33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Automated Unsupervised Respiratory Event Analysis

Carlos A. Robles-Rubio, *Student Member, IEEE*, Karen A. Brown, and Robert E. Kearney, *Fellow, IEEE*

***Abstract*—We recently presented a comprehensive automated off-line method for supervised respiratory event classification from uncalibrated respiratory inductive plethysmography signals. This method required training with a sample of clinical measurements classified by an expert. This human intervention is labor intensive and involves subjective judgments that may introduce bias to the automated classification. To address this we developed a novel method for unsupervised respiratory event classification, named AUREA (Automated Unsupervised Respiratory Event Analysis). This paper describes the algorithm underlying AUREA and demonstrates its successful application to respiratory signals acquired from infants in the postoperative recovery room. The advantages of AUREA are: first, it provides real-time classification of respiratory events; second, it requires no human intervention; and lastly, it has substantially better performance than the supervised method.**

I. Introduction

INFANTS are at increased risk of life threatening apnea following surgery/anesthesia [1]. These Postoperative Apnea (POA) events are rare and so long data records are required to study their relation to other respiratory events. The acquisition of such large data sets has been feasible only rarely because of the need for manual scoring [2], which is labor intensive, expensive, and suffers from low inter-scorer agreement [3].

In [4] we presented a comprehensive, automated, off-line method for supervised respiratory event classification from Respiratory Inductive Plethysmography (RIP) signals. This method provides a reliable and repeatable means of analyzing the large respiratory data records required for the study of POA. It uses several test statistics to classify the respiratory state into one of four categories: Pause, Movement Artifact, Asynchrony, and Quiet Breathing.

The main disadvantage of this method is that it requires a sample of measurements classified by an expert observer to determine the optimum thresholds for the test statistics. This is labor intensive and time consuming since it requires the manual classification of respiratory events from several subjects. Moreover, it involves the subjective judgment of the expert and the low intra-scorer reliability will reduce its accuracy. For these reasons, we have developed an unsupervised event classification method that eliminates the need for human intervention.

The paper is organized as follows: Section II describes the acquisition of infant data, the test statistics used for RIP analysis, and the "gold" standard used to evaluate performance; Section III describes the supervised and unsupervised automated methods developed for respiratory event classification; Section IV reports the performance of these methods when applied to the infant data; and Section V provides concluding remarks.

II. Infant Data Acquisition and Analysis

A. Subjects and Data Acquisition

We acquired data from 16 infants (12 males), postconceptional age 42.8±2.1 weeks, weight 3.7±1.0 kg, in the postoperative period after elective herniorrhaphy with general anesthesia. Written informed parental consent was obtained and the study was approved by the Institutional Ethics Review Board.

Upon arrival at the postanesthesia care unit, infant respibands (Ambulatory Monitoring Inc., Inductobands, Ardsley, NY, USA) were placed around the ribcage and abdomen and interfaced with Respiratory Inductance Plethysmograph (Ambulatory Monitoring Inc., Battery Operated Inductotrace, Ardsley, NY, USA). An infant oximeter probe (Nonin 8600 Portable Digital Pulse Oximeter, Plymouth, MN, USA) was taped to a hand or foot. The analogue outputs were low-pass filtered (cut-off frequency 10 Hz) with an 8-pole Bessel anti-aliasing filter (Kemo, Jacksonville, FL, USA) digitized and sampled at 50 Hz (all respiratory-related information from RIP and all information from the oximeter was observed at frequencies below 5 Hz [5]). The signals were recorded on a computer using MATLAB™ (The MathWorks, Inc., Natick, MA, USA) for off-line analysis. Data were recorded for 6 to 12 hours in accordance with the Montreal Children's Hospital practice guidelines for apnea monitoring in term and former preterm infants. No attempt was made to calibrate the RIP signals in absolute terms.

Manuscript received April 14, 2011; revised June 20, 2011. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. The work of C. A. Robles-Rubio was supported in part by the Mexican National Council for Science and Technology.

