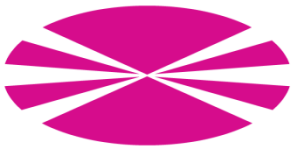


UNIVERSIDADE DA CORUÑA

DOCTORAL THESIS

DEPARTMENT OF COMPUTER SCIENCE

DIAGNOSIS OF THE SLEEP APNEA-HYPOPNEA SYNDROME

A COMPREHENSIVE APPROACH THROUGH AN INTELLIGENT SYSTEM TO SUPPORT MEDICAL DECISION

Diego Álvarez Estévez

2012

Author / Autor: Diego Álvarez Estévez

Title: Diagnosis of the Sleep Apnea-Hypopnea Syndrome: A comprehensive approach through an intelligent system to support medical decision

Título (en español): Diagnóstico del Síndrome de Apneas-Hipopneas del Sueño: Una aproximación integral mediante un sistema inteligente de ayuda a la decisión clínica

Department / Departamento: Computer Science / Computación

Supervisor / Director de tesis: Vicente Moret Bonillo

Year / Año: 2012

D. Vicente Moret Bonillo, Profesor Titular del Departamento de Computación de la Facultad de Informática de la Universidad de A Coruña.

CERTIFICA QUE: La tesis doctoral titulada “Diagnóstico del Síndrome de Apneas-Hipopneas del Sueño: Una aproximación integral mediante un sistema inteligente de ayuda a la decisión clínica”, ha sido realizada por D. Diego Álvarez Estévez bajo mi dirección en el Departamento de Computación de la Universidad de A Coruña y constituye la Tesis que presenta para optar al Grado de DOCTOR en Informática con Mención Internacional.

En A Coruña, Junio de 2012

Fdo: Vicente Moret Bonillo

Director de la Tesis

Acknowledgements

I would really like to thank to all the people who have supported me throughout all this time. And although it would not be possible to name to all the people that I should, it is true that there are certain persons that, for one reason or another, must be specially remarked and included within these lines.

In a first place, I would like to thank to my parents, Rufino and Elvira, and my sister María, which are my family. None of this would be possible without them.

Thanks to my supervisor Vicente, for his patient, his help, his trust…his everything, there would be a large etcetera. Rather than my supervisor, I consider him as my *father-in-science*. Thanks for everything. Also to all the people in the LIDIA lab for their support, especially to Chema who has worked very close with me on the development of the software part and the analysis of the sleep stages.

Thanks to Bob Kemp and to his team at the MCH Westeinde, who have brought me the opportunity to know the hospital part *sat in the front row*. Thanks also to José C. Príncipe and to the people of the CNEL lab for the stimulating research stage at the University of Florida.

I would like to thank also to all my friends, for all and each one of the good moments that I passed with them during these years.

And of course, to you Giusi, *grazie mille, mia piccola cavalletta italiana, per la tua comprensione, il tuo sostegno e la tua compagnia*.

Abstract

This doctoral thesis carries out the development of an intelligent system to support medical decision in the diagnosis of the Sleep Apnea-Hypopnea Syndrome (SAHS). SAHS is the most common disorder within those affecting sleep. The estimates of the disease prevalence range from 3% to 7%. Diagnosis of SAHS requires of a polysomnographic test (PSG) to be done in the Sleep Unit of a medical center. Manual scoring of the resulting recording entails too much effort and time to the medical specialists and as a consequence it implies a high economic cost. In the developed system, automatic analysis of the PSG is accomplished which follows a comprehensive perspective. Firstly an analysis of the neurophysiological signals related to the sleep function is carried out in order to obtain the hypnogram. Then, an analysis is performed over the respiratory signals which have to be subsequently interpreted in the context of the remaining signals included in the PSG. In order to carry out such a task, the developed system is supported by the use of artificial intelligence techniques, specially focusing on the use of reasoning mechanisms capable of handling data imprecision. Ultimately, it is the aim of the proposed system to improve the diagnostic procedure and help physicians in the diagnosis of SAHS.

Keywords

Sleep Disorders, Sleep Apnea-Hypopnea Syndrome, Intelligent Diagnosis, Medical Decision Support Tools, Fuzzy Logic, Machine Learning

Resumen

Esta tesis aborda el desarrollo de un sistema inteligente de apoyo a la decisión clínica para el diagnóstico del Síndrome de Apneas-Hipopneas del Sueño (SAHS). El SAHS es el trastorno más común de aquellos que afectan al sueño. Afecta a un rango del 3% al 7% de la población con consecuencias severas sobre la salud. El diagnóstico requiere la realización de un análisis polisomnográfico (PSG) en una Unidad del Sueño de un centro hospitalario. El análisis manual de dicha prueba resulta muy costoso en tiempo y esfuerzo para el médico especialista, y como consecuencia en un elevado coste económico. El sistema desarrollado lleva a cabo el análisis automático del PSG desde una perspectiva integral. A tal efecto, primero se realiza un análisis de las señales neurofisiológicas vinculadas al sueño para obtener el hipnograma, y seguidamente, se lleva a cabo un análisis neumológico de las señales respiratorias interpretándolas en el contexto que marcan las demás señales del PSG. Para lleva a cabo dicha tarea el sistema se apoya en el uso de distintas técnicas de inteligencia artificial, con especial atención al uso mecanismos de razonamiento con soporte a la imprecisión. El principal objetivo del sistema propuesto es la mejora del procedimiento diagnóstico y ayudar a los médicos en diagnóstico del SAHS.

Palabras clave

Trastornos del Sueño, Síndrome de Apnea-Hipopnea del Sueño, Diagnóstico Inteligente, Herramientas de Apoyo a la Decisión Clínica, Lógica Difusa, Aprendizaje Máquina

Resumo

Esta tese aborda o desenvolvemento dun sistema intelixente de apoio á decisión clínica para o diagnóstico do Síndrome de Apneas-Hipopneas do Sono (SAHS). O SAHS é o trastorno máis común daqueles que afectan ao sono. Afecta a un rango do 3% ao 7% da poboación con consecuencias severas sobre a saúde. O diagnóstico pasa pola realización dunha análise polisomnográfica (PSG) nunha Unidade do Sono dun centro hospitalario. A análise manual da devandita proba resulta moi custosa en tempo e esforzo para o médico especialista, e como consecuencia nun elevado custo económico. O sistema desenvolvido leva a cabo a análise automática do PSG dende unha perspectiva integral. A tal efecto, primeiro realizase unha análise dos sinais neurofisiolóxicos vinculados ao sono para obter o hipnograma, e seguidamente, lévase a cabo unha análise neumolóxica dos sinais respiratorios interpretándoos no contexto que marcan os demais sinais do PSG. Para leva a cabo esta tarefa o sistema apoiarase no uso de distintas técnicas de intelixencia artificial, con especial atención a mecanismos de razoamento con soporte para a imprecisión. O principal obxectivo do sistema proposto é a mellora do procedemento diagnóstico e axudar aos médicos no diagnóstico do SAHS.

Palabras clave

Trastornos do Sono, Síndrome de Apnea-Hipopnea do Sono, Diagnóstico Intelixente, Ferramentas de Apoio á Decisión Clínica, Lóxica Difusa, Aprendizaxe Máquina

Contents

Symbols and abbreviations

κ	Kappa Index
AASM	American Academy of Sleep Medicine
ADI	Adjusted O_2 Desaturation Index
AI	Artificial Intelligence
AgrI	Agreement Index
AHI	Apnea Hypopnea Index
AHI^{NS}	Apnea Hypopnea Index for Non-Supine positions
AHI^{S}	Apnea Hypopnea Index for Supine positions
AHI_{cen}	Central Apnea Hypopnea Index
AHI_{obs}	Obstructive Apnea Hypopnea Index
ANFIS	Adaptive Network Based Fuzzy Inference System
ANN	Artificial Neural Network
ANOVA	Analysis Of Variance
AP	Apneic Pattern
ApI	Apnea Index
ArI	Arousal Index
AUC	Area Under ROC Curve
BMI	Body Mass Index
CBR	Case Based Reasoning
CF	Certainty Factor
CPAP	Continuous Positive Airway Pressure
CSAHS	Central Sleep Apnea Hypopnea Syndrome
CSR	Cheyne-Stokes Respiration
CTCN	Causal Constraint Temporal Networks
CV	Cross Validation
DM	Data Mining
DEEP	Deep Sleep (sleep stage)
DS	Drowsy Sleep (sleep stage)
ECG	Electrocardiogram
EDF	European Data Format
EEG	Electroencephalogram
EMD	Entropy Minimization Discretization
EMG	Electromyogram
EOG	Electrooculogram
EOG_L	Left Electrooculogram
EOG_R	Right Electrooculogram
ES	Expert System
FIS	Fuzzy Inference System
FFT	Fast Fourier Transform
FN	False Negative
FP	False Positive
GA	Genetic Algorithm
GMP	Generalized Modus Ponens
HA	Heavy Artifact

HI	Hypopnea Index
HRV	Heart Rate Variability
ICC	Intraclass Correlation Coefficient
IP	Information Potential
ITL	Information Theoretic Learning
IVS	Independent Validation Set
KBS	Knowledge Based System
KC	K-Complex
KDD	Knowledge Discovery in Databases
LSE	Least Squares Estimator
LMS	Least Mean Square
MIB	Measure of Increasing Belief
MBR	Model Based Reasoning
MCC	Maximum Correntropy Criterion
MID	Measure of Increasing Disbelief
MEE	Minimum Error Entropy
MES	Medical Expert System
MSAHS	Mixed Sleep Apnea Hypopnea Syndrome
MSLT	Multiple Sleep Latency Test
MSE	Mean Squared Error
NFIS	Neuro Fuzzy Inference System
NHLBI	National Heart Lung and Blood Institute
NREM	Non-Rapid Eye Movement
ODI	Oxygen Desaturation Index
OSAHS	Obstructive Sleep Apnea Hypopnea Syndrome
PDF	Probability Density Function
PLM	Periodic Limb Movement
PLMS	Periodic Limb Movement Disorder
PPT	Pulse Transit Time
PSD	Power Spectral Density
PSG	Polysomnogram
r	Pearson's linear correlation coefficient
R&K	Rechtschaffen & Kales
RBF	Radial Basis Function
RBR	Rule Based Reasoning
RDI	Respiratory Disturbance Index
REM	Rapid Eye Movement
RERA	Respiratory Effort Related Arousal
rho	Spearman's rank correlation coefficient
RIP	Respiratory Inductive Plethysmography
RLS	Restless Leg Syndrome
ROC	Receiver Operating Characteristic
SaO_2	Oxygen saturation in arterial blood
SAHS	Sleep Apnea Hypopnea Syndrome
SBS	Sequential Backward Search
SEM	Slow Eye Movement
SFS	Sequential Forward Search
SHHS	Sleep Heart Health Study
SL	Sleep Latency
ST	Sleep Start
STFT	Short Time Fourier Transform
SU	Symmetrical Uncertainty
SVM	Support Vector Machine
SWA	Slow Wave Activity
SWS	Slow Wave Sleep

TN	True Negative
TP	True Positive
TST	Total Sleep Time
UA	Upper Airways
W	Wakefulness (sleep stage)
WA	Weak Artifact
WASO	Wake After Sleep Onset

0. SUMMARY IN SPANISH / RESUMEN EN CASTELLANO

0.1. Introducción

El Síndrome de Apneas-Hipopneas del Sueño (SAHS), es el más frecuente de los trastornos respiratorios que se producen durante el sueño, ya que afecta a alrededor de un 4% de la población adulta. Dicho síndrome se caracteriza por la ocurrencia de paradas respiratorias involuntarias intermitentes durante el sueño.

Las consecuencias inmediatas de esta enfermedad son fatiga e hipersomnolencia diurna, que degeneran en situaciones de irritabilidad, déficit de atención, aumento del riesgo de padecer accidentes laborales y de tráfico (se calcula que hasta 6 veces más que una persona que no padece SAHS), bajo rendimiento, estrés social, y en definitiva, la incapacidad de desarrollo de una actividad diurna satisfactoria. Muchas veces además, las pausas respiratorias vienen acompañadas por ronquido, lo que amplía el problema al entorno colectivo de las personas cercanas al paciente, que tienen que sufrir este problema. A largo plazo, y especialmente en pacientes con un SAHS severo, se relaciona esta enfermedad con un mayor riesgo de padecer enfermedades cardiovasculares, como hipertensión arterial, miocardiopatía isquémica o infarto [1] [2] [3] [4].

En España se estima que entre 2 y 3 millones de personas –entre el 3% y el 6% de la población- padecen esta enfermedad, de las cuales tan solo una de cada diez se encuentra actualmente diagnosticada y tratada. Además se estima que entorno al 25% de estos pacientes padecen un síndrome de tipo grave o muy grave [5]. Estas cifras coinciden con los diversos estudios que se han realizado por todo el mundo en los

últimos años, y que estiman la prevalencia de esta enfermedad en torno al 3% y el 7% de la población adulta [6].

El único procedimiento diagnóstico comúnmente aceptado para determinar la presencia del SAHS consiste en la realización de una prueba polisomnográfica durante una noche de sueño. Esta prueba se lleva a cabo normalmente en las Unidades de Sueño de los centros hospitalarios, y consiste en el registro durante la noche de varias señales fisiológicas del paciente, tanto neumológicas como neurofisiológicas. El registro resultante es analizado posteriormente por parte de un especialista clínico de forma manual. El principal objetivo es determinar una medida del número de pausas respiratorias registradas por hora de sueño: el Índice de Apnea-Hipopnea (IAH), que se utiliza como medida para el diagnóstico del síndrome, además de servir para cuantificar la gravedad del mismo [7].

Este proceso de revisión manual, que equivale al análisis visual de más de 500 metros de papel[1] por paciente, resulta una tarea tediosa para el clínico, y tiene como consecuencia varios factores negativos asociados, como son un descenso en la calidad del diagnóstico debido a la fatiga, un mal aprovechamiento del tiempo del clínico, y como consecuencia, la saturación de las Unidades de Sueño que no pueden absorber el caudal de pacientes debido al tiempo invertido en cada paciente individual. Todo lo anterior deriva en una peor atención al paciente y dispara los costes asociados al diagnóstico de esta enfermedad.

El desarrollo de sistemas que automaticen, al menos en parte, el análisis de la polisomnografía, y que puedan usarse como herramienta de apoyo al diagnóstico, suponen un gran ahorro en tiempo, dinero y esfuerzo, ayudando a focalizar la atención del clínico solamente sobre la información relevante. Se facilita enormemente de esta manera su labor y se optimiza su tiempo. Así, suponiendo un sistema que lograse automatizar completamente -y de forma correcta- el análisis de la polismonografía, la tarea del médico podría verse reducida únicamente a una tarea de comprobación y confirmación de los resultados.

[1] Originariamente se hacía en papel, hoy en día normalmente el registro se encuentra digitalizado en un fichero electrónico y se examina en el ordenador. En cualquier caso, la longitud del mismo no varía.

Esta tesis aborda el desarrollo de un sistema inteligente de apoyo a la decisión clínica para el diagnóstico del SAHS. El problema se aborda desde una perspectiva integral, lo que significa contemplar todas las fases del procedimiento diagnóstico clínico; esto es, realizar primero un análisis de las señales neurofisiológicas vinculadas al sueño para obtener el hipnograma, y seguidamente, realizar un análisis neumológico de las señales respiratorias interpretándolas en el contexto que marcan las demás señales del análisis polismonográfico. Para lleva a cabo dicha tarea el sistema se apoyará en el uso de distintas técnicas de inteligencia artificial, con especial atención a mecanismos de razonamiento que soporten imprecisión en los datos.

El objetivo principal que subyace detrás del sistema desarrollado es la de la búsqueda de una solución al problema del diagnóstico del SAHS, en términos de ahorro en tiempo y esfuerzo al personal médico, y por consiguiente, de dinero a los centros clínicos, siempre con el objetivo primordial de la mejora en la calidad en la atención al paciente. En última instancia, el sistema aspira a convertirse en una herramienta útil de apoyo a la decisión clínica contribuyendo a la mejora del proceso diagnóstico.

0.2. Descripción de la metodología utilizada

Esta tesis tiene como objetivo principal el desarrollo de un sistema de ayuda al diagnóstico del SAHS, que contribuya a dar solución a los problemas de las aproximaciones actuales para el análisis automático de la polisomnografía.

De acuerdo con este objetivo, la metodología seguida para el desarrollo del sistema pretende seguir una filosofía integral en el proceso diagnóstico. En este sentido el sistema desarrollado contempla, no sólo el análisis de las señales respiratorias, sino que incluye también el análisis de las señales electrofisiológicas relacionadas con el sueño. De esta forma el sistema realiza primero un análisis de las señales de electroencefalograma (EEG), electrooculograma (EOG) y electromiograma (EMG) submentoniano para obtener un el mapa del sueño del paciente o hipnograma. Durante el análisis del EEG además se realiza la detección de eventos transitorios tales como sleep spindles, complejos-K y microdespertares de EEG. Una vez finalizada esta fase de análisis, el sistema realiza un procesamiento de las señales respiratorias para la

localización de pausas respiratorias. El conjunto de señales involucradas incluye las de flujo respiratorio y movimientos torácicos y abdominales. También la señal de saturación de oxígeno es utilizada para la localización de las posibles desaturaciones y resaturaciones asociadas. Estos eventos son a continuación correlacionados en el tiempo e interpretados en el contexto de las señales neurofisiológicas. Así por ejemplo, se descartarán como falsos positivos aquellos intervalos apneicos que sucedan durante una períodos estables de vigilia, ya que la apnea debe de ocurrir, por definición, mientras el paciente se encuentra dormido. También por ejemplo será posible vincular los eventos apneicos a los microdesperares asociados y por tanto se podrá determinar la relación causa-efecto entre la ocurrencia de un episodio apneico y el microdespertar asociado correspondiente.

El sistema desarrollado tiene también en cuenta la presencia de posibles artefactos en las distintas señales. Estos artefactos servirán también para mejorar los resultados del análisis y descartar posibles falsos positivos debidos a los mismos. Otro foco de información importante a tener en cuenta dentro de esta aproximación integral es la proveniente de otras señales del contexto como, por ejemplo la posición del paciente. La interpretación de la posición del paciente durante el sueño permite detectar los movimientos del mismo, además de detectar posiciones que pueden favorecer la aparición del evento apneico, como por ejemplo cuando el paciente se encuentra durmiendo en posición supino.

La determinación de los patrones diagnósticos relevantes y la correlación temporal de eventos tiene lugar a través de la implementación de restricciones temporales que definen las relaciones que pueden o deben darse entre los distintos eventos individuales detectados en las diferentes señales que integran la polisomnografía. El objetivo es la detección de patrones diagnósticos relevantes para la identificación de intervalos en la polisomnografía con posible ocurrencia de un evento apneico.

Además de todo lo anterior, una de las principales aportaciones de esta tesis al desarrollo del sistema reside en la incorporación de mecanismos para el soporte de la imprecisión en los datos y la capacidad para establecer mecanismos de razonamiento afectados por incertidumbre. Más concretamente dichos mecanismos están soportados por el uso extensivo de técnicas de inteligencia artificial basadas en el análisis difuso de

la información. Dichas capacidades contribuyen por un lado a la mejora de los resultados en presencia de ruido en las entradas –algo muy común en las señales con alta sensibilidad al ruido-. Por otro lado contribuyen a aumentar las capacidades de generalización del sistema, reduciendo en cierta medida la variabilidad efectiva en los resultados. Esto último se justifica en el sentido de que el sistema tenderá a evitar juicios categóricos y a que los mecanismos de razonamiento estarán basados en la similitud y en la generalidad, más que en el uso de estrictas definiciones cuantitativas.

El paradigma de la lógica difusa facilita asimismo la elaboración de un sistema que permite mostrar los resultados obtenidos en un lenguaje más cercano al experto, a través de la utilización de etiquetas lingüísticas. Por ejemplo en lugar de clasificar un evento categóricamente como de tipo *apnea*, dicha clasificación se lleva a cabo en términos tales como que el evento presenta con *bastante posibilidad* las características de una *apnea*, y que al mismo tiempo dicho evento es *poco posible* que se trate de una *hipopnea*. Este tipo de clasificación de los eventos detectados permite por tanto: (1) presentar los resultados en un lenguaje más propio del ser humano, a través del uso de términos lingüísticos en la clasificación, y así facilitando la explicación de sus resultados, (2) evitar juicios categóricos tasando las clasificaciones en términos de posibilidades y no de certezas absolutas, y (3) facilitar al experto la evaluación final toda la evidencia que apunta a cada una de las diferentes hipótesis ofreciendo, no sólo un único resultado posible, sino una lista de resultados ordenada por su respectivo grado de creencia (grado de pertenencia). De este modo, el experto puede evaluar todos los resultados posibles y decidir sobre el resultado diagnóstico final.

0.3. Conclusiones

Esta tesis doctoral ha abordado el desarrollo de un sistema de ayuda a la decisión clínica para el diagnóstico del SAHS. El objetivo principal ha sido la obtención de un modelo de comportamiento inteligente para lograr un sistema que emule las capacidades de diagnóstico de los expertos clínicos. De este modo se consigue reducir el tiempo y el esfuerzo requerido por el personal médico para inspección visual del registro polisomnográfico.

Las principales limitaciones de los sistemas computacionales de diagnóstico actuales consisten en la escasez de un enfoque integral y el excesivo uso de protocolos fijos y clasificaciones categóricas. Normalmente estos sistemas se limitan a ofrecer soluciones parciales al problema y no son capaces de manejar adecuadamente la variabilidad en los datos y la subjetividad humana. El sistema desarrollado contribuye en este sentido debido a (i) su filosofía integral, en la que la actividad neurofisiológica es usada como contexto para la interpretación de los eventos respiratorios, y (ii) a la implementación de mecanismos para el manejo de datos imprecisos y que imitan los procesos diagnósticos humanos bajos los principios de generalización y juicios aproximados. De hecho, mientras que la discrepancia entre el experto y el sistema computacional causada por la subjetividad y la imprecisión es un problema para la aceptación final de estos sistemas en la práctica real, está claro que los sistemas automáticos que tratan de imitar el examen visual del PSG del clínico no pueden ser mejorados mucho más allá del acuerdo que dos expertos diferentes pueden alcanzar entre sí. Dada la inevitable subjetividad asociada al análisis diagnóstico, una posible vía de mejora debería de considerar el desarrollo de herramientas de apoyo que eviten clasificaciones categóricas y que produzcan juicios basados en criterios de similitud. El sistema desarrollado se sirve del paradigma de la lógica difusa para dar respuesta a los problemas anteriores, sin embargo, sin renunciar a las ventajas que ofrecen los sistemas automáticos de análisis en cuanto al ahorro en tiempo y en esfuerzo para la revisión del polisomnograma.

A pesar de que más investigaciones son necesarias, los resultados obtenidos por el sistema desarrollado están en el nivel de acuerdo general que el que muestran dos expertos humanos entre sí. En este sentido se puede considerar que el sistema se comporta como un experto más en la tarea de diagnóstico. Se puede concluir, por tanto, que se han cumplido los principales objetivos de esta tesis, y que el sistema puede ser usado con efectividad como una herramienta de soporte para el clínico en la tarea de diagnóstico del SAHS.

0.4. Contribuciones más relevantes

Esquemáticamente, las contribuciones más relevantes de este trabajo de investigación son las siguientes:

1. Se ha desarrollado un sistema que modela comportamiento inteligente para ayudar al clínico en el diagnóstico del Síndrome de Apneas-Hipopneas del Sueño

2. El sistema simplifica la tarea de análisis del PSG, reduciendo tanto el tiempo como el esfuerzo requerido por el personal médico

3. La validación del sistema usando pacientes reales ha demostrado que éste se comporta como un experto más respecto a los resultados de su diagnóstico

4. Las limitaciones de las aproximaciones actuales de diagnóstico automático han sido abordadas, específicamente:

    a. El procedimiento de análisis se organiza en el sistema integrando tanto la información neurofisiológica como la respiratoria, dando lugar a una aproximación integral al diagnóstico en la que los eventos respiratorios son interpretados en el contexto del sueño, tanto a nivel de macroestructura como microestructura del mismo, y en función de las demás señales presentes en el registro polisomnográfico

    b. El manejo de información imprecisa y de la subjetividad entre expertos ha sido llevado a cabo a través de la implementación de técnicas de análisis difuso, evitando que el sistema emita juicios categóricos, desarrollando mecanismos de razonamiento basados en la similitud y la aproximación, y proviniendo al sistema con capacidades de explicación de los resultados en un lenguaje cercano al lenguaje natural del clínico

0.5. Referencias

[1] DJ. Gottlieb et al., "Prospective study of obstructive sleep apnea and incident coronary heart disease and heart failure. The Sleep Heart Health Study," *Circulation*, vol. 122, pp. 352-360, 2010.

[2] S. Redline et al., "Obstructive sleep apnea hypopnea and incident stroke: the Sleep Heart Health Study," *American Journal of Respiratory and Critical Care Medicine*, vol. 182, no. 2, pp. 269-277, 2010.

[3] NM. Punjabi et al., "Sleep-disordered breathing and mortality: a prospective cohort study," *Plos Medicine*, vol. 6, no. 8, pp. 1-8, 2009.

[4] K. Monaha et al., "Triggering of nocturnal arrhythmias by sleep-disordered breathing events," *Journal of the American College of Cardiology*, vol. 54, pp. 1797-1804, 2009.

[5] J. Durán-Cantolla et al., "Normativa sobre diagnóstico y tratamiento del síndrome de apneas-hipopneas del sueño," Sociedad Española de Neumología y Cirujía Torácica, ISBN 84-7989-152-1, 2010.

[6] MP. Naresh, "The epidemiology of adult obstructive sleep apnea," *Proceedings of the American Thoracic Society*, vol. 5, pp. 136-143, 2008.

[7] JE. Lawrence et al., "Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults," *Journal of Clinical Sleep Medicine*, vol. 5, no. 3, pp. 263-276, 2009.

1. INTRODUCTION

It is a fact that the human being spends around one third of his life sleeping, approximately 25 years. However it has been an aspect of our life which has not received a sufficient attention so far. Fortunately, this vision has changed over the course of the last years, and nowadays the experts consider a good nocturnal rest as one of the fundamental pillars for a good quality of life. It is clear the relationship between amount and quality of sleep, and health and life expectancy of a person. Sleep is, without any doubt, a basic necessity for the organism, at the same level than the necessity of being fed up or breathing, and its deficiency or deprivation has serious consequences on the individual's health.

Under the common name of Sleep Disorders we group all those pathologies related with the deprivation or the deficiency of sleep, including difficulty for getting or staying asleep (Insomnia), to be sleepy during inappropriate moments along the day, excessive total sleep time, or abnormal conducts related with sleep. Among them the Sleep Apnea-Hypopnea Syndrome (SAHS) is the most frequent of the respiratory diseases occurring during sleep, affecting around the 4% of adult population. It is characterized by the occurrence of intermittent involuntary pauses during sleep.

Immediate consequences of this syndrome are fatigue and daytime hypersomnolence, which degenerate in situations of irritability, deficit of attention, increasing risk of accidents at work and traffic accidents –it has been calculated that up to six times higher than a person not suffering from SAHS-, low performance or social stress and, definitely, incapacity for the development of a satisfactory daytime activity. Even more, lots of times the respiratory pauses come accompanied by snore, which expands the problem to the circle of people surrounding the patient who have to suffer the problem indirectly. In the long term, and especially in the case of patients with a

severe SAHS, the syndrome is related with a higher risk of undergoing into cardiorespiratory problems such as arterial hypertension, myocardial ischemia or stroke [1] [2] [3] [4].

Several studies have been carried around the world during the last years, which estimate that the prevalence of SAHS is between the 3% and the 7% of the adult population [5]. In Spain it is estimated that between 2 and 3 million people –around the 3% and the 6% of the population- suffer from this syndrome, from which only one in every ten is actually diagnosed and treated. Besides it is estimated that around the 25% of these patients experience a severe or a very severe type of the syndrome [6].

The only diagnostic procedure commonly accepted to determine the presence of SAHS requires of a polysomnographic test to be done during the night. This test is normally carried out in the Sleep Labs at the medical centers. It involves the recording of several physiological signals during the night, both respiratory and neurophysiologic. The resulting recording, namely polysomnographic recording or PSG, is then visually analyzed offline by a medical specialist, to determine a measure of the number of respiratory pauses reported per hour of sleep: the Apnea-Hypopnea Index (AHI), which is used as the main measure for the diagnosis of the syndrome as well as to quantify its severity [7]

Manual revision of the PSG, which is equivalent to the visual analysis of more than 500 meters of paper[2] per patient, results in a tedious task for the clinician. As a consequence several negative factors are associated, such as a descent in the diagnosis quality due to fatigue, bad use of clinician's time, and therefore, the saturation of the Sleep Units which cannot absorb patient's demand due to the time invested in the analysis of each individual recording. Finally all the previous results in worse patient care and rising of the associated costs associated with the diagnosis.

In this regard, the development of systems that automate, at least in part, the analysis of the PSG, represents a great saving in terms of time, money and effort, facilitating the clinician's task in a great extent and optimizing his/her time. Indeed, in

[2] Originally it was actually made in paper; nowadays normally the recording is digitalized into an electronic file and is examined in the computer. In any case its length does not change.

the ideal scenario of a system being able to completely –and correctly- automate the analysis of the PSG, the use of supporting tools for the diagnosis should help to focus attention of the scorer only over the relevant information. Ultimately the physician's task could become reduced to the solely task of checking and confirming of the results.

This doctoral thesis deals with the development of an intelligent system for aiding the clinical decision making in the diagnosis of SAHS. The problem is carried out from a comprehensive perspective, which means to consider all the phases regarding the clinical diagnostic procedure; that is, to firstly perform an analysis of the neurophysiological signals bound to the sleep process in order to obtain the hypnogram, and subsequently, to make an analysis of the respiratory signals and interpreting them in the context marked by the remaining signals involved in the polysomnogram. In order to accomplish this task the system makes use of different artificial intelligence techniques, with special attention to the use of reasoning processes being able to deal with imprecise data.

The main objective underlying the developed system is that of looking for a solution to the problem of SAHS diagnosis, in terms of savings in time and effort for the associated medical personnel. Therefore the aim is the design of a useful tool to be effectively used by the clinician. The resulting software is expected to revert in the improvement of the diagnostic procedure, and therefore, to ultimately revert in the general improvement of the quality in patient's care.

1.1. Background

In the revision made by Penzel [8] about the computer systems for the recording and analysis of sleep, there is a reference to the four tasks that a system of such characteristics should fulfill.

The first task is that systems should replace conventional paper chart recorders. The intention of this function is to produce less paper and minimize space requirements for archiving without losing the raw data. The second task is documentation. With the computer based system, a technician should be able to enter all additional notes and

observations made during the nocturnal recording, which have previously been documented on paper. The third task is evaluation of sleep and cardiorespiratory functions. An automatic sleep scoring system should use its computational power to support sleep evaluation. The system should analyze electroencephalography (EEG), electrooculography (EOG) and electromyography (EMG) in terms of sleep stages; respiration, snoring and oxygen saturation, in terms of sleep related breathing disorders; and EMG tibialis in terms of movement disorders. Other parameters recorded in a sleep laboratory may require additional analysis. The system should support visual evaluation as an alternative to automatic analysis, and it should allow editing of the results of automatic analysis. The fourth task is reporting. The computer based system should help in the generation of a final report of the investigation, and it should include an advanced filing system to archive the report as well as the data in a structured way. The former enables the sleep laboratory to keep track of patients and to recall reports when needed. This option also enables the review of former polysomnographic recordings, which is seldom used today due to the difficult access to old paper recordings, in conventional paper archives.

