Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Observations from Chaotic Analysis of Sleep EEGs

John A. Kassebaum, *Senior Member, IEEE,* Brian H. Foresman, *Member, IEEE*, Thomas M. Talavage, *Member, IEEE*, Russell C. Eberhart, *Fellow, IEEE*

***Abstract* — Chaotic analysis and modeling methods were applied to a small collection of healthy individuals' sleeping EEGs in order to establish whether a common representational space could be discovered for use in comparative analysis. Common challenges were encountered and dealt with in a practical manner. The methods used and the choices made in this effort are described herein. The authors observe and describe a common representation, obtained from a delay coordinate embedding, that should be usable for comparative analysis leading to clinical utility.**

***Keywords* — EEG, Chaos, Sleep, Generalized Dimension, Delay-Coordinate Embedding, Quantization, Chaotic Analysis Limitations**

I. Introduction

Chaotic modeling methods were applied to single channel samples electroencephalogram (EEG) data obtained from PhysioNet [1]. Parts of the Sleep-EDF Database [2], which can be found on the internet at http://www.physionet.org/physiobank/database/sleep-edf/, were used for this analysis. This research is intended to discover clinically and analytically useful representations of sleep EEG data acquired from healthy and disordered individuals. The observations described herein comprise some of the initial effort toward this goal, using the methods of chaotic modeling that are generally used for the reconstruction of low-dimensional attractor dynamics [3, 4]. As can be seen, the generalized dimensionality calculated for four different healthy individuals during their sleep periods was high, but the authors do *not* intend to actually reconstruct the dynamics of the underlying dynamical systems giving rise to the EEG signals. Instead, the lesser goal is only to discover regions of interest in the phase space using these representation vectors. The first challenge along this path must be to show some consistency among individuals in order to support the utilization of a single representation for comparative purposes.

II. Data Sets

The Sleep-EDF data set includes eight European Data Format (EDF) records, one for each of eight individuals [5, 6]. The first four individuals were young and considered healthy, while the second four had mild difficulty falling asleep but were otherwise healthy. Only the first four healthy individuals have been studied here. Each record included seven channels and a hypnogram. The first two channels of each record contained EEG data: channel 1 was recorded from the Fpz-Cz location, while channel 2 was recorded from the Pz-Oz location. The EEG data was pre-filtered to the band 0.5 Hz to 100 Hz and sampled at 100 Hz. (The authors cannot correct for this apparent violation of the Nyquist criterion. Although there could be aliasing of frequencies above 50 Hz in the data sets, since the EEG period was identified as sleep, essentially all of the power in the EEG can be expected to be less than 50 Hz. Thus, the authors believe the power above 50Hz should be insignificant and cause little or no aliasing problems.) All channel data was sampled at 12 bits of precision. The data sets included just less than 24 hours of data, including sleep and ambulatory waking periods. For the four data sets studied, the 'Lights Out' time was recorded in the EDF headers, but no 'Lights On' annotation was available. The following table gives meta-information on each of these data sets:

Table 1
EDF Records - Metadata

Record Name	*Gender*	*Age*	*Date Recorded*
sc4002e0	Female	33	25-APR-1989
sc4012e0	Female	33	30-MAR-1989
sc4102e0	Male	25	18-APR-1989
sc4112e0	Male	26	02-MAY-1989

In these studies, the authors utilized the sleep period only (although, for comparison purposes, an analysis of record 'sc4002e0' was performed on the entire ~24h record). The

Final Manuscript received July 6, 2006. This research was partially supported by Stellar Wind Development, LLC.

J. A. Kassebaum, *Senior Member of the IEEE*, is a PhD student at Purdue University and IUPUI, as well as the President of Stellar Wind Development, LLC, 5761 West 73rd St., Indianapolis, IN, 46278, USA (phone: 317-225-4627; fax: 317-328-9442; e-mail: John.Kassebaum@StellarWindDev.Com).

Brian H. Foresman, D.O., M.S., *Member of the IEEE*, is an Associate Professor of Clinical Medicine and Director of the Sleep Medicine and Circadian Biology Program at the Indiana University Medical Center at 550 N. University Blvd. University Hospital/AOC4903, Indianapolis, IN, 46202, USA (phone: 317-274-2136; fax: 317-274-4224; e-mail: bforesma@iupui.edu).

