Clinical Neurophysiology 118 (2007) 2162–2171

www.elsevier.com/locate/clinph

Reliability of quantitative EEG features

Steinn Gudmundsson [a,c], Thomas Philip Runarsson [b], Sven Sigurdsson [a],
Gudrun Eiriksdottir [c], Kristinn Johnsen [c,*]

[a] *Department of Computer Science, University of Iceland, Reykjavik, Iceland*
[b] *Science Institute, University of Iceland, Reykjavik, Iceland*
[c] *Mentis Cura, Grandagardi 7, 101 Reykjavik, Iceland*

Accepted 10 June 2007
Available online 31 August 2007

Abstract

Objective: To investigate the reliability of several well-known quantitative EEG (qEEG) features in the elderly in the resting, eyes closed condition and study the effects of epoch length and channel derivations on reliability.
Methods: Fifteen healthy adults, over 50 years of age, underwent 10 EEG recordings over a 2-month period. Various qEEG features derived from power spectral, coherence, entropy and complexity analysis of the EEG were computed. Reliability was quantified using an intraclass correlation coefficient.
Results: The highest reliability was obtained with the average montage, reliability increased with epoch length up to 40 s, longer epochs gave only marginal improvement. The reliability of the qEEG features was highest for power spectral parameters, followed by regularity measures based on entropy and complexity, coherence being least reliable.
Conclusions: Montage and epoch length had considerable effects on reliability. Several apparently unrelated regularity measures had similar stability. Reliability of coherence measures was strongly dependent on channel location and frequency bands.
Significance: The reliability of regularity measures has until now received limited attention. Low reliability of coherence measures in general may limit their usefulness in the clinical setting.

Keywords: Quantitative EEG; Intra-individual reliability; Power spectrum; Coherence; Complexity; Entropy

1. Introduction

Quantitative EEG is well established for assessing the functional state of the brain. One or more numerical values (features) are calculated from the EEG and used as indicators for the brain state.

In order for a given feature to be clinically useful it must be highly stable in the sense that repeated measurements of a particular feature from a single subject should not exhibit large fluctuations when no systematic change occurs (e.g., drug effects). Stability itself is of course not a guarantee for clinical usefulness, e.g., the parameter may not be relevant for the condition of interest. The variability observed in EEG recordings can be attributed to changes in vigilance and the randomness that is inherent in the EEG. The former can be accounted for to some extent by carefully controlling experimental conditions but the latter is unavoidable. The EEG variability is reflected to a different extent in different features.

A number of studies have been carried out to evaluate the reliability of the resting EEG. Because of different selection of features and reliability measures, difference in choice of channel derivations, subject condition, epoch length, test–retest intervals and artifact handling, comparisons between studies are difficult. Most of the studies have focused on spectral and coherence based measures.

In a study by Grosveld et al. (1973) amplitude, frequency and time-domain parameters were used to discriminate

[*] Corresponding author. Tel.: +354 530 9901.
E-mail address: kristinn@mentiscura.is (K. Johnsen).

1388-2457/$32.00 © 2007 International Federation of Clinical Neurophysiology. Published by Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.clinph.2007.06.018

between subjects (16 subjects, 10 sessions during 1 year). The classification accuracy was 81%. The individual features with the highest discriminating ability were peak frequency in the α band and β power, indicating that inter-individual variation in these parameters is large compared to intra-individual variation.

Intra-individual stability of spectral parameters in 10–13-year-old children (26 subjects, 10-month retest interval) was investigated by Gasser et al. (1985). Their main findings were that for the eyes closed condition, the test–retest reliability was similar for absolute and relative power, it was rather uniform over different derivations but not across frequency bands. The highest reliability was obtained for the α band, then θ and the lowest for δ and β bands. Twenty seconds of data was found to be sufficient, using 40 or 60 s epochs did not improve reliability.

In a later study, Gasser et al. (1987) investigated the test–retest reliability of the coherence for the EEG at rest using the same EEG sample. The reliability was somewhat greater for the two α bands and greater when the coherence itself was large. The reliability was considerably lower than for absolute and relative band powers.

Kondacs and Szabó (1999) studied long-term intra-individual variability of various spectral measures together with coherence in healthy adults (45 subjects, 25–62-month retest interval) in the resting, eyes closed condition. Total power and α mean frequency proved to be most reliable, followed by absolute α and β power. Absolute δ power and α coherence were less reliable. The average montage gave slightly higher reliability than referential and longitudinal bipolar montages. The computation was based on 40 s of EEG.

Corsi-Cabrera et al. (1997) investigated the stability of inter- and intrahemispheric correlation, a measure related to coherence, in young women (9 subjects, 11 sessions during 1 month) in the resting, eyes closed condition. Using 20 s of data, within subject reliability was evaluated by computing the multiple correlation coefficients between all EEG features of the eleven sessions. The correlation measure was found to be a stable characteristic over a 1-month period.

Salinsky et al. (1991) evaluated reliability of spectral parameters in healthy adults (19 subjects, 5 min and 12–16 week retest intervals) in the eyes closed condition while subjects performed an auditory choice reaction time task to stabilize alertness. The peak α frequency and median frequency were the most stable features and there was essentially no difference between absolute and relative band power reliability. Sixty second epochs gave marginally higher averaged reliability score than 40 and 20 s epochs. Montage was found to have a significant effect. No significant association between intra-record and inter-record variability could be demonstrated.

Although the above studies were carried out under different experimental conditions some general conclusions can be made. Stable parameter estimates are obtained with 20–40 s of resting EEG. Absolute and relative band power measures have similar reliability and are considerably more reliable than coherence measures. Power in alpha band has the highest reliability, followed by θ and β bands, with δ being the least reliable. Median alpha and peak frequencies are found to be stable.

Features based on power spectrum decomposition have been the mainstay of qEEG analysis for both clinical and research purposes to this day. Numerous alternative features based on, e.g., autoregressive modelling, source localization, information theory and chaos theory have also been proposed. To our knowledge little is known about the reliability of these alternative measures.

