of this procedure is that all time-varying parameters are reduced to one dimension. Now D_t is a one-dimensional, time-varying function that can be calculated for every time point t on a single-trial basis. Equation (1) is the most informative linear projection (with respect to the class relation) from multiple to one feature; the weight vector $\mathbf{w}_{Tc}$ incorporates the class information and expresses what we are interested in. For the reason, this new feature is the „feature-of-interest". Note, once a weight vector $\mathbf{w}$ is obtained (e.g. from previous recorded data), the TSD can be also calculated on-line [18]. Accordingly, the TSD is the classification output (Figure 1: Part C) and can be used to control a device and provide feedback (Figure 1: Part D).

2.2 *Signal-to-Noise Ratio SNR and Mutual Information MI*

The distance $D_t^{(i)}$ (1) is a measure for the classification. In the ideal case, $D^{(i)} > 0$ if the i-th trial is a left trial and $D_t^{(i)} < 0$ for all right trials. In practice other processes not correlated to the class-relation also influence $D_t^{(i)}$. One can say $D_t^{(i)}$ consists of two types of pro-



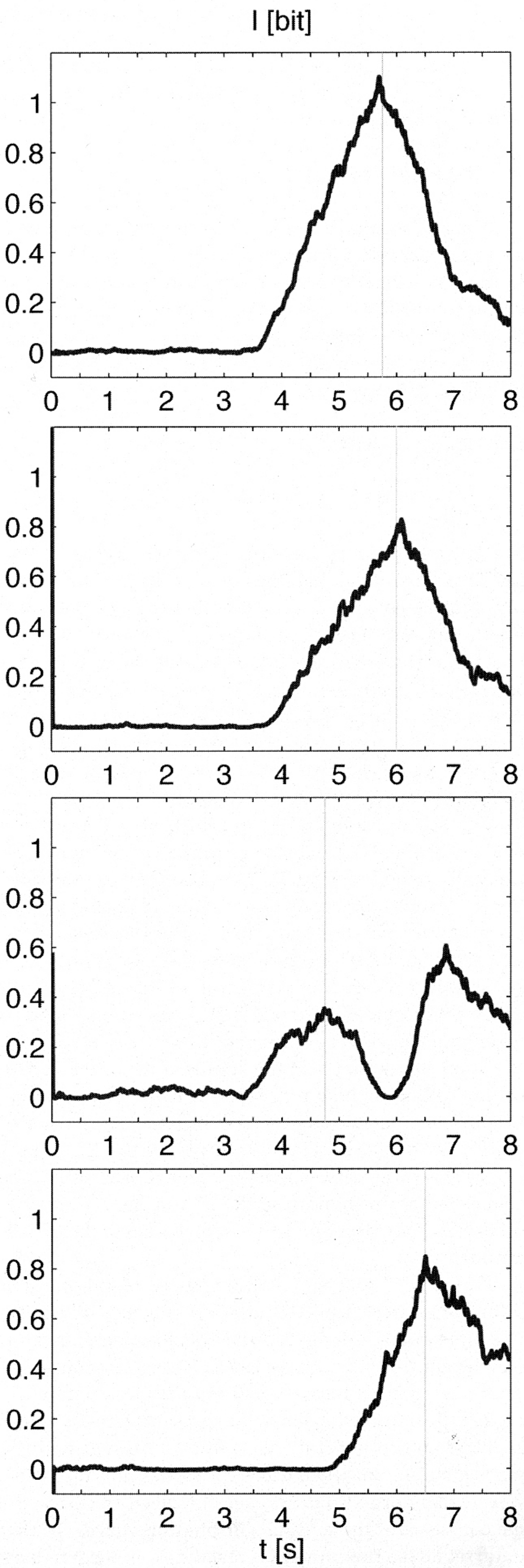


Figure 2. Time courses displaying the separability between two classes obtained in four subjects. The vertical lines indicate the time point used to generate the weight vector. The time point t with the lowest error rate ERR10 was used. (a) In the first column the time courses of the error rates ERR10$_t$ (thick line) and ERR$_t$ (thin line) are shown. ERR(t) gives the classification error with LDA of the EEG-channels C3 and C4 at time t. AAR(10) parameters were used as EEG features. The thick line shows ERR10$_t$ calculated 8 times per second; the thin line shows the time course of ERR$_t$ calculated at every sample. The numbers indicate the lowest ERR10 and the corresponding classification time. (b) The second column shows the averaged TSD for the left and right trials. The TSD is calculated as linear combination of AAR(10)-parameters of the EEG channel C3 and C4. The average TSD curves (thick lines) clearly show a different behavior during imagined left and right hand movement. The thin lines represent the within-class standard deviation (SD) of the TSD and indicate the inter-trial variability of the EEG patterns. (c) The third column shows the mutual information between the TSD and the class relationship. The entropy difference of the TSD with and without class information was calculated every time step. This gives (a time course of) the mutual information in bits/trial.

