WEANING FROM MECHANICAL VENTILATION: FEATURE EXTRACTION FROM A STATISTICAL SIGNAL PROCESSING VIEWPOINT

Pablo Casaseca-de-la-Higuera, Rodrigo de-Luis-García,
Federico Simmross-Wattenberg and Carlos Alberola-López

Laboratorio de Procesado de Imagen. ETSI Telecomunicación. Universidad de Valladolid
Campus Miguel Delibes. Camino del Cementerio s/n, 47011, Valladolid, Spain
phone: + (34) 983423660 (ext. 5590), fax: + (34) 983423667.
emails: casaseca@lpi.tel.uva.es, rodlui@tel.uva.es, fedesim@lpi.tel.uva.es, caralb@tel.uva.es
web: www.lpi.tel.uva.es

ABSTRACT

Clinicians' decision for mechanical aid discontinuation is a challenging task that involves a complete knowledge of a great number of clinical parameters, as well as its evolution in time. Respiratory pattern variability appears as a useful extubation readiness indicator, and thus can be used as an informative feature in a statistical pattern recognition framework. Reliable assessment of this variability involves a set of signal processing techniques that should be carefully evaluated for statistical validity. This paper evaluates different variability extraction techniques aimed to build a Bayesian classifier for *weaning* readiness decision. As a conclusion, Sample Entropy is selected as the best performance extraction method. By calculating it over tidal volume signals, and with mean respiratory rates as additional input patterns, a 2D Bayesian classifier is constructed with principal component analysis selection. The obtained misclassification probability ($P_e = 0.2141$) is acceptable if compared with performance of single feature classifiers.

1. INTRODUCTION

Discontinuation of mechanical ventilation in an Intensive Care Unit (ICU) setup is a process that should not either be delayed unnecessarily or undertaken too early. It is usually done by the gradual removal of the mechanical support as spontaneous breathing is resumed [1], and is referred to as *weaning*.

Weaning readiness assessment may be performed by analysis of several physiological parameters holding information of a potential successful or failed weaning trial. Depending on the variable of interest, the analysis may be performed either by trend evolution observation, punctual pulmonary tests, or breathing pattern variability quantification. We will hereafter refer to these parameters as weaning indices or predictors.

Within this framework, the so-called weaning problem can be formulated from a statistical pattern recognition point of view. Features accounting for pattern variability and predictors' evolution may be extracted from physiological signals, and a posterior feature selection and classification in successful (SW) and failed (FW) weaned groups can be carried out.

Several studies have been developed on classical evolution-based descriptors. Most of them are summarized in a number of review papers such as [2]. Among the physiological variables of interest, both respiratory frequency and its ratio to tidal volume stand out with lower values for successfully weaned patients [3]. Regarding breathing pattern variability, more regular tidal volume series have been found for successfully weaned patients [4]. The same conclusion was reached in [5], both for tidal volume and respiratory peak flow.

The main purpose of the aforementioned studies was the identification of the weaning predictors. This was made by the performance of statistical tests accounting for significant differences among the groups of patients who could maintain spontaneous breathing indefinitely and those unable to do so. Thus, a classifying objective as itself was not established.

In this paper, we propose a simple two-dimensional Bayesian classifier as a clinical aid for weaning readiness assessment. To this end, feature extraction and selection is performed in order to get the best weaning predictors from real breathing patterns. We have payed special attention to pattern variability assessment issues from a signal processing point of view since we believe previous approaces (such as those employed in [4] and [5]) are somehow objectionable. To this end, both classical techniques and nonlinear methods related to chaos theory have been evaluated. The choice of the final pattern variability quantifier was made by prior evaluation of its suitability for the special signals we have dealt with, and its capability of discerning among both groups of patients. A recently developed statistic, Sample Entropy (SampEn) [6] was finally selected as the best candidate for pattern complexity assessment.

The paper is organized as follows: Section 2 presents Sample Entropy and evaluates its suitability as a regularity quantification method for the time series to be studied. Section 3 presents the feature selection issues among the whole set of explored weaning predictors. The Bayesian classifier performance on the selected features is also evaluated in this section. Finally, Section 4 closes the article with the conclusions obtained from the developed work.