The computation of some of the "modern" features is quite involved and often there is a large number of free parameters to be specified, making validation of published results difficult. This is true in particular for features originating in chaos theory such as correlation dimension and Lyapunov exponents.

This study is a part of a larger investigation into the use of qEEG in the diagnosis of Alzheimer's disease (AD). The selection of features and channel derivations is slightly biased towards features which have been found useful for discriminating between healthy and AD subjects. Only features which are relatively simple to implement are included in the study which means that correlation dimension, Lyapunov exponents and several other well-known parameters are omitted. There are still many details that must be taken into consideration during calculation of the qEEG features. The approach here is to duplicate procedures previously found to be useful, not to determine the "proper" way of carrying out the computations.

The aim of the present study was to investigate the reliability of several regularity measures based on entropy and complexity, some of which have recently been introduced in the EEG literature, and compare them to traditional qEEG features. Two important but often overlooked issues in qEEG studies are the selection of montage and epoch length. This study addresses both issues by investigating the reliability of different montages and varying epoch lengths. Most of the work on the quantification of EEG stability to date has been based on data from two recording sessions. In this study, the reliability was quantified on the basis of ten recording sessions.

2. Subjects and methods

2.1. Subjects

Fifteen healthy subjects (13 females and 2 males, mean age 71.7 years, SD 12.2) were recruited by advertisements at local retirement homes. The study was open to staff and residents provided they were over 50 years of age. The subjects received monetary payment for their participation. Each subject underwent 10 EEG recordings over a 2-month period. Written informed consent was obtained from the participants and the study was approved by the National Bioethics Committee.

2.2. EEG recording

The EEGs were obtained with the Nervus system (Taugagreining hf, Iceland). The 10–20 system of electrode placement was used with electrodes placed at Fp1, Fp2, F3, F4, F7, F8, Fz, T3, T4, T5, T6, A1, A2, C3, C4, Cz, P3, P4, Pz, O1, O2 and Oz with Fpz as reference. Two bipolar EOG channels were also recorded to monitor ocular artifacts. The sampling rate was 512 Hz and impedance was kept below 10 kΩ. The EEG was recorded for 3 min in the resting, eyes closed condition. The subjects were alerted in case they became visibly drowsy. The Nervus Reader software was used to manually score the recordings for artifacts. The raw EEG together with artifact data were exported into the Matlab environment (The MathWorks, Natick, MA, USA) where subsequent analysis took place.

2.3. Calculation of qEEG features

In addition to the referential montage (FPZ), the EEG was reformatted to average reference (AVR), source reference (SRC) (Nunez, 1981) and an anterior–posterior bipolar montage (APB): Fp1–F3, Fp1–F7, F7–T3, T3–T5, T5–O1, F3–C3, C3–P3, Fp2–F4, Fp2–F8, F8–T3, T4–T6, T6–O2, F4–C4, C4–P4 and P4–O2. After performing channel derivation, a 50 Hz notch filter was applied, the data band pass filtered between 0.5 and 40 Hz and downsampled to 256 Hz. To investigate the effects of segment length, the features were repeatedly calculated using epoch length of 10, 20, 40, 60, 80, 100 and 120 s.

2.3.1. Power spectral measures

The power spectrum density (PSD) was estimated using Welch's averaged modified periodogram method (Oppenheim and Schafer, 1999) with 2 s blocks, 50% overlap and a Hanning window. Blocks containing artifacts were skipped when averaging the periodograms. Traditionally the EEG power spectrum is partitioned into several frequency bands, this partition is ad hoc in the sense that it has no real biological basis and differs slightly between authors. Here the following definitions were used: δ (0.5–3.5 Hz), θ (3.5–7.5 Hz), α_1 (7.5–9.5 Hz), α_2 (9.5–12.5 Hz), β_1 (12.5–17.5 Hz), β_2 (17.5–25 Hz) and γ (25–40 Hz). For each band, absolute and relative band power were computed together with the total power (TP) in the range 0.5–40 Hz. Peak α frequency (PAF), the frequency with the highest power in the range (7.5–12.5 Hz), median frequency (MF), the frequency below which half of the total power occurs and spectral entropy (SpEn) (Inouye et al., 1991) were also determined. Additionally, the following power ratios were calculated $R_1 = \theta/(\alpha_1 + \alpha_2 + \beta_1)$ and $R_2 = (\delta + \theta)/(\alpha_1 + \alpha_2 + \beta_1 + \beta_2)$ which were found to be useful for discrimination of AD patients from healthy controls (Bennys et al., 2001) and $R_3 = \theta/(\alpha_1 + \alpha_2)$ which was found to be a useful indicator of slow abnormalities (Brunovsky et al., 2003).

2.3.2. Regularity measures

Features that quantify the "regularity" of the EEG have received considerable attention in recent years. The spectral entropy described previously is one measure of regularity (more accurately how sinusoidal the signal is), a sine wave has spectral entropy zero and uncorrelated white noise has spectral entropy one. Various complexity and entropy measures have been used to assess the level of sedation and anesthesia (Ferenets et al., 2006; Zhang et al., 2001), study regularity in epileptic seizures (Radhakrishnan and Gangadhar, 1998) and analyze the EEG background activity in patients with Alzheimer's disease (Abásolo et al., 2005, 2006).

An early attempt to quantitatively describe the EEG are the so-called Hjorth parameters (Hjorth, 1975), activity (A), mobility (M) and complexity (C). They are defined as follows: $A = a_0$, $M = (a_1/a_0)^{1/2}$, $C = (a_2/a_1 - a_1/a_0)^{1/2}$ where a_0 is the variance of the signal, a_1 is the variance of the first derivative of the signal and a_2 is the variance of the second derivative. From Parseval's theorem (Oppenheim and Schafer, 1999) it follows that activity and total power are equivalent features.