Nowadays nobody doubts that tasks one, two and four are perfectly covered already, since the incorporation of the Health Management Information Systems into the hospitals, including the sleep labs. Effectively today almost nobody keeps using a kilometric recording paper to register the signals of the patient during the night. In its place, the signals are recorded into a digitalized file, for example, an EDF file [9]. This file can be subsequently stored occupying no more than just a few megabytes into a digital storage which can fit in a pocket. Moreover the digital file can be sent through the network to the other side of the world, allowing it to be examined by a physician just by launching a computer program being able to read this format, and to project it onto a screen. Besides, this kind of programs allow direct accessing to any part of the recording without any problem, also adjusting the set of signals to visualize, accessing the information about them, making zoom, or adding annotations. Even more, they can reproduce video recorded during the night and synchronizing it in time with the signals, reproduce ambient sounds and snores, or make possible the application of digital filters to get rid of noise and some kind of artifacts [10]. Finally, once the recording has been checked by the technician, it is possible to generate a report taking into account all the annotations made by him/her, and that contains all the relevant information for the

diagnosis. This report can be digitalized as well in any kind of digital format, being then stored into a documentary server and allowing its access, again through the network, from any part of the world just by barely clicking a couple of times.

It is however regarding the task number three, the one relative to the analysis capabilities of the computer programs for the recording and analysis of sleep, where these systems still today present deficiencies. In this regard, even if in one hand sleep medicine can be considered a relatively young discipline within the medical field, it is also true that attempts to automate the sleep analysis task can be considered almost as old as the discipline itself.

In this context, and stepping up to set up a date for the beginnings of these systems, it can certainly established around the year 1968, when A. Rechtschaffen and A. Kales (R&K) published their manual for the standardization of the terminology, techniques and scoring system of sleep stages in human subjects [11]. This manual, which is not absolutely the first work on sleep research, it did became instead, the very first great standardization, which extended for example the previous observations made by Loomis [12] who described sleep phases from A to E, or the observations from Aserinsky and Kleitman [13], who for the first time described the rapid eye movements phase (REM). The set of rules for the sleep analysis described by R&K defined since then the standard method for its characterization. This method has arrived practically unaltered to our days, being only recently modified by the last revision published in 2007 by the American Association of Sleep Medicine (AASM) [14]. In any case, from the advent of the standardized methods for visual evaluation of sleep, several attempts to develop systems for the automatic classification of sleep have followed each other. On the other hand, in the context of SAHS as a disease associated to sleep, the first attempts can be found around the decade of the 70s and beginnings of the 80s [15] [16].

During these last 30 years there have been several the approximations proposed to automate the polysomnographic analysis task. However, research activity in this field still continues today to be very active. Throughout this thesis manuscript, especially in chapter three, a more detailed analysis of these systems will be carried out, specially focusing in those specifically designed for the diagnosis of SAHS. In any case, the

question seems to be obvious: *why the automatic analysis of PSG still constitutes to be a challenge not completely resolved?* We can analyze the causes more in detail:

First of all we can point out to computational and technological reasons. In fact, as it has been already mentioned, a polysomnography of an entire night implies continuous recording lasting around 8 hours of duration, which involves between 10 and 15 different physiological signals. This implies a considerable amount of data which not always the computers have been able to manage. Actually the first algorithms for the PSG analysis made the processing *online* as long as data were generated. The main reason was that they hardly were able to store a file which might weight from 50 MB up to 500 MB. On the top of that, it has to be added the requirements regarding the calculation performance required for the analysis of the different involved signals. Hence, it has been necessary the evolution of the computation capabilities, for the computers to be able to satisfy the necessities to analyze and store a complete PSG.

In any case, and once these limitations were overcame, the complexity of the analysis task continued –and still nowadays- being one of the main challenges preventing the success of this kind of systems. It has to be remarked that polysomnography is composed of a number of signals of different nature. Thus, there is a need for different analysis techniques depending on the concrete signal. On the other hand, many of these signals are very sensitive to noise. Simultaneous recording of the signals favors the possibility of interferences among them. Even more, sleepy patients tend to move or sweat, which causes the easy displacement of the sensors. External factors to the recording such as mains interference can also affect signal registration. In summary, there are several factors that can cause the presence of all kinds of artifacts corrupting the PSG. These artifacts can be reduced somewhat by improving the sensors, the insulating and the application of analog filters at the time of the signal registration. However, the presence of artifacts in the recorded signals is something inherent to the recording itself, and its presence cannot be completely avoided. The occurrence of artifacts in the recording represents a challenge even for the experienced scorer. Thus it does not seem strange that its treatment constitutes, from a computational point of view, one of the main issues that automatic systems for the analysis of PSG should confront [17].

On the other hand, the complexity of the analysis task itself, involving several signals of different nature and their temporal interrelation, makes the development of computer approximations from a comprehensive perspective, something rather scant and difficult to attain. In this respect, and specifically for the case of SAHS, by comprehensive it must be understood those approximations that, besides the analysis of apneic respiratory pauses, carry out an analysis of the neurophysiological activity related to the sleep. Such methodology allows interpreting the respiratory pauses in the context of the patient's sleep structure. The former necessarily involves, at least, the construction of the hypnogram, and the detection of relevant transient events such as, for example, micro arousals which can have an apneic origin. Likewise the detected respiratory events should be interpreted in the context marked by the remaining signals present in the polysomnography. It is of special attention, for example, the analysis of apneic events within the context of the body position of the patient. In this respect, besides the development of individual algorithms for the adequate processing and the extraction of the relevant information on each signal, it grows the importance of temporal correlation mechanisms in charge of accomplishing the identification of significant diagnostic patterns throughout the recording.

All the previous challenges entail that many of the approximations in the scope of the computerized analysis of SAHS, instead of global approaches, rather reduce to the accomplishment of specific subtasks. Within these subtasks, for example, there can be encountered realizations for the analysis of oxygen saturation decays, the identification of respiratory amplitude reductions, or in the best scenario, the performing of a full analysis of the respiratory signals but without interpreting them in the context of the rest of the PSG signals. This situation leads to suboptimal analyses, in many cases, not satisfactory enough for the clinician to consider the use of these programs as valid tools for the diagnosis of SAHS.

However, besides the lack of comprehensive approaches, current systems also suffer from additional drawbacks. To introduce them, let us consider the following. In the domain of SAHS two main types of episodic events can be found: *apneas* and *hypopneas* [14]. An *apnea* is defined as an event of a higher intensity, characterized by a stop in the respiratory airflow during at least 10 seconds. This stop is characterized by a flow reduction of approximately the 90% with respect to the normal breathing.

Obviating the problem of the establishment of the normal breathing, it can be said that previous definition is more or less well characterized and is widely accepted for the different specialists in the field [18]. The event *hypopnea*, on the contrary, is more difficult to measure and its definition is less precise [19]. In this case rather than a *stop* the term *pause* is more commonly found in the literature. The controversy here results mainly at the time of the numerical characterization of such a pause –or reduction, which can vary from one medical center to another from "*discernible*" to a required amplitude reduction of 50% [18]. On the other hand, it is known that both apneas and hypopneas are commonly accompanied by a decrease in the levels of arterial blood oxygen saturation (SaO_2). In this respect, while the apnea definition commonly obviates this desaturation from the requirements, for being an event sufficiently well characterized by other measures, the required levels of desaturation to classify an event as hypopnea may also vary depending on the expert. Desaturation values in the range of 2% and 5% are common. Even sometimes the absence of a desaturation event is permitted. Additionally, the necessity of EEG arousals events accompanying the hypopnea is also an object of controversy. Indeed, even today the AASM in its recent revision of the sleep scoring rules provides two different definitions to score this event [14].

On the top of that, it is necessary again to remark that the analysis of the PSG, which involves the analysis of an entire night of sleep and with many different signals, is a tedious task for the clinician. It can be considered, for example, that the recognition of an apnea event for an experienced specialist is not a difficult task. However recognition of hypopnea events results much more complex due to the less significant airflow reduction occurring in these cases, which therefore makes its quantification more difficult at a glance [20]. Besides, as the time passes the tiredness increases and, as a consequence, the accuracy of the analysis. The scorer may try to go faster and the risk of making mistakes increases. This results in a loss in the final quality of the diagnosis.

All the previous causes that, with the same recording, different diagnoses can be achieved in the presence of different scorers. Even more, the expert itself may show discrepancy on scoring the same recording, but in different instants of time. In definitive, there is variability in the diagnosis due to different definitions of the relevant events, and also because of subjectivity of the human decisions. And although, the use

of automatic systems for the analysis of the PSG should represent an adequate solution to the variability problem, in practice its use from the part of the medical personnel is far from that expected. Thus eventually these programs do not solve the problem for which they were built for.

Taking this situation into account we can extend the reasons which prevent the success of the current computer systems for the analysis of the sleep:

- The discrepancy due to the diversity of criteria existent for the definition of apneic events. As stated before, it results in a generalization problem for these systems since although the results can be correct for an expert, it is possible that another expert does not found them equally correct (expert's *inter-variability*). Moreover even the same expert in different instants of time can disagree with himself (expert's *intra-variability*).

- Categorical decisions. Even in the situation where there is a unique standard criterion, the problems arise in regard with the way a system implemented using such criterion could offer its results. It is a common problem for expert systems to excessively offer its results in categorical form. Let us consider, for example, the following situation in which given two events *A* and *B*, the system associates respective probabilities of occurrence –in percentages- of 90% and 90.1%. Attending to the criterion of the higher probability, a system could express as its resulting output that the most probable event is *B*. However saying nothing about the concurrent high probability of occurrence of *A*. Similarly it can be the case in the context of SAHS of the occurrence of an apneic event which does not present clearly characteristics to be classified without any doubt either as an *apnea* or a *hypopnea*. In this situation the opinion of the clinician expert plays a fundamental role in the final classification. Even if both system and expert use the same standard guideline for identification of the event, the categorical affirmation from the part of the system about the event's classification may not correspond with expert's subjective opinion. This causes discrepancy between the system and the expert which may

eventually result in distrust on the system's results from the part of the clinician.

- The lack of explanation of the results. An expert will give little credibility to a system behaving as a black-box which offers its results without an adequate explanation, even more in the medical field. A good diagnostic system, besides valid outputs, should provide an explanation of its results. Moreover whether this explanation takes place, it should be carried out into a language adequate to the necessities of the expert.

- Usability problems. It is important to design the system thinking on its final user which may do not be a computer technician. The system must *speak* the language of the clinician, fulfill its necessities, and its use should not be too much complicated. It should not offer more of what the physician needs, and it should organize the results through adequate views, making accessible all needed information in an organized manner.

1.2. Scope and objectives

This doctoral thesis has as its main objective the development of a system helping in the diagnosis of SAHS, which contributes to solve the problems of current approaches for the automatic analysis of the PSG.

In order to fulfill that objective, and according to the title of this thesis, first of all, the developed system is aimed at following a comprehensive philosophy over the diagnostic process. In this respect the system carries out not only the analysis of the respiratory signals, but it also includes the analysis of the electrophysiological signals related to sleep. In this manner the system firstly performs an analysis of the electroencephalography (EEG), electrooculography (EOG) and submental electromyography (EMG) to obtain the structure of the patient's sleep or *hypnogram*. Besides, during the EEG analysis a detection of transient events such as sleep spindles, K-complexes or EEG arousals is performed. Once this analysis has finished the system continues by processing respiratory signals for the location of breathing pauses. The set

of involved signals here includes the airflow and the signals recording thoracic and abdominal respiratory movements. Additionally, the oxygen saturation signal is used for the detection of the corresponding desaturations and resaturations. All of these events are subsequently correlated in time, and interpreted in the context of the neurophysiological signals. In this regard, for example, those false positives caused by the occurrence of apneic intervals during stable periods of wakefulness can be discarded. This is because, by definition, the apneic event must occur while the patient is asleep. Another possibility that the comprehensive approach permits, is that of linking the apneic events causing EEG arousals to their corresponding neurophysiologic event, thus allowing the determination of the cause-effect relationship between the apneic event, and its associated microarousal.

The developed system takes also into account the presence of possible artifacts in the monitored signals. Detection and characterization of signal artifacts allows the analysis to detect possible false positives caused by them. Another source of important information in the comprehensive approach comes from the remaining context signals such as, for example, patient's position. Interpretation of the sleeping position allows detecting movements during sleep, or to take into account sleep positions which favor the appearance of the apneic event, like for example when the patient is sleeping in supine position.

The identification of the relevant diagnostic patterns takes place through the implementation of temporal constraints, which define the relationships which can or must occur, among the isolated events detected across the different signals included in the PSG. The objective is to form significant diagnostic patterns to identify recording intervals where the evidence points out to the possible occurrence of an apneic event.

Besides the comprehensive approach, one of the main contributions of this thesis lies in the incorporation of mechanisms supporting imprecision in the data, and the capacity to establish reasoning processes affected by uncertainty. More specifically such procedures are supported by the prominent use of artificial intelligence techniques based on the fuzzy analysis of the information. Such capabilities contribute, on one hand, to the improvement of the results in the presence of noise at the input –something very common among signals with a high noise-to-signal ratio. On the other hand, they

contribute to augment generalization capabilities of the system, thus reducing somewhat the effective variability of the results. This last statement is justified by the fact that the system tends to avoid categorical judgments, besides, reasoning mechanisms operate basing more its decisions on similarity criteria rather than in strict quantitative definitions.

The use of fuzzy approaches to support imprecise information makes it also easier to construct a system being able to express its results in a language closer to the human expert. This is possible by making use of fuzzy linguistic labels. For example, instead of categorically classifying an event as an *apnea*, the classification in the proposed system is carried out in *fuzzy terms*. For example the event may present with *quite a lot of possibility* the characteristics of an *apnea*, but at the same time, the same event is *quite unlikely* to be a *hypopnea*. This kind of classification of the detected events permits the system: (1) to present the results in a more human-like manner through the use of linguistic terms in the classification, also facilitating the explanation of the results, (2) to avoid categorical judgments, evaluating the classifications in terms of possibilities, not in terms of absolute certainties, and (3) to allow the expert to easily evaluate all the evidence pointing out to each one of the considered hypothesis, not by just offering one unique possible result, but a list of possible results sorted by their respective amount of belief[3]. In this manner the expert clinician can evaluate all the possible results and decide out about the final diagnosis.

[3] In this case it may be more correct to say, by their respective *degree of membership*. For further details see Chapter 4

Summarizing, main objectives of this doctoral thesis can be enunciated as follows:

- To construct a system to aid medical decision in the diagnosis of SAHS. The system will make special emphasis in the automatic analysis of the PSG
- To simplify the analysis task of the PSG, reducing both time and effort needed from the medical personnel
- To overcome limitations of current computer methods for the diagnosis of SAHS, and to do so:
    - To construct a system that handles the problem both from the pulmonological and the neurophysiological perspectives
    - To develop an analysis strategy that minimizes the effects of intra and inter experts variability
    - To avoid categorical judgments
    - To develop a system being able to explain its results
- Ultimately, to revert in the society by improving the SAHS diagnostic procedure and therefore the general quality in the patient's care

1.3. Structure of the document

In the following it is given a breakdown of the different chapters which compose the structure of the doctoral manuscript.

The document continues in the next chapter with a general description of the context in which the project is included: the diagnosis of the Sleep Apnea-Hypopnea Syndrome. In this regard, throughout the next chapter the most relevant clinical concepts related with sleep are introduced, focusing in the polysomnographic test and in

the description of the most important signals included in the PSG. After introducing the fundamentals of the sleep analysis, a description of the classical approximation to the diagnosis of SAHS is given. The different types of events associated to this syndrome are described and interpreted in the context of the PSG.

Chapter three is an introduction to the intelligent systems for the support of the clinical decision making. Firstly, historical perspective is given, starting from the beginnings of artificial intelligence in the field of medicine to its evolution reaching the current systems. The different techniques and mechanisms that current intelligent systems implement are described in general terms. In the context of this thesis special attention is made with regard to the different techniques to handle the analysis of imprecise information. Focusing in SAHS domain, a review is performed on the different solutions which have been carried out so far in order to assess the automatic diagnosis of SAHS. To structure the analysis, the different solutions are classified in *commercial systems* and *non-commercial –academic- systems*. Finally, a critical analysis of current approaches is performed in the last part of the chapter, introducing the necessity to carry out new approaches to overcome their limitations.

Throughout chapter four the technological aspects of the developed system are described in relation with the fuzzy analysis of information. Fundamental theoretical aspects of fuzzy logic are firstly described, introducing fuzzy inference mechanisms, to end up with the so-called fuzzy inference systems. Machine-driven parameterization of such systems is outlined by the use of neuro-fuzzy modeling techniques. Discussion on specific techniques used for the construction of the developed system is performed in the last part of the chapter.

Chapter five focus on the description of the system itself. It starts from a software engineering perspective, making reference to the methodology used for its development, analyzing the analysis of the requirements, and making a brief description of the system from the architectonical point of view. In the following sections system's construction is described regarding its design, and explaining and detailing the operation of all its integrating modules. In this respect the processing algorithms are described including, among others, those regarding signal acquisition, artifact detection, respiratory event's

identification, analysis of the neurophysiological signals, hypnogram generation, detection and classification of the apneic events, and final diagnosis generation.

Subsequent chapters deal with the evaluation of the system by carrying out a validation process. The aim is to evaluate if the proposed objectives have been achieved and in which degree. For such purpose a validation process using PSG recordings of real patients is performed. Chapter six introduces the validation process by describing the used validation measures and the design of the validation tests. Chapter seven presents and analyzes the results according to the structure of the proposed validation tests.

Finally, in chapter eight the final discussion is effectuated, in which analysis of the constructed system is performed from a critical point of view. Main conclusions are delivered assessing its lacks and its strengths, commenting possible improvements and future work.

1.4. Summary of this chapter

This chapter introduces the topic of this doctoral thesis. The main objective is to develop a system for the automatic analysis of the PSG to be used as supporting tool for the clinician in the diagnosis of SAHS. SAHS is defined as a syndrome pertaining to the group of diseases affecting sleep. Its diagnosis implies the realization of a polysomnography and the subsequent analysis of the resulting recording. Such a task is costly, thus as a result it is proposed the construction of a system to automate the analysis, helping to reduce diagnostic time and the effort of the involved clinical personnel.

The development of a system of these characteristics presents some difficulties, and throughout the chapter, an analysis of the problematic associated to the automatic analysis of the PSG in the context of SAHS is done. In this respect, some of the factors preventing the satisfactory implantation of these systems in the sleep units of the medical centers are analyzed, which include lack of comprehensive approaches and limitations on the correct analysis of the relevant information.

Main objectives of the doctoral thesis are subsequently described, which mark the guidelines for the development of a system to confront shortages of the current approximations. For such purpose, it is proposed the development of a system which assesses the analysis from a comprehensive perspective. Such system, in addition, makes use of artificial intelligence techniques being able to manage data imprecision and supporting reasoning processes affected by uncertainty. The resulting system is aimed at constituting a possible solution for the shortages of current approximations in the diagnosis of SAHS.

1.5. References

[1] DJ. Gottlieb et al., "Prospective study of obstructive sleep apnea and incident coronary heart disease and heart failure. The Sleep Heart Health Study," *Circulation*, vol. 122, pp. 352-360, 2010.

[2] S. Redline et al., "Obstructive sleep apnea hypopnea and incident stroke: the Sleep Heart Health Study," *American Journal of Respiratory and Critical Care Medicine*, vol. 182, no. 2, pp. 269-277, 2010.

[3] NM. Punjabi et al., "Sleep-disordered breathing and mortality: a prospective cohort study," *Plos Medicine*, vol. 6, no. 8, pp. 1-8, 2009.

[4] K. Monaha et al., "Triggering of nocturnal arrhythmias by sleep-disordered breathing events," *Journal of the American College of Cardiology*, vol. 54, pp. 1797-1804, 2009.

[5] MP. Naresh, "The epidemiology of adult obstructive sleep apnea," *Proceedings of the American Thoracic Society*, vol. 5, pp. 136-143, 2008.

[6] J. Durán-Cantolla et al., "Normativa sobre diagnóstico y tratamiento del síndrome de apneas-hipopneas del sueño," Sociedad Española de Neumología y Cirujía Torácica, ISBN 84-7989-152-1, 2010.

[7] JE. Lawrence et al., "Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults," *Journal of Clinical Sleep Medicine*, vol. 5, no. 3, pp. 263-276, 2009.

[8] T. Penzel and R. Conradt, "Computer based sleep recording and analysis," *Sleep Medicine Reviews*, vol. 4, no. 2, pp. 131-148, 2000.

[9] B. Kemp, A. Värri, AC. Rosa, KD. Nielsen, and J. Gade, "A simple format for exchange of digitized polygraphic recordings," *Electroencephalographic Clinical Neurophysiology*, vol. 82, pp. 391-393, 1992.

[10] B. Kemp and M. Roessen, "Polyman: a free(ing) viewer for standard EDF(+) recordings and scorings," in *Sleep-Wake Research in The Netherlands*, SF Ruijgt et al., Eds.: Dutch Society for Sleep-Wake Research, 2007, pp. 71-73.

[11] A. Rechtschaffen and A. Kales, "A manual of standardised terminology techniques and scoring system for sleep stages of human subjects," U.S. Department of Health, Education and Welfare Public Health Service – NIH/NIND, 1968.

[12] AL. Loomis, EN. Harvey, and GAI. Hobart, "Cerebral stages during sleep, as studied by human brain potentials," *Journal of Experimental Psychology*, vol. 21, pp. 127-144, 1937.

[13] E. Aserinsky and N. Kleitman, "Regularly occurrin periods of eye motility, and concomitant phenomena, during sleep," *Science*, vol. 118, pp. 273-274, 1957.

[14] C. Iber, S. Ancoli-Israel, A. Chesson, and SF. Quan, "The AASM manual for the scoring of sleep and sssociated events: rules, terminology and technical Specifications," American Academy of Sleep Medicine, Westchester, IL, 2007.

[15] P. West and MH. Kryger, "Continuous monitoring of respiratory variables during sleep by microcomputer," *Methods of Information in Medicine*, vol. 22, pp. 198-203, 1983.

[16] JR. Smith, "Computers in sleep research," *CRC Critical Reviews in Bioengineering*, vol. 3, pp. 93-148, 1978.

[17] B. Kemp, "Measurement of sleep," in *Human sleep and cognition, part I: Basic research*., 2010, pp. 21-35.

[18] S. Redline et al., "The scoring of respiratory events in sleep: reliability and validity," *Journal of Clinical Sleep Medicine*, vol. 3, no. 2, pp. 169-200, 2007.

[19] M. Moser, B. Phillips, D. Berry, and L. Harbison, "What is hypopnea, anyway?," *Chest*, vol. 105, pp. 426-28, 1994.

[20] A. Otero, P. Félix, and MR. Álvarez, "Algorithms for the analysis of polysomnographic recordings with customizable criteria," *Expert Systems with Applications*, vol. 38, pp. 10133-10146, 2011.

2. DESCRIPTION OF THE DOMAIN

The word *sleep* designates both the act of sleeping as well as the desire of sleep (to be sleepy). Concerning the act of sleeping, although there are several definitions, it is commonly accepted the one referring to sleep as a uniform restful state from an organism; a complex state quantitatively different from wakefulness, but at the same time closely related to it.

As a restful state, sleep is characterized by a resting posture -varying depending on the animal species- with either an absence or a decrease in the voluntary corporal movements, and with poor response to external stimuli. It is also important to stress its limited duration, which distinguishes it from coma. However at the internal level a series of important biological variations are produced together with a characteristic change of the cerebral activity. This activity is associated with a repairing function for the individual through hormonal, metabolic, biochemical and temperature changes, which are essential for a good resting.

A remarkable phenomenon produced throughout sleep is the dreaming act. Dreaming, as an involuntary mental process which submerges us in a virtual reality involving images, sounds, thoughts and sensations, has only been confirmed in the *Homo Sapiens*. It is though that some mammals could also hold high probabilities of dreaming, however although there are other animals which experiment REM[4] sleep state, their subjective experience is difficult to be determined.

Current advanced scanning systems have detected that in several occasions dreams are cerebral activity loops which repeat night by night. It is also known that every single subject has an unrepeatable and unique form of dreaming. Cerebral activity represented

[4] Period of dreaming during which, as we will see, dreaming is normally produced

by electromagnetic waves on the screen of these scanners presents very similar graphic patterns within each patient, and different between two of them.

What remains clear is that sleeping is essential to develop a normal life. The number of hours the human being dedicates to sleep varies considerably from one person to another, not only because of biological/genetic reasons, but many times because of subject's life habits. Experts establish the recommendable duration of sleep in eight hours, in which it can found a variability range between five and ten hours of sleep. Ideal sleep duration is that which allow us to develop a normal social and working life.

In which all experts agree is that sleep is necessary for life. The lack or the privation of sleep triggers all kinds of disorders affecting behavior, as for example, decreasing in the daytime attention level –thus affecting working and personal performance of the individual- increasing risk of traffic accidents, and cardiovascular dysfunction. The hypothesis that sleep participates in memory consolidation has also been investigated recently. Studies have confirmed the idea that sleep is profoundly implicated in the memory function of both human and animals. In this respect it has been demonstrated its effects in memory consolidation and learning [**1**].

Throughout the last century a great progress has occurred in the scientific studies about sleep. Although even the oldest medical manuals stressed the repairing functions of sleep, the scientific interest has not been developed until beginnings of 20^{th} century. It is in this period when a key publication appears, *The Interpretation of Dreams,* by Freud [**2**]. Nevertheless true scientific research on sleep does not occur until about the middle of the century. It is around the decade of 50s when it is discovered that sleep is not a homogeneous phenomenon, but it fluctuates in a cyclic form between two sequential states. Recent developments in the fields of neurobiology, molecular biology, physiology, neuropsychiatry or cardiology, together with technology advances that have enabled in a great extent what it has been known as *energy of sleep*, have allowed the researchers the study of the sleep details. As a consequence, these developments have enabled the development of a researching movement toward the compression of sleep and its disorders, and the creation at the medical level of the so-called *Sleep Laboratories*.

2.1. Sleep Studies

Given the importance that sleep has in the life of human beings, an increasing interest is raising regarding its study and characterization. Technology advances experienced within the last century have allowed the researchers to define sleep through certain related physiological measures. In this respect the so-called *Sleep Studies* arise.

The purpose of such studies, besides divulging and understanding the structure of sleep, is to diagnose anomalies which may be either direct or indirect responsible of both, nigh sleep problems as well as daytime complications[5]. In fact in the recent years it is common the use of the term *dyssomnias*, avoiding the use of *hypersomnias* or *insomnias*, since many times people sleeping bad during the night may also present abnormal daytime somnolence.

The *Multiple Sleep Latency Test* (MSLT) consists in the study of the input latency of sleep and phase REM; for that purpose the patient undergoes naps (normally five) separated by two hours during a day. With this test it can be known if it exists or not a pathologic hypersomnolence, and to specify if it has to deal with a specific disorder such as narcolepsy.

Another kind of study is *actigraphy*. It consists in the evaluation of the movement, generally of the arm, during several days. It serves as an indirect measure to give an idea of the different sleep periods over patients with sleep problems. In this test only one sensor is used which is an accelerometer placed in the patient's wrist as it was a watch. Common recording periods usually last between 4 and 10 days.

Pulse oximetry is another non-invasive method allowing the monitoring of oxygen concentration in hemoglobin. It also allows measurement of pulse rate. It can be used either as a simple screening test to discard the presence of sleep apneas, or as to control efficacy in the treatment of already diagnosed patients. However it is not an exhaustive diagnostic test and therefore is mostly used for screening purposes.

[5] E.g. insomnia or narcolepsy

There exist different types of studies according to the actual symptoms and the suspected syndrome. Nevertheless the most complete test and the standard procedure to carry out sleep studies in the sleep labs is the polysomnography (PSG).

2.2. The polysomnography

Polysomnography is a diagnostic test in which several sensors are attached to the patient in order to monitor different physiological functions, determine his/her sleep patterns, and to check for possible abnormalities. Depending on the symptoms and the suspected diseases the number of recorded signals can vary. There are studies which uniquely record neurophysiological activity during the night, whereas others –as in the diagnosis of SAHS- usually include many other signals, for example, related to the respiratory function (see Figure 2.1)

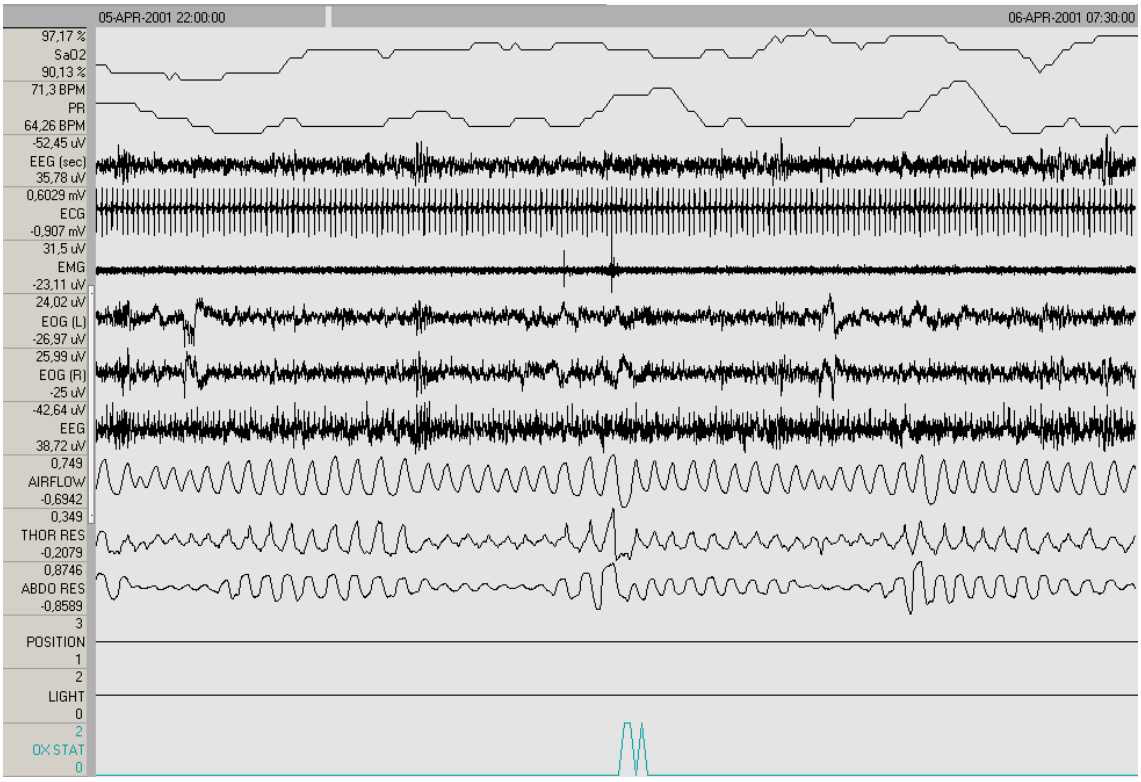


Figure 2.1. Image of a nocturnal digitalized polysomnography

In patients with insomnia, the PSG corroborates and classifies the kind and the severity of the insomnia. It indicates if a difficulty exits either to get or to manage to sleep, and it points out to the presence of nocturnal awakenings, or an early morning awake. For each one of these possibilities the cause and the corresponding treatment differ. In patients with daytime somnolence, the polysomnographic studies help the physician to know whether a nocturnal factor exists that influences this symptom. Actually, patients with presence of obstructive apneas usually comply, not because of the snoring[6], but because they are tired and because they feel somnolent during the day.

