ELSEVIER

Clinical Neurophysiology 110 (1999) 2165–2170

www.elsevier.com/locate/clinph

Quality control of polysomnographic sleep data by histogram and entropy analysis

Alois Schlögl[a], Bob Kemp[b], Thomas Penzel[c], Dieter Kunz[d], Sari-Leena Himanen[e], Alpo Värri[f], Georg Dorffner[g], Gert Pfurtscheller[a,*]

[a]*Institute of Biomedical Engineering, University of Technology, Graz, Austria*
[b]*Holland Sleep Research, Westeinde Hospital, Den Haag, Netherlands*
[c]*Zentrum für Innere Medizin - Schlafmedizinisches Labor, Klinikum der Philipps-Universität Marburg, Marburg, Germany*
[d]*Interdisciplinary Sleep Clinic, Department of Psychiatry, Free University of Berlin, Berlin, Germany*
[e]*Department of Clinical Neurophysiology, Tampere University Hospital, Tampere, Finland*
[f]*Digital Media Institute, Tampere University of Technology, Tampere, Finland*
[g]*Austrian Research Institute for Artificial Intelligence, Vienna, Austria*

Accepted 9 July 1999

Abstract

Objective and methods: Sixteen polysomnographic recordings from 8 European sleep laboratories were analyzed. The histogram analysis was used to introduce quality control of all-night EEG recordings.

Results: It was found that the header information does not always provide the real saturation values of the recording equipment. The entropy measure was used for the quantitative analysis of the dynamic range of routinely used polysomnographic recorders. It was found that the recording equipment provides EEG data with entropy in the range of 8–11 bits.

Conclusion: In the all-night sleep EEG were observed non-linearities. It is recommended that the equipment provide the saturation values in order to apply automated overflow detection.

Keywords: Histogram analysis; Entropy; Polysomnography; Artifact processing; Automatic sleep analysis

1. Introduction

Typical artifacts in biosignal processing are saturation effects caused by the limited dynamic range of the amplifier and/or the analog-to-digital converter (ADC). Overflow artifacts can be caused by electrode movement, failing electrodes, or EMG activity in an EEG channel. If the limit is reached, the true signal is no longer represented, and a saturation effect takes place, which implies that a non-linearity is introduced. In this case, almost all signal processing methods fail or produce large estimation errors.

One way to enlarge the dynamic range and to minimize the probability of reaching the limits is to reduce the amplification gain. However, this would cause the amplitude resolution, or signal-to-noise ratio (SNR), to worsen. In practice, a compromise between the resolution and the dynamic range of the signal is chosen. The dynamic range can be quantified with Shannon's communication theory by applying the concept of the entropy of information to the (coded) EEG data (Shannon and Weaver, 1949).

The analysis of the amplitude histograms is useful for the quantitative analysis of the dynamic range and is often used for testing ADCs. Histograms are useful features, which provide a compressed representation of the data. They were used early in automated analysis systems for different applications. (Elul, 1969; Glass, 1970). Martin-Rodriguez et al. (1982) investigated spiking neurons using interval histograms. Rieke et al. (1997) provided a mathematical framework of entropy analysis for information transmission between spiking neurons.

In this study we calculated the amplitude histograms and entropies of all-night sleep recordings from 8 different sleep laboratories. Polysomnographic signals were measured, amplified and digitized with an ADC. The data were stored in a digital data format for biosignals using 16 bit integer numbers (Kemp et al., 1992). Other technical details besides the technical specifications of the recorders may be important. The following analysis should enlighten the details and

* Corresponding author. Technische Universität Graz, Inffeldgasse 16a, A-8010 Graz, Austria. Tel.: +43-316-873-5300; fax: +43-316-873-5349.
E-mail address: schloegl@dpmi.tu-graz.ac.at (G. Pfurtscheller)

1388-2457/99/$ - see front matter
PII: S1388-2457(99)00172-8

CLINPH 99536

introduce quantitative measures for recordings of biosignals.