Approximate entropy (ApEn) introduced by Pincus (1991) has been widely used in the study of biomedical time series including the EEG (Radhakrishnan and Gangadhar, 1998; Abásolo et al., 2005; Ferenets et al., 2006). It turns out that ApEn has significant weaknesses such as strong dependence on sequence length and poor self-consistency. These shortcomings are described by Richman and Moorman (2000) who proposed an alternative statistic called sample entropy (SampEn). Given parameters m and r, SampEn is the negative log likelihood of the conditional probability that time series of length N having repeated itself within tolerance r for m points will repeat itself for $m + 1$ points (see Appendix A.1). SampEn was calculated using a free C program available from Physionet (www.physionet.org), a research resource for complex physiologic signals. Following Abásolo et al. (2005) the parameter settings were $m = 1$ and $r = 0.2$ times the standard deviation of the time series.

In work on brain–computer interfacing Roberts et al. (1998) suggest a temporal entropy measure (svdEn) based on an embedding space decomposition (see Appendix A.2). Here the embedding dimension m was set to 20 following (Faul et al., 2005; Roberts et al., 1998).

An algorithmic complexity measure introduced by Lempel and Ziv (1976) has been used for analyzing the regularity of oscillations in physiological data. Applications of Lempel–Ziv complexity (LZC) to EEG signals include assessment of the depth of anesthesia (Zhang et al., 2001) and sedation (Ferenets et al., 2006), differentiating between eyes open and eyes closed condition (Watanabe et al., 2003) and analysis of the background activity in Alzheimer's disease (Abásolo et al., 2006). For a given finite symbolic sequence, LZC measures the number of distinct patterns in the sequence. A detailed description of LZC along with an illustrative example is given by Zhang et al. (2001). To apply LZC to EEG data the time series has to be reduced to a symbol sequence. There is no single, correct way to do this but a common strategy is to use a 0–1 sequence and partition around the median, i.e., EEG voltage values that exceed the median voltage get assigned the symbol "1" and "0" otherwise.

The last regularity measure considered here is permutation entropy (PermEn), recently introduced by Bandt and Pompe (2002) and used to study epileptic activity (Keller and Lauffer, 2003; Cao et al., 2004). Following Keller and Lauffer (2003) the parameter settings were $m = 4$ and $\tau = 1$. The time series is converted to a symbolic sequence by counting ordinal patterns which describe up and down movement of the time series. Permutation entropy is defined as the Shannon entropy of the resulting symbolic series (see Appendix A.3).

The complexity and entropy measures were calculated for 5 s blocks (1280 samples) with 50% overlap and the results then averaged. Blocks containing artifacts were excluded from the averaging process.

2.3.3. Coherence measures

The coherence between two EEG signals is a measure of their synchronization and can be interpreted as an indicator of functional relationship between different brain regions. The magnitude squared coherence of signals $x(t)$ and $y(t)$ for frequency f is defined by

$$C_{xy}(f) = \frac{|P_{xy}(f)|^2}{P_{xx}(f)P_{yy}(f)}$$

where $P_{xx}(f)$ and $P_{yy}(f)$ are the power spectral densities of $x(t)$ and $y(t)$ and $P_{xy}(f)$ is their cross-spectral density. The coherence function takes values between 0 and 1 and was estimated using Welch's averaged periodogram method in exactly the same way as the power spectral density, ignoring blocks containing artifacts. The resulting features are the mean coherence in each of the seven frequency bands. Coherence was calculated for the local anterior, local posterior, far intrahemispheric and far interhemispheric brain regions as defined in Brunovsky et al. (2003), both from average and source montages.

2.4. Statistics

To establish that the EEG did not vary systematically between sessions the data were visually inspected as follows: For a fixed feature–channel pair (e.g., total power in P3–O1), the corresponding feature values were plotted against visits for all the subjects ($15 \times 10 = 150$ points in all). This was repeated for all derivations and features. No trend was observed.

The intraclass correlation coefficient ICC(1) (McGraph and Wong, 1996) was used to quantify reliability in this study since it involves ten sessions. The ICC is based on a one-way analysis of variance model and assumes that feature values are normally distributed. It is defined as follows

$$\mathrm{ICC} = \frac{\mathrm{MS}_{\mathrm{between}} - \mathrm{MS}_{\mathrm{within}}}{\mathrm{MS}_{\mathrm{between}} + (k-1)\mathrm{MS}_{\mathrm{within}}}$$

where $\mathrm{MS}_{\mathrm{between}}$ is the mean square error between subjects, $\mathrm{MS}_{\mathrm{within}}$ is the mean square error within subjects and k is the number of sessions. The ICC becomes one when there is perfect agreement between sessions and zero when the between subjects error equals the within subjects error. In rare cases the ICC can become negative, i.e., when the within subject error exceeds the between subjects error. Note that heterogeneity in the subject group will lead to increased between subjects error, hence inflate the ICC. Exact confidence intervals for the ICC are computed as described in McGraph and Wong (1996).

Prior to calculating ICC, the feature values were transformed in order to make them approximately normally distributed. Following (Kondacs and Szabó, 1999) the log transform was applied to absolute band power. Relative band power R was transformed using $\log(R/(1-R))$, magnitude squared coherence C was transformed with $\log(C/(1-C))$ and spectral entropy SpEn using $-\log(1-\mathrm{SpEn})$. The log transform was found to give approximate normality for total power, activity and the power ratios R_1, R_2 and R_3. The remaining features did not require transforms. In all cases, approximate normality was assessed using a normal probability plot.

Following Gasser et al. (1985), the average of reliability scores over all features and derivations was used to measure the effect of epoch length. The montages were evaluated in the same way. Approximate confidence intervals for the averaged reliability values were obtained with the bias-corrected accelerated bootstrap (Efron and Tibshirani, 1994).