Thomas M. Talavage, PhD, *Member of the IEEE*, is an Associate Professor of the School of Electrical and Computer Engineering and the Weldon School of Biomedical Engineering at Purdue University at 465 Northwestern Avenue, West Lafayette, Indiana, 47907-2035, USA (phone: 765-494-5475; fax: 765-494-6440; e-mail: tmt@ecn.purdue.edu).

Russell C. Eberhart, PhD, *Fellow of the IEEE*, is a Professor of Electrical and Computer Engineering and an Adjunct Professor of Biomedical Engineering at the Purdue School of Engineering and Technology at IUPUI, 723 West Michigan Street, SL-160, Indianapolis, IN, 46206-5132, USA (phone: 317-274-9721; fax: 317-274-4493; e-mail: reberhar@iupui.edu).

1-4244-0033-3/06/$20.00 ©2006 IEEE.

sleep period was defined as 'Lights Out' until the last record scored as sleep.

III. Algorithms and Analytical Procedures

The traditional method in chaotic modeling is to use a chaotic attractor reconstruction delay-coordinate embedding discovery method to obtain representation vectors for the data records. The method begins with linear redundancy computations on each EEG time series to find an appropriate delay for our embedding, and then one performs a search calculation for the generalized dimensions of the attractor.

The authors used the bit-interleaved box-counting method of Pineda and Sommerer [4] for the generalized dimension calculations. Two variants of the algorithm were used: The first used 64-bit machine precision numbers that limited the embedding dimension search to less than 6 dimensions at full precision (e.g. 5 dimensions or samples x 12 bits per sample = 60 bits < 64 bits). The second algorithm used arbitrary precision mathematics that was much slower and suffered from some additional numerical inefficiency, but which was not limited to a maximum embedding dimension of five at full precision.

Some interesting challenges encountered in our data analysis were: 1) the availability of high numerical precision samples – e.g. 12 bits per sample; and 2) the amount of data required to reliably compute a high-valued generalized dimension estimate. Algorithms such as the box-counting algorithm are affected by over-quantization with respect to the number of samples available. In a sleep period of eight hours sampled at 100 Hz, there are only about $3\text{x}10^6$ samples available. The limitations on the calculation of redundancy given this number of samples is given by Palus in Appendix 1 of [3]. We show the mathematical result here in equation (1) for reference:

$$N > Q^{n+1} \tag{1}$$

In equation (1), N represents the number of samples available, 'n' is the generalized dimension value, and Q is the number of quantization levels used. Thus, since we have approximately $3\text{x}10^6$ samples available, there are significant limitations on the values of Q and 'n' that we must consider in our computation of redundancy measures. In addition, there are limitations to Grassberger-Procaccia style algorithms [7]. The limitation is that calculated dimension value cannot exceed the limits given in equation (2).

$$d_{\max} < 2\log_{10} N \tag{2}$$

The quantity d_{max} is the maximum generalized dimension that can be produced from the algorithm given that there are N samples available in the time series. For our N of $3\text{x}10^6$ we are therefore limited to a dimension of approximately 12.95. A dimension larger than d_{max} may simply indicate the presence of random noise. Both of the above limitations fundamentally affect the analysis of our problem.

IV. Analysis and Observations

A. Initial Analysis Results

We began by exploring the characteristics of the time series itself. Assuming all the records to be somewhat similar, we selected 'sc4002e0' at random. We did not remove the wakeful periods for this initial analysis. We first estimated an embedding delay based on linear redundancy. The first minimum in the linear redundancy was found at approximately 52 points of delay. As a result, we performed our initial search for the generalized dimension using a delay coordinate of 52. Using the box counting algorithm at a number of embedding dimensions and scales, we observed the saturation of the generalized dimensions at a dimension of approximately 23.

One can see that this value for the generalized dimension violates the limitations in equations (2). In addition, we used full (12-bit) precision in our redundancy calculations, which means we violated the limitations in equation (1) as well.

In order to validate this result, we then reviewed the entire set of data records after removing wakeful periods. We then found that the first linear redundancy minimums (the appropriate delay) ranged from 25 to 36 points for the entire collection. We then performed the generalized dimension calculations using these delays and observed essentially identical dimension values (varying by quantization levels) across all data sets. That is, the generalized dimension at full precision (12-bits) ranged from 21.1 to 21.6 over all data sets, again violating the limitations in equations (1) and (2).