2. Methods and data

The data were recorded according to the protocol of the SIESTA project (Dorffner 1998). Sixteen polysomnographic channels (6 + 1 EEG, 2 EOG, 2 EMG, 1 ECG, 3 respiration and one oxygen saturation (SaO_2) -channels) from 16 all-night recordings (between 6:00 h 26 min 10.0 s and 8:00 h 56 min 00.0 s, see also Table 1) were investigated. The recordings were chosen according to a randomized list. Sleep recorders from the following providers were used: Fa Jaeger (SleepLap 1000P), Nihon Kohden/DeltaMed, Walter Graphtek (PL-EEG), Flaga (EMBLA) and Siemens. Sampling rates of 1, 8, 16, 20, 25, 100, 200, 256 and 400 Hz were used for the various channels. The sampling rates and filter settings were stored in the headers of the data files (Kemp et al., 1992) for further evaluation.

The European data format for biosignals (EDF) (Kemp et al., 1992) uses two byte (16 bit) signed integer numbers; the scaling information is stored in the file header. This format allows values within a range from $-2^{15} = -32\,768$ to $+2^{15} - 1 = 32\,767$ to be represented.

Each voltage value is mapped (coded) to one of these digital values. The smallest voltage difference that gives a different digitized value determines the amplitude resolution. The digitization error is also called the quantification noise.

The digitized data is used to investigate how many samples of one data series (channel) have a certain value. Performing this for every possible value from − 32 768 to + 32 767 gives a function *H(i)*, which is called the histogram of that data series.

Various parameters can be obtained from the histogram: the total number of samples *N* (A1), the entropy of information (A3), the mean value (A6), the variance σ^2 (A7), and the corresponding Gaussian distribution (A8) based on the mean and variance. The latter can be related to the probability (A2) of the occurrence of a certain value. In addition the skewness (A9) and kurtosis (A10, A11) of the data can be calculated. The details of the computation can be found in the Appendix (A1–A11). Note that once the histogram is available, the computational effort – even for an all-night recording – is quite low. Calculations were done using Matlab on an Intel Pentium processor with a Redhat Linux operating system.

Most of the measures, e.g. mean, variance, probability distribution etc, are well known and need no further explanation, with a possible exception of the entropy measure. Shannon and Weaver (1949) introduced the concept of the 'entropy of information' into communication theory (coding, information transmission). The entropy value is a measure for the variability, randomness, the average amount of choices or the average amount of information. The larger the variability, the higher is the entropy. For our purpose is sufficient to know, firstly, how the entropy is defined in discrete systems (see Appendix A3). Secondly, the entropy of a continuous Gaussian process is defined by the variance (A4). One can also estimate the entropy of noise, for example the quantization noise in an ADC or the amplifier noise. Thirdly, the entropy difference between a signal and noise is determined by the SNR (A5) and vice versa. A larger value of the entropy difference means a better SNR and a better resolution of the signal. In the following the entropy of the digitized signal based on the histogram (A3) will be used.

3. Results

In Fig. 1 the histograms of the EEG channel C3-A2 of each recording are shown. The following findings can be observed:

Table 1
The sampling rate, number of samples, maximum and minimum value, mean, standard deviation (SD), skewness, kurtosis and entropy of channel C3-A2 of 16 recordings are shown. The results were obtained from the histograms in Fig. 1. All values (except the entropy) have the unit 1

Record	Sample rate (Hz)	Samples	Maximum	Minimum	Mean	SD	Skewness	Kurtosis	Entropy (bit)
B1	200	5778000	1008	−1851	0	62.24	−1.96E + 05	1.94E + 08	7.7
B04	200	5780400	3395	−1921	−0.9	70.3	3.16E + 05	2.05E + 09	7.9
C07	200	5718000	32767	−32767	−3.4	304.83	−7.26E + 08	7.77E + 13	8.0
C17	200	5802000	32534	−32469	1.1	118.18	3.34E + 06	4.75E + 12	8.2
H02	100	3001000	4080	−1995	−5.9	241.19	5.79E + 06	5.04E + 10	9.8
H09	100	2956000	4080	−2048	−3.4	407.85	−2.56E + 07	3.94E + 11	9.8
M04	200	3216000	32763	−32768	−10.6	954.91	−9.75E + 09	7.72E + 14	8.1
M09	200	5438000	32767	−32768	−2.3	1180.35	−1.10E + 09	1.20E + 15	8.5
N01	256	6815744	2051	−2107	21.8	556.26	5.18E + 07	3.43E + 11	10.8
N04	256	5931520	2030	−2128	11.8	379.26	−4.49E + 06	1.33E + 11	10.4
P01	256	6772736	1610	−1610	25.8	100.09	−1.13E + 05	1.20E + 09	8.5
P02	256	6913280	1562	−1552	25.6	152.93	3.73E + 05	1.46E + 10	8.7
S02	256	7432448	1596	−1643	24	198.41	4.82E + 06	2.26E + 10	9.2
S07	256	7397376	1619	−1642	25.8	159.59	3.00E + 05	1.51E + 10	8.9
UH3	200	5762000	4750	−742	2042	241.61	−3.10E + 06	4.26E + 10	9.5
U06	200	5762000	4600	−676	2060.7	212.69	−1.36E + 06	2.90E + 10	9.6