2.5. Nonlinear associations

The nonlinear association measure (Pijn and da Silva, 1993) was used to assess the correlation between different qEEG features in an attempt to explain why apparently unrelated features exhibited similar reliability. This measure has previously been used to quantify the degree and direction of functional coupling between neuronal populations (Bartolomei et al., 2004), to study the functional dependence between septal and temporal signals in a model of temporal lobe epilepsy (Kalitzin et al., 2005) and to analyze the cortical involvement in the generation of motor seizures (Kalitzin et al., 2007).

The association measure h^2_{XY} quantifies the (nonlinear) relationship between two sequences X and Y by considering Y as a piecewise linear function of X and measuring the reduction in variance obtained by predicting Y according to the fitted curve. Each sequence consists of values of a single feature, aggregated over all subjects, channels and visits. Ten line segments were used to construct the regression curve. Values of h^2_{XY} close to one suggest a strong relationship between X and Y, values close to zero indicate independence. Considering X as a function of Y instead may result in a different value of the association measure (asymmetry). To quantify the association between features X and Y, the average was used, $h^2 = (h^2_{XY} + h^2_{YX})/2$.

3. Results

The reliability values for all power spectral and regularity measures are presented as topographic maps in Fig. 1. The values are based on 40 s epochs and the average montage. The EEGLAB package (Delorme and Makeig, 2004) was used to create the maps. Reliability values for selected channels are presented in Table 1. Table 2 shows 95% confidence intervals for absolute band power in a single channel and indicates the uncertainty in the point estimates.

3.1. Power spectral measures

The effects of epoch length and montage on reliability are illustrated for absolute band power in Fig. 2. Also shown are 95% bootstrapped confidence intervals. The corresponding figures for relative band power and the derived PSD features basically show the same pattern. Reliability was highest for AVR and lowest for SRC and APB with FPZ in between. Reliability increased with epoch length but levels off at 40 s, longer epochs gave only marginal improvement. Reliability across channels was relatively stable, parietal, occipital, Fz and Cz were most reliable. Reliability for absolute and relative band power was similar, highest reliability was observed for the θ band, followed by α and β bands, δ and γ bands were least reliable. Of the derived features, $R_1 - R_3$ had the highest reliability, spectral entropy and MF the lowest.

3.2. Regularity measures

Montage had the same effects as before, AVR was the most reliable montage, followed by FPZ and then by SRC and APB. Reliability increased with epoch length but levelled off at 40 s. Variation in reliability across channels was similar for all the measures,, with parietal, Cz and Fz derivations being most reliable. Reliability of the regularity measures is comparable to that of relative δ and γ band power, i.e., lower than for most PSD measures. To investigate whether this is simply a result of averaging (5 s epochs instead of 2) the reliability calculations were repeated using 2 s epochs. The effect was minimal, suggesting that the difference in reliability compared to PSD measures is not simply due to the effects of averaging. Using different embedding dimensions, $m = 2$ and $m = 5$ for SampEn and $m = 5$ and $m = 10$ for svdEn did not have a noticeable effect on reliability.

3.3. Coherence measures

Reliability levelled off at 40 s with the average montage more reliable than the source montage. Reliability for the mean coherence measures is depicted in Fig. 3. Each channel derivation is represented by a single line, the color indicating the reliability; below 0.4 (blue), 0.4–0.55 (green), 0.55–0.7 (orange) and above 0.7 (red). Reliability was highest for the α bands, followed by θ and β bands, δ and γ bands had lowest reliability. The local posterior area was slightly less reliable than the other areas.

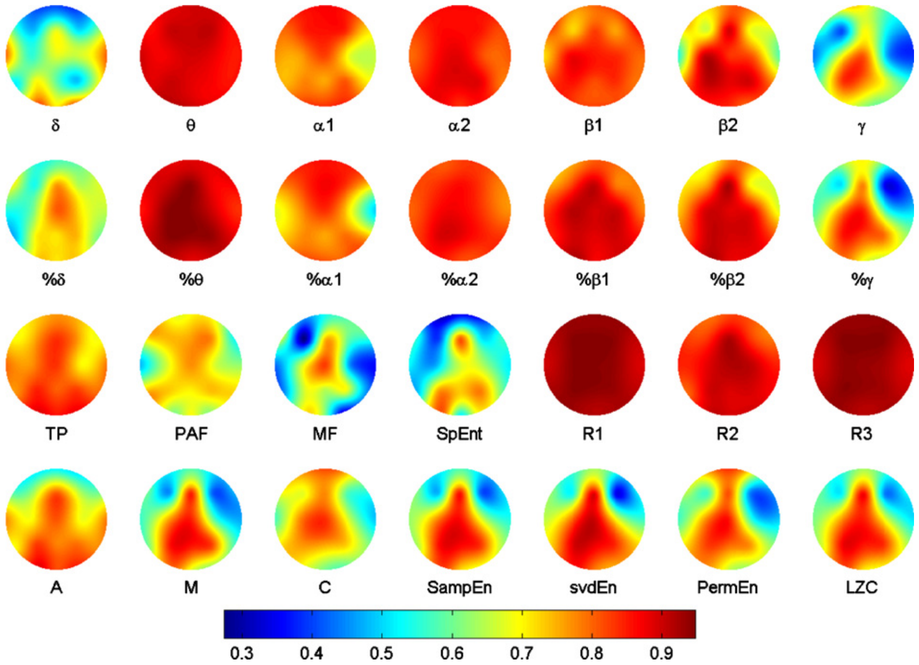

Fig. 1. Topographic maps of reliability across features and channel locations for power spectral measures (top three rows) and regularity measures (bottom row), blue color represents low values and red represents high values.