cesses, one correlated to the movement imagery task containing useful information (signal) and processes not related to the task (noise). The Signal-to-Noise ratio (SNR) (for a fixed time point t within a trial) is defined as follows:

$$SNR_t = \frac{2 * \underset{i\in\{\boldsymbol{L},\boldsymbol{R}\}}{\operatorname{var}}\left\{\boldsymbol{D}_t^{(i)}\right\}}{\underset{i\in\{\boldsymbol{L}\}}{\operatorname{var}}\left\{\boldsymbol{D}_t^{(i)}\right\} + \underset{i\in\{\boldsymbol{R}\}}{\operatorname{var}}\left\{\boldsymbol{D}_t^{(i)}\right\}} - 1 \qquad (2)$$

where {$\boldsymbol{L}$} and {$\boldsymbol{R}$} are the sets of left and right trials, and var{.} is the variance over all trials $\boldsymbol{i}$.

The aim is to extract information about the kind of motor imagery (class) by observing output D. According to Shannon and Weaver [13], the mutual information I is an entropy difference, whereas the entropy H of a Gaussian process x is

$$H(x) = 0.5 * \log_2(2\pi e\ \sigma_x^2) \tag{3}$$

The mutual information between the BCI output D and the class relationship can be determined by calculating the difference between the entropy for the total variance and the within-class variance of D (for more details see Appendix A). This leads to the following formula for the mutual information for each possible classification time t:

$$I_t = 0.5 * \log_2(1 + SNR_t) \tag{4}$$

3 Results

The time courses of the error rate $\mathrm{ERR10}_t$ and ERR_t, the TSD (mean and standard deviation) for both classes, and the entropy obtained with 4 subjects are shown in Figure 2. The first column in (a) shows the time courses of the error rate. In all 4 cases the decrease of the error rate starts during presentation of the cue stimulus (3.0–4.25 s). This means that the EEG patterns become more separable. For every subject, the time point with the best discrimination can be found. These time points were used for calculating the classifier w_{Tc}.

The second column in Figure 2b shows the time courses of the TSD. It can be seen clearly, that after presenting the cue, the TSD is different for the two tasks. Furthermore, a different temporal behavior for every subject can be found. Especially in subject f7, a crossing of the TSD can be observed. These results (based however on an AAR(6) instead of an AAR(10) model) were discussed in detail in [17] and can be explained by the „post-imaginary beta ERS“.

A novelty is that also the standard deviation (S.D.) of the TSD is included in Figure 2(b). This shows the within-class variability of the output. It is caused mainly by the inter-trial variability of the EEG patterns. The ratio between the average TSD of the two classes on the one hand, and the S.D. of the TSD on the other hand, yields an impression of the SNR of the BCI output.

Figure 2c displays the time courses of the mutual information between TSD and the cue (left or right). The time course of the mutual information shows at which point of time the patterns are most separable. As expected, no information about the class relationship can be obtained prior to cue presentation (starting at 3.5 s) and in general, the mutual information displays a maximum when the error rate is minimal. However, there are also differences between the error time courses and the mutual information curves. In the data set f310, the mutual information is larger than 1 bit per trial, although the error rate is not zero. Data set f76 shows a larger mutual information at about $t = 7$ s; this is surprising because the error rate is smaller at $t = 5$ s. The unexpected result in g37 is that (the time course of) the mutual information does not increase prior to $t = 5$ s although the error rate already decreases at 4 s. One explanation for these differences might be that the EEG patterns at 4–5 s and 6–8 s are different and the error rate is not the best criteria for selecting the optimal discrimination.

The optimal time point for classification is 5.75, 6.0, 4.74 and 6.5 s for f310, f56, f76 and g37, respectively. The task-related imagination starts at approx. 3.5 s. Hence, it can be said that the maximum information rate is 1.1 bits/2.25 s, 0.8 bits/2.5 s, 0.4 bits/1.25 s and 0.8 bits/3.0 s. Overall, approx. 0.3–0.4 bits per second were obtained in these experiments.

4 Discussion and conclusion

Basically, a *single-trial analysis* of EEG patterns was applied considering intra-trial (time-course) as well as inter-trial variability. Furthermore, a *measure of information* was introduced to analyze the BCI output (= control signal). This measure estimates the channel capacity (bit-rate) of a BCI. A limitation of the presented method is that the weight vector used for calculating the TSD and the derived measures like SNR and MI were calculated from the same data set. This means that no cross-validation was applied, and therefore the results might be biased. This must be considered when the presented entropy analysis is applied to BCI data. In the presented case, the number of trials (151–160) is much larger than the number of classification parameters (2*10 = 20). For this reason, the bias due to overfitting not to large.