C. A. Robles-Rubio is with the Department of Biomedical Engineering, McGill University, Montreal, Quebec, H3A 2B4, Canada. (e-mail: carlos.roblesrubio@mail.mcgill.ca).

K. A. Brown is with the Department of Anesthesia, McGill University Health Center, Montreal, Quebec, H3A 2B4, Canada. (e-mail: karen.brown@mcgill.ca).

R. E. Kearney is with the Department of Biomedical Engineering, McGill University, Montreal, Quebec, H3A 2B4, Canada. (e-mail: robert.kearney@mcgill.ca).

B. Test Statistics for RIP Analysis

Test statistics were computed from the RIP data for use in detecting pauses, movement artifact, and asynchrony. The statistics used here are similar to those we presented in [4], with some modifications to permit their use in real-time.

The pause test statistic quantifies the power of quiet breathing in either the ribcage *(rc)* or the abdomen *(ab)* RIP signal. Pauses are defined by a lack of respiratory effort and so the RIP signals are expected to have low power in the quiet breathing band, which we determined to be from 0.4 – 2.0 Hz [4]. The original version of this statistic used the entire data record to determine the median power associated with quiet breathing, and so could only be used off-line. To eliminate this constraint, we modified the statistic to estimate the power associated with quiet breathing from the previous N_Q samples, instead of the complete recording. The modified pause test statistic for *rc* was defined by:

$$p^{rc}[n] = \sqrt{\frac{\wp^{rc}_{QB}[n]}{\underset{l\in[n-N_Q+1,n]}{\text{median}}\left\{\wp^{rc}_{QB}[l]\right\}}} \qquad (1)$$

where N_Q is the length of the window used to estimate the median quiet breathing power at each sample n, and

$$\wp^{rc}_{QB}[n] = \frac{1}{N_P}\sum_{k=n-(N_P-1)/2}^{n+(N_P-1)/2} rc^2_{bp}[k]$$

is the power in the quiet breathing band over a window of length $N_P << N_Q$. Here rc_{bp} is the band-pass filtered ribcage signal (using a band-pass filter with cut-off frequencies at 0.4 Hz and 2.0 Hz). The values of p^{rc} are expected to be close to 1 during quiet breathing and lower during pauses. A similar statistic p^{ab} is computed for the abdomen.

The movement test statistic we used previously compares the power in the movement artifact band (i.e., 0 - 0.4 Hz) to that in the quiet breathing band. It is calculated using the outputs of a filter bank spanning the frequencies 0 – 2 Hz; each filter with a 0.2 Hz bandwidth. The filters were grouped into two sets, $J=\{1,2\}$ and $I=\{3,4,...,13\}$ to span the Movement Artifact and Quiet Breathing bands respectively. Thus, the movement test statistic for *rc* was defined as:

$$m^{rc}[n] = \frac{\max_i\left\{\wp^{rc}_i[n]\right\}_{i\in I} - \max_i\left\{\wp^{rc}_i[n]\right\}_{i\in J}}{\max_i\left\{\wp^{rc}_i[n]\right\}_{i\in I} + \max_i\left\{\wp^{rc}_i[n]\right\}_{i\in J}} \qquad (2)$$

where

$$\wp^{rc}_i[n] = \frac{1}{N_M}\sum_{k=n-(N_M-1)/2}^{n+(N_M-1)/2} rc^2_i[k], \qquad i=1,2,...,13$$

is the power of rc_i, the output of the i^{th} filter in the bank, computed over a window of length N_M. The values of m^{rc} are close to 1 during Quiet Breathing and shift towards -1 during Movement Artifacts. A similar movement artifact detection statistic (m^{ab}) was defined for the abdomen.