2. ON FEATURE EXTRACTION: SAMPLE ENTROPY AS A BREATHING PATTERN VARIABILITY QUANTIFIER

The study was carried out on breathing pattern signals obtained from 40 intubated patients who underwent a spontaneous 30 minutes weaning trial once practitioners decided readiness for extubation had been achieved. A higher variability of tidal volume patterns in patients who had failed such a trial was expected as outcome, as well as a higher mean respiratory frequency value. The total number of successfully weaned patients (those who passed the trial and did not need reintubation) was 20. The remaining 20 did not pass the trial and needed reintubation within 48 hours after the experiment.

With such a clinical setup, 500 samples of inspiratory (VTI) and expiratory (VTE) tidal volume patterns were obtained for variability assessment. Additionally, respiratory rate (RR) patterns of the same number of points were also recorded for trend evolution analysis. We will focus on complexity assessment since trend evolution was just analyzed from the mean value of the RR time series.

The authors acknowledge the Comisión Interministerial de Ciencia y Tecnología, Spain, for research grants TIC2001-3808-C02-02 and TEC2004-06647-C03-01. Acknowledgments are extended to the European Commission for the funding associated to the Network of Excellence SIMILAR (FP6-507609). Additionally, the work was partially supported by grant number PA01/03 from the Junta de Castilla y León. Special thanks are also given to Mrs. Ana Domínguez Berrot for providing the experimental data and for her technical support in clinical issues.

2.1 Descriptive Statistics

One of the most widely used statistics in biomedical applications is the coefficient of variation (CV), obtained as the ratio of the standard deviation (std) to the mean value of the time series. In [5] it appeared as a useful complexity quantifier and relevant differences between both groups of patients were obtained. Nevertheless, though it is a simple, easy to interpret (in physiological terms) and accessible method, it presents insensitivity to the orderliness of the data, and thus, does not yield information on the system's inherent dynamics. An assumption of static time series is taken, and any variation in measurement is considered to be random sampling error around a "true" mean. We checked out its suitability by extracting the coefficient of variation from each VT pattern and performing a Student's t-Test [7] to look for significant differences among groups. Mean and standard deviation values for each group are summarized in Table 1, jointly with the *p*-value obtained from the statistic. The conclusion was that no significant differences ($p < 0.05$) were obtained by means of the CV. Indeed, a higher mean value for the SW (see table caption) group was obtained from the test, which was opposite to the expected.

2.2 Testing the Degree of Stationarity

Since a static time series can not be assumed, it seems natural to think that a less stationary behavior may be expected for breathing patterns extracted from patients that are not successfully weaned. So, testing for stationarity might be a useful task for detecting differences between groups. In this context, stationarity must be understood as maintenance of the statistical properties invariant along time. There exists a number of stationarity tests for time series in the literature, both parametric and non-parametric. Those in the first group, assume a model for the time series to be analyzed and check if the model is followed along time, whilst the non-parametric approaches do not take any basic assumptions about the nature of the time series.

One of the most widely used parametric test is the KPSS [8], which considers the time series as the sum of a deterministic trend, a random walk, and a stationary error. Both the hypothesis of trend-stationarity (i.e., the signal follows a linear trend, with some random noise), and level-stationarity (random variance around a mean level) are tested. A statistic is set to check for these hypothesis, and higher values of the statistic account for an increasing degree of non-stationarity.

The non-parametric group includes the so-called *runs test* (RT) [9], which looks for the number of *runs*[1] in a time series and compares it to the distribution of runs which should appear if the process followed a random stationary pattern. Once again a higher value of the statistic used for comparison is an index of a higher degree of non-stationarity.

Our aim is not to describe both tests, but to test whether a higher non-stationary behavior can be found for the tidal volume patterns in the FW group. Thus, we have performed both tests over each of the tidal patterns, and have taken the value of the statistic as a quantifier of the degree of nonstationarity. Only level-nonstationarity has been checked with the KPSS, which it is indeed the hypothesis tested by the RT. The Student's t-Tests were repeated in order to check for differences between both groups. The obtained values are also summarized in Table 1. No significant differences between groups were found except for the RT performed over VTE patterns. The result of the test was opposite to the expected, i.e., more regular patterns appear in the FW (see table caption) group in terms of non-stationarity. But these results should be carefully considered. The RT checks whether the runs distribution in the time series is similar to the expected distribution for a random i.i.d process. However, the signals we are dealing with, can be modeled as auto-regressive 1^{st} order –AR(1)– processes [10] with some degree of correlation, and thus, this assumption can not be made.

[1] A run can be defined as a succession of one or more identical symbols, which are followed and preceded by a different symbol or no symbol at all. In our context, we assign symbol "1" if a value in the time series is above the median, and "0" otherwise.