Another purpose of the polysomnographic studies is that of diagnosing peculiar episodes which come out during the night, as they are for example awakenings, somniloquy[7], sleep walking, bruxism[8] or night terrors.

Polysomnography is carried out in the sleep labs of the medical centers. It is the most precise diagnostic method for the majority of the most common sleep disorders. Manual revision of the PSG is, however, a laborious task, complex and expensive. On the other hand, in many cases PSG is not well tolerated by the patients, mainly because it may result an uncomfortable test due to the presence of wires and equipment. Also because of the hospital environment in which the recording is carried out which differs from the usual sleep place of the patient. The increasing healthcare demand that the specialized centers in sleep disorders are experiencing, causes the necessity of alternative methods to PSG of being simpler and with less associated cost. In this respect, development of portable systems which allow a first analysis to be performed ambulatory is an increasing field of interest. In chapter 3 some examples of portable systems are analyzed. In any case, these devices are normally limited to fulfill with a screening function.

In the case of polysomnography applied to SAHS, the analysis principally consists in the detection of apneic events (involuntary respiratory pauses), mainly through the information coming from the interpretation of the respiratory signals, and its posterior analysis in the context of the neurophysiological activity, which determine the sleep

[6] The partner is the one who usually complains about snoring
[7] Sleep talking
[8] Involuntary mandibular movements

state of the patient. Other sources of contextual information are used in addition to refine the analysis, for example, to detect the presence of false positives due to situations such as changes in body position of the patient, or the occurrence of artifacts in the signals.

In accordance, it can be considered that in the case of SAHS, the polysomnographic montage usually involves three different types of signals:

1. Pulmonological signals for the localization of the respiratory pauses –apneic events- mainly comprising respiratory movements, oxygen saturation in arterial blood and airflow.

2. Neurophysiological signals related with the sleep function mainly including electroencephalogram (EEG), electrooculogram (EOG) and electromyogram (EMG). The main objective is the assessment of the sleep map of the patient – hypnogram- which serves as the fundamental context for the interpretation of the respiratory events. Besides, there are useful for the localization of additional events of importance for the diagnostic, such as for example EEG micro-arousals, sleep spindles or K-complexes.

3. Additional contextual information signals, as they may be the recording of lights state, patient's body position throughout the test, electrocardiogram (ECG) or the snore signal.

In the following subsection a more detailed analysis of the most relevant signals within the PSG is performed regarding the diagnosis of SAHS.

2.2.1. Signals related to sleep of the patient

Electrooculogram

Around the middle of the 19th century it has been discovered a difference in the potential between the cornea and the retina of human eye. In this regard, when cornea is positively charged, retina is charged negatively, and therefore it allows the electrical

characterization of the human eye as a rotary dipole. This potential difference is used to monitor the ocular movements, and to distinguish the different patterns of eye movements which occur during some sleep periods. It is common to work in a range between 20 and 250 µV (peak to peak) with sampling frequencies between 100 to 500 Hz (see Figure 2.2). Analysis range of EOG is normally below 50 Hz.

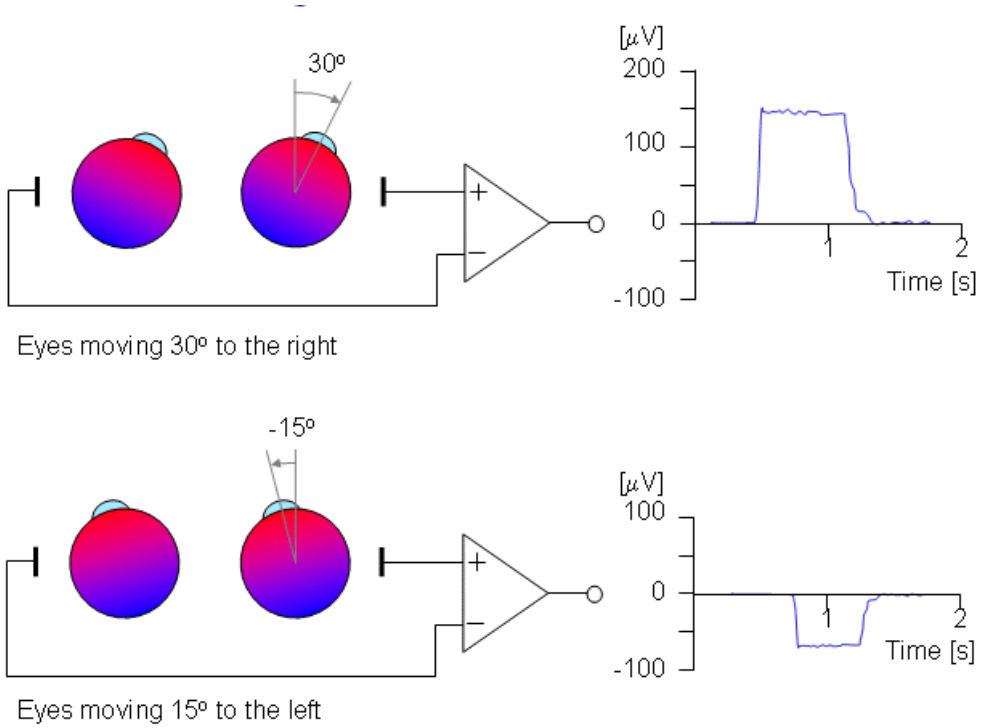


Figure 2.2. Image of the mechanism to record eye movements. Figure from the web version of the book by Jaakko Malmivuo & Robert Plonsey [**3**], chapter 28

Electromyogram

The EMG monitors the neuromuscular activity associated with muscle contraction. It is recorded using electrodes situated over the skin surface of the muscle which is intended to be monitored. In the scope of sleep the most commonly used derivation is the submental EMG, because the muscles in this area provide good quality signals that reflect changes produced in the normal progression of sleep (see Figure 2.3). Additionally, two supplementary tibialis derivations are normally used to gather muscle activity from the legs. Monitoring of leg movements is useful to keep track of arousals caused by leg movements, and it is of special interest in the case of suspecting from Restless Leg Syndrome (RLS) or Periodic Limb Movement Disorder (PLMS). Usual recording amplitudes within this signal are in the range from -100 to 100 µV, and with

sampling frequencies which oscillate between 100 and 500 Hz (analysis frequencies usually range between 10 and 100 Hz).

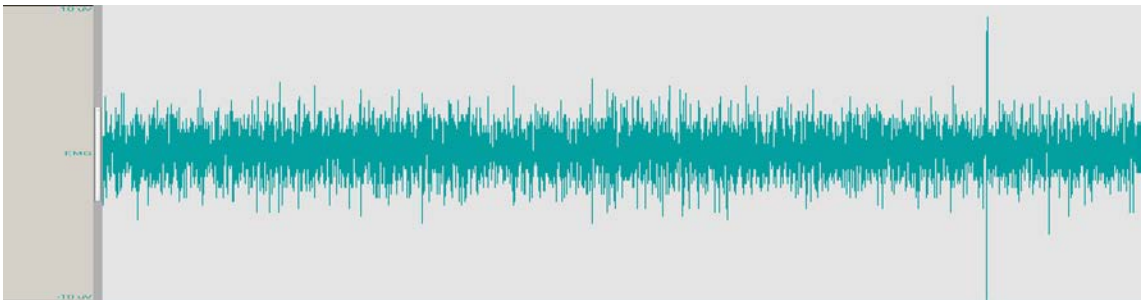


Figure 2.3. Image of a submental EMG

Electroencephalogram

Brain cells communicate producing little electrical impulses. For the recording of EEG several electrodes are placed in the scalp over multiple areas to detect and record electrical activity patterns (see Figure 2.4). Such electrodes are usually arranged in different strategic points along the skull, normally in a bipolar setting, in which one extreme is stick to a certain specific region and the other to a common reference region for all the electrodes. This reference region is usually over the mastoid (M) or the ear lobule (A). The result of the measurement of the potential difference between both electrodes results in the recorded EEG signal. In general the setting up of the electrodes for the monitoring of EEG follows the standard system 10-20 which is shown in Figure 2.5.

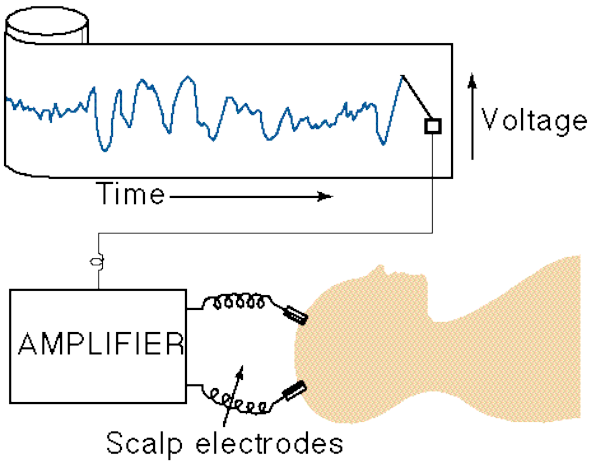


Figure 2.4. General schema of an EEG recording system

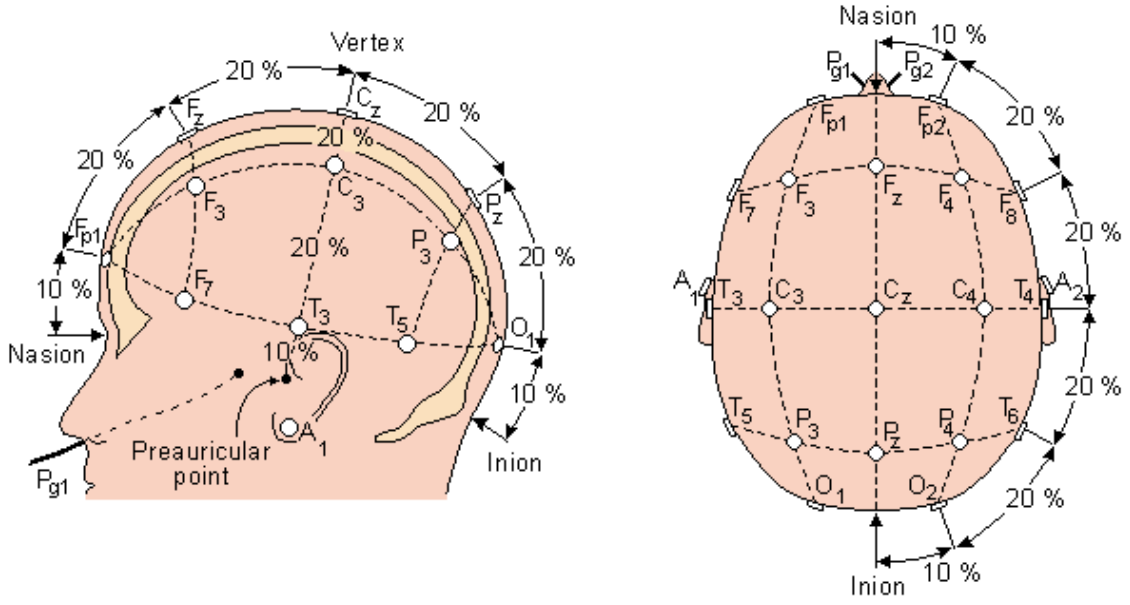


Figure 2.5. The standardized 10-20 electrode system. Figure from the web version of the book by Jaakko Malmivuo & Robert Plonsey [**3**], chapter 13.3

The EEG is, by far, the most complex of the neurophysiological signals involved in the characterization and structuring of the sleep states due to its non-linear nature, non-stationarity, and its low signal-to-noise ratio [**4**]. Usual unavoidable sources of noise in the recording of non-invasive EEG are, for example, cancellation effects produced among the firings of the distinct neurons under the influence of the same electrode, or the effect of the skull which acts as low pass filter between the origin of the signal and the sensor.

Analysis of the EEG during sleep is often based on the analysis of the spectra of dominant frequencies within the signal. In general the sleep spectrum can be broke down into four different categories. The frequency bands designated by these categories are namely alpha (α), beta (β), theta (θ) and delta (δ) (see Table 2.1). Each one of these bands shows different types of activity depending on the current sleep phase of the patient. On the other hand, besides the predominant frequencies, it is also of interest in the EEG the detection of transient events such as K-complexes, sleep spindles or micro-arousals. These events help to classify the different sleep states that the patient undergoes throughout the night.

Table 2.1. Summary of main frequencies of the sleep EEG

	Frequency range	**Amplitude range**	**Observations**
Beta	>16 Hz	5 – 50 µV	-It can be observed during excited EEG and REM phases
Alpha	8-13 Hz	5 – 50 µV	-Localized in the occipital region *Alpha rhythm:* -Is seen in the relaxed waking state with eyes closed -Attenuates with eye opening, anxiety or mental activity such as calculations
Theta	4-8 Hz	50-100 µV	-Appears in low amplitude patterns during drowsiness and phase REM
Delta	<4 Hz	100-200 µV	-It can be observed in deep sleep states

2.2.2. Signals related to respiratory function

Respiratory airflow

Airflow occurs when there is a difference of pressure between the external ambient pressure and the pressure inside the lungs. Air circulates from the region with higher pressure toward the one with lower pressure. The higher the difference, the higher the velocity, being *flow* a physical measure which quantifies volume per unit of time.

Recording of airflow signal has as its objective to obtain a measure of the volume of air inhaled and expelled from the lungs. The result is a sinusoidal signal which reflects respiratory rhythm, usually with a positive slope during inhalation, and with a negative slope during expiration[9] (see Figure 2.6). In the past to obtain this signal a thermocouple was used, which is a transducer formed by the union of two different metals producing a voltage in function of the temperature between the two metals. However the use of thermocouple in the clinical field has been replaced lately by the use of thermistor. The thermistor works as a resistive sensor offering much more precision and it is also simpler to calibrate than the thermocouple. In any case both devices act as temperature sensors and are placed close to the upperway respiratory orifices of the subject, normally in the mouth or in the nostrils. In this respect because

[9] Ultimately this will depend on the used amplifier (if it inverts or not the signal)

of the difference in temperature between inhaled and expelled air, an indirect measure of the air volume breathed in and breathed out is respectively obtained. Both devices – thermocouple and thermistor- are valid for the detection of total absence of respiratory flow –apnea- because of the total absence in the variation of temperature.

However the necessity of measuring also partial pauses –hypopneas- introduced the necessity of using new sensors because, in this case, the relation between temperature change and the actual airflow cannot be so easily measured. That is why according to guidelines proposed by the American Academy of Sleep Medicine (AASM), in order to measure airflow both, nasal cannula with pressure transducer and thermistor, should be simultaneously used [**5**]. Pressure transducer produces a more direct and reliable airflow estimation, which is more sensitive to little changes, and therefore it is better for the detection of hypopneas. The reason why AASM still recommends the use of a thermistor for the detection of apneas may be probably related to legacy reasons and its use will presumably be excluded in future revisions.

On the other hand in patients with Continuous Positive Airway Pressure (CPAP), airflow is commonly measured through a pressure sensor included right in the mask attached to the patient in order to control the CPAP pressure.

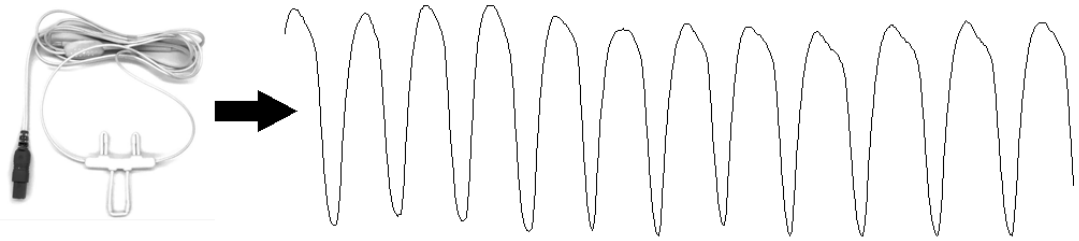


Figure 2.6. Thermistor and airflow signal

Oxygen saturation

Pulse oximeters monitor oxygen saturation in arterial blood in a non-invasive manner. They calculate the percentage of arterial hemoglobin by measuring changes in light absorption resulting from beats in the arterial blood flow (see Figure 2.7). The probe of the pulse oximeter is applied to a body region, normally to a finger or to a toe,

emitting both a red and an infrared light through the skin. The corresponding wavelengths are absorbed respectively by the deoxyhemoglobin and the oxyhemoglobin. Oxygen blood saturation can then be derived using the ratio between the absorbed red light and infrared light.

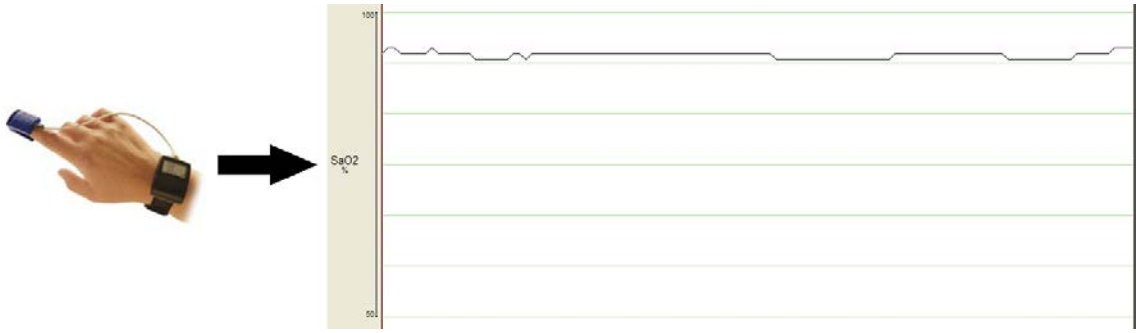


Figure 2.7. Pulse oximeter and oxygen saturation signal

Thoracic and abdominal movements

Non-invasive recording of thoracic and abdominal movement is performed using impedance pneumography, or more recently, through inductive plethysmography [**6**]. The former employs two surface electrodes to record volume changes in the rib cage by the voltage drop across the electrodes. A weak alternating electrical current is passed through the electrodes allowing the impedance to be measured, which increases during inspiration and decreases during expiration. However more recently the use of inductive plethysmography is preferred. Respiratory Inductive Plethysmography (RIP) employs sensors to measure changes in the cross-section area of the rib cage and the abdominal region. The sensors are electrically stimulated generating a magnetic field by the movement occurring during respiration. Alterations in the cross-sectional area change the shape of the magnetic field, thus inducing an opposing current that can be measured. This measure is gathered to generate the movement signal through an oscillator and subsequent frequency demodulation. The generated signal follows a sinusoidal pattern representing ascending and descending movements in each one of the monitored zones (see Figure 2.8). Advantages of RIP under pneumography impedance include less noise sensitivity, better calibration, or the possibility to quantify phase difference between the rib cage and the abdominal region, therefore allowing the classification of apneic events as obstructive, central or mixed.

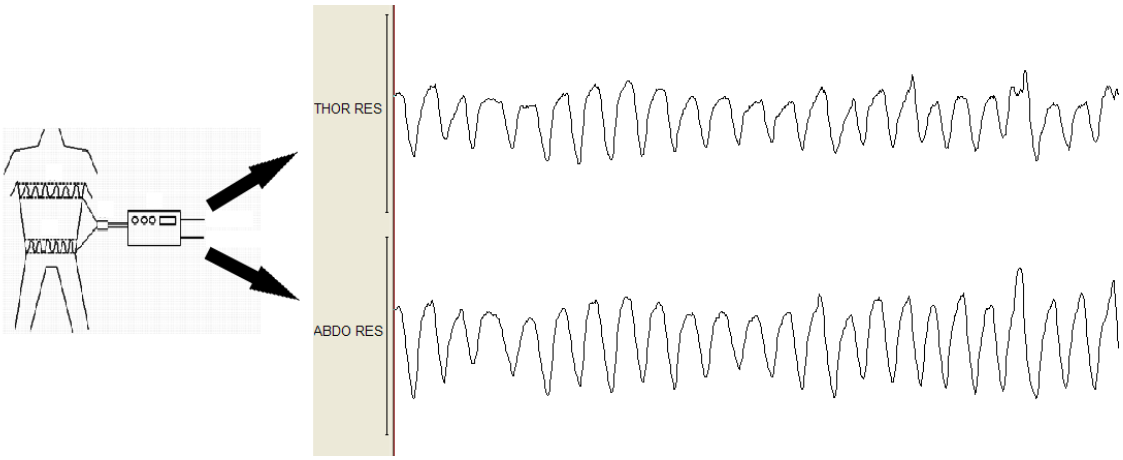


Figure 2.8. Inductive plethysmography and respiratory movements signals. Thor Res = thoracic respiration; Abdo Res = abdominal respiration

2.2.3. Additional contextual signals

Body position

Recording of body position is normally performed through an accelerometer sensor attached to patient's trunk. The continuous signal from the accelerometer is then discretized segmenting it in intervals representing the different sleeping positions of the patient. Normally four fundamental positions are considered: (i) supine, (ii) prone, (iii) left lateral, and (iv) right lateral. However, depending on the desired degree of detail, in some occasions intermediate positions can be differentiated.

Lights control

Lights control signal is mainly used in the sleep labs to keep track of periods in which the patient is already in bed and is about to sleep. That is, it is used to discard intervals in the PSG in which the recording is active but cannot be computed as valid sleep periods. These intervals may obey to different reasons such as periods of bad calibration of the device, interruptions caused by the presence of the technician in the monitoring room or moments in which the patient stands up to go to the bathroom. All these non-valid recording periods are marked by setting the lights recording channel to OFF.

In this respect, lights recording could be considered more as a set of markers or annotations than actually a signal itself. However there also exist approximations implementing lights recording through an ambient light sensor, normally as a function of the present lux, thus resulting in a continuous signal. In these cases a posterior processing of the resulting signal can be performed to segment it in intervals of lights ON and lights OFF according to predefined threshold values.

Snore sound signal

Recording of ambient sound in PSG recordings comes out because of the high snore prevalence in apneic patients, especially in the case of obstructive syndrome, where occlusion in the upper airways produces an obstruction in the airflow causing snoring. Therefore sound recording –normally by an ambient microphone- can be used for the localization of respiratory pauses associated with snoring.

Although recording montages that include the snore sound signal are increasing both in ambulatory and overnight hospital monitoring, the use in practice of the snore signal as an apneic event detection procedure is rather scant. Main reasons include its high sensitivity to noise which decreases its discriminative capabilities[10], and its limitation to be only applied in cases with an obstructive origin, thus not being valid to detect apneic events with a central origin.

Electrocardiogram

ECG signal is usually included among the default recorded signals in the PSG. It does not have a direct implication in the detection of apneic events, moreover its recording may obey more to historical motivations: the ECG is par excellence the vital monitoring signal, with a long tradition, the most studied, with a relative ease of acquisition, and good signal-to-noise ratio. On the other hand it can be used as an indirect measure reflecting the occurrence of apneic events. Effectively in 1984 Guilleminault et al. [**7**] described the cardiac cyclic variation as a pattern reflecting

[10] For example it is difficult to distinguish between the sound caused by an apneic event from that produced by a patient's movement

fluctuations in the heart rate, repeating during each apneic episode, and characteristic of obstructive sleep apnea. Its calculation from the cardiac rhythm is a relatively easy procedure, and in this respect heart rate can be used to detect increases during the hyperventilatory compensation phase prior to respiration recovering. In any case, and although its use for screening purposes is under research because of the nice properties of the ECG signal, its incapacity to reflect central events, as well as the fact of being an indirect measure of the respiratory pause, prevents in practice the use of ECG alone to be enough to obtain a reliable diagnosis of SAHS.

2.3. Structural analysis of sleep

As previously outlined in the introduction, the beginning of modern sleep research dates back to the 1930s, and it is closely related to the invention of the electroencephalography. In 1937, Loomis et al. [**8**] were the first to observe that sleep is not a homogeneous state during the whole night and they described different stages of sleep based on the EEG. In 1953, Aserinsky and Kleitman observed a special state of sleep during which rapid, binocularly symmetrical eye movements occur. It was denominated the rapid eye movement (REM) sleep. During REM state EEG pattern is similar to the one observed during wakefulness, and both respiratory and heart rates are increased in contrast to other sleep stages. Their experiments resulted in a relationship between REM sleep and dreaming: majority of people awakened from REM sleep reported dreams, whereas people awakened during non-REM (NREM) sleep did not recall dreams [**9**]. From the overnight recording of EEG and electrooculogram (EOG), Kleitman and Demet specified the cyclic pattern of REM-NREM sleep [**10**]. Aserinsky and Kleitman also divided NREM sleep into four stages: 1 through 4, ranging from the lightest sleep in stage 1 to the deepest sleep in stage 4.

Traditionally structural analysis of sleep is carried out from the PSG based on three fundamental sources of information which defines it from a physiological point of view: EOG, EMG and EEG, which in the literature can also be found abbreviated as EXG. The popularization of this set of signals dates back to 1968 when a committee co-chaired by Rechtschaffen and Kales (R&K) published "*A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*" [**11**].

The manual comprises parameters, techniques and wave patterns of polysomnographic recordings, and it involved the first standardization on the sleep analysis, lasting almost until today. The purpose of these uniform and standard criteria was to increase the comparability and replicability of the results from different laboratories.

According to R&K criterion, sleep is basically divided in two great stages: REM and NREM. At the same time NREM can be further classified into four stages 1 through 4 according to the findings of Aserinsky and Kleitman. In the R&K manual default EEG derivations are C4/A1 or C3/A2 taking as reference the 10-20 electrode placement system. The potentials for the recording of eye movements are measured from 1cm above and slightly lateral to the outer canthus of one eye, and 1cm below and lateral to the outer canthus of the second eye. The reference electrodes for both eyes are placed on the same ear lobe or mastoid. The EMG is recorded beneath the chin (mental, submental). R&K manual also recommended an epoch-by-epoch approach to scoring, using epochs of 20 or 30 seconds. Table 2.2 summarizes the R&K scoring criteria.

Structural analysis of sleep proposed by R&K was kept unaltered and it was considered the only standard method for around 40 years. It was only recently when the AASM proposed a modification of the scoring method [**5**]. The new manual was published in 2007 aimed at giving a response to the advancing of sleep science, incorporating evolutionary changes as well as newer technical methods and capabilities. In general, the rules and specifications for the scoring of sleep retain much of the framework of R&K, based on the accumulated validity and reliability of this scoring system, with some new definitions and rule modifications, as well as with new rules for pediatric visualization. Arousals, movements, respiratory events and cardiac events are now included into the standardized scoring system. However, the new AASM manual can be considered in many senses more as a specification over the R&K method, rather than a new method itself. Perhaps major modification regarding sleep macrostructure involves the fusion of the old stages S3 and S4 into a new unique stage N3 representing deep sleep. Some other modifications include new recommended derivations for the EEG scoring. For example, now frontal (F4-M1), central (C4-M1) and occipital (O2-M1) derivations with backup electrodes (F3-M2, C3-M2, O1-M2) should be used in order to register the EEG activity. Also three electrodes should be placed in order to record chin EMG, one above and two below the mandible, choosing between one of the

two inferiors to be referenced to the one above, keeping the second as backup electrode. Further specification on recording requirements can be found in [**5**].

Table 2.2. Summary of Rechtschaffen and Kales sleep staging criteria

Sleep Stage	**Scoring Criteria**
Wakefulness	>50% of page (epoch) consists of alpha (8-13 Hz) activity or low voltage, mixed (2-7 Hz) frequency activity
Stage 1	50% of the epoch consists of relatively low voltage mixed (2-7 Hz) activity, and <50% of the epoch contains alpha activity. Slow rolling eye movements lasting several seconds often seen in early stage 1
Stage 2	Appearance of sleep spindles and/or K-complexes and <20% of the epoch may contain high voltage (>75μV, <2 Hz) activity. Sleep spindles and K-complexes each must last >0.5 seconds
Stage 3	20%-50% of the epoch consists of high voltage (>75μV), low frequency (<2 Hz activity)
Stage 4	>50% of the epoch consists of high voltage (>75 μV) <2 Hz delta activity
Stage REM	Relatively low voltage mixed (2-7 Hz) frequency EEG with episodic rapid eye movements and absent or reduced chin EMG activity

In the following main characteristics of each one of sleep stages are enunciated. For its description current classification following AASM manual (W, N1, N2, N3, REM) is used:

- **Stage W**. It represents the waking state, ranging from full alertness through early stages of drowsiness. Electrophysiological and psychophysiological markers of drowsiness may be present during stage W and may persist into stage N1. In stage W, the majority of individuals with eyes closed will demonstrate alpha rhythm: trains of sinusoidal 8-13 Hz activity recorded over the occipital region which attenuates with eye opening. The EEG pattern with eyes open consists of low amplitude activity (chiefly beta and

alpha frequencies) without the rhythmicity of alpha rhythm. The EOG during wakefulness may demonstrate rapid eye blinks at a frequency of about 0.5-2 Hz. As drowsiness develops, the frequency of blinking slows, and eye blinks may be replaced by slow eye movements, even in the presence of continued alpha rhythm. If the eyes are open, voluntary rapid eye movements or reading eye movements may be seen. The chin EMG during stage W is of variable amplitude, but is usually higher than during sleep stages. Figure 2.9 shows the typical picture of the PSG during wakefulness.

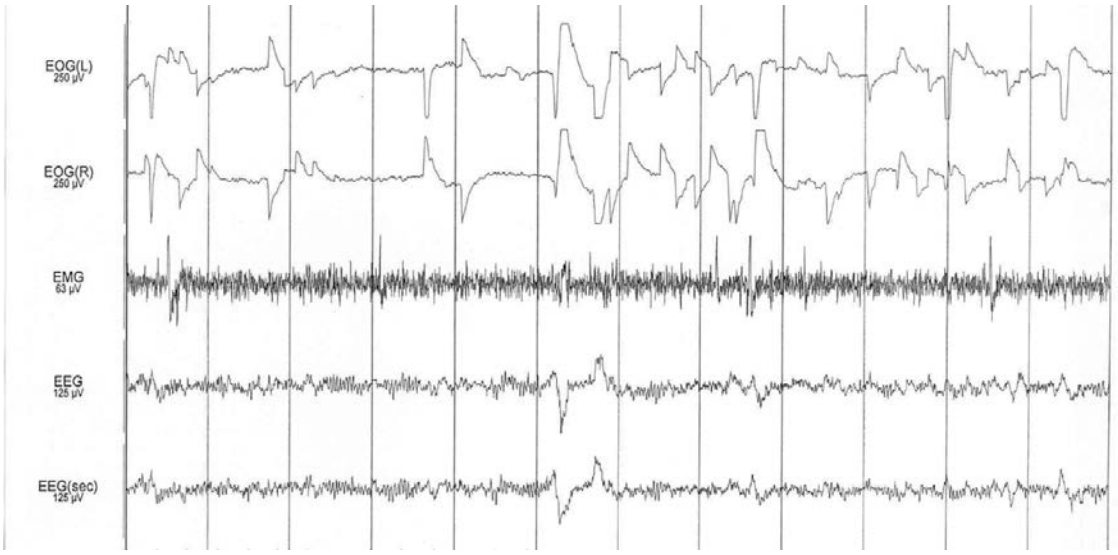


Figure 2.9. Figure shows an example of wakefulness stage. Eyes are open with movement. EOG channels show vertical (in phase) and lateral (out of phase) eye movements. Chin EMG is high. Traces of alpha rhythm can be observed in the EEG. Image adapted from SHHS's manual of operations [**12**].