An alternative movement detection test statistic [6] is the sum of the root mean square (RMS) of the ribcage *(rc)* and abdomen *(ab)* signals. It may be a useful complement to the frequency information obtained with (2). It was defined as:

$$r^+[n] = \sqrt{\frac{1}{N_R}\sum_{k=n-a}^{n+a} rc^2[k]} + \sqrt{\frac{1}{N_R}\sum_{k=n-a}^{n+a} ab^2[k]} \qquad (3)$$

where $a=(N_R-1)/2$, and N_R is the length of the window used to calculate the RMS value.

The asynchrony test statistic estimates the phase between *rc* and *ab* using selectively filtered RIP signals (rc_S and ab_S) to improve the signal-to-noise ratio. These signals are obtained by making $rc_S[n] = rc_{imax}[n]$ and $ab_S[n] = ab_{imax}[n]$, where the subscript *imax* indicates the number of the filter whose output has the highest power for *ab*.

The ribcage signal was then converted to a binary signal as $rc_S^b[n] = 1$ if $rc_S[n] \geq 0$ and $rc_S^b[n] = 0$ otherwise. The abdomen signal was converted similarly to $ab_S^b[n]$. Then the Exclusive-OR (XOR) was computed between these binary signals as $u[n] = rc_S^b[n]\ XOR\ ab_S^b[n]$, and the result was used to define the asynchrony test statistic as

$$\phi[n] = \frac{1}{N_A}\sum_{k=n-(N_A-1)/2}^{n+(N_A-1)/2} u[n-k] \qquad (4)$$

where N_A is the length of the window used to estimate the phase in the range [0, 1], corresponding to [0°, 180°].

For a detailed analysis and graphical examples of the test statistics, refer to [4].

TABLE I
MANUALLY CLASSIFIED EVENTS

Event	Number of Events
Pause	1 605
Movement Artifact	2 523
Asynchrony	839
Quiet Breathing	7 105

C. Manual Classification

To provide a "gold" standard for comparison with automated results, one of the investigators (KAB) manually classified events as described in [4], in a random sample of epochs from the first 9 infant data sets acquired. Only a subset of the 16 recordings was manually classified due to the time consuming and labor intensive nature of the task. Events were classified into one of the following categories: pause, movement artifact, asynchrony, and quiet breathing. Table I shows the number of events classified in this way.

III. AUTOMATED EVENT CLASSIFICATION

This section describes the supervised and unsupervised methods we devised to classify automatically the respiratory data from these statistics.

A. Supervised Classification

In [4] the presence of events was determined using detectors for pause (D_P), movement artifact (D_M), and asynchrony (D_A). The detectors compared test statistics to thresholds and were defined as: $D_P=1$ if $p^{rc} \leq \gamma_P^{rc}$ *AND* $p^{ab} \leq \gamma_P^{ab}$, $D_P=0$ otherwise; $D_M=1$ if $m^{rc} \leq \gamma_M^{rc}$ *AND* $m^{ab} \leq \gamma_M^{ab}$,

$D_M=0$ otherwise; and $D_A=1$ if $\phi \geq \gamma_A$, $D_A=0$ otherwise. Here γ_P^{rc}, γ_P^{ab}, γ_M^{rc}, γ_M^{ab} and γ_A were the thresholds for p^{rc}, p^{ab}, m^{rc}, m^{ab} and ϕ respectively. A value of 1 corresponded to event detected.

To determine the thresholds, we first had an expert clinician manually classify recordings from a representative sample of subjects. These classifications were then used to estimate two nonparametric probability density functions (PDFs) for each test statistic: one for samples classified as containing the event (Pause, Movement Artifact or Asynchrony), and one for samples classified as Quiet Breathing. Considering the test statistics as event detectors, the PDFs were used to generate the Receiver Operating Characteristics (ROC) curves relating the probability of detection (P_D) to the probability of false alarm (P_{FA}) as a function of the threshold. The value of the threshold for each test statistic was selected to provide the best tradeoff between P_D and P_{FA}, that is the point in the ROC curve farthest from the chance line $(P_D = P_{FA})$. Note that on average, each recording contained 10 hrs of respiratory data so the manual analysis was very labor intensive.