We have seen how simple regularity quantifying methods appear as not suitable for the particular kind of signals we are working with. Thus, it is time to look for a non-linear statistic that can account for subtle variability behavior not detected with the methods evaluated above. The first potential candidate is the approximate entropy (ApEn), which was the one used in [4]. But we shall experimentally show, that it is not suitable for AR(1) models. Instead, sample entropy (SampEn) may be a more appropriate candidate.

2.3 SampEn as a regularity quantifier

Sample Entropy was introduced as a quantification of regularity in data. Regular signals are expected to have low SampEn, while complexity is related to higher SampEn values. Additionally, SampEn may also be used to determine the occurrence probabilities of specific patterns in a time series [6].

For the calculation of SampEn we first take the original time series $x[i], i = 1,...,N$, and construct vector sequences of size m, $\mathbf{u}[1]$ through $\mathbf{u}[N-m+1]$, defined by $\mathbf{u}[i] = \{x[i],...,x[i+m-1]\}$. The vectors length m, is known as the embedded dimension. The constructed vectors represent m consecutive x values commencing with the ith point.

The distance $d(\mathbf{u}[i],\mathbf{u}[j])$ between vectors $\mathbf{u}[i]$ and $\mathbf{u}[j]$ is defined as $d(\mathbf{u}[i],\mathbf{u}[j]) = max(|\mathbf{u}[i;k]-\mathbf{u}[j;k]|, 0 \le k \le m-1)$ where k accounts for the vector component index. The probability of finding another vector within distance r from the template vector $\mathbf{u}[i]$ is estimated by

$$C_i^{\prime m}(r) = \frac{1}{N-m}\{\text{number of } j, j \neq i, \\ 1 \le j \le N-m+1, \text{such that } d(\mathbf{u}[i],\mathbf{u}[j]) \le r\} \quad (1)$$

SampEn(m,r,N) is the negative natural logarithm of the conditional probability that two sequences similar for m points remain similar at the next point. In order to estimate this, the previous expression must be averaged over i

$$\Phi^{\prime m}(r) = (N-m+1)^{-1}\sum_{i=1}^{N-m+1} C_i^{\prime m}(r) \quad (2)$$

and finally

$$\text{SampEn}(m,r,N) = -\ln[\Phi^{\prime m}(r) - \Phi^{\prime m+1}(r)] \quad (3)$$

M. Engoren used ApEn as a weaning predictor in [4]. The calculation of ApEn is very similar to that of SampEn and can be found elsewhere (as an example, the reader may refer to [11]). The main differences stand on that ApEn does count for self-matches in the probability estimation, and this implies a bias to lower values of the statistic, if it is compared with its theoretical counterpart ApEn(m,r). Additionally, the average in Eq. 2 is performed over natural logarithm values of $C_i^{\prime m}(r)$. In contrast, SampEn does not count self-matches, and has been demonstrated to present relative consistency in some cases where ApEn does not [6]. Additionally, computational load is also reduced with SampEn.

Both ApEn and SampEn have a strong dependence on three parameters, the embedded dimension m, the distance filtering parameter r, and the length of the time series N. With a minimum value of $N = 1000$, and an embedded dimension of $m = 2$ ApEn values have been reported to produce good statistical validity for r in a range from 0.1 to 0.25 of the signal standard deviation. This has been tested both theoretically and in several clinical applications (see [11] and references therein). This reason justifies the use of ApEn in [4]. However, we will show that this is not true for AR models, but first, we will pay attention to the m and r parameters selection.

VTI	CV	KPSS statistic	RT statistic	**SampEn**$(2,20\%\sigma,500)$	**ApEn**$(2,20\%\sigma,500)$
SW	0.228 ± 0.097	0.64 ± 0.70	8.58 ± 3.07	1.62 ± 0.29	1.20 ± 0.10
FW	0.195 ± 0.076	0.43 ± 0.30	7.13 ± 2.77	1.98 ± 0.37	1.16 ± 0.15
p	0.23	0.25	0.13	*0.0016*	0.38
VTE	**CV**	**KPSS statistic**	**RT statistic**	**SampEn**$(2,20\%\sigma,500)$	**ApEn**$(2,20\%\sigma,500)$
SW	0.228 ± 0.100	0.51 ± 0.49	9.45 ± 2.56	1.57 ± 0.29	1.18 ± 0.13
FW	0.195 ± 0.079	0.41 ± 0.31	7.64 ± 2.20	1.93 ± 0.29	1.19 ± 0.14
p	0.21	0.47	0.021	*0.0004*	0.77

Table 1: Student's t-Test results (mean $\pm$ std) on different pattern variability quantifiers. Significant differences between groups are considered for $p < 0.05$. SW and FW stand for Successful Weaning and Failed Weaning respectively.