Table 1
Reliability, 40 s epochs and AVR montage

	F3	F4	C3	C4	P3	P4	O1	O2	Mean
δ	0.49	0.51	0.56	0.67	0.69	0.48	0.81	0.79	0.62
θ	0.90	0.91	0.87	0.87	0.90	0.87	0.91	0.88	0.89
α_1	0.85	0.84	0.76	0.67	0.78	0.81	0.83	0.84	0.80
α_2	0.86	0.86	0.86	0.81	0.88	0.88	0.87	0.86	0.86
β_1	0.73	0.76	0.87	0.84	0.86	0.84	0.82	0.82	0.82
β_2	0.67	0.69	0.90	0.80	0.91	0.89	0.85	0.81	0.82
γ	0.41	0.55	0.65	0.46	0.83	0.69	0.69	0.60	0.61
Mean	0.70	0.73	0.78	0.73	0.84	0.78	0.83	0.80	0.77
$\%\delta$	0.54	0.65	0.59	0.71	0.70	0.74	0.74	0.74	0.67
$\%\theta$	0.91	0.88	0.94	0.88	0.95	0.94	0.91	0.90	0.91
$\%\alpha_1$	0.83	0.83	0.77	0.70	0.78	0.79	0.82	0.80	0.79
$\%\alpha_2$	0.84	0.83	0.86	0.82	0.89	0.85	0.85	0.82	0.84
$\%\beta_1$	0.83	0.79	0.91	0.89	0.90	0.90	0.92	0.89	0.88
$\%\beta_2$	0.76	0.70	0.88	0.86	0.90	0.89	0.92	0.88	0.85
$\%\gamma$	0.52	0.36	0.75	0.53	0.85	0.79	0.79	0.66	0.66
Mean	0.75	0.72	0.81	0.77	0.85	0.84	0.85	0.81	0.80
TP	0.76	0.77	0.77	0.70	0.85	0.83	0.89	0.88	0.80
PAF	0.74	0.77	0.66	0.66	0.73	0.74	0.61	0.62	0.69
MF	0.27	0.61	0.65	0.44	0.67	0.69	0.73	0.35	0.55
SpE	0.40	0.52	0.46	0.63	0.76	0.78	0.79	0.68	0.63
R_1	0.95	0.94	0.94	0.93	0.94	0.93	0.93	0.92	0.93
R_2	0.83	0.83	0.87	0.90	0.89	0.89	0.91	0.90	0.88
R_3	0.95	0.94	0.93	0.92	0.94	0.93	0.93	0.92	0.93
Mean	0.70	0.77	0.75	0.74	0.83	0.83	0.83	0.75	0.77
A	0.66	0.67	0.69	0.68	0.83	0.80	0.88	0.85	0.76
M	0.46	0.43	0.73	0.56	0.86	0.83	0.80	0.63	0.66
C	0.68	0.58	0.78	0.65	0.77	0.73	0.67	0.58	0.68
SampEn	0.49	0.42	0.76	0.64	0.86	0.83	0.83	0.70	0.69
svdEn	0.55	0.36	0.77	0.66	0.89	0.84	0.86	0.75	0.71
PermEn	0.57	0.45	0.75	0.52	0.83	0.77	0.75	0.63	0.66
LZC	0.53	0.45	0.74	0.65	0.87	0.83	0.82	0.70	0.70
Mean	0.56	0.48	0.75	0.62	0.84	0.80	0.80	0.69	0.69

4. Discussion

4.1. Power spectral measures

Compared to AVR the other montages had lower overall reliability which is consistent with Kondacs and Szabó (1999). This finding is not surprising since the potential difference between a single electrode and the average over all channels will have lower variance than the potential difference between two electrodes (FPZ and ABP) or the average of only 3–5 (SRC). The largest reliability differences between AVR and the other montages were found in the fronto-polar and frontal derivations. The FPZ reference montage was found to have slightly higher reliability than the longitudinal APB montage which is in agreement with Salinsky et al. (1991). For the FPZ montage, reliability was low frontally but increased over the parietal and occipital areas.

Table 2
95% Confidence intervals on reliability illustrated for channel F3 and absolute band power

F3-AV	δ	θ	α_1	α_2	β_1	β_2	γ
Value	0.49	0.90	0.85	0.86	0.73	0.67	0.41
Lower	0.30	0.82	0.73	0.76	0.56	0.49	0.23
Upper	0.72	0.96	0.93	0.94	0.87	0.84	0.66

Reliability of the PSD measures increased for up to 40 s epochs whereas Gasser et al. (1985) found practically no improvement after 20 s. In Salinsky et al. (1991), 20 s was found to be nearly as reliable as 60 s for test–retest correlations but a different criteria gave markedly higher variability for 20 s epochs than for 60 s epochs. In both Gasser et al. (1985) and Salinsky et al. (1991) elaborate methods were used to reduce the effect of artifacts. Absolute and relative band power had similar reliability which is in agreement with Gasser et al. (1985) and variability over channels is also modest. The δ and γ bands were less reliable than the other bands.

The low reliability of MF and moderate reliability of PAF are in contrast with Salinsky et al. (1991); Kondacs and Szabó (1999) which found both measures to be highly

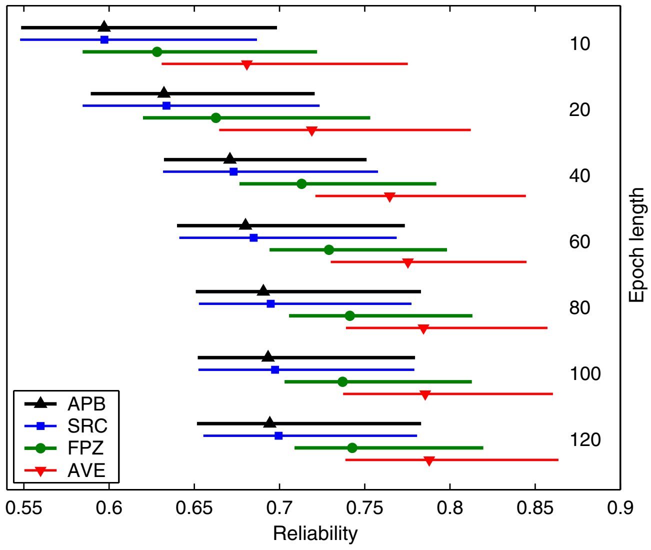

Fig. 2. Effects of epoch length and montage on reliability for absolute band power (averaged across channels and frequency bands). The point estimates are denoted with ▲ (APB), ■ (SRC), ●(FPZ) and ▼(AVR), horizontal lines indicate corresponding 95% confidence intervals.

reliable. The discrete nature of the measures may be a contributing factor here, small variation in power may lead to relatively large (0.5 Hz) jumps in the frequency estimates.