The respiratory state was determined by combining the output of the detectors D_P, D_M and D_A with the following precedence: Pause had the highest priority so when a pause was detected the other states were forced to zero. Movement Artifact was assigned the second level of precedence with the output of asynchrony and quiet breathing states forced to zero when movement was detected. Asynchrony detection had the third level of precedence. Samples not assigned to any categories were scored as Quiet Breathing.

B. Unsupervised Classification

The objective of the present work was to eliminate the need for human intervention by developing an unsupervised event classification procedure. To do so we chose K-means [7] clustering which automatically partitions a data set into k clusters, using f inputs.

The Euclidian distance, frequently used for k-means, is an appropriate metric for numeric inputs. Considering this, the decision boundary between clusters $j=1,2,\ldots,k,$ and $m=1,2,\ldots,k \neq j$, forms a hyperplane containing the point $\boldsymbol{\gamma}_{jm} \in \mathbb{R}^f$ with normal vector $\mathbf{v}_{jm} \in \mathbb{R}^f$, both determined by:

$$\begin{aligned} \mathbf{v}_{jm} &= \mathbf{c}_m - \mathbf{c}_j \\ \boldsymbol{\gamma}_{jm} &= w_{jm}\mathbf{v}_{jm} + \mathbf{c}_j \end{aligned} \tag{5}$$

where $\boldsymbol{c}_j \in \mathbb{R}^f$ and $\boldsymbol{c}_m \in \mathbb{R}^f$ are the centers of clusters j and m respectively, and $w_{jm}=0.5$ is the decision boundary weighting factor that determines the proportion of the Euclidean space covered by each cluster. The association of each instance $\boldsymbol{x}_i$ to cluster j (i.e., set C_j) is defined as

$$\mathbf{x}_i \in C_j \;\leftrightarrow\; \mathbf{v}_{jm} \cdot \left(\mathbf{x}_i - \boldsymbol{\gamma}_{jm}\right) < 0, \forall m \neq j\,. \tag{6}$$

We first applied k-means to our data set using inputs $ln(p^{rc})$, $ln(p^{ab})$, m^{rc}, m^{ab} and ϕ, to obtain 4 categories (the statistics p^{rc}, p^{ab} and r^+ were logarithmic-transformed to provide a more evenly scaled input space). This provided acceptable results when the number of samples for each event was similar (i.e., balanced). However, for unbalanced sampling the clusters were biased towards the category with the most samples. This is a problem with respiratory events, where Pause events are rare (< 10% of the events) while Quiet Breathing is common (> 50%). To correct for this, we modified the decision boundary weighting factor to reflect the relative proportion between clusters j and m as

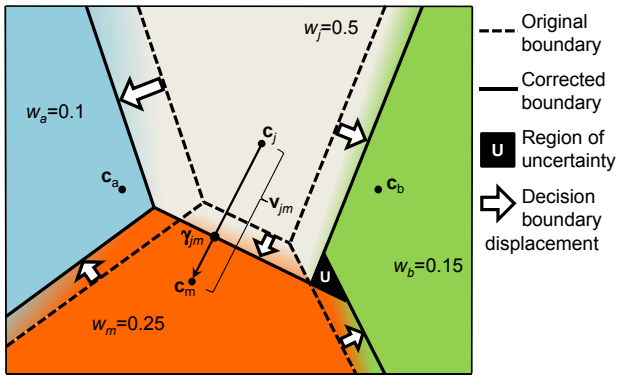

Fig. 1. Illustration of the k-means input space showing the decision boundary correction for unbalanced sampling for $k = 4$ clusters.

$$w_{jm} = \frac{w_j}{w_j + w_m} \tag{7}$$

where w_j and w_m are the proportion of samples that belong to cluster j and m respectively once k-means has converged.