- **Stage N1**. It is the lightest sleep state in which the subject can still perceive the majority of stimuli which happen around. Sleep in stage N1 is not practically restful at all. In subjects who generate alpha rhythm, N1 stage is scored when alpha rhythm is attenuated and replaced by low amplitude, mixed (4-7 Hz) frequency activity for more than 50% of the epoch. Other hallmarks of N1 sleep stage are the presence of vertex sharp waves and slow eye movements. Vertex waves are sharply contoured waves with duration <0.5 seconds maximal over the central region and distinguishable from the background activity. Slow eye movements are characterized by reasonably regular, sinusoidal eye movements with an initial deflection usually lasting

>500 msec. During stage N1 the chin EMG is variable but often lower than in stage W. An example of PSG recording during N1 is shown in Figure 2.10.

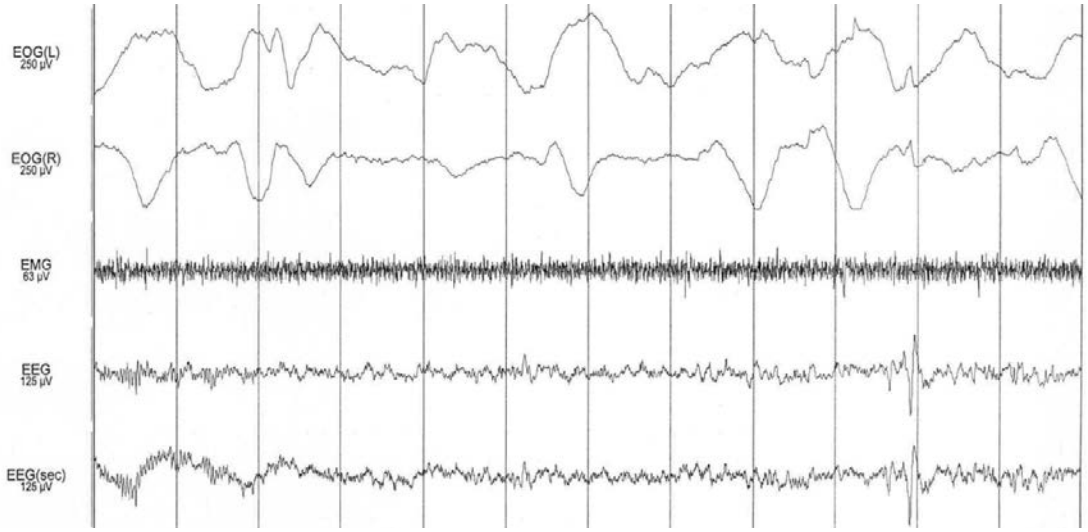


Figure 2.10. Example of N1 sleep state. Slow rolling eye movements are seen on the EOGs. High voltage vertex sharp waves can be observed intermixed with 2-7 Hz activity in the last part of the epoch. Burst of alpha rhythm may still be present in some parts of the epoch. Image adapted from SHHS's manual of operations [**12**].

- **Stage N2**. In this stage a blocking of sensorial inputs at the thalamus level is produced. This blocking entails a disconnection from the environment which facilitates the sleeping process. Sleep in stage N2 is partially recovering which suggest that it is not enough to rest completely. EEG activity during N2 is characterized by low amplitude and mixed frequency with predominance of theta frequency but also delta activity increases with respect to stage N1. However main physiological activity characterizing stage N2 comprises the occurrence of transient sleep spindles events and K-complexes. A sleep spindle is defined as a train of distinct waves with frequency 11-16 Hz (mostly commonly 12-14 Hz) with duration ≥0.5 seconds, usually maximal in amplitude using central derivations. K-complexes are defined as well-delineated negative sharp waves immediately followed by a positive component standing out from the background EEG, with duration ≥0.5 seconds, usually maximal in amplitude when recorded using frontal derivations. EOG usually shows no eye movement activity during stage N2 sleep, but slow eye movements may persists in some

subjects. On the other hand, the chin EMG is of variable amplitude, but is normally lower than in stage W or N1. Figure 2.11 shows an example of typical signal trends during N2.

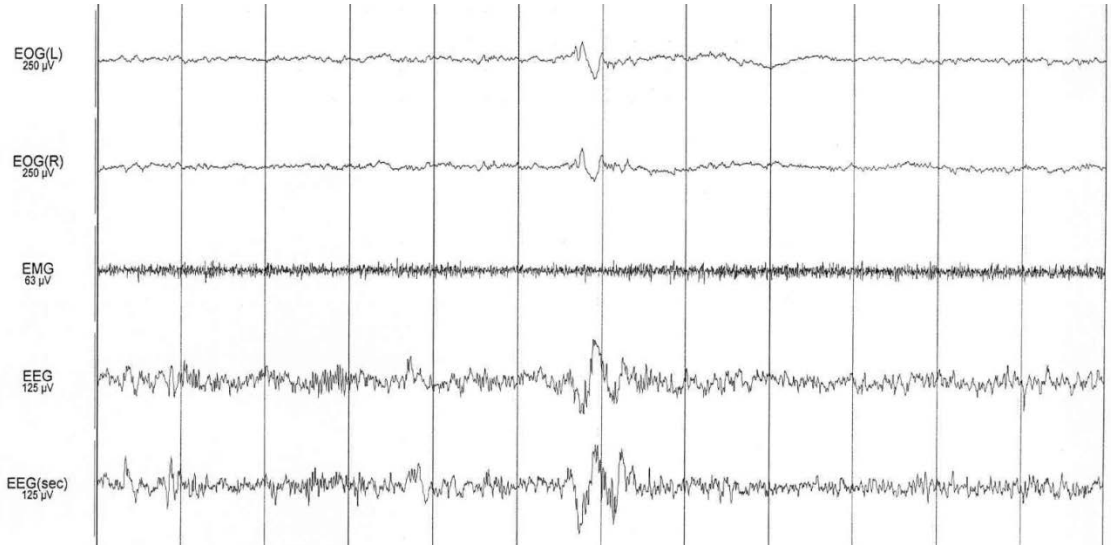


Figure 2.11. Example of N2 sleep state. Background EEG similar to N1 with presence of K-complexes (center) and sleep spindles (first third of the epoch). Chin EMG is more relaxed than in N1. Absence of eye movements. Image adapted from SHHS's manual of operations [**12**].

- **Stage N3**. Sensorial blocking intensifies in this stage in respect to N2 which indicate a deeper sleep. If the subject wakes up in this state he/she will probably experiment confusion and disorientation. Sleep in stage N3 is essential for a restful sleep. In this state EEG activity is characterized by the presence of slow waves with predominance of delta frequency. Slow wave activity comprises waves of frequency 0.5-2 Hz and peak-to-peak amplitude >75 µV, measured over the frontal regions. Normally stage N3 is scored when 20% or more of an epoch consists of slow wave activity, irrespective of age. Sleep spindles may persist in stage N3. Eye movements are not typically seen during stage N3 and they might reflect the EEG pattern (which can also happen in N2). In stage N3, the chin EMG is often lower than in stage N2 and sometimes as low as in stage REM. An example of PSG recording during N3 can be seen in Figure 2.12.

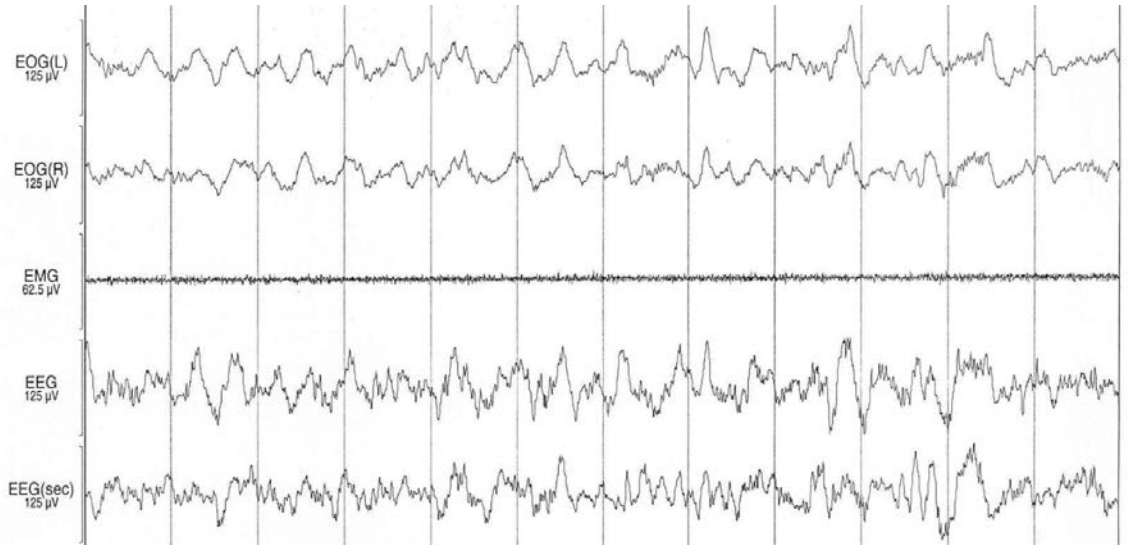


Figure 2.12. Figure shows a typical epoch of N3 state. High amplitude delta waves are present in the EEG which reflect in the EOG (do not confuse with eye movements). Chin EMG is even more relaxed than in N2. Image adapted from SHHS's manual of operations [**12**].

- **Stage REM**. It is the phase where we typically dream. Cerebral activity in REM stage is fast, with low amplitude and mixed frequency with predominance of theta activity and possible presence of beta bursts. It resembles activity seen in stage N1. A typical transient pattern of EEG activity is the presence of sawtooth waves. A sawtooth wave is a train of sharply contoured or triangular, often serrated, 2-6 Hz waves maximal in amplitude over the central head regions. In some individuals a greater amount of alpha activity can be seen in stage REM than in stage N1, however alpha frequency in stage REM often is 1-2 Hz slower than during wakefulness. In the EOG rapid eye movements are characteristic of this phase which can be identified as conjugate, irregular, sharply peaked eye movements with an initial deflection usually lasting <500 msec. Transient muscle activity is also usual in the EMG which on the other hand normally reaches its lowest amplitude levels. The transient muscle activity appears as short irregular bursts of EMG activity usually with duration <0.25 seconds superimposed on low EMG tone. This activity is maximal in association with rapid eye movements. It is interesting in the scope of this thesis to comment that because of the absence of muscle tone, the possibility of occurrence of an obstruction of the upper airway tract increases during REM. Thus it is a period of special relevance for the diagnosis of SAHS. Figure 2.13 shows an interval of PSG during REM.

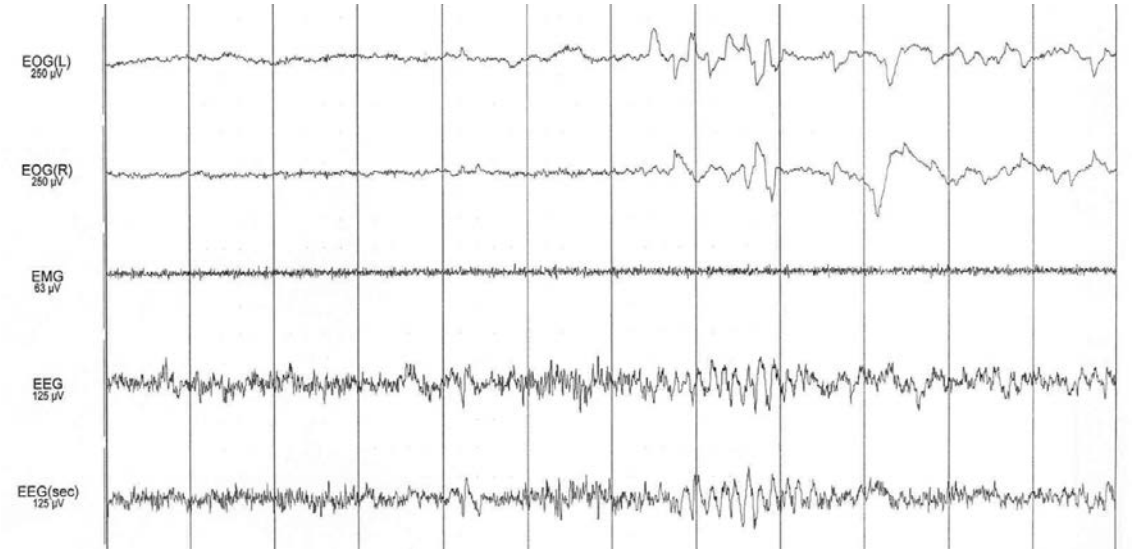


Figure 2.13. Figure shows an example of REM stage. Mixed frequency EEG activity, including saw-tooth waves (between seven and eight vertical bars). Chin tone is at its lowest level. Saccadic, rapid eye movements are apparent on the EOG channels. Image adapted from SHHS's manual of operations [**12**].

2.3.1. The normal sleep cycle

From the overnight recording of EEG and EOG, Kleitman and Demet specified the cyclic pattern of REM-NREM sleep [**10**]. During the sleeping process a cyclic alternation of the sleep states takes place. In the sleep onset the normal individual progressively develops the four phases of NREM and then the first REM block appears. This process constitutes the first sleep cycle. One cycle of NREM lasts about 90-100 minutes and during the night, 4-5 of these cycles occur.

Normal adult sleep oscillates between seven and ten hours. Along this period an opposite distribution exists between REM state and phases of slow wave sleep (SWS): in the early hours of sleep SWS dominates, whereas REM sleep occurs more often in the second part of sleep. On the other hand in the second half of the night the contrary takes place: REM sleep is abundant with periods of progressive longer duration, being the last one about 25-30 minutes, whereas there is almost an absence of slow waves. The portion of REM sleep during night also alters with age: in new-born babies REM sleep lasts for 50% of the night, in adults approximately for 20%.

Stage N1 concentrates at the beginning of sleep and after awakenings in-between. It is a transition stage whose normal latency is about 5 and 25 minutes. In relation with stage N2, the later manifests throughout all the night occupying approximately 45% of the total time. Deep sleep stages on the other hand, concentrates in the first half and they occupy between 15% and 20% of total sleep time. REM sleep cyclically repeats throughout the night, approximately every 90 minutes, involving around 20% or 25% of the total sleep time in the adult. Finally, within the sleep time the amount of wakefulness does not often exceed the 5% in normal conditions and it normally happens because of brief awakenings that the person does not even notice about.

On the other hand, it is interesting to comment that normal sleep structure alters when a person has slept less than usual in the preceding nights. In these cases sleep does not recover in amount but in quality. In this respect what is normally modified with respect to the normal sleep pattern is the increasing proportion of both slow wave sleep and REM.

The results obtained after visual analysis of the recordings allows the construction of a graphical representation of the different states and sleep phases throughout the night. This representation receives the name of hypnogram, and it facilitates the study of normal and pathologic sleep by giving a general vision of the sleep macrostructure (see Figure 2.14). The temporal axis is segmented according to arbitrary units called epochs with duration of 30 seconds each.

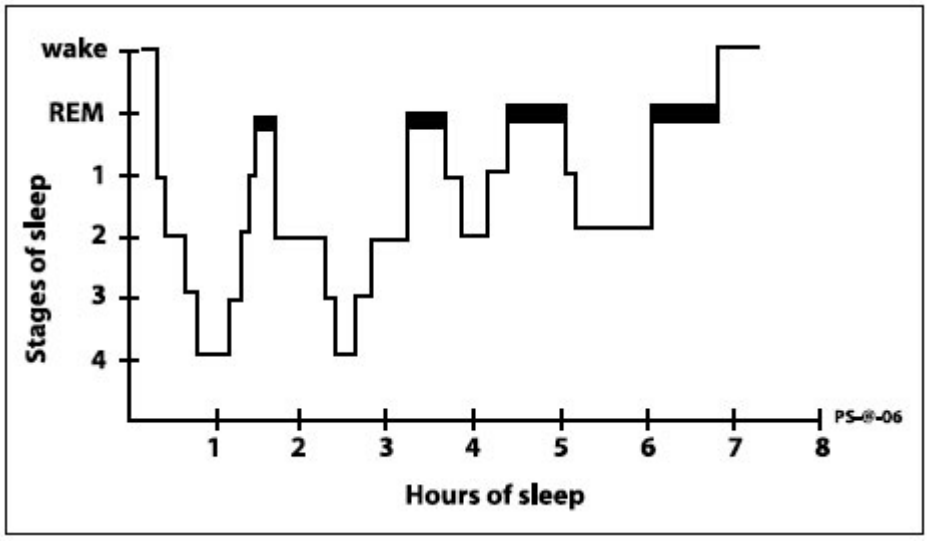


Figure 2.14. Example of R&K hypnogram showing a normal sleep pattern.

It is necessary to point out that representation of the sleep structure as previously described corresponds with that proposed by R&K-like standards. However although such structuring has been widely used as the standard criterion for more than 40 years, it has also been criticized many times [**13**] [**14**]. One of its main limitations relies on its organization based on the scoring of discrete states every 30 seconds. It is well-known that biological processes do not usually occur under discrete phenomena, but by means of continuous transitions between the different states. Therefore sleep structuring according to epoch-based classification rules results somewhat unnatural. Moreover characterization of the different sleep states can be especially difficult because of the variability present among the different subjects, or because of the high sensitivity of the signals to the presence of noise. In addition standard R&K-like rules are not always precise at the time of performing the classification in a number of situations, leading to subjective interpretations by the clinician. There has been a lot of research under the topic of improving the current sleep scoring standards. Further discussion on this topic is addressed throughout the subsequent chapters of this doctoral thesis. In fact with respect to this problem, the proposed system carries out its own tentative solution (see Chapter 5, *Hypnogram generation*).

2.3.2. Alterations of the normal sleep cycle

In the previous section normal sleep structure has been described. In this respect by normal one may understand the medium age adult subject, with healthy habits and without apparent sleep pathologies.

However sleep structure is not a fixed pattern that repeats itself every night in the same manner. On the contrary several variations alter the ideal structure depending on a number of factors, even among non pathological subjects. A clear example of the previous are the variations produced with age. In this respect, it is known for example that increasing age carries out associated a reduction in the electroencephalographic slow-wave activity (SWA) and in spindle frequency activity. Increasing in the number of involuntary awakenings during sleep has been also reported to represent one of the hallmarks of age within human sleep alterations. Thus, all the previous imply a general decrease in the NREM sleep consolidation [**15**] [**16**] [**17**].

It has been discussed a lot about the gender factor as another possible parameter influencing sleep structure. Several studies point out to differences both at the microstructure [**18**] and at the macrostructure levels [**19**]. These differences usually account for females having higher proportion of slow wave sleep EEG than males [**20**]. On the other hand there are studies pointing out to the fact that gender differences in the sleep structure are rather caused by current limitations in the analysis methods and not because of mechanisms inherent to the proper sleep physiology. According to Kemp et al, for example, current methods for both visual and computer analysis of sleep -which are based on quantification of frequency power and/or amplitude, e.g. using Fourier analysis- are influenced by factors unrelated to the sleep process [**21**]. These non-sleep-related factors include brain anatomic orientation, thickness of the skull, gender or headside. However the same study supports the age to actually be a sleep-related factor.

In any case, differences caused by these purely *circumstantial* factors can also be included within the parameters of normality, i.e. they are not considered as pathological. It is in the pathological patient where, on the other hand, the associated alterations have important implications. Figure 2.15 shows an example of the typical hypnogram of a patient with sleep problems in which there can be observed clear differences with respect to the structural pattern described in the previous section.

Figure 2.15. Figure shows image of a fragmented hypnogram. The patient experiences continuous awakenings that break up the normal sleep structure.

Indeed, the patient presenting pathological alterations in the sleep tends to exhibit a high fragmented hypnogram, with constant awakenings preventing a restful sleep to be conciliated. In the apneic patient, for example, constant respiratory pauses follow each other causing a biological response that increments the alertness level. This situation ultimately produces an abnormal incidence of micro-arousals breaking up the normal sleep pattern. Ondze et al. concluded that even in mild disordered breathing subjects,

sleep architecture is characterized by a high degree of fragmentation, resulting in a different time course of SWA and a decreased sleep spindle index when compared to controls [**22**].

In patients with congestive heart failure it is also common the presence of a form of periodic breathing with a crescendo-decrescendo alteration in tidal volume separated by periods of apnea. It receives the name of Cheyne-Stokes Respiration (CSR). Patients with CSR usually show fragmented sleep with frequent arousals, and nocturnal oxygen desaturations leading to poor sleep efficiency [**23**].

Alterations in the normal sleep can be associated to psychiatric problems as well. There are reports of sleep spindles and K-complexes varying together, for example, in the case of dementia where both spindles and K-complexes are reduced. Increase of arousal index or abnormal REM activity can also be influenced among patients suffering from depression [**24**].

Other sources of alteration are related to person's life style and habits, which can influence then sleep cycle even within the same normal subject. For example people who work in nocturnal shift or alternate frequently in their work shift show alterations in their normal sleep-wake cycles. These modifications have much to do with homeostatic and circadian regulation of human sleep. The homeostatic process takes control of the amount of sleep and wakefulness [**25**]. It increases fatigue and sleep propensity during wakefulness and it decreases them during sleep. The indicator of homeostatic regulation is the slow wave sleep (SWS) which is significantly enhanced during the recovery night after sleep deprivation, which in contrast is attenuated by daytime naps. The circadian process on the other hand reflects the influence of external events which oscillate with the circadian rhythm. It represents the alternation of sleep propensity within 24 hours [**26**]. Therefore circadian rhythm sleep disorders usually occur when there is an alteration of the internal timing mechanism or a misalignment between sleep and the 24-h social and physical environment.

As it has been shown, alterations of the normal sleep cycle can be due to a number of causes. In the next section sleep disorders are introduced which, as it has been mentioned already, constitute an important source of alterations in the normal sleep

pattern. In fact normal sleep architecture is the exception in the pathological patient. The former is an important fact to be taken into account when dealing with the construction of an automatic method to characterize the sleep process. Indeed as it will be shown throughout the next chapters, the complication at the time of constructing an automatic sleep analyzer lies mostly in the complexity and variability of the sleep patterns among the different subjects.

2.4. Sleep disorders

A good nocturnal rest is essential to carry out a full and healthy life. The time period necessary for a good rest ultimately depends on the particular person; however it is situated somewhere in-between seven or ten hours. Even though some people can get used to sleep fewer hours than what would be normally necessary, this ends up by affecting their judgment, reaction time and many other fundamental functions during daytime [**27**].

Many studies have proved that the lack of sleep is dangerous. Persons with lack of sleep have been exposed to driving simulators and manual and to ocular coordination test. The results showed performance as bad as –or even worse than- people in an inebriation state. Lack of sleep has also been reported to make worse alcohol effects [**28**].

According to the National Heart Lung and Blood Institute (NHLBI) an estimated 50-70 million Americans chronically suffer from sleep or circadian disorder [**29**]. An estimated 25-30% of the general adult population and a comparable percentage of children and adolescents are affected by decrements in the sleep health, which is a proven contributor to disability, morbidity and mortality. Studies carried out in Europe conclude that over 30% of the population is affected by sleep problems [**30**]. Concretely Spain is one of the countries more affected by sleep problems. Prevalence is higher than 30% of the population, however only around 5% of the cases are actually diagnosed and receive treatment in the sleep labs [**31**] [**32**] [**33**].

There are more than 70 different sleep disorders [**34**]. The most common symptom is insomnia, and as a consequence, daytime sleepiness. Severity associated to the different sleep disorders is uneven and many of them can be satisfactorily controlled once they are diagnosed.

Without the aim to be exhaustive some of the most common sleep disorders are subsequently outlined. Further details as well as the complete list of officially recognized sleep disorders can be found in [**34**]:

- **Insomnia**: More than a disorder itself the general term insomnia rather refers to a common symptom characterized by the difficulty in falling or staying asleep. In fact depending on the subjacent cause several specific types of insomnia can be recognized. However in general difficulty to fall asleep is more common among the young population, whereas in old persons the difficulty lies in staying asleep. Prevalence is usually higher among females. In addition smokers often tend to present sleep architecture with fewer deep stages and less REM sleep – they even may experience awakenings every 3 or 4 hours due to nicotine abstinence. For short-term insomnia, good sleep habits have demonstrated to be useful to relieve or even to cure sleep deprivation. In severe cases some other – more experimental- treatments include light therapy and medication.

- **Periodic Limb Movement (PLM) Disorder:** It is characterized by periodic episodes of repetitive and highly stereotyped limb movements that occur during sleep. The movements usually occur in the legs and they consist of extension of the big toe in combination with partial flexion of the ankle, knee, and sometimes hip. Similar movements can occur in the upper limbs. The movements are often associated with a partial arousal or awakening; however, the patient is usually unaware of the limb movements or the frequent sleep disruption. Individuals with restless legs syndrome[11] usually have periodic leg movements detected during polysomnography monitoring. PLM can accompany narcolepsy and the obstructive sleep apnea syndrome. Prevalence is found to be rare in children and progresses with advancing age to become a common finding in up to 34% of

[11] A different sleep disorder characterized by disagreeable leg sensations that usually occur prior to sleep onset and that cause an almost irresistible urge to move the legs

patients over age of 60 years. No sex differences have been reported. In PSG PLMs can appear immediately with the onset of non-REM N1 sleep, are frequent during N2 sleep stage and decrease during deep sleep. PLMs are usually absent during REM sleep.

- **Narcolepsy**: Narcolepsy is a disorder of unknown etiology that is characterized by excessive sleepiness that typically is associated with cataplexy and other REM-sleep phenomena, such as sleep paralysis and hypnagogic hallucinations. The excessive sleepiness of narcolepsy is characterized by repeated episodes of naps or lapses into sleep of short duration (usually less than one hour). The narcoleptic patient typically sleeps for 10 to 20 minutes and awakens refreshed but within the next two to three hours begins to feel sleepy again. The patients can often tolerate the sleepiness if, with much effort and attention, they make a strong attempt to stay awake. Eventually, however, it is impossible to combat the recurrent daily sleepiness. A history of cataplexy is a characteristic and unique feature of narcolepsy. It is characterized by sudden loss of muscle tone provoked by strong emotion. Narcolepsy can be inherited but sometimes is associated with brain damage. Medication such as stimulants and antidepressants can help to control the symptoms.

- **Night terrors**: Also known as sleep terrors, they are characterized by a sudden arousal from slow-wave sleep with a piercing scream or cry, accompanied by autonomic and behavioral manifestations of intense fear. Sleep terrors manifest as a severe autonomic discharge which can include tachycardia, tachypnea, flushing of the skin or increased muscle tone. The patient usually sits up in bed, is unresponsive to external stimuli, and, if awakened, is confused and disoriented. Night terrors are typically observed in children between the ages of 4 and 12 and tend to resolve spontaneously during adolescence. They are more frequent in males than in females. Night terrors begin in deep sleep (N3), usually in the first third of the major sleep episode. However, episodes can occur in slow-wave sleep at any time.

- **Sleep Apnea-Hypopnea Syndrome:** It is by far the most common sleep disorder and it is characterized by the occurrence of involuntary respiratory pauses during sleep. Its diagnosis is the main objective of this doctoral thesis,

and due to its relative importance, its characteristics are further discussed in the subsequent section.

2.5. The Sleep Apnea-Hypopnea Syndrome (SAHS)

The general name of Sleep Apnea-Hypopnea Syndrome (SAHS) is usually used in the literature to refer to a syndrome which is characterized by the repeated occurrence of episodes of total or partial reduction in patient's respiration during the night. As stated before, SAHS is by far the most common of the disorders affecting sleep. Several studies have been carried around the world during the last years, which estimate that the prevalence of SAHS is between the 3% and the 7% of the adult population [**35**] [**36**]. In Spain it is estimated that between 2 and 3 million people –around the 3% and the 6% of the population- suffer from this syndrome, from which only one in every ten is actually diagnosed and treated. Besides it is estimated that around the 25% of these patients experience a severe or a very severe type of the syndrome [**31**].

Historical origin of SAHS is attributed to two different European researching groups which independently described its symptoms for the first time in 1965: Gastaut et al. [**37**] in France and Jung and Khuhlo [**38**] in Germany who informed of their respective findings about the Pickwickian Syndrome of Sleep Apneas. Such a name was given from the term coined by Burwell et al. [**39**] in 1956, in honor to the character Joe, the sleepy boy from *The Posthumous Papers of the Pickwick Club* by Charles Dickens [**40**]. However it was not until 1973 when Guilleminault et al. [**41**] formally characterized the apneic event[12] regarding its duration and type.

Common terminology with respect to SAHS is often inconsistent and confusing, including different terms like *sleep apnea*, *obstructive apnea*, *upper airway apnea*, *hypersomnia sleep apnea syndrome*, *sleep hypopnea syndrome* or *obesity hypoventilation syndrome*, among others. Actually many of the previous terms rather refer to more specific denominations of the same syndrome. Definitions used within this doctoral manuscript are aimed to be the most widespread, and in this respect, from now

[12] Term that will be used to refer the individual occurrence of an episode of respiration reduction

on the term SAHS will be used to refer to the disease from a general point of view. More specific terms will be used throughout this section when assessing subclassification of the different types of SAHS.

Throughout the rest of the chapter SAHS is introduced in more detail, describing first its physiopathology and diagnostic procedure. The different types of apneic events are then introduced that lead to the classification of SAHS according to the predominant type of apneic event in the patient. Interpretation of the apneic events in the context of the remaining signals of the PSG is subsequently assessed. Discussion ends up with a brief introduction to the different treatment options once SAHS has been diagnosed.

2.5.1. Physiopathology and diagnosis

Patients suffering from sleep apnea present involuntary respiratory pauses that repeats throughout the night. Its duration is variable and it depends on the concrete patient, however such duration must be of at least 10 seconds and it should not exceed the 2 minutes. Typical duration is about 20 to 40 seconds. A distinction is made within the apneic event, mainly attending to the associated degree of reduction in the airflow. In this respect baseline breathing should be determined which is defined as a period of regular breathing with stable oxygen levels [**12**]. In a broad sense a hypopnea is defined as a respiratory pause meeting the duration criteria with an associated reduction around 30%-50% respect to baseline breathing. The exact definition however highly depends on the concrete reference [**42**]. In the case of an apnea the associated reduction is more pronounced and it usually situates about 90% or total breathing cessation. Exact definitions by the AASM can be consulted in [**5**]. In Figure 2.16 a respiratory polygraphy is shown where these two types of events can be identified. It can also be shown in Figure 2.16 that the pauses are usually accompanied by a drop in the oxygen saturation levels, which is proportional to the reduction associated to the causing airflow reduction event. The lack of oxygen in arterial blood usually triggers an autonomic response increasing the alertness level of the individual which often causes neurophysiological awakening [**43**] [**44**]. These associated micro-arousals break up the normal sleep structure preventing the refreshing rest. As a consequence daytime

sleepiness is usual in apneic patients, impacting in their social, working and family life, as well as causing neuropsychiatric and cardiorespiratory disorders.