4.2. Regularity measures

The regularity measures were found to be slightly less reliable than PSD features in most cases. To the best of our knowledge, the reliability of regularity measures has not received much attention in the EEG literature. Spectral entropy was studied in Kondacs and Szabó (1999) and found to have moderate stability compared to other PSD features.

When selecting the epoch length there is a trade-off between obtaining sufficiently reliable feature estimates and fluctuations in alertness of the subjects which will have severe effects on the parameter values. We recommend that for the regularity measures studied here at least 40 s epochs are used, although this value may depend on the block size and whether block overlap is used or not.

The reliability of mobility, SampEn, svdEn and LZC was quite similar. This is somewhat surprising. Although they are all measures of "regularity", the measures have different theoretical underpinnings and the algorithms for their computation do not seem to have a lot in common at first glance.

Scatter plots can be used to reveal relationship between two features. The symmetric scatter plot matrix in Fig. 4 contains all pairwise scatter plots for the regularity features and the corresponding values of the association measure h^2. Each scatter plot was generated by pooling feature values for all subjects, visits and channels ($15 \times 10 \times 20 = 3000$ points). Perfect agreement between features i and j would show up as a straight line in row i, column j. If there were no correlation between the two features, the corresponding scatter plot would display a "cloud" of points. Activity appears to have the least in common with the other measures. Complexity seems to be mostly unrelated with the other measures except PermEn. On the other hand, mobility, sample entropy, svd entropy and Lempel–Ziv complexity appear to be quite related. These four measures were quite strongly associated with relative γ power ($h^2 > 0.8$), more so than with spectral entropy.

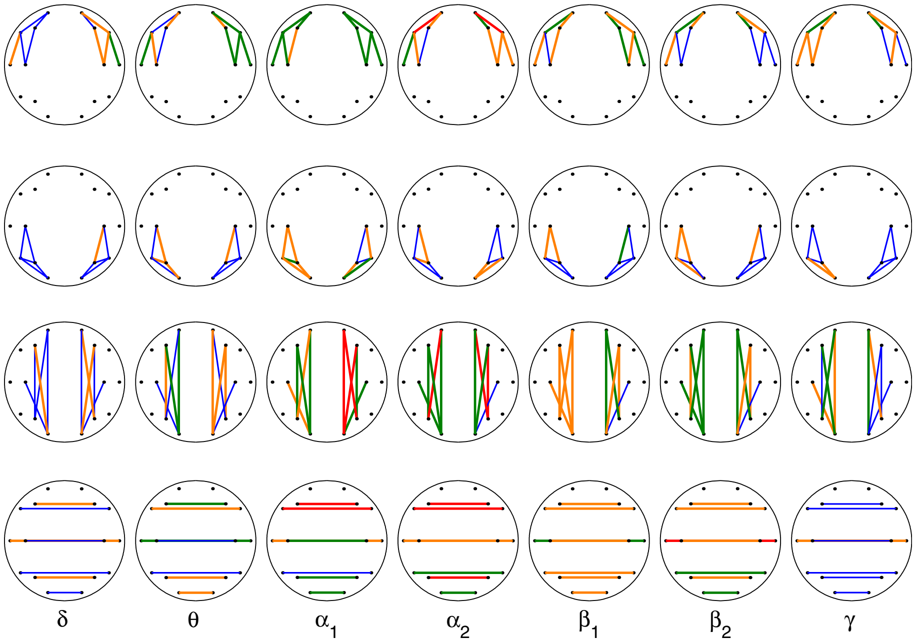

Fig. 3. Coherence reliability: below 0.4 (blue), 0.4–0.55 (green), 0.55–0.7 (orange) and above 0.7 (red). From top to bottom; local anterior, local posterior, far intrahemispheric and far interhemispheric.

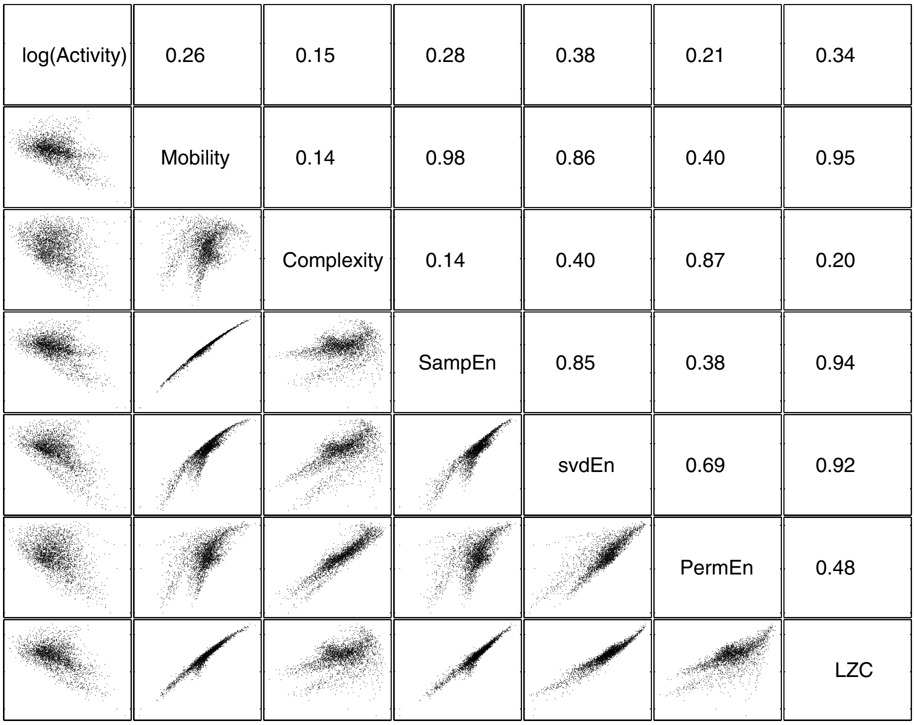

Fig. 4. Scatter plot matrix of the complexity and entropy features and corresponding values of the nonlinear association measure h^2.