Fig. 1 shows the boundaries for an example with $f = 2$ inputs and $k = 4$ clusters, before and after applying this correction. Unfortunately, shifting all the decision boundaries at once using the unbalanced sampling correction generates an uncertainty region in the input space, where instances are not assigned to any cluster. To avoid this, we defined the Automated Unsupervised Respiratory Event Analysis method (AUREA) which used a precedence-based correction for each category individually as follows:

1) Apply k-means with $k=4$ (i.e., Pause, Movement Artifact, Asynchrony and Quiet Breathing) and the inputs: $ln(p^{rc})$, $ln(p^{ab})$, m^{rc}, m^{ab}.
2) Correct the decision boundaries of the Pause cluster, identify the instances that belong to the corrected cluster and remove them from the working data set.
3) Apply k-means with $k=2$ (i.e., Movement Artifact and Breathing [Quiet+Asynchronous]) and the inputs: $ln(p^{rc})$, $ln(p^{ab})$, m^{rc}, m^{ab}.
4) Correct the boundary of the Movement Artifact cluster, identify the instances that belong to the corrected cluster and remove them.
5) Apply k-means with $k=2$ (i.e., Asynchrony and Quiet Breathing) and ϕ as input.
6) Correct the boundary between Asynchrony and Quiet Breathing, identify the instances accordingly.

This precedence was determined on the basis of our previous work [4] and an exploratory analysis performed to determine the optimal combination of inputs for each step. This revealed that m^{rc} and m^{ab} distinguish well between

Breathing (Asynchrony and Quiet Breathing) and Non-breathing (Pause and Movement Artifact), while $ln(p^{rc})$ and $ln(p^{ab})$ ameliorate this separation while improving the separation of Pause from Movement Artifact. The asynchrony test statistic ϕ distinguishes between Asynchrony and Quiet Breathing. Table II shows suggested test statistic values for initial cluster centers, obtained from representative values of the four categories.

TABLE II
TEST STATISTICS VALUES FOR K-MEANS INITIAL CLUSTER CENTERS

Cluster	$ln(p^{rc})$	$ln(p^{ab})$	m^{rc}	m^{ab}	ϕ	$ln(r^{+})$
Pause	ln(0.1)	ln(0.1)	0	0	N/A	ln(0.1)
Movement	0	0	-1	-1	N/A	1
Asynchrony	0	0	0	0	1	ln(0.1)
Q. Breathing	0	0	1	1	0	ln(0.1)

N/A= Not applicable.

IV. PERFORMANCE COMPARISON OF SUPERVISED AND UNSUPERVISED CLASSIFICATION

We evaluated the performance of the methods in terms of the agreement between the manually scored state and that estimated automatically. This was measured on a sample-by-sample basis using Fleiss' kappa (κ) statistic for inter-rater reliability [8]. A value of $\kappa=1$ indicates perfect agreement, while $\kappa=0$ reflects the performance expected by chance. We computed the overall κ value, and also the category specific agreement for each class: Pause, Movement, Asynchrony and Quiet Breathing. We also evaluated the agreement between Breathing (i.e., Asynchrony and Quiet Breathing) and Non-breathing (i.e., Pause and Movement).

Automated classification was performed on all 16 data sets using both the supervised and unsupervised (AUREA) methods (all data sets were classified to obtain a more accurate estimation of the respiratory state with AUREA). For AUREA we first used the five basic inputs $ln(p^{rc})$, $ln(p^{ab})$, m^{rc}, m^{ab} and ϕ, and then evaluated the inclusion of $ln(r^{+})$ to steps 1) and 3). The window length parameters were set to $N_P=51$, $N_M=251$, $N_A=251$, $N_Q=6001$ and $N_R=251$.