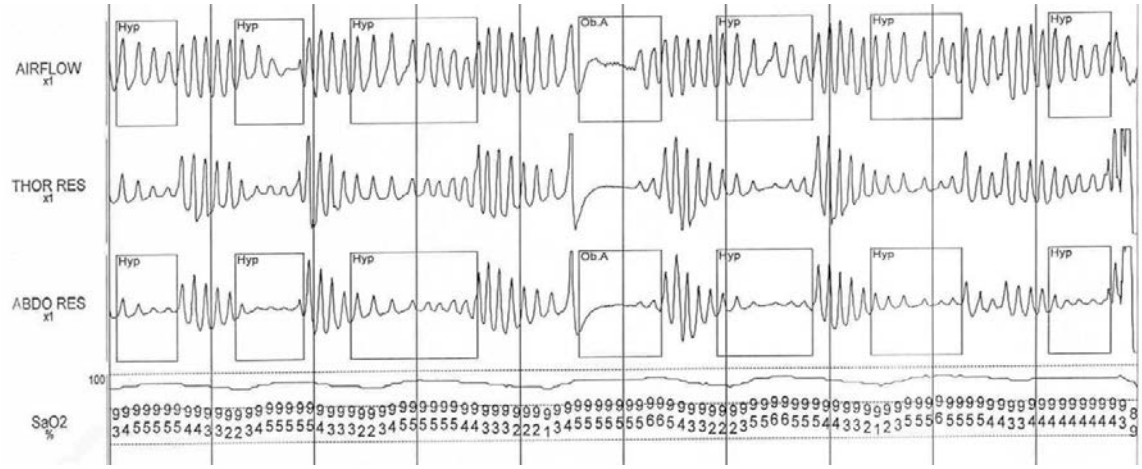


Figure 2.16. PSG where apneic events are marked over respiratory signals; Hyp = Hypopnea; Ob.A = Obstructive Apnea. Image adapted from SHHS's manual of operations [**12**].

Upon awakening, patients typically feel unrefreshed and they may describe feelings of disorientation, grogginess, mental dullness, and incoordination. Severe dryness of the mouth is common and often leads the patient to get something to drink during the night or upon awakening in the morning. Morning headaches, characteristically dull and generalized, are often reported. Excessive sleepiness is one of the most frequently symptoms. One of the main problems of this disease is that patients are usually unaware of their own symptoms. That is the main reason because most of the patients are currently underdiagnosed. Snoring is characteristic in the obstructive patient and it can be so loud that it disturbs the sleep of bedpartners or others sleeping in close proximity. That is why in many times it may be useful to ask to familiars for symptoms, especially to the patient's partner [**34**].

In 1983 Guilleminault et al. [**45**] described cardiac arrhythmias and behavior disorders related to SAHS. This description was followed by several reports searching for cardiac arrhythmia, hypertension, cerebrovascular accidents and sudden death as sequels of SAHS. Nowadays SAHS is associated with an increasing risk of suffering cardiac and cerebral infarct, high arterial pressure, arrhythmias, and in general, several dysfunctions of the cardiorespiratory system [**46**] [**47**] [**48**] [**49**].

Standard diagnostic procedure to determine the presence of SAHS requires of a polysomnographic test to be done during the night. This test is normally carried out in the sleep units of the medical centers, and it involves the recording of several physiological signals during the night, both respiratory and neurophysiological. The resulting recording, namely polysomnographic recording or PSG, is then visually analyzed offline by the medical specialists. Main parameter used for syndrome diagnosis is the Apnea-Hypopnea Index (AHI), which is calculated as the number of apneic events (either apnea or hypopnea) present in the PSG per hour of sleep. Normally an AHI > 10 is of clinical significance although this number can vary according to the followed reference. When no hypopneas are included in the index it is called Apnea Index (ApI) and it usually is relevant when ApI > 5. Assessment of AHI implies manual revision of the PSG recording, considering all evidences present in the respiratory signals and interpreting them according to contextual information of remaining PSG signals. This represents a high cost in double sense: (i) in time, for the length of a full night polysomnographic record which, if printed all continuously, may easily achieve half a kilometer long, and (ii) in effort, for the complexity of the analysis, as well as the amount and diversity of signals involved.

The high cost associated to the visual review of the PSG can eventually degenerate in a loss in the quality of analysis due to the accumulated tiredness throughout the revision task. From the point of view of hospital's administration, it may also represent a waste of personnel resources which could be dedicating their time to other affairs. On the other hand there is a saturation of the sleep units, not being able to support analysis demand. All the previous carries as a consequence the elevation of the economic costs associated to the diagnosis of SAHS.

2.5.2. Apneic event classification

In the previous section a first classification of the different types of apneic events has been performed –i.e. apneas or hypopneas- according to the associated degree of reduction in the respiratory airflow. How this first classification should be actually accomplished may itself represent a source of controversy. Several definitions regarding apnea/hypopnea differentiation can be found in the literature leading to variability in the criteria (see Chapter 1, "*Background*").

Putting aside this discussion, once the apneic event has been detected, it can be further classified according to the nature of its underlying physiological cause. Three types of apneic events can be differentiated in this respect: *obstructive*, *central* and *mixed*. Prevalence of a certain type of event conditions the diagnosis, leading to specification of the concrete syndrome of the patient, respectively, Obstructive Sleep Apnea-Hypopnea Syndrome (OSAHS), Central Sleep Apnea-Hypopnea Syndrome (CSAHS) and Mixed Sleep Apnea-Hypopnea Syndrome (MSAHS). Differences between the different types are subsequently addressed.

Obstructive SAHS

In this case origin of the apneic event is found in the partial or the total obstruction of the upper airways (UAs). An increase in respiratory effort of the patient can be observed as a consequence of the obstruction. In many times such an effort eventually causes the unconscious awakening of the patient.

In Ramirez et al. [**50**] three factors are mentioned which generally determine the proper operation of UAs during sleep: its size, activity and neuromuscular coordination. Pharynx plays an important role in respiration. In pharynx constricting and dilating muscles can be found, the later preventing its collapse during inspiration, together with the action of other non-intrinsic muscles exerting similar function. The role of these muscles lay in the regulation of existing balance between pharynx opening and the negative pressure provoked by the thoracic muscles in the inspiratory act, thus allowing permeability of UAs.

It has been observed that during rapid eye movement phases in sleep there is a decrease in the dilating muscular activity of the pharynx. In certain circumstances that favors the partial or the complete obstruction of UAs. Among them, it can be highlighted the presence of benign or malignant tumefaction, palatal or lingual hypertrophy, macroglossia or vocal chords dysfunction. However obesity is the most often predisposing factor associated with obstructive sleep apnea since overweight favors occlusion of UAs. As a consequence of this occlusion snore is predominant among obstructive patients [**50**].

Some studies have suggested that sex or familial history can also influence as predisposing factors. In adults the male to female ratio is about 2:1. Usual consuming of alcohol and smoking also favors OSHAS [**51**]. With respect to age factors, obstructive sleep apnea can occur at any age, from infancy to old age. However severity tends to increase with age, reaching its peak around the ages of 40 or 60 [**52**]. Women are more likely to develop obstructive sleep apnea after menopause.

Detection of obstructive apneic events through the PSG is carried out by examining the context of the apneic event, searching for signs of presence of respiratory effort. Respiratory effort points out to organism reaction because of obstruction of UAs. As it has been previously described, the widest used method to keep track of respiratory effort is the recording of thoracic and abdominal respiratory activity by inductive plethysmography. In presence of effort the sinusoidal wave representing respectively rib cage and abdominal movements should be discernible in these signals. Even though, similar reduction should be observed in the amplitude of thoracic-abdominal movements with respect to main airflow derivation. In addition, it is important to point out that, because of the obstruction, synchronization of thoracic and abdominal movements usually presents certain phase lag. Figure 2.17 shows an example of an obstructive event in the PSG.

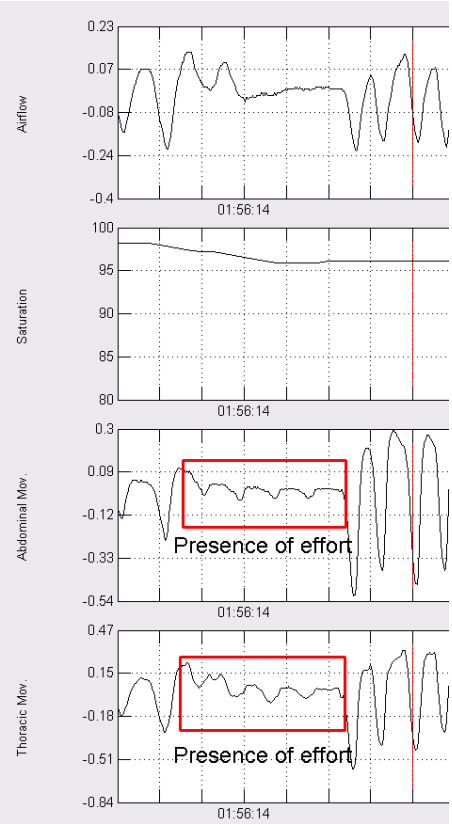


Figure 2.17. Example of an obstructive event, presence of respiratory effort can be seen in both thoracic and abdominal derivations.

Central SAHS

In this case the respiratory pause has normally a neurological origin which is derived from the way the brain controls respiration. The repetitive central sleep apneas appear to be related to the oscillations of a physiologic feedback loop from lung to brain. Therefore the origin is not of obstructive nature but for a temporal loss of the respiratory effort. This circumstance clearly differences central events with respect to obstructive events.

Central apneic events general occur in patients affected by severe lesions in the inferior part of the brain stem which controls the respiratory function. Therefore it generally occurs among seriously diseased people. It may manifest, for example, in patients with bulb poliomyelitis, encephalitis of cerebral stem, neurodegenerative diseases or cerebrovascular accident [**53**].

Patients with pure central syndrome, i.e. almost without presence of obstructive or mixed events, rarely complain about daytime hypersomnolence or necessity of taking naps. On the other hand, its main complain usually includes insomnia and fragmented sleep accompanied by continuous awakenings during the night. Depression is another common symptom among these patients [**34**].

Central apnea can be asymptomatic; therefore, its exact prevalence is unknown. What is known is that prevalence of this kind of apnea within the population is lesser than that of obstructive type (some studies point out to about 12% of total apneic patients [**54**]). This circumstance favors the existence of very few studies with sufficient number of central events, which causes knowledge about this disease to be scant. Central sleep apnea is observed with increasing frequency in the general population as a function of age. In adults, central apneic events appear to be more prevalent in men than in women. After menopause, this difference is less apparent [**34**].

Detection of central events in the PSG is similarly based on the examination of respiratory movements to assess the presence of respiratory effort. In this case both derivations thoracic and abdominal show an almost flat signal, evidence of no respiratory effort (see Figure 2.18). However sometimes little oscillations may be observed due to interference of ECG.

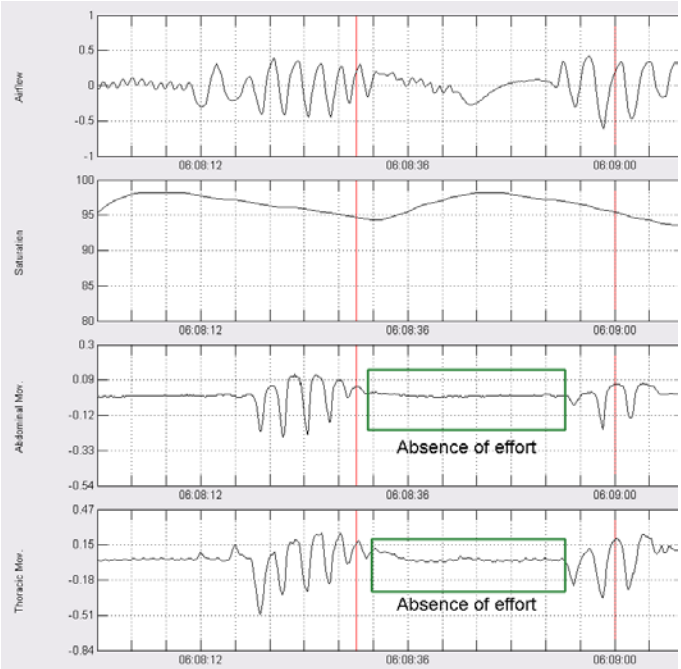


Figure 2.18. Example of central event: both thoracic and abdominal excursions show almost complete absence of movement during occurrence of the apneic event.

Mixed SAHS

Mixed events appear in the PSG as a combination between a central and an obstructive event. Normally the event starts as central and subsequently an obstructive respiratory effort is produced (see Figure 2.19). Muscles of UAs, behaving also as respiratory muscles, dilate the pharynx during inspiration allowing air to come in. If a decrease in their activity occurs as well as in the diaphragm due to a central event, the drop in the muscle tone of the dilating muscles of the pharynx may produce the occlusion of UAs, and as a consequence, the presence of respiratory effort in the resuming of diaphragmatic activity. Mixed events can also be caused artificially by external factors, for example, by the use of CPAP.

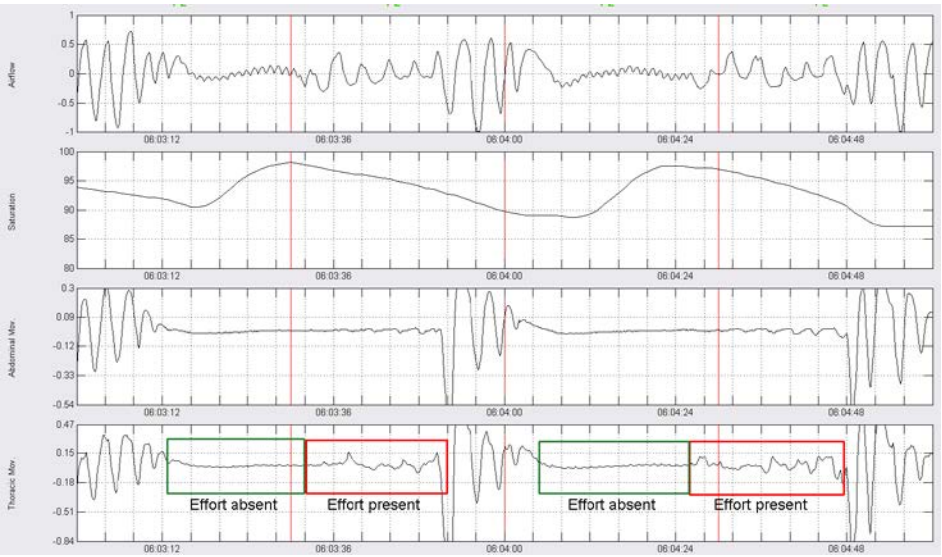


Figure 2.19. Example of mixed events. Events start as central and end with presence of respiratory effort. Classification of the events is clearer from the thoracic derivation.

Declaration of mixed SAHS as a unique clinical syndrome *per se* may generate controversy. In fact mixed apnea type is not recognized by standard guidelines [**34**] as a differentiated syndrome category. Mixed events do not usually appear isolated but often are accompanied by central and obstructive events. The diagnosis for a patient presenting mixed events tends to be either OSAHS, when obstructive events predominate, or CSAHS when most of the events are central. Moreover many clinicians do not even consider the mixed event detection but classify it as obstructive [**12**]. On the other hand some researchers claim about complex sleep apnea syndrome to be identified as a new unique clinical syndrome. Recent studies have been conducted on patients with apparent OSAHS that after elimination of obstructive events using CPAP, have emerged central apneas or Cheyne-Stokes breathing pattern previously unseen [**55**]. For the aims of this doctoral thesis, mixed SAHS will not be considered a separated syndrome category. On the other hand, the designed system will detect mixed events leading the clinician the final decision in the diagnosis (for more details, see Chapter 5).

Figure 2.20 shows a summary table where main characteristics of the different types of apneic events are displayed.

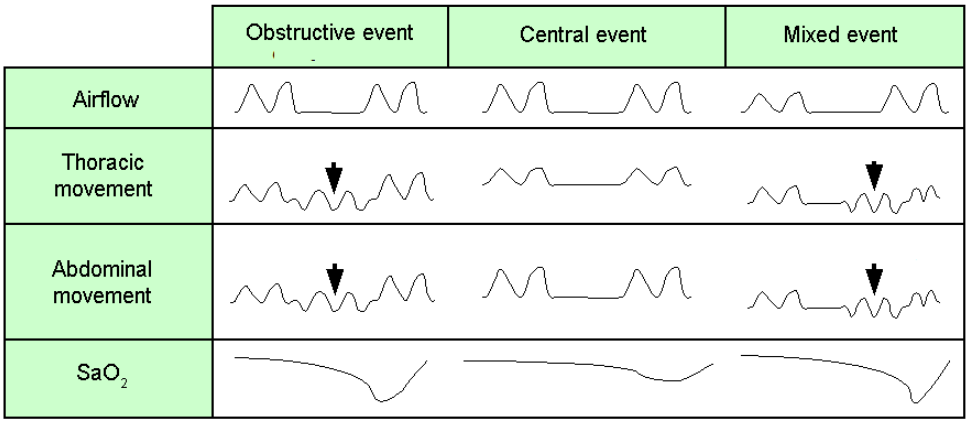


Figure 2.20. Summary of characteristics of the different classes of apneic events

2.5.3. Contextual interpretation of apneic events

In the previous sections a general description of the domain of sleep studies has been given, comprising the most important signals included in the PSG, and describing the different types of apneic events related with SAHS. It has also been pointed out that localization of these events, together with its quantification and its classification, determines the fundamental pieces of information in order to issue a diagnosis and assess its severity. As it has been shown up to this point, principal involved signals in this task are the so-called pulmonological or respiratory signals, mainly comprising airflow, oxygen saturation, and thoracic and abdominal respiratory signals.

However as it has been previously advanced, the solely examination of these signals is not enough for a correct syndrome characterization: interpretation of detected apneic events in the respiratory signals must be always carried out in the context of the remaining signals of the PSG. That comprises the other two previously characterized groups of signals: the neurophysiological signals related with the patient's sleep structure, and the remaining contextual signals of the PSG. In the following the most important relationships that can be established among them are described, as well as how these relationships can influence the correct interpretation of the events detected in the respiratory signals.

First, an important source of information for the interpretation of the apneic events is given through the hypnogram of the patient. In this context an important rule of

interpretation to be taken into account is that, *by definition* [**5**], the occurrence of an apneic event must occur while the patient is asleep. Therefore airflow reductions localized in the periods of stable wakefulness (stage W in the hypnogram) should be discarded as being effectively caused by the occurrence of an apneic event. Besides, it is important to bear in mind that during wakefulness the monitored patient tends to make the normal movements of a waking person. That introduces lots of possible artifact sources which may look like respiratory airflow reductions but, logically, they do not actually have an apneic origin. Thus, it is evident that if these false positives are not discarded, then the final AHI will be considerably increased, and therefore, the presence of SAHS in the patient will be overestimated.

Another important context of interpretation derived from the hypnogram of the patient, has to do with the consideration of the different sleep stages in which the reduction in the airflow takes place. In this respect, it is known that respiratory activity becomes more unstable during REM sleep [**56**]. Additionally, it has to be added that muscular relaxation associated to REM stage favors the collapse of the upper airways [**57**], which makes REM sleep to be a period with a special prevalence of apneic events. In the case of hypopneas, for example, because of the previous, observation of abnormally long periods of hypopnea can be interpreted as normal within periods of REM stage [**12**]. For this reason, which in a different situation could be considered as an abnormal behavior of the respiratory signals, it can be normalized by taking into account the context of REM occurrence in the hypnogram. There are some other cases, such as in transition form a light sleep state to deep sleep, where a slightly reduction in the respiratory signal is normally produced. In this case however the reduction should be related with the phase transition, not being attributed to occurrence of an apneic event. Therefore, in case of being detected, the possible hypopnea should be discarded [**57**].

Also related with sleep structure, but rather related to a microstructure level, the occurrence of EEG arousals constitutes another factor of interest for the detection of apneic events. As it has been introduced already, hypoventilation associated with occurrence of the apneic event produces in many cases a body response increasing alertness level. This is reflected at the level of microstructure by the triggering of a transient arousal event in the EEG. In this regard detection of micro-arousal events may

result in neurophysiological evidence of the immediately previous occurrence of an apneic event in the respiratory signals [**44**]. Moreover, according with the last review of standards for detection and classification of apneic events, the presence of EEG arousal may be required in order to confirm the existence of a hypopneic event type [**5**].

In addition to the context offered by the neurophysiological signals, there are many other factors to be considered that may refine the consideration of an airflow reduction as having an actual apneic origin. Among them, body position of the patient during sleep is one of the most important. Similarly to the case of hypnogram, several studies have been performed that determine that certain positions favor blocking or obstruction of UAs. Sleep in supine position stands out in this regard as a position especially favoring occurrence of apneic events with an obstructive origin [**58**] [**59**]. That is so that literature often refers to the term *positional OSAHS*, in order to quantify the influence that body position exerts over syndrome severity. In this regard to make a classification of apneic events according to body position, often the different positions are grouped in two classes: supine (on the back) and non-supine (any other sleep position), thus defining two positional parameters: the apnea-hypopnea index for non-supine positions (AHI^{NS}), and the apnea-hypopnea index for supine positions (AHI^{S}). Definition of positional categories –positional OSAHS and non-positional OSAHS- is often based on the relationship between both parameters according to Table 2.3. Hence a patient is assigned to the positional group when relationship $\frac{AHI^{NS}}{AHI^{S}} \leq 0.5$ is true. On the contrary if $0.5 < \frac{AHI^{NS}}{AHI^{S}} \leq 2.0$ then the patient is assigned to the non-positional group.

Table 2.3. Definition of positional SAHS categories

$OSAHS_{POSITIONAL} \Leftrightarrow \frac{AHI^{NS}}{AHI^{S}} \leq 0.5$
$OSAHS_{NON\_POSITIONAL} \Leftrightarrow 0.5 < \frac{AHI^{NS}}{AHI^{S}} \leq 2.0$
$OSAHS_{NON-SUPINE\_POSITIONAL} \Leftrightarrow \frac{AHI^{NS}}{AHI^{S}} > 2.0$

As a result, although AHI for non-supine positions is up to two times higher than supine AHI, the patient is still classified in the non-positional group. Only when this ration is higher than two, the possibility of a contrary positional effect is considered [**60**].

With regard to body position, on the other hand, it is important to bear in mind that when an airflow reduction takes place in the context of a sleep position change, this should not be considered as with an apneic origin, but as a consequence of change in the position.

Another contextual signal which may help the interpretation of PSG events is the recording ambient light during the monitoring. Presence of lights on helps localization of recording periods susceptible of being discarded as valid for its scoring. These periods often correspond with periods of sensor calibration or moments in which, for example, the patient lays reading in bed, watching TV or when he/she has gone to bathroom.

The previously mentioned situations may be included within a more general set: the set of intervals in the recording with presence of artifacts. The concept of artifact can be susceptible of interpretation and it is discussed in posterior sections (see Chapter 5, "*Handling of artifacts*"). In any case artifacts can be enunciated from a general point of view, as being those signal intervals influenced by external interferences, hence not being entirely caused by the monitored entity. Examples of artifacts include patient movements, sweating, presence of electrical or magnetic fields, equipment failure, etc. which, ultimately, alter biological measures recorded by the sensor. Localization and interpretation of such intervals can therefore contribute to modify the judgment about the actual origin of data reflected in the signal. An abrupt drop in desaturation levels reaching zero levels for example, should be rather attributed to a sensor failure than to a respiratory obstruction –i.e. at least if the patient is still alive.

Another possible source of contextual information are the cardiorespiratory indexes, derived from the fact that normally pulse rate decreases during the apneic event (brabdycardia) whereas an increase in heart rate is observed near the end of the apnea (tachycardia). This increase peaks during the few breaths after the apnea. The cyclic

behavior of heart rate has been called cyclical variation of heart rate, and it is considered as being specific for the sleep apnea [**7**]. Estimation of the RR intervals from an ECG derivation, or from pulse oximetry, can be therefore performed, and used to calculate the heart rate variability (HRV), which may help to the detect the apneic event.

2.5.4. SAHS treatment

As it has been previously outlined, an apnea index higher or equal to five[13] is considered relevant in order to diagnose SAHS and to start its treatment. It has been proved that ApI > 20 increases associated mortality for the non treated patients, reason why patients under these circumstances are considered as a special risk group which should be urgently treated [**61**].

The class of SAHS with more treatment possibilities is that with an obstructive origin. In this respect, for example for OSAHS, it has been demonstrated that the ambulatory patient may experience considerable improvement by making some changes in his/her daily life [**62**]:

- Several studies have reported about the importance of losing weight as a therapeutic measure in cases of OSAHS obese patients [**63**]. A moderate loose (around 10 Kg) resulted in 50% decrease of computed ApI, improving daytime sleepiness, results of multiple latency test, and stability of oxygen saturation levels in the arterial blood. In cases with drastic weight loose (higher than 50 Kg) total abolition has been even reported. Unfortunately most of the patients are not able to loose or even to keep their weight, hence additional measurements are needed.

- Suppression of alcohol ingestion before going bed is also an important part of the therapy. Alcohol contributes to pharynx muscle relaxation, precipitating upper airways obstruction.

- It is known that supine position favors appearance of the apneic event with respect to lateral positions. Different strategies have been used to prevent

[13] Or ten if we consider AHI

supine position in OSAHS patients. However, although useful in the short term, they do not exhibit clear benefits on the long term.

Treatment of hypertension using medication in patients with OSAHS has not presented excessive good results [**62**].

CPAP is by far the most common therapy applied for treatment of SAHS. It is specially recommended in cases of severe SAHS, and when there have been detected life risk situations such as progressive chronic respiratory failure or severe CO_2 retention. CPAP consist of a ventilatory unit which generates and supervises airflow toward the patient, hence exerting a positive pressure over UAs. Collapsing airways region is therefore pneumatically dilated allowing its opening.

Nasal CPAP has proved to be effective in elimination of both obstructive and mixed apneas. In the case of patients with central syndrome, although not so effective, CPAP is also applied because normally central events rarely appear isolated. In this respect, the improvement is obtained over the suppression of obstructive and mixed events. Complains of the patients regarding use of CPAP usually refer to excessive airflow pressure, or uncomfortability about attaching mechanism of the mask. Severe complications in the therapy with nasal CPAC are extremely rare and they represent isolated cases through literature. In the cases of patients with CPAP intolerance because of increased resistance of UAs while inspiration, second generation mechanisms such as BiPAP can be used, allowing independent adjusting of inspiratory and expiratory pressures.

In extreme OSAHS cases, surgery of pharyngeal region may become an alternative [**64**]. However surgery is not always recommended, being therapy with continuous positive pressure the most extended treatment in most of the cases.

Central sleep apnea still remains considered as relatively rare syndrome whose etiology is not completely defined. Given the range of physiopathologic factors contributing to the varied forms of CSAHS, treatment approaches also vary considerably [**53**]. In any case, the ideal treatment should consider the particular context of each patient, and ultimately it should be supported by results of the global clinical

examination, results of the polysomnographic test, severity of the symptoms, and patient's life habits.

2.6. Summary of this chapter

This chapter carries out the necessary description of the medical context related with the Sleep Apnea-Hypopnea Syndrome. The chapter starts by introducing sleep as a resting state with a repairing function for the human being, and sleep science as the discipline that studies its fundamentals and its biological mechanisms. Sleep science is presented as a relatively young discipline and of increasing interest, thus an object of a great research activity within the last years.

The so called sleep studies are subsequently presented as the clinic tool for the study of sleep and its alterations. The discussion especially focuses on the analysis of the PSG as the *par excellence* clinical test and gold standard for the analysis of sleep. Within the mentioned test, a description of the most important signals involved in its recording is performed, classifying them into three blocks: (1) signals related with patient's neurophysiological sleep, (2) signals for the analysis of the respiratory function, and (3) additional contextual signals.

Once the most important signals involved in the sleep analysis have been described, the chapter continues by introducing the fundamental concepts for the analysis of sleep. From the neurophysiological perspective, the features that determine the sleep structure are described in order to determine the hypnogram. The hypnogram is interpreted as an epoch-based chart that allows representation of the sleep structure throughout the night to be done segmenting it into a series of *sleep phases*.

The main diseases that break up the normal sleep cycle are afterwards outlined prior to fully center the discussion on the particularities of SAHS. In this regard SAHS physiopathology and diagnostic procedure are firstly described. Then a description is given of the different types of apneic events occurring in the context of this syndrome and their classification. Localization of apneic events involves the analysis of the respiratory signals, mainly including airflow, thoracic and abdominal movements, and

arterial blood oxygen saturation. The apneic events with a respiratory origin must be interpreted in the context of the hypnogram and the rest of the PSG signals. The analysis of such relationships is carried out next.

The taxonomy caused by the subclassification of SAHS into more concrete subtypes is an important factor in the diagnostic procedure. Such classification is done according to the main event type taking place in each case. Classification of SAHS is of interest since the nature of each SAHS subclass can have consequences on the posterior treatment. An overview regarding SAHS treatment is given in the last part of the chapter.

2.7. References

[1] P. Wozniak. Good sleep, good learning, good life. [Online]. http://www.supermemo.com/articles/sleep.htm

[2] S. Freud, *La interpretación de los sueños*.: Biblioteca Nueva, 1973.

[3] J. Malmivuo and R. Plonsey, *Bioelectromagnetism - Principles and applications of bioelectric and biomagnetic fields*. New York, US: Oxford University Press, 1995.

[4] W. Klonowski, "Everything you wanted to ask about EEG but were afraid to get the right answer," *Nonlinear Biomedical Physics*, vol. 3, no. 2, 2009.

[5] C. Iber, S. Ancoli-Israel, A. Chesson, and SF. Quan, "The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications," American Academy of Sleep Medicine, Westchester, IL, 2007.

[6] GG. Mazeika and R. Swanson, "Respiratory inductance plethysmography. An introduction," Pro-Tech Services Inc., 2007.

[7] C. Guilleminault, SJ. Connolly, R. Winkle, K. Melvin, and A. Tilkian, "Cyclical variation of the heart rate in sleep apnoea syndrome. Mechanisms and usefulness of 24 h electrocardiography as a screening technique," *The Lancet*, vol. 1, pp. 126-131, 1984.

[8] AL. Loomis, EN. Harvey, and GAI Hobart, "Cerebral stages during sleep, as studied by human brain potentials," *Journal of Experimental Psychology*, vol. 21, pp. 127-144, 1937.

[9] E. Aserinsky and N. Kleitman, "Regularly occurrin periods of eye motility, and concomitant phenomena, during sleep," *Science*, vol. 118, pp. 273-274, 1957.