Note that apparent relationships (or lack thereof) between different measures may depend strongly on the choice of parameters (e.g., values of m and τ in case of PermEn). Mobility is such a simple parameter to compute and understand it is therefore recommended as a benchmark in further studies involving these features.

4.3. Coherence measures

We recommend to use average reference and at least 40 s epochs when computing coherence. Reliability of coherence was found to be lower than for absolute and relative band power (and in fact lower than most of the features included in the study). Coherence was most reliable in the α bands and least reliable in the δ and γ bands. These findings are consistent with previous studies (Gasser et al., 1987; Kondacs and Szabó, 1999). The lower reliability of coherence when compared to PSD measures can be explained in part by the greater statistical variability of coherence and the possibility that synchronization between brain regions is significantly affected by the mental state of the subject (Gasser et al., 1987). The ability of coherence to detect brain coupling may be offset by the apparent low reliability in clinical applications.

Recently, numerous synchronization measures have been proposed in the EEG literature (for an overview, see Quiroga et al., 2002). The low reliability of coherence observed in this study would suggest that further studies into the stability of these new parameters are needed.

Acknowledgments

The project was supported by The Icelandic Center for Research (RANNIS). We thank Jóhannes Helgason, Gísli Hólmar Jóhannesson and Nicolas Blin for their assistance with the EEG recordings.

Appendix A

A.1. Calculation of sample entropy

Given a scalar time series $x(t)$ of length N, a time-delay embedding of $x(t)$ is obtained by forming delay vectors

$$\mathbf{x}_m(i) = [x(i), x(i+\tau), \ldots, x(i+(m-1)\tau)]^{\mathrm{T}}$$

for $i = 1, \ldots, N-(m-1)\tau$ where m is the embedding dimension and τ is the time delay.

To compute SampEn let $\tau = 1$ and assume that parameters m and r are fixed.

Define distance between two vectors as

$$d[\mathbf{x}_m(i), \mathbf{x}_m(j)] = \max_{0 \leqslant k \leqslant m-1} [|x(i+k) - x(j+k)|]$$

Compute the probability that two sequences will match for m points

$$B^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} B_i^m(r)$$

where $B_i^m(r) = (N-m-1)^{-1}$ times the number of vectors $\mathbf{x}_m(j)$ within distance r of $\mathbf{x}_m(i), j = 1, \ldots, N-m, j \neq i$.

The conditional probability $A^m(r)$ that two sequences will match for $m+1$ points is defined analogously

$$A^m(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} A_i^m(r)$$

where $A_i^m(r) = (N-m-1)^{-1}$ times the number of vectors $\mathbf{x}_{m+1}(j)$ within distance r of $\mathbf{x}_{m+1}(i), j = 1,\ldots,N-m$, $j \neq i$. Now

$$\mathrm{SampEn}(m,r) = -\ln\frac{A^m(r)}{B^m(r)}.$$

Commonly used values for the parameters are $m = 1,\ldots,5$ and $r = 0.1$–0.2 times the standard deviation (STD) of the original time series.

A.2. Calculation of svd entropy

Computation of svdEn proceeds as follows: First an embedding matrix is constructed from the m-dimensional time delay vectors with $\tau = 1$

$$X = [\mathbf{x}_m(1), \mathbf{x}_m(2), \ldots, \mathbf{x}_m(N-(m-1))]^{\mathrm{T}}$$

Compute the singular value decomposition $X = \mathrm{USV}^{\mathrm{T}}$. The diagonal matrix S contains the singular values, $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_m \geqslant 0$. The entropy of the singular value spectrum is defined as

$$\mathrm{svdEn} = -\sum_{i=1}^{m} \hat{\sigma}_i \log \hat{\sigma}_i$$

where $\hat{\sigma}_i$ are the normalized singular values $\hat{\sigma}_i = \sigma_i / \sum_{j=1}^{m} \sigma_j$.

A.3. Calculation of permutation entropy

For a given embedding dimension m, time delay τ and time i, arrange the elements of the time delay vector $\mathbf{x}_m(i)$ in increasing order

$$\tilde{\mathbf{x}}_m(i) = [x(i+j_0\tau) \leqslant x(i+j_1\tau) \leqslant \cdots \leqslant x(i+j_{m-1}\tau)]^{\mathrm{T}}$$

where $(j_0, j_1, \ldots, j_{m-1})$ is called an ordinal pattern and is a permutation of $(0,1,\ldots,m-1)$. To ensure a unique result in case of equalities, set $j_{k-1} < j_k$ when $x(i+j_{k-1}\tau) = x(i+j_k\tau)$. Now any $\tilde{\mathbf{x}}_m(i)$ is uniquely mapped onto $(j_0, j_1, \ldots, j_{m-1})$. Each ordinal pattern can be considered as one of $m!$ distinct symbols. Denote the relative frequency of the distinct symbols by $P_1, P_2, \ldots, P_K$ where $K \leqslant m!$. The (normalized) permutation entropy for $x(t)$ is defined as

$$\mathrm{PermEn} = -\frac{1}{\log(m!)}\sum_{j=1}^{K} P_k \log P_k$$

and takes values between 0 and 1. Different values of the time delay provide different details about the time series. Bandt and Pompe (2002) use $\tau = 1$ and recommend $m = 3,\ldots,7$.