In the case of the two-class problem, 100 % accuracy (0 % error) would provide 1 bit of information. Therefore, one might argue that the amount of information is described sufficiently by the error rate. However, the measure for the mutual information can also give larger values than 1. Then the class-related variability is larger than the background activity. More information (than 1 bit) can be obtained, even within a two-class paradigm. A large SNR and MI opens the possibility of increasing the bit rate of a BCI and using efficient coding schemes for e.g. letter selecting, beyond binary decisions.

From a communication theory point-of-view [13], a BCI experiment can be seen as a communication channel with added noise. It extracts relevant (useful) information, but might be the source of additional noise too. First, the input represented by the cue (one of two classes, 1 bit) is mentally processed by the user. Further, the EEG amplifier, AD converter, feature extraction method (e.g. AAR estimation) and the classifier (linear combiner) contribute to the one-dimensional output signal, which contains a part related to the task (useful signal) and one that is not task-specific (noise). The ratio between the signal and the noise corresponds to the mutual information between the input (cue) and the output.

The MI is also a measure for the quality of the feedback provided to the subject. Therefore, analyzing the mutual information of the feedback might be an important step in BCI research. It can be used to measure the improvement over time and for comparing different BCI systems. In summary, the proposed method introduces a measure for the maximum amount of mutual information and provides an estimate about the amount of information transmitted by single trial EEG patterns. The quantification of the transmitted information is an important criteria for a BCI. It is suggested that the entropy should be analyzed in all BCI experiments.

Appendix A

Assuming that an observable process v_k is given that consists of a signal process u_k and an additional noise process n_k

$$D^{(k)} = v_k = u_k + n_k \tag{A1}$$

Without loss of generality and for the purpose of simplicity, it can be assumed that the signal and noise process are distributed normally and completely independent; one can also say u_k and n_k are un-correlated.

$$u_k = N(\mu_u, \sigma_u^2) \tag{A2}$$

$$n_k = N(\mu_n, \sigma_n^2) \tag{A3}$$

$$\sum_k \left(u_k - \mu_u\right) \cdot \left(n_k - \mu_n\right) = 0 \tag{A4}$$

In case of linearity and independence, the variance σ^2 of the observed process v_k is

$$\sigma_v^2 = \sigma_u^2 + \sigma_n^2. \tag{A5}$$

Next, we can define a signal-to-noise ratio (*SNR*)

$$SNR = \sigma_u^2/\sigma_n^2 = \sigma_v^2/\sigma_n^2 - 1 \tag{A6}$$

The entropy H of a Gaussian process is:

$$H(x) = 0.5 * \log_2(2\pi e \sigma_x^2) \tag{A7}$$

The entropy difference between the variability of x and the variability of x under the condition class c is known, gives the maximum mutual information between x and the class information c.

$$I = H(x) - H(x,c) \tag{A8}$$

This results in the following equation

$$I = 0.5 * \log_2 \left\{ \frac{\underset{i \in \{\boldsymbol{L}, \boldsymbol{R}\}}{\operatorname{var}} \left\{ \boldsymbol{D}_t^{(i)} \right\}}{\frac{1}{2} \left(\underset{i \in \{\boldsymbol{L}\}}{\operatorname{var}} \left\{ \boldsymbol{D}^{(i)} \right\} + \underset{i \in \{\boldsymbol{R}\}}{\operatorname{var}} \left\{ \boldsymbol{D}^{(i)} \right\} \right)} \right\} \tag{A9}$$

where $\{\boldsymbol{L}\}$ and $\{\boldsymbol{R}\}$ are the sets of left and right trials, t is the (fixed) time point and var{.} is the variance over all trials $\boldsymbol{i}$ of set {.} as shown in equation (A9). The maximum mutual information is obtained if the noise process is Gaussian, otherwise the mutual information is smaller [14].

The denominator in (A9) is the variance of signal plus noise, the denominator is the variance of the noise (average of the within-class variance). Hence, the relationship between the *SNR* and maximum mutual information is

$$I = 0.5 * \log_2(1 + SNR) = 0.5 * \log_2(\sigma_v^2/\sigma_n^2) \tag{A10}$$

One can estimate the variance σ_s^2 from the obtained data D with

$$\sigma_s^2 = \underset{i \in \{S\}}{\operatorname{var}} \left\{ D^{(i)} \right\} = \tfrac{1}{N} \sum_i \left(D^{(i)} - \mu_S \right) * \left(D^{(i)} - \mu_S \right) \tag{A11}$$

where $\{S\}$ is a set of trials and N the number of elements within the set $\{S\}$. To estimate the variance of the noise process σ_n^2, we took the average variation within classes. The total variance σ_v^2 is easily obtained by calculating the variance over all trials.