[10] WC. Dement and N. Kleitman, "Cyclic variations in EEG during sleep and their relation to eye movements, body motility and dreaming," *Electroencephalography and Clinical Neurophysiology*, vol. 9, no. 4, pp. 673-690, 1957.

[11] A. Rechtschaffen and A. Kales, "A manual of standardised terminology techniques and scoring system for sleep stages of human subjects," U.S. Department of Health, Education and Welfare Public Health Service – NIH/NIND, 1968.

[12] Sleep Heart Health Study SRC, "Sleep Heart Health Study. Reading center manual of operations," Case Western Reserve University, Tech Report VMLA-039-02, 2002.

[13] SL. Himanen and J. Hasan, "Limitations of Rechtschaffen and Kales," *Sleep Medicine Reviews*, vol. 4, no. 2, pp. 149-167, 2000.

[14] S. Kubicki, WM. Herrmann, and L. Höller, "Critical comments on the rules by Rechtschaffen and Kales concerning the visual evaluation of EEG records," in *Methods of sleep research*, S. Kubicki and WM Herrmann, Eds. Stuttgart New York: Fischer, 1985, pp. 19-35.

[15] M. Boselli, L. Parrino, A. Smerieri, and MG. Terzano, "Effect of age on EEG arousals in normal sleep," *Sleep*, vol. 21, no. 4, pp. 361-367, 1998.

[16] JC. Principe and JR. Smith, "Sleep spindle characteristics as a function of age," *Sleep*, vol. 5, no. 1, pp. 73-84, 1982.

[17] C. Cajoche, M. Münch, V. Knoblauch, K. Blatter, and A. Wirz-Justice, "Age-related changes in the circadian and homeostatic regulation of human sleep," *Chronobiology International*, vol. 23, no. 1 and 2, pp. 461-474, 2006.

[18] F. Latta, R. Leproult, E. Tasali, E. Hofmann, and E. Van Cauter, "Sex differences in delta and alpha EEG activities in healthy older adults," *Sleep*, vol. 28, no. 12, pp. 1525-1534, 2005.

[19] S. Redline et al., "The effects of age, sex, ethnicity, and sleep-disordered breathing on sleep architecture," *Archives of Internal Medicine*, vol. 164, pp. 406-418, 204.

[20] DJ. Dijk, DGM. Beersma, and GM. Bloem, "Sex differences in the sleep EEG of young adults: visual scoring and spectral analysis," *Sleep*, vol. 12, pp. 500-507, 1989.

[21] MS. Mourtazaev, B. Kemp, AH. Zwinderman, and HAC. Kamphuisen, "Age and gender affect different characteristic of slow waves in the sleep EEG," *Sleep*, vol. 18, no. 7, pp. 557-564, 1995.

[22] B. Ondze, F. Espa, Y. Dauvilliers, M. Billiard, and A. Besset, "Sleep architecture, slow wave activity and sleep spindles in mild sleep disordered breathing," *Clinical Neurophysiology*, vol. 114, pp. 867-874, 2003.

[23] AJ. Quaranta, GE. D'Alonzo, and SL. Krachman, "Cheyne-Stokes respiration during sleep in congestive heart failure," *Chest*, vol. 111, pp. 467-473, 1997.

[24] A. Muzet, "Alteration of sleep microstructure in psychiatric disorders," *Dialogues in clinical neuroscience*, vol. 7, pp. 315-321, 2005.

[25] P. Achermann and AA. Borbély, "Mathematical models of sleep regulation," *Frontiers in Bioscience*, vol. 8, pp. 683-693, 2003.

[26] C. Cajochen and DJ. Dijk, "Electroencephalographic activity during wakefulness, rapid eye movement and non-rapid eye movement sleep in humans: Comparison of their circadian and homeostatic modulation," *Sleep and Biological Rhythms*, vol. 1, pp. 85-95, 2003.

[27] J. Orzel-Gryglewska, "Consequences of sleep deprivation," *International Journal of Occupational Medicine and Enviormental Health*, vol. 23, no. 1, pp. 95-114, 2010.

[28] NB. Powell and JK. Chau, "Sleepy driving," *The Medical clinics of North America*, vol. 94, no. 3, pp. 531-540, 2010.

[29] National Center on Sleep Disorders Research (NCSDR), "2011 National Institutes of Health Sleep Disorders Research Plan," National Heart Lung and Blood Institute (NHLBI), 2011.

[30] D. Léger, B. Poursain, D. Neubauer, and M. Uchiyama, "An international survey of sleeping problems in the general population," *Current Medical Research and Opinion*, vol. 24, no. 1, pp. 307-317, 2008.

[31] J. Durán-Cantolla et al., "Normativa sobre diagnóstico y tratamiento del síndrome de apneas-hipopneas del sueño," Sociedad Española de Neumología y Cirujía Torácica, ISBN 84-7989-152-1, 2010.

[32] JA. Pedregal et al., "Alteraciones del sueño: evolución en una zona básica de salud de Oviedo," *Revista Electrónica de Psiquiatría*, vol. 2, no. 1, 1997.

[33] J. Blanco and R. Mateos, "Prevalencia de trastornos del sueño en el anciano," *Psquiatria.com*, vol. 9, no. 2, 2005.

[34] American Academy of Sleep Medicine, *International classification of sleep disordersm 2nd ed: Diagnostic and coding manual*, 2nd ed. Westchester, IL: American Academy of Sleep Medicine, 2005.

[35] MP. Naresh, "The epidemiology of adult obstructive sleep apnea," *Proceedings of the American Thoracic Society*, vol. 5, pp. 136-143, 2008.

[36] T. Young et al., "The occurrence of sleep-disordered breathing among middle-aged adults," *The New England Journal of Medicine*, vol. 328, pp. 1230-1235, 1993.

[37] H. Gastaut, C. Tassarini, and B. Duron, "Etudes polygraphiques des manifestations episodiques (hypniques et respiratoires) du syndrome de pickwick," *Rev. Neurol.*, vol. 112, pp. 568-579, 1965.

[38] R. Jung and W. Kuhlo, "Neurophysiological studies of abnormal night sleep and the pickwickian syndrome," *Prog. Brain Res.*, vol. 18, pp. 140-159, 1965.

[39] C. Burwell, E. Robin, R. Whaley, and A. Bickelmann, "Extreme obesity associated with alveolar hypoventilation: a pickwickian syndrome," *Am. J. Med.*, vol. 21, pp. 811-818, 1956.

[40] C. Dickens, *Los documentos póstumos del club Pickwick*.: Editorial Planeta 1963, 1836.

[41] C. Guilleminault, F Eldridge, and W. Dement, "Imsomnia with sleep apnea: a new syndrome," *Science*, vol. 181, pp. 856-858, 1973.

[42] M. Moser, B. Phillips, D. Berry, and L. Harbison, "What is hypopnea, anyway?," *Chest*, vol. 105, pp. 426-28, 1994.

[43] DJ. Pitson and JR. Stradling, "Autonomic markers of arousal during sleep in patients undergoing investigation for obstructive sleep apnoea, their relationship to EEG arousals, respiratory events and subjective sleepiness," *Journal of Sleep Research*, vol. 7, pp. 53-59, 1998.

[44] RJ. Thomas, "Arousals in sleep-disordered breathing: patterns and implications," *Sleep*, vol. 26, no. 8, pp. 1042-1047, 2003.

[45] C. Guilleminault, S. Connolly, and R. Winkle, "Cardiac arrhythmia and conduction disturbances during sleep in 400 pacients with sleep apnea syndrome," *Am. J. Cardiol.*, vol. 52, pp. 490-494, 1983.

[46] DJ. Gottlieb et al., "Prospective study of obstructive sleep apnea and incident coronary heart disease and heart failure. The Sleep Heart Health Study," *Circulation*, vol. 122, pp. 352-360, 2010.

[47] S. Redline et al., "Obstructive sleep apnea hypopnea and incident stroke: the Sleep Heart Health Study," *American Journal of Respiratory and Critical Care Medicine*, vol. 182, no. 2, pp. 269-277, 2010.

[48] NM. Punjabi et al., "Sleep-disordered breathing and mortality: a prospective cohort study," *Plos Medicine*, vol. 6, no. 8, pp. 1-8, 2009.

[49] K. Monaha et al., "Triggering of nocturnal arrhythmias by sleep-disordered breathing events," *Journal of the American College of Cardiology*, vol. 54, pp. 1797-1804, 2009.

[50] J. Ramírez, I. Jiménez, and C. Barrego, "Aspectos odontológicos y médicos del síndrome de apnea obstructiva del sueño," *CES Odonto.*, vol. 5, no. 2, pp. 159-167, 1992.

[51] K. Strohl and S. Redline, "State-of-the-art: recognition of obstructive sleep apnea," *Am. J. Respir. Crit. Care Med.*, vol. 154, pp. 279-289, 1996.

[52] I. Khawaja and B. Phillips, "Obstructive sleep apnea: diagnosis and treatment," *Hosp. Med.*, vol. 34, pp. 33-36, 1998.

[53] DJ. Eckert, AS. Jordan, P. Merchia, and A. Malhotra, "Central Sleep Apnea: pathophysiology and treatment," *Chest*, vol. 131, pp. 595-607, 2007.

[54] EO. Bixler, AN. Vgontzas, TT. Have, K. Tyson, and A. Kales, "Effects of age on sleep apnea in men," *American Journal of Respiratory and Critical Care Medicine*, vol. 157, pp. 144-148, 1998.

[55] TI. Morgenthaler, V. Kagramanov, V. Hanak, and PA. Decker, "Complex Sleep Apnea Syndrome: Is it a unique clinical syndrome?," *Sleep*, vol. 29, no. 9, pp. 1203-1209, 2006.

[56] RD. Cartwright, F. Diaz, and S. Lloyd, "The effects of sleep posture and sleep stage on apnea frequency," *Sleep*, vol. 14, no. 4, pp. 351-353, 1991.

[57] M. Cabrero-Canosa et al., "An intelligent system for the detection and interpretation of sleep apneas," *Expert Systems with Applications*, vol. 24, pp. 335-349, 2003.

[58] R. Wietske et al., "The role of sleep position in obstructive sleep apnea syndrome," *Eur Arch Otorhinolaryngol*, vol. 263, pp. 946-950, 2006.

[59] T. Penzel, M. Möller, HF. Becker, L. Knaack, and P. Jörg-Hermann, "Effect of sleep position and sleep stage on the collapsibility of the upper airways in patients with sleep apnea," *Sleep*, vol. 24, no. 1, pp. 90-95, 2001.

[60] DA. Pevernagie and JW. Shepard, "Relations between sleep stage, posture and effective nasal CPAP levels in OSA," *Sleep*, vol. 15, no. 2, pp. 162-167, 1992.

[61] J. He, M. Kryger, and F. Ziric, "Mortality in apnea index in obstructive sleep apnea: experience in 385 male patients," *Chest*, vol. 94, pp. 9-14, 1988.

[62] G. Cardona-Arango and G. Hincapie, "Tratamiento médico del síndrome de apnea obstructiva del sueño," *Acta de Otorrinolaringología y Cirugía de Cabeza y Cuello*, vol. 27, no. 3, 1999.

[63] C. Browman, M. Sampson, and S. Yolles, "Obstructive sleep apnea and body weight," *Chest*, vol. 85, pp. 435-438, 1984.

[64] E. Kezirian and A. Goldberg, "Hypopharryngeal surgery in obstructive sleep apnea: an evidence-based medicine review," *Arch. Otolaryngol Head Neck Surgery*, vol. 132, pp. 206-213, 2006.

3. INTELLIGENT SYSTEMS IN THE DIAGNOSIS OF SAHS

One may define the *clinical decision aiding systems* as computer-based tools providing support for the diagnosis within a certain medical domain. Ultimately, the aim is that of helping the professional clinician who is in charge of making the final decision on the diagnosis. However, within these systems, two different classes can be established with respect to the actual presence of *intelligence* in the supporting task:

- In one hand, one may group under the common class of *indirect supporting systems*, to those systems that base their support by providing routines to help data acquisition and management of the information. These systems usually make extensive use of databases in order to organize information, allowing its posterior access through the use of querying tools, and procedures for automatic report generation. This kind of systems does not necessarily entail the use of artificial intelligence.

- On the other hand, it is inside the group of *direct supporting systems*, where artificial intelligence techniques are extensively used to aid the clinician in medical decision. These systems are knowledge intensive, and as the basic mechanism to support medical decision, they usually try to mimic human's reasoning processes. In some manner, it can be said that within these systems, the computer plays in part the role of clinician, being the main objective the partial or the full automation of the diagnostic task.

As it was introduced in during the preceding chapters, it is the interest of this doctoral thesis to focus in the automatic diagnosis of SAHS. Such a task requires emulation of clinician procedures to examine the PSG, and therefore this second group of systems is preferred. It is the object of this chapter to introduce these systems, firstly

from a general perspective, and afterwards to carry out a specific analysis with regard to the current approaches in the domain of SAHS diagnosis. The corresponding start-of-the-art can be therefore assessed which will lay the foundations for the design of a comprehensive system to aid the physician in the diagnosis of SAHS.

3.1. Artificial intelligence and medicine

Even though historical origins of artificial intelligence –from now on AI- could be dated back to the Greek ages though mythologies of the first androids, it is only around the middle of 20th century when one properly can start talking about AI.

It was in 1943 when three fundamental articles settled the theoretical bases around what nowadays is known as *Cybernetics*. In the first one, by Wierner, Rosenblueth and Bigelow, it is suggested how goals and purposes can be conferred to machines [**1**]. In the second one, McCulloch and Pitts prove how any input-output law can be modeled by using artificial neural networks [**2**]. The last one, owned to Craik, proposes the use of models in order to allow machines to be used as problem solvers [**3**].

However it is necessary to wait until middle of 50s for these bases to abandon the framework of theory and speculation. This coincides with an increment in the processing capabilities of computers, being by this time capable of accomplishing sufficiently complex programming tasks. It was in the meeting celebrated at the Darmouth College in 1956 when the term *artificial intelligence* was coined for the first time. The most outstanding researchers of the moment in the field were cited in this meeting, such as Samuel, McCarthy, Minsky or Newell. They would pass to history as the actual fathers of modern AI.

There have been several definitions of AI through literature, however here the one given by Minsky [**4**] is preferred: *the science of allowing machines to carry out things that would require intelligence if they were made by a human being*. Let us analyze its meaning:

First of all, as a science, AI tries to develop the vocabulary and the concepts necessaries to help understanding, and occasionally mimicking, intelligent behavior. That is, it has as its objective the study of intelligent behavior, which on the other hand may seem a controversy since, actually, intelligence itself is still a concept not fully understood.

In a second place AI can also be interpreted, from an engineering point of view, as a set of methods which allow us to acquire knowledge at the high level, formalize it, represent it efficiently, and use it to solve problems in concrete application domains. And, in this respect, to accomplish more or less complex tasks which might be attributed to an intelligent being. It is precisely with regard to the adjective *intelligent*, where AI can be differentiated from the automatic processes for problem solving, which are merely based on mass processing and intensive calculation attributable to conventional computing systems. In contraposition, the term *heuristic*, bound to the AI program, denotes the capability of guiding the search for the solution to the problem similarly as it does the human expert, that is, without the necessity of evaluating all the possible states of the problem. In other words, in the view of the combinatory explosion of exploring all the possible cases, the success is not based anymore on the mere *brute force* analysis, but on the use of an adequate search strategy according to the actual problem definition.

In this context, during a first period of research the interest on AI focused on the construction of general purpose mechanisms which attempted to concatenate elementary reasoning steps for the constitution of complete solutions. These mechanisms receive the name of *weak methods*, and they are characterized by the fact that domain information is very low, or even inexistent. However as the time passed, application fields of AI diversified, and more specific areas emerged, for example to implement the guiding system of a robot, construction of manufacturing systems, or medical diagnosis. These applications represent more complex domains where specific knowledge is needed. Nevertheless, it has been precisely in these tasks, where AI has experienced its greatest success as an applied science. This kind of AI systems, which make intensive use knowledge as its fundamental, have been known as *Knowledge Based Systems* (KBSs).

The so-called *Expert Systems* (ESs) which are aimed at giving solutions in more specific environments can be interpreted as specializations of the more general KBSs. These systems attempt to act as one more expert in the domain, for which special effort is performed over knowledge acquisition and learning tasks. ESs constitute elegant solutions in well-known, structured and –up to certain point- restricted domains.

With regard to the application of AI in the field of medicine, the beginnings can be dated back to appearance, in 1956, of the first articles mentioning the possibility of constructing automatic systems to aid medical decision [**5**]. Later on, in 1964, the first experimental prototypes began to show its utility in the field [**6**]. The actual successes, however, have to wait for the evolution of AI toward systems being able to manage high level symbolic knowledge. In this respect, the work of Gorry et al. [**7**] analyzes the reasons that motivated the evolution from the first conventional approximations to the new systems managing symbolic knowledge. These systems have been denominated *Medical Expert Systems* (MESs) and are characterized by the use of symbolic reasoning techniques, which produce qualitative judgments based on expertise. This expertise is coded in the form of heuristic rules.

Even though in the middle of 60s the first MES used in the real practice was DENDRAL, the first successful system in the field of medicine, more specifically in the diagnosis of infectious blood diseases, was MYCIN [**8**]. This system, inspired by DENDRAL, and developed by Edward Shortliffe at University of Stanford, has leaved its footprint on history for several reasons. Firstly, it marked the guidelines with regard to separation of the knowledge base from the control structures in expert systems. In addition, it faced the handling of uncertainty and imprecision by means of a novel reasoning schema, which although *ad hoc*, it eventually transcended as a referent for the developing of future systems because of its simplicity and effectiveness. This schema is actually known as the model of *certainty factors* of Shortliffe and Buchanan [**9**].

Decade of 70s was a period of great optimism about the role that MES might develop in the medicine field. Other successful systems of this time were for example CASNET [**10**] or the system of Leeds [**11**].

Nevertheless, during the next three decades, definitive settlement of MESs in the clinical context still not completely occurred. The work of Engle [**12**] analyzes the problematic in the creation and in the development of this kind of systems, especially applied to the diagnostic task, throughout 30 years of experience. Among the main factors causing the rejection of these systems from the part of the clinician, he points out to *non critical* factors such as usability problems or the high consideration the clinicians have over their own capacities. In this respect, it is important to remark that, even with the use of these systems, the last decision on the diagnostic should ultimately remain on the hands of the physician. In other words, it is not about to substitute the clinician in the diagnostic procedure, but aiding him/her to perform the task. On the other hand Engle also points out to *critical factors*, as they are, in his opinion, the impossibility of developing an adequate database or the problematics for the construction of an effective set of decision rules which act within these systems.

In spite of the previous difficulties, nowadays AI in medicine has become a growing and attractive discipline. Construction of hybrid systems which integrate not only the symbolic perspective, but also connectionism techniques of AI such as artificial neural networks or new techniques from the field of machine learning like support vector machines or data mining, open new paths in the field of intelligent monitoring and clinical decision aiding systems. An overview of the different approaches in this regard in carried out in the following section.

3.2. Knowledge and intelligent systems in medicine

Knowledge based systems (KBS) are widely used in the areas where knowledge is predominant instead of data, and where heuristic and logic is required in reasoning to infer new knowledge.

According with definition of Ackoff [**13**], data is classified as raw which simply exists and has no significance beyond its existence, whereas knowledge is the appropriate collection of information which is useful. In the medical field, data and knowledge proportionally integrate for detection, diagnosis, interpretation and treatment

of diseases. Actual proportion between data and knowledge ultimately depends on the problem in hand.

From the point of view of the different computational techniques related to artificial intelligence and its application in the medical field, according to the previous distinction, Pandeley y Mishra [**14**] establish the following groups as pertaining to the class of directed by knowledge: rule based-reasoning (RBR), case-based reasoning (CBR) and model-based reasoning (MBR). On the other hand, data governed methods include artificial neural networks (ANNs), genetic algorithms (GAs) or data mining (DM).

KBS are general purpose problem solvers that depend on a rich base of knowledge to perform difficult tasks. Knowledge is stored in a knowledge base separated from the control and inference mechanisms. Blackboard based architecture is a kind of KBS which uses a form of opportunistic reasoning [**15**]. Knowledge in a KBS may be represented in several forms, for example, trough frames [**16**], Bayesian networks [**17**] or production rules [**18**].

In rule-based systems knowledge is represented by symbolic rules. The inference in the system is performed by a process of chaining through rules recursively, either by backward or forward reasoning [**19**].

Regarding CBR, domain knowledge that is needed to group diagnoses in episodes (events) is eminently implicit. It lends itself for reasoning more based on analogy than in formulation of domain rules or the construction of a model [**20**]. Knowledge is stored in form of cases, and given the presence of a new problem, this is solved in the basis of analogy reusing the past cases. The process within CBR can be divided into five phases: case representation, indexing, matching, adaptation and storage [**21**]. When a new problem arrives, the situation is identified by case representation phase. After that, the features of the new case are assigned to represent it during the indexing phase, and those indexes are then used in the matching phase. According to the similarity of the indexes, the matching phase retrieves similar cases to the base case. Adaptation phase takes advantage of the solutions for similar cases and some suitable adaptations are applied to

solve the new problem. Finally the new case is stored with the base case after the new problem and its solution are confirmed by the user via the storage phase.

In MBR knowledge base is represented through a set of models (satisfying assignments, examples) of the world rather than a logical formula describing it. When a query is presented, reasoning is performed by evaluating the query on these models [**22**].

All the above KBS has some advantages and disadvantages. In this respect, for example, rules of RBRs have the advantage of their ability to express the general knowledge, naturalness of representation, modularity and its disposition for providing explanations. However disadvantages include bottleneck of knowledge acquisition, brittleness of rules, inference efficiency problem, difficulty in maintenance of large rule bases, inability of exploiting problem-solving experience, and interpretation problem. CBRs have advantage of expressing specialized knowledge, naturalness of representation, modularity, easier knowledge acquisition, self updatability and handling of unexpected or missing values; however they also suffer from problems such as inability to express general knowledge, knowledge acquisition problem, in some cases efficiency problem, and inability of explanation [**23**]. MBRs offer enhanced interpretation and explanation power, principled approach that provides the reference for model manipulation and reasoning, provision for the generation or treatment of all cases within a well-defined framework, and handling of unexpected cases. They face however problems such as difficult modeling, lack of model-builders, need for reusable libraries and the need for integration with other methods [**24**].

Due to the different advantages and disadvantages among RBR, CBR and MBR, in the medical domain sometimes it is difficult to solve problems using a unique approximation. On the other hand, by the combination of the different approaches for a concrete problem one can exploit their advantages and minimize the shortages of the individual models. Examples of integration of different KBS approximations can be found in [**25**] and [**26**], which combine RBR and CBR; [**27**] and [**28**] integrate CBR and MBR; in [**29**] combination of RBR, CBR and MBR is carried out.

However, in general, the translation of implicit knowledge into explicit rules – problem of knowledge acquisition- leads to loss and distortion of information content. An alternative to this kind of inference is the use of statistical inference, derived from Baye's theorem, which sets a probabilistic value for each considered output[14]. Examples can be found in [**30**] and [**31**]. This type of expert system can be successfully used for mutually exclusive diseases and independent symptoms, but it fails when some symptoms have the same cause or the patient may suffer from more than one disease. To deal with uncertainty and inexact information artificial intelligence techniques such as theory of evidence [**32**], certainty factors model [**9**] or fuzzy logic [**33**] are rather preferable. More discussion on handling of imprecise information is assessed throughout the next section.

On the other hand there exist many cases in which it is very difficult or directly impossible to implement human intelligence using KBS[15]. This is mainly the field in which connectionist approximations such as ANNs have been developed. ANNs have been widely utilized and accepted methods for the diagnosis in data intensive domains. They are special kind of machine learning models that mimic how the biological neurons work. Basically they are composed of a set of interconnected nodes. Each connection has a weight which is a measure of the relative importance of this connection. Different models of ANNs are available throughout the literature depending on the architecture that mainly differs on the topology and in the activation functions of the processing elements [**34**]. ANNs possess some advantages over RBR: they present a complementary approach with respect to the numeric knowledge representation by the network weights, and the adaptive capability of adjusting the weights based on training is widely regarded as learning-like. Although ANNs have been successfully used in many areas of medicine [**35**], they also possess some drawbacks. In this respect, structure of ANNs is not transparent, approaching behavior of a black-box that simply maps the input to the output; additionally, often expert knowledge cannot be used in order to initialize net parameters for better initializing and improve convergence.

GAs are efficient methods based on principles of natural selection and genetics in which operators of selection, mutation and crossover are applied over a population of

[14] For example to each possible disease in the context of medical diagnosis
[15] Again, the bottleneck of knowledge acquisition

individuals that represent possible solutions to a problem [**36**]. Optimization is carried out by minimizing a fitness function so that on each iteration –generation- best individuals are obtained, thus obtaining better solutions according to the objective function. In the field of medicine, GAs have been used mainly for evaluation of features in diagnosis and treatment, including many applications such as detection of cancer cells in blood and bone marrow [**37**], three-dimensional radiation therapy treatments planning [**38**], stereotactic radiosurgery and radiotherapy [**39**], or detection of mass lesions on digital mammography [**40**]. More applications can be found in the review of Pandey and Mishra [**14**]. Combination of ANNs and GAs is used for optimization of parameters or topology of ANNs since GAs are global searching methods, thus helping to reduce the possibility to obtain suboptimal trainings because of falling into local minima [**41**].

DM is an emerging area of computational intelligence that offers new theories, techniques and tools for the analysis of large datasets [**42**]. In this respect, rather than a concrete technique, DM consists of several approximations within the field of machine learning such as Support Vector Machines (SVMs), decision trees, clustering techniques, and many other related approaches aimed at performing knowledge discovery in datasets. Within the range of DM techniques, feature selection represents a subset of special interest in medicine which is devoted to determine the best set of descriptors of a dataset that identifies to a certain disease [**43**]. DM has been used in the diagnosis and treatment of a number of diseases in the medical domain including diabetes [**44**], pulmonary [**45**], Alzheimer [**46**], heart diseases prediction [**47**] and OSAHS [**48**] [**49**].

As it has been shown, there is a diversity of methods and modeling techniques within the field of intelligent systems and its application to medicine. In general, two great groups of techniques where distinguished specially suited, respectively, for knowledge or data driven domains, each one with their corresponding advantages and disadvantages. However, combination of the two great approximations (knowledge and data) is also possible leading to what is known as hybrid systems [**50**]. Indeed, hybrid systems can exploit the advantages while minimizing the shortages of each of the approximations. The system developed throughout this doctoral thesis, in fact, can be classified as pertaining to this category, since it combines approaches from both

knowledge and data based approaches. For a detailed description of the concrete techniques used within the developed system and their integration, the reader is referred to consult the chapters 4 and 5.

3.3. Handling of imprecise information

Development of computational systems in medicine entails several difficulties due to uncertainty associated to medical domain. Indeed, it is possible to identify different sources of imprecision, which may represent a problem when one has to deal with the construction of automatic reasoning systems in environments involving expert knowledge in medicine:

- Diversity of criteria for the identification of the same event. Although there usually are standards for the identification of the different events and relevant patterns in the task of interpretation, many times the involved definitions are not precise, sometimes implying a series of variable limits around the range of possible values. This kind of variability is usual among experts from the same discipline but pertaining to *different schools*. Moreover, even in the case in which the standard criteria followed for the interpretation is perfectly defined (experts pertain to the *same school*) there is still variability because of the subjectivity of the human expert, and the compiled nature of its knowledge. With the term *inter-expert variability*, one refers to the present variability derived from the discrepancy among different experts within the same area of expertise.

- Subjective variability among experts. In this case variability is associated to the fact that nature of the human being is not of an exact nature. Application of the same knowledge at the time of interpretation is affected by additional factors, such as the time the expert has available to carry out such a task, the degree of tiredness or the state of humor. One talks about *intra-expert variability* referring to the discrepancy that exists between the expert and himself/herself in two different instants of time.

- In general, expert reasoning processes do not obey to exact numeric calculations, but often make use of generalization, expressions containing approximate linguistic terms and ambiguity. Besides, there is a component of imprecision which is bound to the compiled character of the expert knowledge. Experts usually reason on the basis of identification of patterns, trends, and similarity criteria, with respect to previous experience in similar cases. They develop subjective rules which hardly can be quantified in exact or numeric terms.

- In the domains involving diagnosis over biomedical signals –as in the case of SAHS- the variability issue is additionally emphasized by problems inherent to the nature of the recording process: imprecision due to limited sensitivity of the measuring devices, loss of information in the signal digitalization process, as well as the presence of artifacts in the signals such as noise, punctual interferences, bad calibration or loss of focus.

In this context, the use of Bayesian techniques for the managing of probabilities arises as one of the first successful approximations in order to handle imprecision [**51**]. Based on the famous Theorem of Bayes [**52**], the main advantage of the Bayesian methods lies on the use of the probability theory, to avoid the -up to that time- mandatory use of categorical interpretations. However, although the use of the Bayesian approximation constituted an authentic revolution in the development of intelligent systems, the theoretical basis of the Bayesian perspective carries out some limitations. Its main difficulty has to do with the great amount of probabilities that is necessary to obtain in order to build a knowledge base. Additionally the correct application of the Bayesian model requires mutually exclusive hypotheses and conditionally independent evidences. Unfortunately, assumption of conditional independency is rarely valid, moreover, in practice, exclusivity and exhaustivity between hypotheses is often false, being the most current situation the appearance of concurrent and superimposed hypothesis. Another disadvantage is that Bayesian methods do not permit a clear explanation of their conclusions to take place, allowing at the same time a single evidence to support both, one hypothesis and its negation[16].

[16] One should take into consideration that according to axioms of probability, given an event A and its complementary $\neg A$, then $p(A) = 1 - p(\neg A)$

One approximation that tries to solve the problems of the Bayesian method is *belief networks*. A belief network is a special type of influence diagram in which the nodes represent random variables. Pearl [**53**] proved that the use of belief networks allows probabilistic knowledge bases to be constructed without the necessity to impose assumptions on the conditional independence. These networks also guarantee that the evidence favoring to a certain hypothesis does not necessarily imply the partial support of its negation, and that consistent explanations can be obtained by tracking of the different beliefs until the initial points of the network. When the network is constructed in form of non directed graph, these networks receive the name of *Markov networks*. In these cases the association between the variables is considered to be a form of correlation instead of a form of causality.

In parallel to the Bayesian *perspective*, several other models came out allowing handling of uncertainty. The model of *certainty factors* created by Shortliffe and Buchanan [**9**] is a model designed *ad hoc* for the construction of the expert system MYCIN. This model deals with the problem of uncertainty and imprecise knowledge defining two independent measures: the *measure of increasing belief* (MIB) and the *measure of increasing disbelief* (MID). The first one is a dynamic index representing the increment of confidence in one hypothesis given certain evidence. The second is another index representing the increasing amount of disbelief. The two measures can be formally summarized by means of the so-called *certainty factor (CF)*. The CF is conceptually different to the respective conditional probabilities, because corresponding certainty factors to hypotheses H and $\neg H$ are not complementary to the unit, but opposite between them. In this respect, the idea is that if the support provided by the evidence with respect to certain hypothesis is low, accordingly, the corresponding support regarding the negation of the same hypothesis due to the same evidence should not be high.