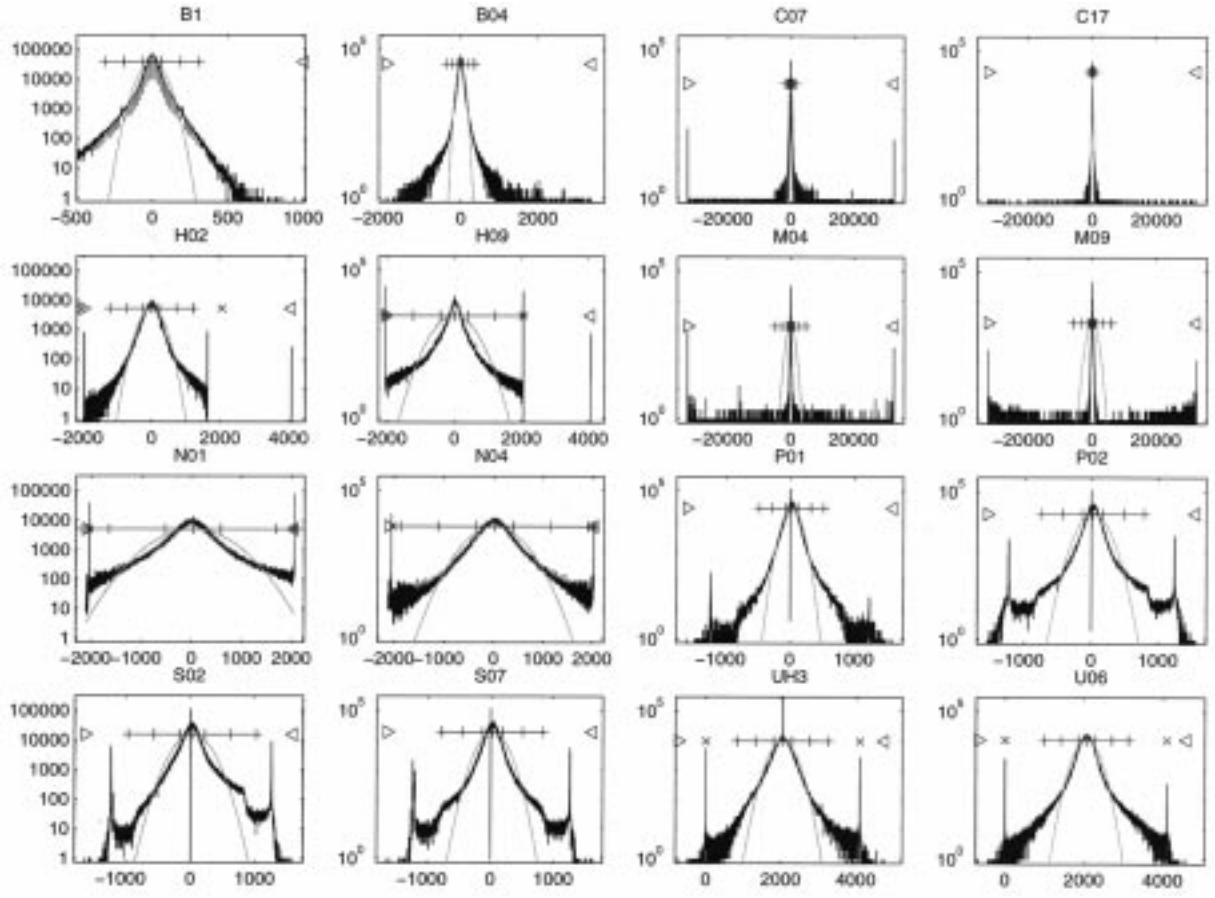

Fig. 1. Histograms of EEG channel C3–A2 from 16 all-night recordings. The units at the horizontal axes are digits and can range from −32 768 to 32 767. The horizontal line with the vertical ticks displays the mean ± the 1, 3 and 5 times of the standard deviation. The markers ▷ and ◁ indicate the real maxima and minima found in the recordings. The markers × indicate the digital minima and maxima as stored in the header information; if invisible the values are outside the scope. The light parabolic lines display the Gaussian distribution with the corresponding mean and variance.

1. The histogram of B1 is not smooth. Certain values have much lower probability than their neighboring values. The pattern is also quite periodic. One explanation is that the integer values obtained by the analog-digital-converter (ADC) were scaled with some non-integer values. Effects of rounding effects yield this kind of histogram.
2. (a) In recordings H02, H09, N01 and N04 large upper and lower 'sidelobes' in the histograms can be seen. These 'sidelobes' are clipping peaks and indicate the saturation of the amplifier and/or the ADC. The sample value does not represent the signal; an overflow took place. These peaks can be found in almost all histograms. They represent the saturation effect and are not unexpected. More surprising is that values in the EDF-file-header (indicated with marker ×) rarely correspond to these saturation values. (b) In some recordings (e.g. P02, S02, S07 and UH3) the saturation values were diffuse rather than sharp. This indicates that the digitized signal was filtered and/or re-sampled to different sampling rate without considering the overflow effects. Alternatively, the amplifier with drifting saturation thresholds caused the clipping.
3. Recordings P01, P02, S02 and S07 display a singular peak with a missing neighboring value close to the sample-value zero. It is obvious that the ADC has a 'strange' behavior with one value in the input range apparently missing. The corresponding value is assigned to the neighboring value.