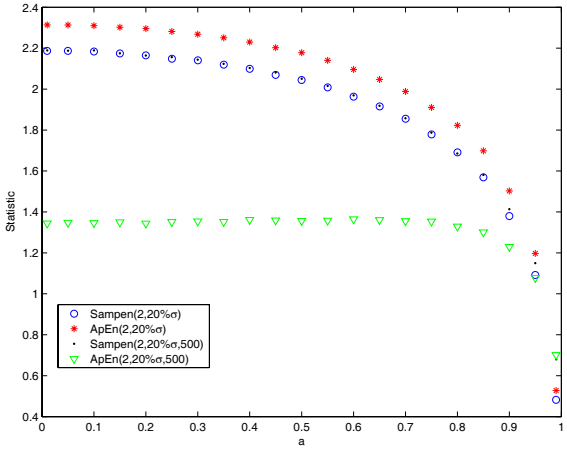

Figure 1: SampEn$(2,20\%\sigma)$ and ApEn$(2,20\%\sigma)$ values compared to mean values from 100 estimates of SampEn$(2,20\%\sigma,500)$ and ApEn$(2,20\%\sigma,500)$for AR(1) models with a values ranging from 0.01 to 0.99.

The embedded dimension m should be as high as possible, but the following condition has to be met: $10^m \leq N_{min} \leq 30^m$. Since we have limitations on the time series length (N=500), the maximal available value for m is 2. For r value selection, we have proceeded as in [12]. A r selection of 20% the std of the signal was obtained as the better trade off between an acceptable number of matches (low variance of the estimator), and an excessive higher probability estimation (low discrimination capability).

Once a proper selection of the statistic parameters is achieved, the theoretical values of ApEn$(2,20\%\sigma)$ and SampEn$(2,20\%\sigma)$ can be calculated and compared with the estimated values for $N = 500$. Figure 1 presents the mean values obtained from 100 simulations of both ApEn$(2,20\%\sigma,500)$ and SampEn$(2,20\%\sigma,500)$ for an AR(1) model obtained from the output of the IIR filter $H(z) = \frac{1}{1-az}$ with a zero-mean, unitary power white Gaussian noise sequence $w[n]$ as input. The statistics have been obtained for different values of the filter coefficient a. Taking into account that a more correlated signal is obtained as a grows from 0 to 1, both SampEn and ApEn values should decrease with increasing a, and this is not true for ApEn due to the bias to low values. This can be easily observed by just looking at the theoretical values, also sketched in Fig. 1. For their calculation, we can start from the basis in [13] and after some algebra, we arrive at

$$\text{SampEn}(m,r) = -\ln\left(\int_{\mathbb{R}^{m+1}} \frac{P(B_{m+1})}{P(B_m)} f_{\mathbf{u}^{m+1}}\left(\mathbf{u}^{m+1}[i]\right) d\mathbf{u}^{m+1}[i]\right) \quad (4)$$

and

$$\text{ApEn}(m,r) = -\int_{\mathbb{R}^{m+1}} \ln\left(\frac{P(B_{m+1})}{P(B_m)}\right) f_{\mathbf{u}^{m+1}}\left(\mathbf{u}^{m+1}[i]\right) d\mathbf{u}^{m+1}[i] \quad (5)$$

with

$$P(B_m) = \int_{\mathbf{u}^m[i]-r}^{\mathbf{u}^m[i]+r} f_{\mathbf{u}^m}\left(\mathbf{u}^m[j]\right) d\mathbf{u}^m[j] \quad (6)$$

where $f_{\mathbf{u}}\left(\mathbf{u}^m[j]\right)$ is the joint m-dimensional probability density function (pdf) of $\mathbf{u}^m[j]$. Equivalently, evaluation of $P(B_{m+1})$ can be performed just substituting m by $m+1$ in Eq. 6. For the AR process we are dealing with, $f_{\mathbf{u}}\left(\mathbf{u}^m[j]\right)$ is a m-dimensional Gaussian distribution with zero mean and a Toeplitz covariance matrix whose elements are defined as $\{\overline{\overline{\mathbf{C}}}_{\mathbf{u}^m}\}_{i,j} = \frac{\sigma^2}{1-a^2} a^{|i-j|}$, $1 \leq i,j \leq m$.