Theory of evidence of Dempster and Shaffer [**32**] represents a more formal skeleton to conveniently handle both the inexact knowledge and the lack of knowledge. The theory works over a set of hypotheses, without the necessity to distribute the confidence contributed by the individual evidences among the individual hypotheses of the considered set, which receives the name of *universal set*. When several evidences exist,

supporting different groups of contradictory hypotheses, the model normalizes the corresponding results. For that purpose an index is defined –*degree of conflict*- that quantifies the amount of conflict between beliefs assigned to the different subsets of hypothesis. The uncertainty problem is managed using the concept of *confidence interval*, which is constructed using another two measures, *credibility* and *plausibility*, respectively, indicating the minimum and the maximum confidence that can be deposited on a concrete set of hypotheses, in a certain instant of time. The confidence interval dynamically evolves with the appearance of new evidences. A particularly interesting result of the evidential theory is that it contains the model of Shortliffe and Buchanan in the cases in which that model works well, but also in the cases where model of certainty factors present difficulties.

On the other hand, from the knowledge perspective, decision making implies the creation of a list with all the possible strategies and actions, the evaluation of the result of the application of each one, and the selection of the most adequate solution for the concrete case. However, in medicine the previous procedure does not happen in that manner. Physicians almost always work with possibilities rather than with certainties, possibilities over which it is though more in qualitative than in quantitative terms. In this regard expert knowledge often refers to expressions on the style of: certain combinations of symptoms *usually* point out to a certain disease, or *it is known* that *sometimes* a concrete drug causes certain secondary effect but it *rarely* causes another. Always a physician takes a decision, he/she makes a choice between a number of alternatives which are obscured by such qualifications. Hence, it emerges the necessity of designing systems being able to manage such inexact information also from a qualitative point of view, and allow it to be taken into account in the medical decision making.

In this context, the fuzzy logic paradigm [**54**] seems to be more appropriated, because of its capabilities to represent imprecise concepts, as well as allowing propagation of uncertainty in the reasoning processes through the use of fuzzy rules. Fuzzy logic has been used already for the resolution of great variety of problems, and it has been proved successfully in different domains including control of industrial processes, and decision systems in general. Systems based in fuzzy handling of information have also shown their applicability managing imprecision and uncertainty

to aid the physician in the clinical decision making [**55**] [**56**]. The paradigm of fuzzy logic will be used in the construction of the proposed system for the diagnosis of SAHS. In this respect, further detailed description of its fundamentals and underlying mechanisms is performed throughout Chapter 4.

3.4. State-of-the-art in the diagnosis of SAHS

Although compared with other automatic detection systems, computer approaches for the analysis of sleep structure and diagnosis of its diseases may not had had a great historical demand in medicine, this trend has radically changed over the last years. That is in great part because of the emerging importance that today is given to a good nocturnal rest, as a fundamental factor for the developing of a full and satisfactory daytime life. As a consequence, in the last years an increasing demand has been produced in the diagnosis of SAHS and its associated treatment. Accordingly, the number of performed PSGs has also increased, which in the lack of a simpler and more precise test, still today continues to be the only standard procedure for the diagnosis of SAHS. On the other hand, as it was commented already, visual analysis of the PSG is a costly, tedious and long duration procedure. This situation impacts on the daily working life of the specialists in the field, since they have to devote great part of their working time to the analysis. With this precedent, it emerges a great interest in the development of automatic systems for diagnosis of SAHS.

Nowadays, the increasing interest in sleep medicine demands for an important technology support, which due to the complexity of data analysis, it necessarily involves the introduction of automatic analysis systems for aiding in the decision making. However, the developed systems up to this time still present certain drawbacks which have caused their use in practice yet to be low and very restricted (see Chapter one, "*Background*"). These two factors (increasing demands in the analysis and lacks of the current developments) favor that research lines in the development of these systems continue today to be an open area of interest.

In the following an analysis of the current state-of-the-art in the field of automatic diagnosis of SAHS is performed. The analysis takes into consideration both commercial as well as more academic -or research- related realizations.

It is necessary to clarify that in this chapter only those systems or approaches exclusively centered in the diagnosis of SAHS are considered. Relevant references with regard to related PSG analysis tasks –but not exclusively associated to SAHS diagnosis- such as, for example, classification of sleep stages of the patient, or the detection of transient events in the EEG, are discussed in the corresponding sections of Chapter five.

3.4.1. Commercial systems

Even though as it was previously outlined (see Chapter 1, "*Background*") current automatic SAHS diagnostic systems still present some shortages, from the commercial point of view, there has been some time since sleep labs already have systems allowing the digitalization of the PSG signals. For the clinician, the simple fact of being able to carry out an offline analysis over the digitalized signals yet represents an important evolution. Advantages include not only the amount of saved paper[17], but many others such as the possibility to visualize the signals over different time and amplitude scales, easiness in the annotation of the detected events, or the possibility of incorporating supporting tools that automate, at least in part, the diagnostic task.

In the following some of the most relevant commercial systems are introduced, together with a brief description of their capabilities. It does not attempt to be much less an exhaustive enumeration, but to present a summary of the current possibilities of these systems.

[17] In the old paper recordings, a PSG could easily achieve around one kilometer length

PolySmith, Neurotronics, Nihon Kohden, USA

PolySmith is software for recording and reviewing sleep data files. Over thirty years of research has been engineered into the PolySmith automated analysis [**57**] [**58**]. Automatic analysis capabilities include automated analysis of sleep stage, respiratory events, desaturations, leg movements and microarousals. The software allows users to customize preferences including colors, workspaces, filters, and scan settings. In its last version (PolySmith 8.0 released December 2010) it includes new features to support paperless patient questionnaires and new interfaces to integrate portable sleep devices.

No validation studies were able to be found regarding the reliability of its automatic analysis capabilities in the general population. The only found reference is available through its website [**59**] and refers to validation study carried out over 11 pediatric records with ages from 4 to 14. After removal of artifacts and assessing of good signal quality, validation was performed including sleep stage, respiratory event, microarousal and O_2 desaturation agreement. According to published results, agreement indexes regarding respiratory, microarousal and desaturation are comparable to human scorers. For the automated sleep stage, reported agreement of PolySmith with human scorers for pediatric records is 77.1%.

Somnolyzer 24x7, Royal Philips Electronics, The Netherlands

Software Somnolyzer 24x7 is the resulting software produced by the The Siesta Group initiative [**60**] [**61**] which has recently been acquired by Philips. Software is integrated into an e-Health solution so that involved centers can upload the resulting digital recording to perform a centralized automatic analysis, receiving back the resulting scored PSG and the corresponding report. According to information available through Philips related website [**62**] Somnolyzer 24x7 provides with automated analysis of sleep staging and detection of EEG arousal, respiratory events and leg movements. Description of the system and validation results according to reliability in automatic sleep scoring capabilities regarding R&K criteria can be found in [**63**]. Posterior adaptation and validation regarding the new AASM criteria was published in [**64**]. Human quality control over the automatic classification is provided in order to ensure correctness of the results. In this respect sleep experts receive different quality check data, and on the basis of this information, the human expert has to decide whether or not, and if so to which extent, the automatic scoring has to be edited and correctly

visually. According to published validation only around 4% of epochs were changed by experts during quality control, resulting in a general 82% agreement (Cohen's kappa: 0.76) for the AASM validation study.

Not many information has been found regarding analysis of respiratory signals and SAHS diagnosis. However according to [**65**], the system has been tested in a database of 51 subjects to assesss agreement in resulting AHI in comparison to human scorers. The results show correlation index of $r = 0.92$ during the first night, and $r = 0.94$ during the second night.

Aura Lab-based PSG, Grass Technologies, USA

Equipment Aura of company Grass Technologies allows the technician the realization of polysomnographic tests either ambulatory or hospital attended. It includes the possibility of performing a wireless recording of up to 33 channels (15 AC referential, 10 respiratory plus an additional module with 8 DC channels). Workflows allow data analysis to be performed automatically, manually or by means of a combination of the two previous. Thus once the recording has been finalized, one can opt to carry out an automatic analysis or not. The automatic analysis is also flexible and it can comprise respiratory signals, neurophysiological signals or both. The software also permits manual annotation of events and modification of the automatic analysis results [**66**].

Within Aura PSG series the add-on FASS, which is in charge of the construction of the sleep map in two phases. In the first one the hypnogram is determined as a function of the presence or absence of events obtained from EEG, EOG and EMG signals. In a second phase, FASS applies a series of procedures with the objective of correcting the hypnogram, as for example, to check for improbable phase changes or the application of the 3-minutes rule to enter in stage 2. On the other hand the module Twin provides capabilities for the detection of apneas and hypopneas, analysis of flow curves, capnography, and pulse transit time (PPT) and detection of microarousals and K-complexes. As far as the author knows the software lacks from explanation capabilities although it possess a powerful report generator which allows data customization. No validation studies were able to be found regarding its automatic analysis capabilities.

Somnostar 4100, Sensormedics Corporation, USA

It records signals of EEG, EOG, EMG, ECG, digital pulse oximetry, thoracic and abdominal movements, body position and oro-nasal flow. The system is complemented with the Chephalo Pro amplifier, also from the same company, for signal acquisition. With Somnostar software one can review either part of or the full night study. It is quite interactive and it possesses widely configurable interfaces, favoring event's annotation, sleep staging and signal visualization under several views. The software also provides with automatic analysis of respiratory function and classification of sleep stages from the EEG, however validation studies have shown low concordance between results of the automatic analysis and those of expert's manual revision, specially in the detection of hypopnea events and in the classification of sleep stages [**67**]. Similarly to the most of the systems, for the EOG and EEG, Somnostar works in both frequency and amplitude domains. For the respiratory signals respiratory baseline is established by calculating average respiration during the two previous minutes to the occurrence of the apneic event. Classification of apneic events and construction of the hypnogram are based on default established –fixed- parameters: in the case of apneas detection, for example, a true positive is considered if there is a reduction of 80% with respect to the calculated baseline.

SomtéPSG, Compumedics Limited, Australia

SomtéPSG is presented as a family of products of the company Compumedics which is offered in three versions: laboratory, portable and ambulatory. Among all versions the product is subdivided into a hardware part in charge of signal acquisition, and software (ProFusion PSG) which allows both the manual assisted as well as automatic analysis, to be performed over the recording [**68**].

The set of supported signals includes main ones: EEG, EOG, EMG, ECG, airflow, snore, thoracic and abdominal movements, arterial oximetry, pulse and body position. ProFusionPSG possess capabilities for automatic analysis of sleep structure and detection of apneic events. Nevertheless both approaches are based on the application of "aseptic" protocols, which implies lack of contextual data interpretation or patient dependent handling of information. On the other hand, it has not been possible to find any published study on the validation of these capabilities, thus it is difficult to perform a detailed analysis of the product. It is noteworthy the possibility of establishing a

bluetooh connection from the signal acquisition device to the data capturing workstation. In this manner, there is a gain in autonomy and comfort, because it avoids the wired connection of the patient to the computer, favoring his/her relaxation and helping obtaining a more reliable recording.

Polyman, Bob Kemp and Marco Roessen, The Netherlands

Polyman is not strictly speaking commercial software for the automatic analysis of sleep but an EDF/EDF+ viewer and sleep scoring supporting program that was created by Marco Roessen and Bob Kemp [**69**]. However it includes several aiding tools that can perform automatic analysis of some sleep scoring subtasks. It is available in two versions, free and licensed.

Basic free version enables the display of any number of EDF(+) files from one subject and allows the possibility of different configurations for visual analysis of the digital recording. Each signal can be freely filtered, adjusted and automatically analyzed. Automatic analysis of frequency content (FFT), threshold crossings, neuronal feedback analysis [**70**], and rectified EMG can be applied to the signals on the screen. It also supports manual scoring of sleep stages, apneas, leg movements and arousals according to standard R&K or AASM rules. The scorings are kept in standard EDF+ files and a report of standardized sleep quality parameters is produced. Licensed version includes additional modules to automatic scoring of limb movements, respiration, body position, pulse rate and oxygen saturation.

Table 3.1 summarizes characteristics of the analyzed commercial systems according to their main advantages and disadvantages.

Table 3.1. Summary of advantages and disadvantages of the analyzed commercial systems

PolySmith	
Advantages	✓ Performs automatic sleep staging ✓ Performs detection and classification of apneic events ✓ Presentation of results in both text and graphics formats
Disadvantages	✓ Poor sleep states classification ✓ No interrelation between hypnogram and apneic events
Somnolyzer 24x7	
Advantages	✓ Performs automatic sleep staging ✓ Performs detection and classification of apneic events ✓ Integrated into an e-Health solution ✓ Provides quality control
Disadvantages	✓ No interrelation between hypnogram and apneic events ✓ Use of fixed protocols in order to classify the detected events
Aura Lab-based PSG	
Advantages	✓ Performs automatic sleep staging ✓ Performs detection and classification of apneic events ✓ Powerful report generator with customizable data
Disadvantages	✓ No interrelation between hypnogram and apneic events ✓ Use of fixed protocols in order to classify the detected events ✓ No explanation capabilities
Somnostar 4100	
Advantages	✓ Performs automatic sleep staging ✓ Performs detection and classification of apneic events ✓ Presentation of results in both text and graphics formats ✓ Possesses intuitive interfaces and several configurations
Disadvantages	✓ Poor sleep states classification ✓ No interrelation between hypnogram and apneic events
Polyman	
Advantages	✓ Flexible and fully customizable ✓ Performs detection of apneic events ✓ Analysis of EEG neuro-feedback loop ✓ Includes a free version
Disadvantages	✓ It does not perform classification of sleep stages

Besides the previously mentioned devices, it is important to point out to the growing interest toward the design of portable devices for the in-home monitoring. Although these devices do not offer the full functionality of the previous ones -aimed at performing the complete PSG- they save the patient the nuisance of the displacement to the hospital in order to carry out the polysomnographic test. Indeed the patient can be monitored in his/her own domicile, from which a screening diagnosis can be issued serving as first alert or control. In addition, the cost of in-home studies is much lesser than those carried out at the hospital, even if a supporting technician is required to attend [**71**]. Many studies have shown that the patient gains in sleep quality during ambulatory PSG recording. Actually from a general point of view, some studies have reported reliable results for the screening function [**72**] [**73**].

Some examples of this kind of devices are analyzed subsequently:

Apnea Risk Evaluation System (ARES), Advanced Brain Monitoring, Inc, USA

The product is composed of a hardware part that carries out the signal acquisition, ARES Unicoder, and software performing analysis of the recorded information, ARES Insight software. ARES Unicoder is a device aimed at the monitoring of the ambulatory patient. The device records the following signals by means of a casing attached to the patient's head: oxygen saturation, pulse, snore (through a microphone), body and head position (using accelerometers) [**74**]. Last versions also include measuring of airflow by nasal cannula connected to a pressure transducer [**75**]. The system is able to record up to 14 hours of sleep which can be structured in two different parts (two nights, 7 hours each).

From one recording, ARES Insight software analyzes the signals to compute an indirect measure of the number of apneic events per hour of sleep or Respiratory Disturbance Index (RDI). Calculation of this measure is based on the analysis of the oxygen saturation signal searching for desaturation and resaturation patterns to obtain an estimation of the number of apneic events. Incorporation of airflow measure also allows computation of AHI based on the airflow channel. Measures comprising pulse, snore and sleep position serve as contextual information to discard invalid desaturations, and also to detect awakenings that confirm the presence of the apneic

event. These awakenings should not be confused with EEG arousals since the software does not provide any analysis capabilities over the neurophysiological signals.

Additionally, the system complements with a questionnaire. The software also allows the physician the visual evaluation of the recorded respiratory signals and the revision of the recording in order to add additional events or to discard false positives. Validation results compared against attended in-lab PSG monitoring with manual scoring showed Intraclass Correlation Coefficient (ICC) of 0.8, with mean difference of 4.1 events per hour, when using a 4% desaturation threshold. The study was performed on a set of 86 patients and the resulting sensitivity/specificity when applying AHI cut-off of 10 was 0.86/0.82. Quality control was previously performed excluding non-valid PSG periods from the analysis time [**74**].

MicroDigitrapper-S, Synetics Medical, UK

The system consist of a portable device that measures patient's body position, intensity of snore, oronasal airflow, thoracic and abdominal efforts, heart rate and oxygen blood saturation. It does not provide of any mechanism to record signals related to sleep structure such as EEG, EOG and EMG; therefore analysis limits to the respiratory function which can influence analysis results of MicroDigitrapper-S.

Analysis results of MicroDigitrapper-S return a positive case when the patient suffers from SAHOS. No distinction is done between apneas or hypopneas, neither to classify events as obstructive, central or mixed. All detected events are classified as obstructive apnea; therefore it should only be used in patients with an obstructive apneic origin.

An experiment carried out on 30 patients at the sleep disorders center of Milan San Raffaele Hospital, verified that while index of severity associated to SAHS remained relatively low (AHI over 10) validation results on estimation of final AHI showed good precision (sensitivity of 1). However as long as severity indexes increase (AHI > 40) the associated precision decreased considerably (sensitivity 0.55), hence recommending the manual revision of the results [**76**].

Somnocheck, Weinmann GmbH, Germany

Sommocheck is a device addressing portable monitoring of patients in risk of SAHOS. It consists of a device attached to the chest of the patient that records the signals of airflow (thermistor), snore (microphone), oxygen saturation and pulse (finger sensor) and body position (using an integrated sensor in the main unit). Software provided with the unit allows the physician visualization and automatic analysis of the recorded signals. The analysis counts the number of desaturations and apneic events (apneas and hypopneas) and calculates basic parameters of Oxygen Desaturation Index (ODI), Apnea Index (ApI), Hypopnea Index (HI) and Apnea-Hypopnea Index (AHI), respectively, as the number of desaturations, apneas, hypopneas and apneas/hypopneas per hour of sleep. Validation results in [**77**] showed 83% sensitivity and 87% specificity for AHI $\geq$ 5 cut-off for in-lab attended PSG. Sensitivity decreased until 61% while 100% was obtained using AHI $\geq$ 40 cut-off. In a second, more recent study [**78**], values of sensitivity/specificity obtained for unattended conditions were respectively 96% and 65% (AHI $\geq$ 5) and 80% and 92% (AHI $\geq$ 30). According to information available through related website [**79**], capabilities for recording of thoracoabdominal movements and classification of apneic events as obstructive or centrals are also provided.

WristOx 3100, Nonin Medical Inc., USA

The product consists of two parts: a pulse oximeter hardware (Nonin WristOx 3100) and software for the analysis of the recorded SaO_2 signal (nVision 5.0). SaO_2 signal can be recorded at several sampling frequencies (1 Hz, 0.5 Hz y 0.25 Hz). Analysis algorithms then count the number of desaturations to compute the amount of apneic occurrences per hour of sleep.

The goal of this product is to detect or to discard the presence of SAHS in the patient by means of counting the number of desaturations. In this respect the software focus on the establishment of a first preliminary diagnosis regarding SAHS prevalence; hence the results may be used to proceed with preventive treatment. However, the former does not exempt the patient from eventually carrying out the complete PSG test at the hospital in order to establish the correct diagnosis and the corresponding treatment. On a study carried out over 154 patients comparison of adjusted O_2 desaturation index (ADI) was performed using several AHI cut-off values. Obtained

sensitivity/specificity percentages were 89%/94% for AHI≥5 (ADI2>19.3), 88%/94% for AHI≥10 (ADI3>10.5), and 88%/90% for AHI≥15 (ADI3>13.4) [**80**].

Subsequently Table 3.2 summarizes main characteristics of the portable systems previously analyzed.

Table 3.2. Summary of advantages and disadvantages of the analyzed portable systems

Apnea Risk Evaluation System (ARES)	
Advantages	✓ Performs automatic detection of apneic events ✓ Presentation of results in both text and graphics formats
Disadvantages	✓ It does not include analysis of neurophysiological activity ✓ It does not record respiratory effort ✓ It does not perform construction of the hypnogram
MicroDigitrapper-S	
Advantages	✓ Automatic detection of apneic events ✓ Automatic calibration and verification of the signals ✓ Configurable analyses
Disadvantages	✓ No distinction between different event types ✓ Only accounts for obstructive events ✓ It does not include analysis of neurophysiological activity ✓ Precision decays with increased OSAHS severity
Somnocheck	
Advantages	✓ Performs detection and classification of apneic events ✓ It allows manual override of automatic analysis
Disadvantages	✓ It does not include analysis of neurophysiological activity ✓ It does not perform construction of the hypnogram
WristOx 3100	
Advantages	✓ Only finger pulse oximeter sensor is required ✓ Automatic detection of desaturations
Disadvantages	✓ No distinction between different event types ✓ It does not perform classification of apneic events ✓ It does not include analysis of neurophysiological activity ✓ High sensitivity to presence of artifacts

3.4.2. Academic systems

In the previous section some of the current commercial PSG systems used at the sleep labs and at the diagnostic centers have been presented. These systems support the clinician during the SAHS diagnostic procedure. Nowadays it is difficult to find a center with a certain relevance that continues to perform the purely manual revision of data on paper. In this respect, at least with regard to information managing tasks, this kind of systems are already quite implanted. Nevertheless, with regard to their data analysis capabilities, such systems still continue to present several shortages, and in most of the cases, the manual method is yet used to perform the analysis of the PSG. There exists therefore an intense interest in the development and in the continuous improvement of the analysis capabilities of these systems for aiding in the diagnosis of SAHS. Besides, additional research lines come up attempting to simplify the standard analysis method or to make it more comfortable for the patient.

In this context, there can be found the approximations here referred as *non commercial* or *academic* approximations, which in general, are aimed at overcoming the drawbacks of current commercial systems, either augmenting their features or innovating in the use of new diagnostic techniques. The name was given since normally these approximations follow a more researching or academic philosophy, and in this respect, it is rare the research line that pursues the construction of a complete system from the scratch. Instead, at this level, one usually finds attempts of innovation focusing on specific tasks within the general diagnostic approach or in the development of alternative methods aimed at substituting –normally by simplifying- the current PSG analysis procedure. On the other hand, it is also true that the dividing line between commercial and academic systems is certainly fuzzy in many senses. In fact the research line, when successful, uses to be either a predecessor of the commercial product, or be promoted by continuous improvement processes of products that already exist in the market.

Subsequently several approximations included under the *academic* label are analyzed in contraposition to the preceding *commercial systems*. Ultimately, the objective of the discussion is to reflect current research state on the different methods and techniques developed for the computer assisted diagnosis of SAHS. In order to

structure the different approximations the discussion is divided into three subsections: SAHS estimation systems, apneic events detection and classification systems, and comprehensive diagnostic systems.

SAHS estimation systems

Due to the complexity and the elevated cost of the polysomnographic test, one of the main research lines is focused toward reducing the necessity of the patient to conduct such a test. For that purpose the resulting approximations are normally based in the substitution of undergoing PSG by means of the calculation of a supplementary measure based on historical or statistical risk factors. The indirect derived measure is then interpreted as preliminary evaluation of the *–a priori-* possibility that the patient may suffer from SAHS.

In this respect, predicting models based on questionnaires identify several physical and clinical features as there can be age, gender, body mass index (BMI) or presence of snore, indicative of a high risk of SAHS. For example yet in the beginnings of nineties there are studies intending the previous, however without reaching considerable good results [**81**] [**82**]. The study of Crocker et al. [**83**] considers as independent variables respiratory pauses, hypertension, BMI and age, and shows by means of a statistical model, how the necessity to carry out sleep studies can be reduced by one third. More contemporary Rodsutti et al. [**84**] continuing with this approach, attempt to find a decision rule to prioritize the waiting list to carry out the polysomnography at the Newcastle Sleep Disorders Center in Australia. In his work Rodsutti uses a regression model that considers age, gender, BMI, snore and presence of respiratory pauses, and it classifies the output as low, moderate or high regarding the risk of suffering from SAHS. Although the approach shows good results, it is reprehensible the fact that the study only considered patients already suspected of SAHS. In this respect, the work of Sharma et al. [**85**] results more interesting, since in the selection for the study patients with evident symptoms of OSAHS were previously rejected. In this work in order to predict OSAHS the model considers the variables gender, BMI, snore index, and obstruction factor. Another interesting work for SAHS prediction is the one of Sweere et al. [**86**], which is based on statistical models constructed over parameters gathered on patient's questionnaires previous to effectuation of the polysomnographic test.

ANNs based approximations for the same prediction task can be found in the works of Kirby et al. [**87**] and de El-Sohl et al. [**88**]. In the last one, besides, the first validation of this kind of RNA-based systems is performed. A more recent approximation is the one by Victor-Marcos et al. [**89**], which makes use of radial basis networks in order to classify the patients as OSAS or not OSAS using non-linear analysis of the oxygen saturation signal. This approximation differs from the rest, because instead of using statistical data, it implies the monitoring of the patient's SaO_2 signal. In the same line of proposing a minimal monitoring requirement, Caseiro et al. [**90**] propose the application of the Hilbert-Huang decomposition over 5 minutes of oronasal airway pressure signal. Results gathered from 41 patients show a sensitivity of 80% for a specificity of 95%. It has to be underlined, however, that approximations based on ANNs experience the problem of absence of any explanation capability of their results since these systems behave as black-boxes.

A recent work by Eiseman et al. attempts to classify individuals as patients (AHI > 5) or normals using features including demographics, polysomnogram and electrocardiogram and testing several classifiers: *k*-nearest neighbor, naïve Bayes, and support vector machines. The support vector machine performed similarly to naïve Bayes for predicting sleep apnea class. Reported sensitivity/specificity results are in the range 57.5-59%/73.7-74.5% using clinical features, and 39-43.4%/82.7-83.5% using spectrographic features [**91**].

Apneic events detection and classification systems

This category includes those approximations pursuing the localization of the individual event in patient's biosignals, instead of, as in the previous group, seeking for a global statistical characterization of the recording. In this respect, whereas estimative approaches are oriented toward obtaining a supplementary measure assessing the necessity of the patient to undergo nocturnal polysomnography, these ones rather focus on the direct localization, measurement and classification of the actual apneic events. Calculation of AHI/RDI is then directly derived by counting the number of individual detected events and dividing the result by the number of hours of sleep/processed recording.

One of the first works in this field is the one by West and Kryger [**92**], who designed a microcomputer for the monitoring of respiratory variables including expired carbon dioxide, respiratory inductance plethysmography, oxygen saturation and heart rate. The related work of Smith along the decade of 70s and 80s can also be considered within the group of the precursors [**93**].

A little bit later George et al. [**94**] developed a simple strategy that uses the oxygen saturation signal as a trigger for the localization of apneic events. In the said work, and only by means of the analysis of the saturation signal, the authors attempt to localize and even to quantify the duration of the apneic events. The fundamental hypothesis is that each apneic event induces a drop in the arterial blood oxygen concentration levels. Validation carried out using 9 polysomnographic recordings corresponding to 6 different OSAHS patients showed good sensitivity results. Main limitation of the method lies in the incapacity to distinguish between apneas and hypopneas, nor to carry out a classification of the event as obstructive, central or mixed. Besides, the method does not result very useful in the cases where the apneic event does not cause a significant drop in the saturation levels.

Using a similar hypothesis, the work of Rauscher et al. develops a system based on the searching of rapid resaturations [**95**]. In this case the assumption is that after the apneic event that causes a respiratory insufficiency, it follows an episode of compensatory hyperventilation inducing a fast increase of oxygen saturation levels. To test the validity of the method, the authors performed an experiment comparing the results of their algorithm with respect to a method based on the detection of desaturations. Although through their experiments they obtain better results, it is also evident that no resaturation can be produced without a previous desaturation, thus the relative improvement is, at least, matter of opinion. In any case, precisely because of their similarity, detection of apneic events via localization of resaturations presents the same drawbacks than the work of George et al.

In general, more reliable systems for detection of SAHS often carry out an analysis of the respiratory signals including a subset of the following: airflow, oxygen saturation and/or respiratory movements. In the work of Taha et al. analysis starts with the detection of desaturation and then the sum of RIP is analyzed to detect periods of no

breathing [**96**]. The work of Salmi et al. [**97**] develops a system that measures respiratory movements indirectly by using a static charge sensitive bed. A recent and interesting work is the one by A. Otero et al. in which fuzzy detection methods are used for marking apneic intervals in the respiratory signals. The method additionally facilitates knowledge acquisition by using an interface to model the membership functions, thus allowing system parameterization [**98**]. Other approximations that use fuzzy logic based decision algorithms for the analysis of the respiratory signals can be found in Al-Ashmouny et al. [**99**] and in Pittman et al. [**100**].

In Nakano et al. [**101**] a different strategy is followed by attempting the detection by means of analysis of the tracheal sound. For that purpose the manual computed index (AHI) is compared with respect to a measure of the number of falls per hour produced in the time series of the spectral power of the tracheal sound. Although both measures show good correlation, the obtained values cannot be considered enough to consider it a precise approach. Estimation of AHI through snore has been carried out in other studies such as in [**102**], which also combines the analysis of the oxygen saturation signal. In practice, the snore signal used as the only method to detect apneic events, does not receive a great support by clinicians, mainly because of its high sensitivity to noise that limits its discriminative capabilities. However snore sound can be used as approximate evidence, for example, in the development of portable systems for ambulatory screening.

It has been widely discussed about the validity of electrocardiogram (ECG) as a method to detect the apneic event. In 1984 Guilleminault et al. [**103**] described the cyclic heart rate variability (HRV) as a fluctuating pattern in the cardiac rhythm, characteristic of obstructive sleep apnea, which repeats during each apneic episode. Since then this variation in the heart rate has been studied as potential detector of apneic events [**104**] [**105**]. With the motivation of exploring this approximation, in the year 2000 a competition was celebrated sponsored by Computers in Cardiology[18] (CINC) and PhysioNet[19]. The contest was aimed at evaluating the validity of the use of the ECG signal in order to detect apneic events. The event consisted of a dual challenge:

[18] International anual conference celebrated since 1974
[19] Database of biomedical signals supported by the US National Institutes of Health's National Center for Research Resources (NIH NCRR)

- To classify a total of 30 recordings determining which of them pertained to OSAHS patients and which do not

- To annotate in one minute intervals, the presence of the apneic event or not

To accomplish the tasks only the use of the ECG signal was allowed. The first task was successfully solved, whereas in the second one the best algorithms obtained a precision around 90%. However it has to be considered that detection is done only on a one minute time scale which could to lead to imprecise results and subestimation of the actual number of apneic events. Besides, from the perspective of a comprehensive approach, the ECG approximation is far from being considered valid enough, mainly because of its deficiency at the time of classifying the different types of causing events. Other approximations based on the use of ECG for apneic event detection are under research, as in Figliola et al. [**106**], Maier and Dickhaus [**107**], or in Amir et al. [**108**], however until now, obtained results can endorse the use of ECG only as an approximate diagnostic technique.