4. A similar phenomenon with some histogram peak can be found in H09, N01 and is exceptionally large in UH3. In contrast to previous cases, no neighboring values are missing. This phenomenon can be caused by a very busy computer, in which the complete recording system does not fulfill the real-time conditions. In the investigated cases, the ADC was off for some time; e.g. the channel was switched to ground. During that time no data was recorded and the values were set to zero. A detector might be useful to indicate these periods. The remaining data can be used for further analysis.

5. It can be seen that the distributions of the EEG amplitudes deviate from a normal distribution. Partly, artifacts can explain it. It is important to remember that the *y* axis is scaled logarithmically; so the deviation concerns only relatively few samples. Despite that fact, the non-linear behavior can be clearly observed.
6. The value range on the *x*-axis illustrates whether a 16 bit (a better quantization with an e.g. 22 bit-ADC cannot be represented by the EDF data format) or a 12 bit ADC converter was used. Recordings C07, C17, M04 and M09 have a value range of ±32 767 (16 bit), all others have a value range of about ±2048 (12 bit). The standard deviation of the 16 bit data (C07, C17, M04 and M07) is quite small compared to the total value range. Therefore, it can be recommended to increase the amplification gain in these cases. Overflow effects would not be greatly increased, but the signal-to-noise ratio – between EEG power and quantization noise – would be improved.
7. Recordings H02 and H09 show a histogram peak at 4096. This is caused by a bug in the ADC that appears in the last few seconds before terminating the recording. Between 0 and 1000 samples have this value which is less than 10 s.

Measures obtained from the histograms displayed in Fig. 1 are summarized in Table 1. Comparing the entropy with the histogram, it can be said that histograms like those in N01 and N04 produce large entropy values (10.4–10.8 bits), whereas B1, B04 and C07 result in low entropy (7.7–8.0 bits).

Fig. 2 shows that the entropy of the EEG, EOG and ECG channels is in the range of 7.7–11 bits (channel 1 in B1 is a non-typical outlier). Hence, a 16 bit data format, such as EDF, provides a dynamic range from 32 (2^5) to 315 ($2^{8.3}$) times of the standard deviation. The typical entropy values of EMG, respiration (airflow, chest, abdomen) and SaO_2 - channels range from 5 (2) to 10, 5 to 14, and 2 to 5 bits, respectively.