Table III demonstrates that that AUREA (case c) performed substantially better than the supervised method (case a), increasing the overall agreement by more than 23%, and the agreement on each category by at least 20%. The improvement for Pause classification was even higher (46%). It is also evident that using $ln(r^{+})$ as an additional input (case c) for AUREA provided better classification than the five basic inputs (case b); the most notable improvement was for the Movement Artifact (19%).

TABLE III
AGREEMENT (κ) BETWEEN AUTOMATED AND EXPERT SCORER

Case	B-NB	P	M	A	Q	O
(a)	0.60	0.39	0.60	0.45	0.57	0.55
(b)	0.65	0.54	0.61	0.56	0.64	0.61
(c)	0.75	0.57	0.73	0.54	0.70	0.68

B-NB = Breathing vs. Non-breathing, P = Pause, M = Movement, A = Asynchrony, Q = Quiet breathing, O = Overall.

(a) Supervised method; (b) AUREA with inputs $ln(p^{rc})$, $ln(p^{ab})$, m^{rc}, m^{ab} and ϕ; (c) AUREA with inputs $ln(p^{rc})$, $ln(p^{ab})$, m^{rc}, m^{ab}, ϕ and $ln(r^{+})$.

V. DISCUSSION

We have presented a novel completely automated unsupervised respiratory event classification method for RIP signals, and successfully applied it to data from infants recovering from surgery/anesthesia. Although the performance evaluation was limited to the agreement with a single expert, our method performed substantially better than our previously supervised procedure. For the specific case of Pause classification, the new method had a performance improvement of 46%. This is important for the study of POA, where Pause events are very relevant. The new method eliminates the shortcomings of human intervention and had very good overall agreement with an expert scorer $(\kappa=0.68)$, contrasted to that observed between expert technologists in sleep laboratories $(\kappa=0.31)$, for a respiratory index consisting on the total number of Pauses [3]. Moreover, the method can be implemented in real-time starting with a classification based on the population results and then adjusting adaptively as data are recorded for a subject. Finally, the new classification scheme makes it possible to use new test statistics without the need for additional manual scoring. Future work will exploit this capability to identify test statistics that improve the performance of respiratory state classification, and will aim to fully validate AUREA using a larger data set manually classified by several experts.

REFERENCES

[1] C. D. Kurth, A. R. Spitzer, A. M. Broennle, and J. J. Downes, "Postoperative Apnea in Preterm Infants," *Anesthesiology*, vol. 66, pp. 483-488, 1987.

[2] G. Little, R. Ballard, and J. Brooks, "National Institutes of Health consensus development conference on infantile apnea and home monitoring. September 1986," *Pediatrics*, vol. 79, pp. 292-299, 1987.

[3] N. A. Collop, "Scoring variability between polysomnography technologists in different sleep laboratories," *Sleep Med.*, vol. 3, pp. 43-47, 2002.

[4] A. A. Aoude, R. E. Kearney, K. A. Brown, H. L. Galiana, and C. A. Robles-Rubio, "Automated Off-Line Respiratory Event Detection for the Study of Postoperative Apnea in Infants," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 6, pp.1724-1733, 2011.

[5] S. M. Semienchuk, A. L. Motto, H. L. Galiana, K. A. Brown, and R. E. Kearney, "A Portable, PC-Based Monitor for Automated, On-line Cardiorespiratory State Classification," in *Proc. 27th IEEE Ann. Int. Conf. Eng. Med. Biol. Soc.*, 2005, pp. 4420-4423.

[6] A. L. Motto, H. L. Galiana, K. A. Brown, and R. E. Kearney, "Detection of movement artifacts in respiratory inductance plethysmography: performance analysis of a Neyman-Pearson energy-based detector," in *Proc. 26th IEEE Ann. Int. Conf. Eng. Med. Biol. Soc.*, 2004, pp. 49-52.

[7] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Fifth Berkeley Symposium on Math, Statistics, and Probability*, 1967, pp. 281-297.

[8] J. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bulletin*, vol. 76, pp. 378-382, 1971.