The only reasonable choice was to reduce the number of quantization levels being used in each of these calculations since we cannot increase the number of samples available. Fortunately, we can satisfy these two equations *separately* since the actual dimension of the data is dependant on the number of quantization levels utilized in the analysis.

B. Quantization and Redundancy

The problem with using too many quantization levels for redundancy calculations is that the measure will ignore underlying data and become determined primarily by the quantization level as the number of quantization levels increases [3]. Thus, one should avoid over-quantization while performing redundancy calculations. To calculate an appropriate redundancy value we decided to limit our consideration to only 8-bits of precision so that equation (1) would be satisfied. Since there are approximately $3\text{x}10^6$ samples, and since we observed that the generalized dimension value saturates consistently at approximately 4.3 for 8-bits of precision (see Table 3) for our entire sample set of EEGs (e.g. $3\text{x}10^6 > 8^{5.3} = 6.2\ \text{x}10^5$), the appropriate precision for the redundancy calculations should be 8-bits.

Assuming 8-bits of precision, one then uses Takens' theorem [8] to determine an embedding dimension (e.g. 2 x 4.3 + 1 = ~10). Unfortunately, we are limited by our algorithm implementation that uses 64 bit numbers so we chose an embedding dimension of 8 (as close as possible to 10). Again using Takens' theorem [8], we find that at 6-bits of precision the generalized dimension of 3.02 leads to an embedding dimension of 8. Then, using a full non-linear redundancy calculation with an 8 dimensional embedding, we observed some interesting variations in the first minimum of the redundancy calculation for several quantization levels.

Table 2
Observed Redundancy Minima for each Data Set

Record Name	Delay (Q=6)	(Q=7)	(Q=8)
sc4002e0	20	18	6
sc4012e0	29	29	10
sc4102e0	>40	>40	29
sc4112e0	11	11	3

As can be observed, the set of minima has a wide range for this collection. However, they are not too different from our original linear redundancy calculation of 25 to 36 except for the 'sc4112e0' record. In addition, most minima are nearly multiples of each other (10, 20, 30). This is significant because we expect to see reappearances of minima in the redundancy at delays corresponding to multiples of the first minimum. For example: if the first minimum is found at a delay of '3', then we would expect to see evidence of this minimum at or near delays that are approximately multiples of '3'. One should realize that an observed minimum does not precisely specify the choice of delay because the data (an EEG time series) is sampled rather than continuous in nature. One looks for a delay choice for the model that will maximize the amount of information available from the underlying attractor. The technique accomplishes this by minimizing the shared information between collections of points from the time series at different delays. Of course, the mutual information could be minimized by choosing very long delays, but this approach leaves us with states having less specific time locality (state vectors whose components are spread over longer periods of time). The shortest delay available near a minimum in the mutual information should work best. Thus, since we have only minimal support for the short delay of '3', a better choice over this collection of data sets might be a delay near 10.

C. *Quantization and Generalized Dimension*

Once we establish a delay coordinate, we must also find an estimate for the dimension of the attractor. Using our earlier observations, we observed the saturation of the generalized dimensions at a variety of values dependent on the quantization levels as may be seen in Table 3.

Using these values, and the limits from equation (2), one may choose to ignore the high dimension values shown for full precision in favor of the lower dimension values for lesser quantization levels. Since there are approximately $3x10^6$ samples, one should choose a quantization level that gives a generalized dimension less than the limit of 12.95 given by equation (2). In effect, one is then considering the information in the extra bits of precision to be inconsequential. This seems reasonable since we do not have sufficient samples in each data set to justify a high dimensional value near 21.5, which would be required using 12-bits of precision.

Table 3
Generalized Dimension vs. Quantization Level

Quantization (bits)	Gen. Dimension Value
6	2.69 – 3.02
7	3.08 – 3.60
8	3.59 – 4.33
9	5.28 – 5.40
10	7.04 – 7.20
11	10.6 – 10.8
12	21.1 – 21.6

Because of these observations and the limitation of equation (2), we chose to use the dimension range of 10.6 to 10.8 for 11-bits of precision for our embedding for the entire set of data.

D. *Delay-Coordinate Embedding*

To obtain a system model for the EEGs under study, we apply Takens' theorem [8] to our selected generalized dimension of approximately 10.8 to obtain an embedding dimension of 23 (e.g. 2 x 10.8 + 1 = ~23). We also use the delay (delay = 10) chosen in section B.