References

Abásolo D, Hornero R, Espino P, Poza J, Sánchez CI, de la Rosa R. Analysis of regularity in the EEG background activity of Alzheimers disease patients with approximate entropy. Clin Neurophysiol 2005;116:1826–34.

Abásolo D, Hornero R, Gómez C, López M. Analysis of EEG background activity in Alzheimer's disease patients with Lempel–Ziv complexity and central tendency measure. Med Eng Phys 2006;28:315–22.

Bandt C, Pompe B. Permutation entropy: a natural complexity measure for time series. Phys Rev Lett 2002:88.

Bartolomei F, Wendling F, Régis J, Gavaret M, Guye M, Chauvel P. Pre-ictal synchronicity in limbic networks of mesial temporal lobe epilepsy. Epilepsy Res 2004;61:89–104.

Bennys K, Rondouin G, Vergnes C, Touchon J. Diagnostic value of quantitative EEG in Alzheimer's disease. Neurophysiol Clin 2001;31:153–60.

Brunovsky M, Matousek M, Edman A, Cervena K, Krajca V. Object assessment of the degree of dementia by means of EEG. Neuropsychobiology 2003;48:19–26.

Cao Y, Tung W, Gao J, Protopopescu V, Hively L. Detecting dynamical changes in time series using permutation entropy. Phys Rev E 2004;70:046217.

Corsi-Cabrera M, Solís-Ortiz S, Guevara M. Stability of EEG inter- and intrahemisphereic coherence in women. Electroencephalogr Clin Neurophysiol 1997;102:248–55.

Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial eeg dynamics. J Neurosci Methods 2004;134:9–21.

Efron B, Tibshirani R. Introduction to the bootstrap. Boca Raton: Chapman & Hall/CRC; 1994.

Faul S, Boylan G, Connolly S, Marnane W, Lightbody G. Chaos theory analysis of the newborn EEG: is it worth the wait? In: Proceedings of the IEEE international symposium on intelligent signal processing 2005. p. 381–6.

Ferenets R, Lipping T, Anier A, Jäntti V, Melto S, Hovilehto S. Comparison of entropy and complexity measures for the assessment of depth of sedation. IEEE Trans Biomed Eng 2006;53:1067–77.

Gasser T, Bacher P, Steinberg H. Test–retest reliability of spectral parameters of the EEG. Electroencephalogr Clin Neurophysiol 1985;60:312–9.

Gasser T, Jennen-Steinmetz C, Verleger R. EEG coherence at rest and during a visual task in two groups of children. Electroencephalogr Clin Neurophysiol 1987;67:151–8.

Grosveld F, Jansen B, Hasman A, Visser S. La reconnaissance des individus a l'interieur d'un groupe de 16 sujets normaux. Rev Electroencephalogr Neurophysiol Clin 1973;6:297.

Hjorth B. Time domain descriptors and their relation to a particular model for generation of EEG activity. In: CEAN – Computerized EEG analysis. Stuttgart: Gustav Fischer Verlag; 1975. p. 3–8.

Inouye T, Shinosaki K, Sakamotor H, Toi S, Ukai S, Iyama A, et al. Quantification of EEG irregularity by use of the entropy of the power spectrum. Electroencephalogr Clin Neurophysiol 1991;79:204–10.

Kalitzin SN, Derchansky M, Velis DN, Parra J, Carlen PL, da Silva FL. Amplitude and phase synchronization in a model of temporal lobe epilepsy. In: Proceedings of the third European medical and biological engineering conference; 2005. p. 1–6.

Kalitzin SN, Parra J, Velis DN, da Silva FL. Quantification of unidirectional non-linear associations between multidimensional signals. IEEE Trans Biomed Eng 2007;54:454–61.

Keller K, Lauffer H. Symbolic analysis of high-dimensional time series. Int J Bifurcat Chaos 2003;13:2657–68.

Kondacs A, Szabó M. Long-term intra-individual variability of the background EEG in normals. Clin Neurophysiol 1999;110:1708–16.

Lempel A, Ziv J. On the complexity of finite sequences. IEEE Trans Inf Theory 1976;22:75–88.

McGraph K, Wong S. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30–46.

Nunez PL. Electric fields of the brain. New York: Oxford; 1981.

Oppenheim A, Schafer R. Discrete-time signal processing. New Jersey: Prentice Hall; 1999.

Pijn J, da Silva FL. Propagation of electrical activity: nonlinear associations and time delays between EEG signals. In: Zschocke S, Speckmann EJ, editors. Basic mechanisms of the EEG. Boston: Birkauser; 1993. p. 41–61.

Pincus S. Approximate entropy as a measure of system complexity. Proc Natl Acad Sci USA 1991;88:2297–301.

Quiroga RQ, Kraskov A, Kreuz T, Grassberger P. Synchronization measures in real data: a case study on electroencephalographic signals. Phys Rev E 2002;65:041903.

Radhakrishnan N, Gangadhar B. Estimating regularity in epileptic seizure time-series data. IEEE Eng Med Biol Mag 1998;17:89–94.

Richman J, Moorman J. Physiological time-series analysis using approxi mate entropy and sample entropy. Am J Physiol Heart Circ Physiol 2000;278:2039–49.

Roberts SJ, Penny W, Rezed I. Temporal and spatial complexity measures for EEG-based brain–computer interfacing. Med Biol Eng Comput 1998;37:93–9.

Salinsky M, Oken B, Morehead L. Test–retest reliability in EEG frequency analysis. Electroencephalogr Clin Neurophysiol 1991;79:382–92.

Watanabe T, Cellucci C, Kohegyi E, Bashore T, Josiassen R, Greenbaun N, et al. The algorithmic complexity of multichannel EEGs is sensitive to changes in behavior. Psychophysiology 2003;40:77–97.

Zhang X, Roy R, Jensen E. EEG complexity as a measure of depth of anesthesia for patients. IEEE Trans Biomed Eng 2001;48:1424–33.