4. Discussion

The noise of an EEG recorder consists of the quantization noise of the ADC and the amplifier noise. For an entropy analysis of the amplifier, the amplifier noise has to be known. To provide a high SNR, both types of noise must be low. In this study only results of the histogram-based entropy were analyzed. The entropy analysis of the amplifier was outlined theoretically.

The phenomena (1) and (3) may be considered unimportant, because the error is only 1 digit. However, in terms of signal-to-noise ratio it implies that the quantization noise is larger (sometimes twice as large), the dynamic range is

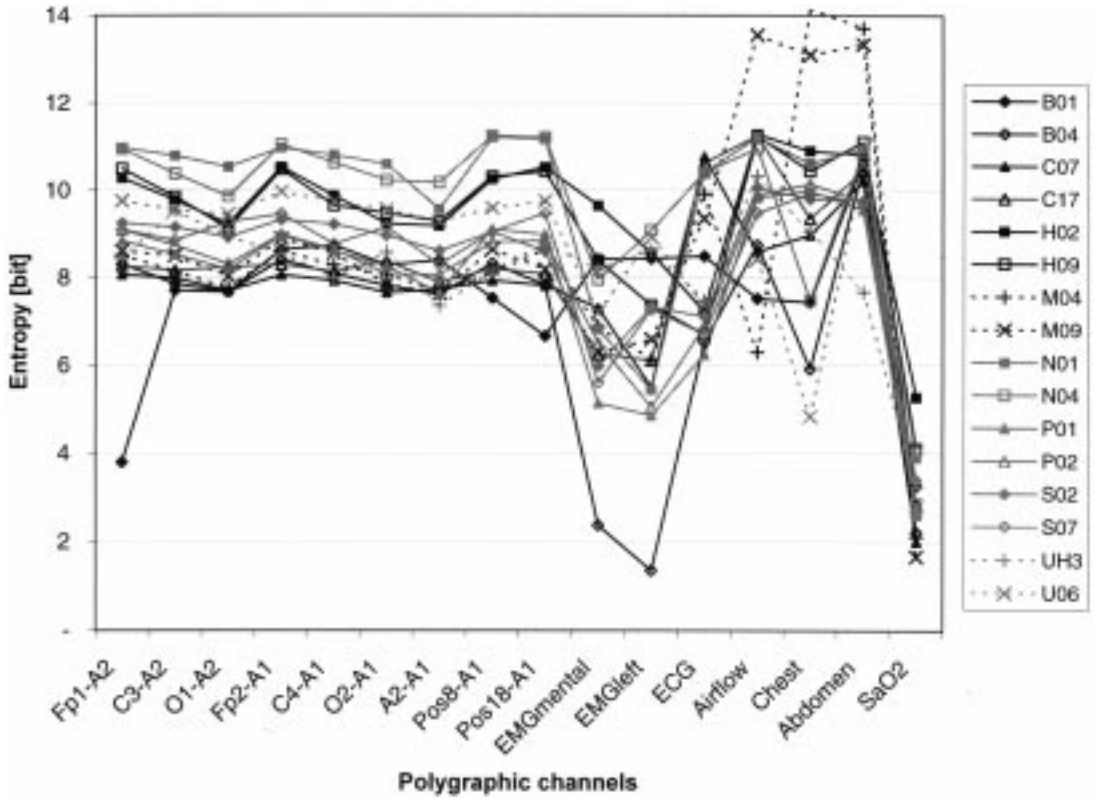

Fig. 2. Entropy values from 6 + 1 EEG, 2 EOG, 2 EMG, 1 ECG, 3 respiratory and 1 oxygen saturation channels of 16 polysomnographic all-night recordings.

reduced (only half) and the entropy is reduced (by 1 bit). To avoid effect (1) it is recommended not to multiply the digitized value but rather, to consider the scaling factors in the header information. The origin of effect (3) might be caused by the difficulties with the two's-complement representation of negative integer values.