V. Discussion

It is significant that the essential character of the delay-coordinate embeddings found for four distinct individuals is so similar. Our aim is to establish a single embedding that might be useable for comparative analysis between individuals in order to discover new approaches with clinical utility. The biggest difficulty we see is in the variety of delays that may be appropriate for the comparison. That the dimension of the embedding is identical over all four individuals is very helpful to this endeavor. What remains to do is to carry out the analysis described herein on EEGs of sleep-disordered individuals and to explore comparisons, in the space described by this common embedding, among healthy and sleep-disordered individuals.

There is an assumption that underlies the theoretical foundation of delay-coordinate embedding and our use of it for EEG analysis. Delay-coordinate embedding via Takens' theorem [8] is based on the evolution of a solution to a single dynamical system (a set of differential equations). The brain, however, is actually constructed from cooperating and competing systems overlain on top of one another, which when operating independently function physiologically as distinct dynamical systems. In particular, it is known that each stage of sleep utilizes specific and different physiological brain structures to perform their unique functions during sleep [9]. The EEG measured from this composite system is a combination of the components' distinctions and interactions. It is possible and reasonable that these separate dynamical systems may occlude each other in our measured EEG and prevent our use of a single comparative model via delay-coordinate embedding. This is likely one reason why delay-coordinate embedding has not been particularly successful in brain state prediction. The authors have avoided this pitfall by declining to do explicit reconstruction of the dynamics of the composite system. Instead, the authors intent is only to generate a model representation to be used for comparatives analysis over multiple individuals. The observations made in this paper show that such a space is readily obtainable since the parameters of the model for multiple individuals were observed to be similar. What remains to discover is whether this kind of model, which surely contains information related to specific physiological contexts (the sleep stages), can be interpreted in a clinically useful manner. The results of that research will be discussed in future publications.

Acknowledgment

John A. Kassebaum would like to thank Dr. Leonid Rubchinsky and Dr. Robert Worth of the Department of Mathematics at IUPUI for discussions of and mathematical feedback in regard to this effort.

References

[1] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet : Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, pp. 215e-220, 2000.

[2] B. Kemp, "The Sleep-EDF Database: Sleep Recordings and Hypnograms in European Data Format (EDF)," PhysioNet, MIT Room E25-505A, 77 Massachusetts Avenue, Cambridge, MA 02139 USA, 2002.

[3] M. Palus, "Identifying and Quantifying Chaos by Using Information-Theoretic Functionals," in *Time Series Prediction: Forecasting the Future and Understanding the Past*, vol. XV, *NATO Advanced Research Workshop on Comparative Time Series Analysis*, A. S. Weigend and N. A. Gershenfeld, Eds. Santa Fe, New Mexico: Addison-Wesley Publishing Co., 1993, pp. 387-413.

[4] F. J. Pineda and J. C. Sommerer, "Estimating Generalized Dimensions and Choosing Time Delays: A Fast Algorithm," in *Time Series Prediction: Forecasting the Future and Understanding the Past*, vol. XV, *NATO Advanced Research Workshop on Comparative Time Series Analysis*, A. S. Weigend and N. A. Gershenfeld, Eds. Santa Fe, New Mexico: Addison-Wesley Publishing Co., 1994, pp. 367-385.

[5] M. S. Mourtazaev, B. Kemp, A. H. Zwinderman, and H. A. C. Kamphuisen, "Age and gender affect different characteristics of slow waves in the sleep EEG," *Sleep*, vol. 18, pp. 557-564, 1995.

[6] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *Biomedical Engineering, IEEE Transactions on*, vol. 47, pp. 1185-1194, 2000.

[7] J.-P. Eckmann and D. Ruelle, "Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems," *Physica D*, vol. 56, pp. 185-187, 1992.

[8] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence*, vol. 898, *Lecture Notes in Mathematics*, D. A. Rand and L. S. Young, Eds. Berlin: Springer, 1980, pp. 366-381.

[9] C. Kaufmann, R. Wehrle, T. C. Wetter, F. Holsboer, D. P. Auer, T. Pollmacher, and M. Czisch, "Brain activation and hypothalamic functional connectivity during human non-rapid eye movement sleep: an EEG/fMRI study," *Brain*, vol. 129, pp. 655-667, 2006.