Coming back to our respiratory signals, the Student's t-Tests were performed on ApEn and SampEn values calculated over the VTI and VTE patterns. Results from the tests are also summarized in Table 1. No significant differences were found for ApEn values. Indeed, higher complexity values were obtained for the SW group when evaluating VTI. This can be justified just by looking at Fig. 1. Due to the bias, low ApEn estimates may be obtained from very irregular signals. On the other hand, results with SampEn were as expected from the clinicians, obtaining more irregular patterns in the FW group (higher SampEn mean value with statistical significance). Moreover, the p-values obtained from both tests were quite low (both below 0.01).

3. A DISCUSSION ON FEATURE SELECTION AND CLASSIFICATION PERFORMANCE

The evaluation of descriptors presented on Section 2 may serve as a guideline for a first feature selection approach in order to get the best affordable Bayesian classifier [14]. Thus, it seems logical to select the best meaningful descriptors provided they follow the clinical expectations. This is the same as the selection of SampEn statistics obtained over the VTI and VTE patterns. Additionally, a classical evolution parameter, the RR may serve as a useful feature. Information from it can be extracted by just averaging the RRs time series. The evaluation of the Student's t-Test over the mean RR data resulted in a higher averaged respiratory rate for the failed weaned group with a significance level $p < 0.0073$, 24.72 ± 4.05 breaths/min for the successfully weaned group and 28.78 ± 4.95 breaths/min for the failed (mean $\pm$ std).

Thus, starting from the SampEn$(2,20\%\sigma,500)$ of VTI and VTE, and the mean value of the RR, we can evaluate 1D, 2D and 3D Bayesian classifiers assuming Gaussian distributions for each class. Since we do not have many patients to establish representative training and test sets, we train our classifier with the 40 samples, and then perform a leave-one-out test [14] for performance evaluation. Performance has been evaluated in terms of probability of misclassification (P_e), i.e., the ratio of misclassified patterns to the total number of them. Table 2 presents the average P_e values from the leave-one-out test for all the possible combinations of selected features.

Feature 1	Feature 2	Feature 3	P_e
SampEn(VTE)			0.2564
SampEn(VTI)			0.2750
Mean(RR)			0.3955
SampEn(VTE)	SampEn(VTI)		0.2282
SampEn(VTI)	Mean(RR)		0.2705
Mean(RR)	SampEn(VTE)		0.2250
SampEn(VTE)	SampEn(VTI)	Mean(RR)	0.2276
Principal Component Analysis: 3D→2D			**P_e=0.2141**

Table 2: P_e values obtained with 1D, 2D, and 3D Bayesian classifier. SampEn values were obtained for $m = 2$ and $r = 20\%\sigma$ over 500 samples. Performance from PCA is also presented.

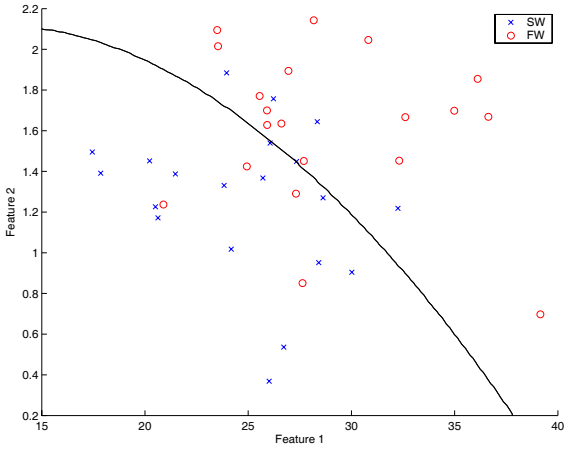

Figure 2: Bayesian 2D classification boundary on the whole data set after PCA.