Looking at phenomenon (2a) and (2b), it is obvious that an overflow check is important. Any filtering, resampling (changing the sampling rate), or re-referencing (linear combination of EEG channels) which does not consider the saturation effect, yields incorrect results. An overflow check would also detect phenomenon (7). The phenomenon (4) requires a detector if the input is off or at ground. If such events are encoded with a number larger than or equal to the saturation value, this case could also be handled automatically by a simple overflow detector.

One possibility is, also, to store these events in a separate event-marking channel. A method for implementing events in the EDF format has been presented by Van de Velde et al. (1998).

In real-world applications an infinite dynamic range is never available. A tradeoff must always be made between the amplitude range and amplitude resolution. In other words, an overflow can rarely be avoided. On the other hand, the occurrence of an overflow is a technical artifact. The sample value does not represent the true value of the biosignal; a non-linear effect is introduced.

The crucial point in overflow detection is finding the correct upper and lower threshold. Three possibilities of an automatic detection were considered. First, using the header information, but the minimum and maximum in the header is rarely associated with the saturation values. Second, the real minimum and maximum found in the recording are not useful in case of a diffuse overflow. Third, the threshold values could be set to the mean $\pm$ a multiple of the standard deviation, but in this case no consistent factor could be derived. A fourth possibility is choosing the threshold visually, but this has the disadvantage that it can not be automated. None of these methods is completely suitable for a general-purpose method for automated overflow detection. Therefore, it is highly recommended that the providers of recording equipment supply a reliable overflow threshold in the header information. Furthermore, it should be encoded if the input is off or at ground (e.g. as described by phenomenon 7). These simple improvements would provide an important tool for the quality control of the data.

Besides the saturation effect another non-linearity in the recordings was observed (phenomenon 5). The histograms deviate from a Gaussian distribution. It seems to be a non-linear superposition of one or more Gaussian processes. Several reasons might cause this effect. First, it might be caused by technical artifact e.g. the slow decay after an electrode overflow or a non-linearity of the ADC; second, the EEG might actually be really non-linear (Elul, 1969); thirdly and most likely the time-varying behavior of the sleep EEG causes this non-normal distributed histogram.

With the method of histogram analysis, we can, firstly, calculate the mean, variance, skewness, and kurtosis with low computational effort. The skewness and the kurtosis describe the deviation from Gaussianity of the data. That might be of some theoretical importance for the understanding of the underlying brain processes. Secondly, quality control by means of overflow detection can not currently be automated because the threshold (i.e. saturation) values are not available. The histograms can be used to identify the saturation values of the amplifier. Thirdly and finally, the histograms are the basis for calculating the entropy measure. The question is how much information about the brain processes we can gain from observing the EEG. Improving the technology (e.g. 16 bit or even 22 bit instead of 12 bit) may have important consequences for the EEG recording technique. But it is also shown that it is not sufficient to use a 16 bit ADC instead of a 12 bit ADC; one must also use the full dynamic range. Otherwise, the same low entropy measures (circa 8 bit) are obtained. In this way the histogram and entropy analysis is one important tool for the quality control of EEG and other biomedical recordings.

Acknowledgements

This work was funded by the European Commission, DG XII - Project Biomed-2 BMH4-CT97-2040. We would like to thank the Free University of Berlin (W. Herrmann), Tampere University Hospital (J. Hasan), University Clinic of Neurology, Vienna (J. Zeitlhofer), Department of Psychiatry, Medical School University of Vienna (B. Saletu), Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau, Area d'Investigacio Farmacologica, Barcelona (M. Barbanoj), Department of Psychiatry, University of Mainz (M. Grötzinger), Holland Sleep Research, Westeinde Hospital, The Hague, and Klinikum der Philipps-Universität Marburg for providing the data and Britta Ortmayr for proofreading.

Appendix A. Entropy

Given is a time series Y consisting of N elements $[y(1)\ldots y(N)]$. N is the number of available samples, e.g. for an 8 h recording with 200 Hz sampling rate we obtain $N = 8 * 60 * 60 * 200 = 5\,760\,000$ samples. Each element $y(k)$ is digitized, i.e. it is assigned to one value out of $2^{16} = 65\,536$ possible ones. The value range goes from $-32\,768$ to $+32\,787$. A histogram H_Y is a function of i; for each possible value i is $H_Y(i)$ the number of samples of Y, which have the value i. The index Y indicates that H_Y is the histogram of the signal Y. The total number of samples is