References:

[1] J. R. Wolpaw, D. J. McFarland: Multichannel EEG-based brain-computer communication. Electroenceph. and clin. Neurophysiol. 90: 444–449, 1994.

[2] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, H. Flor: A brain-controlled spelling device for the completely paralyzed. Nature 398: 297–298, 1999.

[3] T. M. Vaughan, J. R. Wolpaw, E. Donchin: EEG-based communication: prospects and problems. IEEE Transactions on Rehab. and Engng. 4 (4): 425–430, 1996.

[4] G. Pfurtscheller, J. Kalcher, C. Neuper, D. Flotzinger, M. Pregenzer: On-line EEG classification during externally-paced hand movements using a neural network-based classifier. Electroenceph. Clin. Neurophysiol. 99: 416–425, 1996.

[5] G. Pfurtscheller, Ch. Neuper, D. Flotzinger, M. Pregenzer: EEG-based discrimination between imagination of right and left hand movement. Electroenceph. Clin. Neurophysiol. 103: 642–651, 1997.

[6] M. Pregenzer, G. Pfurtscheller, D. Flotzinger: Automated feature selection with a distinction sensitive learning vector quantizer. Neurocomputing 11: 19–29, 1994.

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

[8] D. J. McFarland, L. M. McCane, J. R. Wolpaw: EEG-based communication and control: short-term role of feedback. IEEE Trans Rehab. Engng. 6: 7–11, 1998.

[9] A. Kübler, B. Kotchoubey, T. Hinterberger, N. Ghanayim, J. Perelmouter, M. Schauer, C. Fritsch, E. Taub, N. Birbaumer: The thought translation device – A methodology for communication in total motor paralysis. Exp. Brain Research 124: 223–232, 1999.

[10] N. Birbaumer, T. Elbert, B. Rockstroh, W. Lutzenberger: Biofeedback of event-related potentials of the brain. Int. J. Psychophysiol. 16: 389–415, 1981.

[11] A. Kübler, B. Kotchoubey, H.-P. Salzmann, N. Ghanayim, J. Perelmouter, V. Hömberg, N. Birbaumer: Self-regulation of slow cortical potentials in completely paralyzed human patients. Neurosci. Letters 252: 171–174, 1998.

[12] A. Schlögl, B. Kemp, T. Penzel, D. Kunz, S.-L. Himanen, A. Värri, G. Dorffner, G. Pfurtscheller: Quality control of polysomnographic sleep data by histogram and entropy analysis. Clin. Neurophys. 110 (12): 2165–2170, 1999.

[13] C. E. Shannon, W. Weaver: The mathematical theory of communication. University of Illinois Press, Urbana, 1949.

[14] F. Rieke, D. Warland, Rob de Ruyter van Steveninck, W. Bialek: Spikes – Exploring the neural code. MIT Press, Cambridge, 1997.

[15] A. Schlögl, D. Flotzinger, G. Pfurtscheller: Adaptive Autoregressive Modeling used for Single-trial EEG classification. Biomed. Tech. 42 (6):162–167, 1997.

[16] A. Schlögl: The electroencephalogram and the adaptive autoregressive model: theory and applications. Shaker Verlag, Aachen, 2000.

[17] G. Pfurtscheller, C. Neuper, A. Schlögl, K. Lugger: Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters. IEEE Trans. on Rehab. 6 (3): 316–325, 1998.

[18] A. Schlögl, K. Lugger, G. Pfurtscheller: Using adaptive autoregressive parameters for a brain-computer-interface experiment, Proc. 19th Int. Conf. IEEE/EMBS, S. 1533–1535, Chicago, 1997.

[19] C. Neuper, A. Schlögl, G. Pfurtscheller: Enhancement of left-right sensorimotor EEG differences during feedback-regulated motor imagery. J. Clin. Neurophys. 16 (2), 1999.

[20] K. Lugger, D. Flotzinger, A. Schlögl, M. Pregenzer, G. Pfurtscheller: Feature extraction for on-line EEG classification using principal components and linear discriminants. Med. Biol. Eng. Comput. 36, 309–314, 1998.

990

Address of correspondence:
Dr. Alois Schlögl
Phone: +43 316 873 5300
Fax: +43 316 873 5349
E-mail: schloegl@dpmi.tu-graz.ac.at