As shown in the table, SampEn(VTE) is the feature that minimizes the P_e as itself. Nevertheless, an improvement if combined with some of the other two can be obtained. Though Mean(RR) does not provide a good performance in a 1D classification problem, the best results are obtained when joined to SampEn(VTE). Indeed, the P_e values obtained from the 3D classifier are higher than its 2D counterpart when these two features are combined. A further improvement is obtained if principal component analysis (PCA) [14] is applied to the 3D problem in order to get a 2D classifier with better performance. The P_e value after PCA is also presented in Table 2. An improvement has been achieved if compared with those results in [3] (Table 3). This work obtained an overall performance of $P_e = 0.2535$ after applying a discriminant analysis over 12 features extracted from patients with diseases of different etiologies. Since our patients suffered from heterogeneous diseases, better performance results in [3] obtained for specific disease groups ($P_e = 0.0606$ for patients suffering from Chronic Obstructive Pulmonary Disease –COPD–) may not be compared to ours. The final selection ended up with 6 different features. The only one accounting for pattern variability among them was the CV.

4. CONCLUSIONS

In this paper, we have formulated the so-called weaning problem under a statistical pattern recognition framework. Regarding feature extraction, we have evaluated different variability quantification techniques in order to get as much information as possible on weaning readiness. Once meaningful features have been identified, a selection process and Bayesian classification has been performed, with results presented in terms of probability of misclassification.

It is important to take into account that the population under study is not high due to clinical constraints, and this issue poses limits on the classifier performance. Moreover, clinical issues also impose airflow signals not longer than 500 points. Of course, if longer signals were analyzed, more reliable complexity quantification would be obtained. Nevertheless, we have taken into account this issue and selected the best method for our particular problem.

Further lines of research are to be open while additional patients are included in the study. We will explore additional variability assessment techniques such as extraction of spectral features, and try to find the best fitting model to our particular signals. Once a model is fitted, suitability of the different techniques can be checked by means of theoretical developments as we have shown in Section 2.3. Additionally, we believe that multimodal approaches and, in particular, the use of EEG together with respiratory signals would include additional information on the weaning problem resolution, opening up new processing frontiers in this field. This should be, in our opinion, explored.

REFERENCES

[1] B. Baumeister, M. F. El-Khatib, P. G. Smith, and J. L. Blumer, "Evaluation of predictors of weaning from mechanical ventilation in pediatric patients," *Pediatr. Pulmonol.*, vol. 24, pp. 344–352, 1997.

[2] T. Vassilakopoulos, C. Roussos, and Z. Spyros, "Weaning from mechanical ventilation," *J. Crit. Care*, vol. 14, no. 1, pp. 39–62, 1999.

[3] I. Vallverdú, N. Calaf, M. Subirana, A. Net, S. Benito, and J. Mancebo, "Clinical characteristics, respiratory functional parameters, and outcome of a two-hour T-piece trial in patients weaning from mechanical ventilation," *Am. J. Respir. Crit. Care Med.*, vol. 158, pp. 1855–1862, 1998.

[4] M. Engoren, "Approximate entropy of respiratory rate and tidal volume during weaning from mechanical ventilation," *Crit. Care Med.*, vol. 26, pp. 1817–1823, 1998.

[5] M. El-Khatib, G. Jamaleddine, R. Soubra, and M. Muallem, "Pattern of spontaneous breathing: potential marker for weaning outcome," *Intensive Care Med.*, vol. 27, pp. 52–58, 2001.

[6] J. S. Richmann and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Am. J. Physiol. Heart Circ. Physiol.*, vol. 278, pp. 2039–2049, 2000.

[7] B. Rosner, *Fundamentals of Biostatistics*. Pacific Grove, CA, USA: Duxbury Thomson Learning, 2000.

[8] D. Kwiatkowski, P. C. B. Philips, P. Schdmit, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root," *Journal of Econometrics*, vol. 54, pp. 159–178, 1992.

[9] J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures*. New York: Wiley & Sons, 1986.

[10] M. Modarreszadh, N. Bruce, and B. Gothe, "Nonrandom variability in respiratory cycle parameters of humans during stage 2 sleep," *J. Appl. Physiol.*, vol. 69, pp. 630–639, 1990.

[11] S. M. Pincus and A. L. Goldberger, "Physiological time-series: what does regularity quantify?" *Am. J. Physiol. Heart Circ. Physiol.*, vol. 266, pp. 1643–1645, 1994.

[12] D. E. Lake, J. S. Richman, M. P. Griffin, and J. R. Moorman, "Sample entropy analysis of neonatal heart rate variability," *Am. J. Physiol. Regul. Integr. Comp. Physiol.*, vol. 283, pp. 789–797, 2002.

[13] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proc. Natl. Acad. Sci. USA*, vol. 88, pp. 2297–2301, 1991.

[14] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern. Anal. Machine Intelligence*, vol. 22, no. 1, pp. 4–37, January 2000.