$$N_Y = \sum_i (H_Y(i)) \tag{A1}$$

For large N the histogram corresponds to the probability density function of the time series Y. In other words, the

probability p that a certain value of the digitized signal Y has the value i is

$$p_Y(i) = H_Y(i)/N_Y \tag{A2}$$

It is known from coding theory that the entropy of information in binary digits (bits) is

$$I_Y = -\sum_i (P_Y(i))\log_2(P_Y(i))) \tag{A3}$$

Assuming, that all of the 16 bits are significant, the entropy provides the amount of information inherent in each sample of the signal Y. The largest entropy in a discrete system is if the histogram is flat. In case of a 16 bit system, $2^{16} = 65\,536$ values are possible; if the probability $p(i)$ is 1/65 536 for each value i, the entropy is 16 bit. The entropy is smaller if $p(i)$ is not equally distributed.

The entropy of a continuous Gaussian process Y is determined by the variance σ_Y^2 (e denotes Euler's constant 2.718…)

$$I_Y = 0.5\log_2(2\pi e\sigma_Y^2) \tag{A4}$$

The entropy difference ΔI between signal Y and noise N is determined by the SNR

$$\Delta I = 0.5\log_2(\sigma_Y^2 + \sigma_N^2)/\sigma_N^2 = 0.5\log_2(1 + SNR) \tag{A5}$$

For more details see also Shannon and Weaver (1949) and Rieke et al. (1997).

The mean (μ) and variance (σ^2) of signal Y can be obtained from the histogram

$$\mu_Y = E\{Y(t)\} = \sum_i (i\ H_Y(i))/N_Y \tag{A6}$$

$$\sigma_Y^2 = E\left\{(Y(t) - \mu_Y)^2\right\} = \sum_i \left((i - \mu_Y)^2 H_Y(i)\right)/N_Y \tag{A7}$$

The corresponding Gaussian distribution (gd) is

$$gd(x) = N\left(2\pi\sigma^2\right)^{-1/2}\exp\left(-(x - \mu)^2/\left(2\sigma^2\right)\right) \tag{A8}$$

whereby x corresponds to i and ranges from $-\infty < x < \infty$. Furthermore, the skewness $\gamma 3$ and kurtosis $\gamma 4$ of the data series Y are defined (Nikias and Petropulu, 1993)

$$\gamma_Y^3 = \sum_i \left((i - \mu_Y)^3 H_Y(i)\right)/N_Y \tag{A9}$$

$$\gamma_Y^4 = \sum_i \left((i - \mu_Y)^4 H_Y(i)\right)/N_Y - 3\left(\sigma^2\right)^2 \tag{A10}$$

References

Dorffner G. Towards a new standard of modeling sleep based on polysomnograms – the SIESTA project. Proc. ECCN 98, Ljubljana. Electroenceph clin Neurophysiol 1998;106(Suppl. 1001):28.

Elul R. Gaussian behaviour of the electroencephalogram: changes during performance of mental task. Science 1969;164:328–331.

Glass A. Factors influencing changes in the amplitude histogram of the normal EEG during eye opening and mental arithmetic. Electroenceph clin Neurophysiol 1970;28:429–430.

Kemp B, Värri A, Rosa AC, Nielsen KD, Gade J. A simple format for exchange of digitized polygraphic recordings. Electroenceph clin Neurophysiol 1992;82:391–393.

Martin-Rodriguez JG, Buno Jr W, Garcia-Austt E. Human pulvinar units, spontaneous activity and sensory-motor influences. Electroenceph clin Neurophysiol 1982;54:388–398.

Nikias CL, Petropulu AP. Higher-order spectra analysis, Englewood Cliffs, NJ: Prentice Hall, 1993.

Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W. Spikes – exploring the neural code. Cambridge: MIT Press, 1997.

Shannon CE, Weaver W. The mathematical theory of communication. Urbana: University of Illinois Press, 1949.

Van de Velde M, van den Berg-Lenssen MM, van Boxtel GJ, Cluitmans PJ, Kemp B, Gade J, Thomsen CE, Varri A. Digital archival and exchange of events in a simple format for polygraphic recordings with application in event related potential studies. Electroenceph clin Neurophysiol 1998;106:547–551.