More centered in the classification task of the apneic event according to the nature of its origin –obstructive, central or mixed- there can be found the first approximations based on artificial neural networks in [**109**] [**110**]. In these cases a backpropagation method was employed, however, the classification rates did not exceed 60%. In [**111**] a radial basis function neural network is applied for an integrated detection-classification task obtaining an accuracy of 64±3.4% for adults and 62.6±3.4% for infants. More recently it can be found the work of O. Fontenla-Romero et al. [**112**] developed at the LIDIA lab of University of A Coruña. In the said work the detection of apneic events is performed from the airflow signal, and once detected, a wavelet processing is applied to the corresponding intervals of thoracic effort signal. An artificial neural network is then in charge of finally classifying the interval as central, obstructive or mixed. The work of Tagluk et al. [**113**] is also based on the use of wavelets but with worse classification results with respect to obstructive and mixed events. Finally, also within the LIDIA group it has been recently developed a method that combines machine learning techniques and expert knowledge that improves previous results of Fontenla-Romero et al. on the classification task [**114**].

Comprehensive diagnostic systems

With the term *comprehensive diagnostic systems* we here refer to those systems that group –as a minimum- classification of sleep stages and analysis of the respiratory signals, in order to offer an integrated and complete diagnosis. In this respect, in the line as an expert would do, approximations under this category involve the analysis of the polysomnographic signals in the context of a comprehensive sleep analysis. This characteristic differentiates the comprehensive diagnostic systems from the previous commented approaches which, instead, specialized in the accomplishment of specific subtasks, such as the detection of a concrete type of event, or the classification of the detected events. The ones referenced in this subsection are, therefore, more oriented toward the realization of the full diagnosis, and ultimately, they are aimed at constituting a global solution in the form of *clinical decision supporting systems* in the context of SAHS.

Within this group the system ISAS [**115**] can be cited as one of the precursors. Developed at INSEC[20] this integrated system for sleep patient monitoring system has been validated at hospital Santo Antonio de Oporto. Its architecture includes the following modules:

- HIDRA [**116**] is the responsible for the detection and parameterization of activity related with the sleep function (EEG, EOG, EMG).
- SAIAS [**117**] is the module responsible for the detection of apneas occurring throughout sleep. It also carries out their classification as obstructive or central.
- OSCAR determines the oxygen saturation in blood and the cardiac frequency from data supplied by a commercial oximeter.
- TIMEMAKER is an element that works as a clock in order to synchronize information coming from the different detecting devices.

The output information provided by the system at the end of the sleep exam consists of a set of tables with statistical information of the different episodes and types

[20] Instituto de Engenharia de Sistemas e Computadores, Departamento de Electrónica e Telecomunicacoes de Aveiro, Portugal

of apneas suffered by the patient. It also provides graphics of the oxygen saturation and the cardiac rhythm signal, and a table with the duration of the different cerebral activities exhibited by the patient throughout the night.

One of the main problems presented by this approximation refers to the classification of sleep stages. Since the system does not achieve an adequate discrimination capability between EMG and EOG, it is not able to correctly classify REM and wakefulness. Another difficulty arises with regard to the sleep staging supervision strategy –performed in a minute by minute basis. In this case it often identifies the same sleep phase, repetitively and with short duration intervals, when however it is actually the same phase that maintains along a much longer period.

However, the most important problem with ISAS resides in the apneas detection method, which is carried out by using fixed thresholds applied over the respiratory function signals. In this respect, it is known that throughout sleep amplitude of signals varies, and therefore it is necessary a recalibration of the thresholds to perform a correct apnea detection. The fundamental problem in this regard is the system incapability to determine if an amplitude reduction was caused by an apneic event or due to a transition in the sleep state. That is, classification of sleep stages and detection and classification of apneic events are performed independently without any kind of interaction between the two processes.

Another interesting development is the system PSG-EXPERT [**118**], which is presented as the particularization for the case of SAHS of a more general integrated environment for the development of diagnostic expert systems. From an application point of view, PSG-EXPERT is defined as an auxiliary diagnosis system for sleep disorders based on polysomnographic data. Such data are extracted through a series of processings and are inserted in a database organized according to the following categories: clinical history, hypnogram data, sleep parameters, spectral data, EEG time related activity and non EEG activity. The system is developed using an extension of CLIPS [**119**] and it supports handling of imprecise information through the use of the model of certainty factors [**9**]. It also includes a validation module which allows testing of concrete patient's cases by comparing the results of the analysis with those of medical experts. Main limitation of the system is due to its general purpose philosophy,

thus it just operates at the symbolic level. In other words, processing and segmentation of the raw signal has to be performed separately and the resulting data has to be inserted in the database afterwards. In addition, despite of its built-in validation module, no validation results have been reported assessing the actual performance of the system.

The design and implementation of an intelligent diagnostic system for SAHS aimed at solving the problems of their preceding systems, as well as improving their capabilities, was the object of the doctoral thesis of Elena Mª Hernández Pereira, member of the LIDIA group. The result was the system MIDAS [**120**] which later evolved into the SAMOA system [**121**] [**122**]. SAMOA, unlike other approaches, integrates both artificial intelligence techniques for the development of reasoning processes over well-known rules, and classical techniques of signal processing and software engineering, for the development of an integrated product which, in addition, it is able to provide explanation of its results. The previous converts SAMOA in a valuable tool to aid the physician in the diagnosis decision in the context of SAHS.

Architecture of the system is displayed in Figure 3.1.

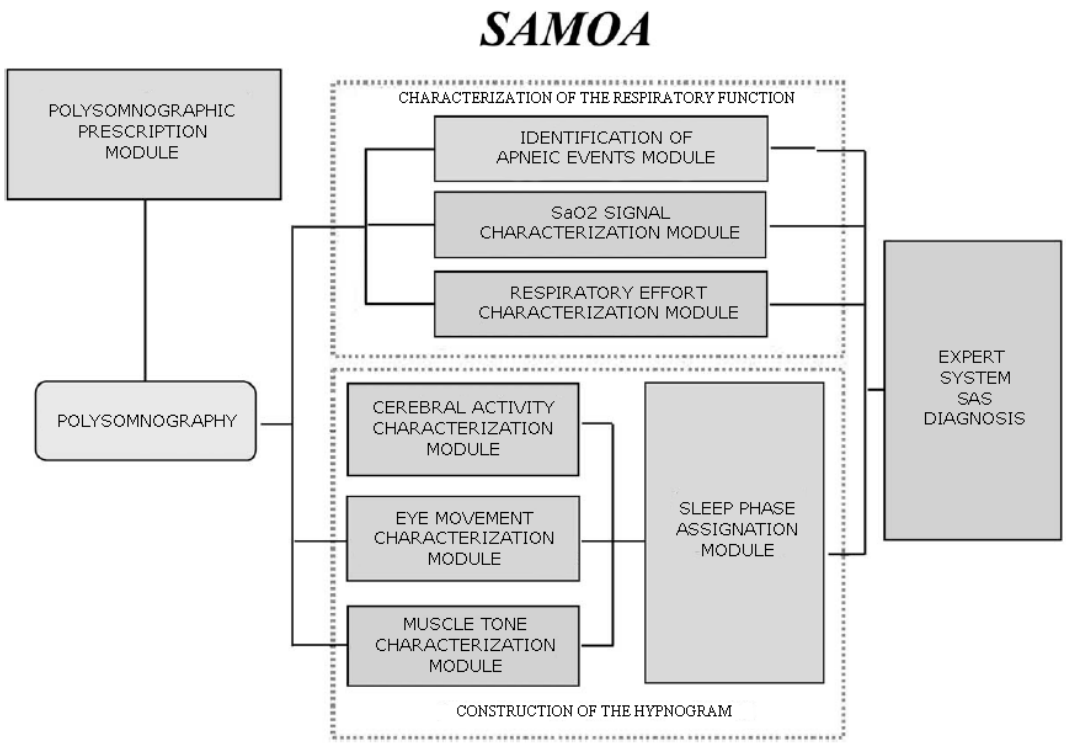


Figure 3.1. Modular representation of the SAMOA system

Subsequently a summary of the related functionality of its main modules is given:

- *Polysomnographic prescription module*: Its objective is the evaluation of the convenience of submitting the patient to carry out the polysomnographic test. It acts as a screening tool in order to optimize hospital resources, avoiding realization of PSG in the unnecessary cases and/or to prioritize the cases where the patient is suspected to suffer from SAHS. For that purpose a questionnaire is filled by the patient and a numerical index indicating the convenience to carrying out the test is obtained using a model of certainty factors.

- *Characterization of respiratory function*: Its objective is the analysis of the respiratory signals, which in SAMOA comprises airflow, oxygen saturation and presence of thoracic-abdominal effort. Eventually, detected apneic intervals, together with the information provided by the remaining modules, are confirmed as apneas or hypopneas, or on the contrary, are discarded as apneic events.

- *Construction of the hypnogram*: Analysis of the neurophysiological information for the construction of the sequence of sleep phases of the patient. The resulting hypnogram is segmented according to discrete time units –epochs- of 30 seconds following guidelines of R&K protocol.

- *Diagnostic module*: It is the responsible to integrate the results obtained from the three previously described tasks. In this respect data is interpreted taking into account contextual information and the final diagnosis is issued. Final report includes a summary of all detected events together with their classification, as well as a diagnostic evaluation of the presence of SAHS and its severity in the patient.

The philosophy of the SAMOA system established the initial point that inspired the current doctoral thesis. SAMOA approximation constitutes a comprehensive approach to the problem of SAHS diagnosis; it covers evaluation of the adequacy of carrying out the polysomnographic test, automatic analysis of the resulting recording, and diagnosis of SAHS integrating both respiratory and cerebral activity. It also incorporates

capabilities to explain the obtained results. On the other hand, although the system in general offered good results, there were several points susceptible to be improved.

One of these points is related with the procedure to detect apneic events in the respiratory signals. In this respect SAMOA base the detection in the identification of intervals in the airflow signal (using thermistor) which present amplitude reductions. For that purpose each respiratory cycle is classified according some predefined thresholds in one of the following labels: *normal* (*N*), *slightly reduced* (*SR*), *clearly reduced* (*CR*) and *totally reduced* (*TR*). On the basis of the corresponding labeling (*N, SR, CR, TR*), the possible event is then considered in a different form on the subsequent analysis. The previous procedure presents several drawbacks:

- It uses fixed protocols in order to classify the respiratory cycles, therefore it cannot account for variability in data due to imprecision and inter-subject differences. Analysis *at a glance* of the human eye does not have the same precision as the computer analysis at the time of evaluating the signal amplitude. On the contrary, as it has been commented already, subjectivity in the human evaluation often causes discrepancy in the measurements between clinicians and system.

- SAMOA only searches for amplitude reductions in the airflow signal. However, the solely analysis of airflow with a thermistor may be insufficient, especially in the presence of artifacts or for the detection of hypopnea events [**123**]. A better characterization of the respiratory pauses can be done by considering evidences from additional respiratory signals such as the thoracic and abdominal movements.

- The detection algorithm itself excessively depends on the value established for the normal respiration. In cases with subsequent repeatedly occurrences of airflow reductions –common in severe SAHS patients, establishment of a normal amplitude reference value represents a difficulty. In this cases SAMOA does not correctly detect all the respiratory cycles and, consequently, it fails to detect all the apneic event occurrences.

Aside from the processing of the respiratory signals, another aspect that required of special improvement in SAMOA refers to its capabilities regarding the analysis of the neurophysiological activity. Indeed, besides counting with mechanisms for generation of the hypnogram, the system does not actually carry out an analysis of the EEG activity. Instead, numerical values corresponding to the different rhythms and transitory components are fed into the system by previous spectral analysis carried out by supplementary systems [**124**]. Besides, SAMOA does not provide of any procedure for the detection transient microstructure events such as sleep spindles, K-complexes or EEG arousals.

Table 3.3 summarizes the different academic approximations analyzed in this section.

Table 3.3. Summary of analyzed academic approximations

Academic approximations	
SAHS estimation systems	▪ *Based on questionnaires* Viner et al. [**81**], Flemons et al. [**82**], Crocer et al. [**83**], Rodsutti et al. [**84**], Sharma et al. [**85**], Sweere et al. [**86**] ▪ *Based on ANNs* Kirby et al. [**87**], El-Solh et al. [**88**], Victor-Marcos et al. [**89**] ▪ *Based on Hilbert-Huang decomposition* Caseiro et al. [**90**] ▪ *Based on machine learning classifiers* Eiseman et al. [**91**]
Apneic events detection and classification systems	▪ *Detection of AHI* West and Kryger [**92**], Smith [**57**] ▪ *Analysis of saturation signal* George et al. [**94**], Rauscher et al. [**95**] ▪ *Analysis of saturation signal and RIP sum* Taha et al. [**96**] ▪ *Analysis of respiratory movements by static charged bed* Salmi et al. [**97**] ▪ *Fuzzy analysis of respiratory signals* Otero et al. [**98**], Al-Ashmouny et al. [**99**], Pittman et al. [**100**] ▪ *Analysis of tracheal sound* Nakano et al. [**101**], Yadollahi et al. [**102**] ▪ *ECG-based* Penzel et al. [**104**], Mendez et al. [**105**], Figliola et al. [**106**], Maier and Dickhaus [**107**], Amir et al. [**108**] ▪ *Classification of apneic events using ANNs* Cabian et al. [**110**] [**109**], Zemen et al. [**111**], Fontenla-Romero et al. [**112**] ▪ *Classification of apneic events using wavelets* Tagluk et al. [**113**], Peteiro-Barral et al. [**114**]
Comprehensive diagnostic systems	ISAS [**115**] [**117**] [**116**] PSG-EXPERT [**118**] MIDAS [**120**] SAMOA [**121**] [**122**]

3.5. Critical analysis

Throughout the previous sections an analysis of the state-of-the-art has been performed with respect to the automatic systems for the diagnosis of SAHS. As it has been shown the referred context constitutes a field with a certain maturity –despite the relatively youth of the sleep science- and where the number of approximations is significant. In fact the interest in the development of SAHS diagnostic systems is increasing, and in the last times, such systems are yet bypassing the researching frontiers, up to the point that currently several solutions of a certain consideration are already present in the market. However acceptation of these systems is still low, as well as it is their real use in practice by the medical specialists. Hence, it can be said that implantation of SAHS diagnostic systems in the clinical routine is still in a preliminary stage.

Complexity of the analysis task is one of the main responsible for the slow transfer of these systems into the clinical domain. Such complexity causes that many of the currently available tools limit to the realization of partial tasks within the diagnostic process. There exist systems providing capabilities for the analysis of a certain subset of polysomnographic signals, for example, regarding the respiratory signals, but present deficiencies from the neurophysiological point of view, and vice versa. Other shortages found among the different analyzed approaches include lack of explanation of their results, or excessive sensitivity to the presence of artifacts and to the variability of the signals among the different patients.

Devices destined for ambulatory monitoring, or techniques based on questionnaires for estimation of SAHS, are principally aimed at holding a screening function. They simplify in a great extent the required input signals with respect to the clinical PSG, at the prize of limiting validity of their diagnoses. Among such systems many times the screening task reduces to a dichotomous response (yes/no) that does not provide any gradation with respect to the associated severity of the disease. Others, on the other hand, attempt to include an estimative prediction of the severity. In both cases they may be used to make a first selection, allowing prioritization of waiting lists in order to carry out hospital PSG. In this regard screening systems save the patient the inconvenience of

moving to the hospital in order to carry out the polysomnographic tests, permitting a first diagnosis to be performed at home which can be used as alert or control[21]. In any case the use of portable ambulatory devices for preliminary screening does not exempt the patient from an attended in-hospital PSG for a reliable diagnosis.

In spite of all, a common critic to the majority of the current systems (whether they are portable, based on clinical record information or in the case of comprehensive approaches) is that they might be defined as fixed or protocolar, in the sense that they accomplish their tasks by means of a set of algorithms that excessively depend on the appropriate setting of critical values and thresholds. Within the analysis processes of the previously analyzed approximations, it is common that decisions are carried out through a discriminant analysis, partitioning the possible outputs into disjoint sets (*is* or *not is*). Then, by checking if the value associated to a certain feature is, or not, included within the valid range of pre-established values, categorical classification of the event is performed. In this regard, for example it can be assessed if the related airflow reduction is in the range 50%-80% with respect to normal respiration, and then it can be classified as a *clear reduction*. However, the former leads to non-realistic situations such as that an associated reduction of 49% may not mean anything at all, whereas a reduction of 51% rather has clinical relevance. It can also be considered the case of the patient suffering from continuous oxygen desaturations at 3%, which lacks of significance when the minimum desaturation percentage required to score an event is set to 4%. The former certainly differs from the situation of a healthy patient that maintains stable saturation levels during the whole night. However the use of *all or nothing* frontiers causes that by using the former criteria both situations to be equally significant for the final diagnosis.

On the other hand, it has been previously enunciated in this chapter that reasoning processes in the medical domain contain a high component of imprecision. This imprecision comes from different sources which are summarized in Table 3.4.

[21] For example in chronic patients already diagnosed

Table 3.4. Sources of imprecision in clinical contexts

- Data variability
- Presence of noise
- Data dependence and redundancy
- Limited sensitivity of the transducer
- Loss of information due to analog-to-digital process
- Interference
- Expert's intra and inter variability
- Expert's subjectivity
- Reasoning on the basis of qualitative terms and similarity criteria

In this respect, a common drawback of current available systems is that analysis processes are handled from an excessively quantitative perspective, not accounting for methods to handle imprecision. In addition, it is usual within the clinical language, to express opinions in terms of possibilities rather than in terms of certainties, possibilities over which it is thought more in qualitative than in quantitative terms. Clinicians often express their knowledge using sentences of the type, for example, that a certain combinations of symptoms *usually* point out to a particular disease, or that a drug *is known* to *sometimes* cause a particular side effect but it *rarely* causes another. Indeed when a physician takes a decision, he/she chooses from several alternatives which almost always are affected by such possibility expressions. Given this context, it emerges the necessity of constructing systems being able to handle this inexact information, to carry out decision making in medicine from a qualitative point of view. Sleep medicine is not an exception.

Let us consider the following example in which we have two definitions *A* and *B* of a same type of event. Let us say that definition *A* identifies the event when there is a reduction in the saturation levels of 3% and a reduction in the airflow amplitude of 30%. On the other hand, based on the criteria of *B*, which sustains a more strict definition, the event must present a descent in the saturation levels of at least a 3% and an airflow reduction of at least of 50%. Now, during the signal recording it is identified a possible event that the computer determines –with high precision- that it has associated a decrease in the saturation levels of 3.5%, and that its related airflow reduction is of 40%. According to definition *A* this event would be identified as a true positive because it fulfills all the definition requirements. On the other hand, according

to the definition *B*, the event could not be classified like that, although it satisfies the requirements regarding the associated desaturation levels. This is because its associated airflow is slightly less than the 50%. Thus, this situation according to strictly quantitative criteria of *A* and *B* would generate a discrepancy between both definitions. However, it seems logical to think that in approximate terms the identified pattern looks similar to both the definition *A* as well as the definition *B*. Therefore it can be delivered a similarity judgment –perhaps more appropriately- so that the system might determine that the identified pattern *is similar to A* but, at the same time, that it *is similar to B*, although without matching the exact definition in none of the cases. Moreover, it may be possible to quantify such a deviation, for example, it could be said that the pattern is *quite similar* to *A*, but at the same time it is *somewhat similar* to *B*.

In addition, if the sensor handling the registration of the airflow has an associated sensitivity of the ±10% -it would be a very bad sensor in this case- then the associated uncertainty increases, and this situation should be taken into account. In this manner it seems more logical to admit that the recorded event might also match the definition *B*, because its discrepancy regarding the airflow requirements can be caused by the presence of noise in the data. And after all, *it is not true that the human being tends to make its judgments on the basis of approximation and similarity criteria?*

Most of the current available systems fail in dealing with situations of this kind, as it has been mentioned, because of their inappropriate handling of data imprecision and their excessive dependence on fixed thresholds and categorical classifications. Implementation of capabilities to handle approximate reasoning, in this respect, should help to deal with these limitations. For this reason, in the context of the development of the proposed solution to aid in the diagnosis of SAHS, besides a comprehensive philosophy, the implementation of analysis mechanisms to manage data imprecision will also have an important weight. Specifically, for such a task, the fuzzy logic paradigm is used because of its nice properties that include, among others, the possibility to model imprecision both using quantitative and qualitative terms, as well as allowing propagation of uncertainty in the reasoning processes through the use of fuzzy rules. The next chapter introduces fuzzy systems and the modeling framework that allows implementation of mechanisms for an effective handling of imprecision within the developed system.

3.6. Summary of this chapter

This chapter is an introduction to the intelligent systems for clinical decision support. The chapter starts by making a description, from a historical perspective, to the origins of artificial intelligence and its relationship with the field of medicine. This relationship emerged together with the advent of Expert Systems along the decade of the 60s and it continued evolving until today. This evolution has lead the step from the strongly knowledge based systems to the use, nowadays, of the so-called hybrid systems. Besides knowledge modeling and symbolic processing mechanisms, hybrid systems make use of the latest computational intelligence techniques, including artificial neural networks, genetic algorithms or fuzzy analysis of information. An overview of the main approaches and techniques used by modern AI systems in the field of medicine is given in this regard.

The discussion continues putting special emphasis in an area of special interest within the scope of the intelligent diagnosis: the handling of imprecise information. Dealing with imprecision is of interest due to the uncertainty associated to medical domain which includes subjective variability, imprecision caused by the limited sensitivity of the measurement devices, loss of information in the signal digitalization processes, or the presence of artifacts such as noise, punctual interferences, poor signal calibration or focus loss.

The chapter continues then focusing in the particular domain of SAHS, by analyzing the state-of-the-art regarding the context of the *automatic analysis of sleep recordings for SAHS diagnosis*. This context is faced from different points of view.

A first distinction can be made based on the degree of settlement that the technology presents in the market. In this respect it has been differentiated between *commercial systems* and *academic systems.* Within the first group there have been enclosed those systems that are currently used by the sleep labs at the medical centers, or that at least have a certain impact as commercial products. On the other hand *academic systems* –or non-commercial- can be understood as those systems bound to a more research philosophy, whose main objective is to propose alternative analysis methods and, in general, to improve features of current commercial systems.

Special mention has been done with regard to the increasing interest in the development of portable systems. Even though the features of such systems do not achieve the performance of their precedent versions, they do improve patient's comfort at the time of recording, lowering also the associated costs of the diagnosis. Such systems can be used as devices for a follow-up of the patient at home, or as alarm mechanisms to assess the necessity of a deeper PSG test to be done in the hospital. Portable devices partially contribute in this way to the alleviation of congestion suffered by medical centers due to the high ratio between the analysis requests from the population and the available resources.

Analysis of non-commercial systems has been classified in: (1) *systems for SAHS prediction*, (2) *systems for the detection and classification of apneic events*, and (3) *comprehensive approaches*. The first ones are aimed at predicting the necessity of performing the polysomnographic test, based on the evaluation of possible risk factors from the patient. This is done mainly in order to fully exploit current available resources. The second ones are focused in the analysis of the respiratory signals for the detection and the classification of the apneic event. Finally, the so-called *comprehensive approaches* are claimed to become global solutions, evaluating both the respiratory and neurophysiological signals, to produce a full diagnosis in the context of SAHS.

Our system, which is included in this last category, has the ultimate goal of improving capabilities of current available systems. In this regard the last part of the chapter performs a critical analysis on this topic, introducing the necessity to carry out new approaches to overcome their limitations. Besides the lack of comprehensive approaches, an adequate handling of data imprecision and human subjectivity is also identified as an important feature to be improved. The use of fuzzy theory is proposed in this regard as the framework to develop new analysis strategies to deal with such imprecision. Next chapter introduces fuzzy theory and the modeling techniques allowing the construction of fuzzy inference systems. These systems are then integrated in the proposed solution of an automatic system to help the physician in the analysis of SAHS.

3.7. References

[1] A. Rosenblueth, N. Wierner, and J. Bigelow, "Behaviour, purpose and teleology," *Philosophy of Sciences*, vol. 10, pp. 18-24, 1943.

[2] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biology*, vol. 52, pp. 95-115, 1943.

[3] K. Craick, *The nature of explanation*.: Cambridge University Press, 1943.

[4] M. Minsky, *Semantic information processing*.: MIT Press, 1968.

[5] R. Ledley and L. Lusted, "Reasoning foundations of medical diagnosis," *Science*, vol. 130, pp. 9-21, 1959.

[6] H. Warner and A.Veasy, L. Toronto, "Experience with baye's theorem for computer diagnosis of congenital heart disease," *Annals of the New York Academy of Science*, vol. 152, 1964.

[7] G. Gorry, J. Kassirer, A. Essig, and W. Schwartz, "Decision analysis as the basis for computer-aided management of acute renal failure," *American Journal of Medicine*, vol. 55, pp. 473-484, 1973.

[8] E. Shortliffe, *Computer-based medical consultations: MYCIN*.: Elsevier Science Ltd., 1976.

[9] EH. Shortliffe and B. Buchanan, "A model of inexact reasoning in medicine," *Mathematical Biosciences*, vol. 23, pp. 351-379, 1975.

[10] S. Weiss, C. Kullowski, and S. Amarel, "A model based method for computer-aided medical decision making," *Artificial Intelligence*, vol. 11, pp. 145-172, 1978.

[11] FD Dombal, D. Leaper, and J. Staniland, "Computer aided diagnosis of acute abdominal pain," *British Medical Journal*, vol. 2, pp. 9-13, 1972.

[12] E. Engle, "Attempts to use computers as diagnostic aids in medical decision making: a thirty-year experience," *Perspect. Biol. Med.*, vol. 35, pp. 207-219, 1992.

[13] RL. Ackoff, "From data to wisdom," *Journal of Applied Systems Analysis*, vol. 16, pp. 3-9, 1989.

[14] B. Pandeley and RB. Mishra, "Knowledge and intelligent computing system in medicine," *Computers in Biology and Medicine*, vol. 39, pp. 215-230, 2009.

[15] DW. Patterson, *Introduction to Artificial Intelligence and Expert System*. Englewood Cliffs, NJ, USA: Prentice-Hall, Inc., 1990.

[16] C. Hernandez-Sande, V. Moret-Bonillo, and A. Alonso-Betanzos, "ESTER: an expert system for management of respiratory weaning therapy," *IEEE Transactions on Biomedical Engineering*, vol. 36, no. 5, pp. 559-564, 1989.

[17] M. Suojanen, S. Andreassen, and KG. Olesen, "A method for diagnosing multiple diseases in MUNIN," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 5, pp. 522-532, 2001.

[18] PH. Winston, *Inteligencia Artificial*.: Addison-Wesley Iberoamericana, 1992.

[19] S. Russell and P. Norvig, *Artificial Intelligence: a modern approach*, 2nd ed.: Pearson Eductaion, Inc., 2003.

[20] MCJ. Biermans et al., "Development of a case-based system for grouping diagnoses in general practice," *International Journal of Medical Informatics*, vol. 77, no. 7, pp. 431-439, 2007.

[21] G. Finnie and Z. Sun, "R-five model for case-based reasoning," *Journal of knowledge-based systems*, vol. 16, pp. 59-65, 2003.

[22] R. Khardon and D. Roth, "Defaults and relevance in model-based reasoning," *Artificial Intelligence*, vol. 97, pp. 169-193, 1997.

[23] J. Prentzas and I. Hatzilygeroudis, "Categorizing approaches combining rule-based and case-based reasoning," *Expert Systems*, vol. 24, no. 2, pp. 97-122, 2007.

[24] A. Lewis. RIMSAT DSS Project: Integrating model-based and case-based reasoning. [Online]. DSSResources.com

[25] E. Lopez and E. Plaza, "Case-based learning of plans and medical diagnosis goal states," *Artificial Intelligence in Medicine*, vol. 9, pp. 29-60, 1997.

[26] A. Khan and A. Hoffman, "Building a case-based diet recommendation system without knowledge engineer," *Artificial Intelligence in Medicine*, vol. 27, pp. 155-179, 2003.

[27] BW. Porter and ER. Bareiss, "PROTOS: an experiment in knowledge acquisition for heuristic classification tasks," in *Proceedings of the 1st International Meeting on Advances in Learning (IMAL)*, Les Arcs, France, 1986, pp. 159-174.

[28] P. Koton, "Reasoning about evidence in causal explanations," in *Proceedings of the 7th National Conference on Artificial Intelligence*, Menlo Park, CA, 1988, pp. 256-263.

[29] S. Montani et al., "Integrating model-based decision support in a multi-modal reasoning system for managing type 1 diabetic patients," *Artificial Intelligence in Medicine*, vol. 29, pp. 131-151, 2003.

[30] AD. Bernstein, CJ. Chiang, and V. Parsonnet, "Diagnosis and management of pacemaker-related problems using an interactive expert system," in *IEEE 17th Annual Conference on Engineering in Medicine and Biology Society*, 1995, pp. 701-702.

[31] S. Vinterbo and L. Ohno-Machado, "A genetic algorithm approach to multi-disorder diagnosis," *Artificial Intelligence in Medicine*, vol. 18, pp. 117-132, 2000.

[32] G. Shafer, *A mathematical theory of evidence*.: Princeton University Press, 1976.

[33] LA. Zadeh, "Outline of a new approach to the analysis of complex systems and decision processes," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, no. 1, pp. 28-44, 1973.

[34] JC. Príncipe, N. Euliano, and C. Lefebvre, *Neural Systems: fundamentals through simulations*.: John Wiley, 2000.

[35] JP. Lisboa, "A review of evidence of health benefit from artificial neural networks in medicine intervention," *Neural Networks*, vol. 15, pp. 11-39, 2002.

[36] DE. Goldberg, *Genetic Algorithms in search, optimization and machine learning*. Boston, MA: Kluwer Academic Publishers, 1989.

[37] HJ. Gross, B. Verwer, D. Houck, RA. Hoffman, and D. Recktenwald, "Model study detecting breast cancer cells in peripherical blood mononuclear cells at frequencies as low as 10^-7," in *National Academy of Sciences of USA 92*, 1995, pp. 537-541.

[38] GE. Ezzell, "Genetic and geometric optimization of three-dimensional radiation therapy treatment planning," *Medical Physics*, vol. 23, no. 3, pp. 293-305, 1996.

[39] Y. Yu, MC. Schell, and JBY. Zhang, "Decision theoretic steering and genetic algorithm optimization: application to stereotactic radiosurgery treatment planning," *Medical Physics*, vol. 24, no. 11, pp. 1742-1750, 1997.

[40] MA. Kupinski and ML. Giger, "Feature selection and classifiers for the computerized detection of mass lesions in digital mammography," in *International Conference on Neural Networks*, 1997, pp. 2460-2463.

[41] LM. Brasil, FM. Azevedo, and JM. Barreto, "Hybrid expert system for decision supporting in the medical area: complexity and cognitive computing," *Journal of Medical Informatics*, vol. 63, pp. 19-30, 2001.

[42] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, pp. 37-54, 1996.

[43] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507-17, 2007.

[44] JL. Breault, CR. Goodall, and PJ. Fos, "Data mining a diabetic data warehouse," *Artificial Intelligence in Medicine*, vol. 26, pp. 37-54, 2002.

[45] A. Kusiak, JA. Kern, KH. Kernstine, and BTL. Tseng, "Autonomous decision-making: a data mining approach," *IEEE Transactions on Information Technology in Biomedicine*, vol. 4, no. 4, pp. 274-284, 2000.

[46] PR. Walker et al., "Data mining of gene expression changes in Alzheimer brain," *Artificial Intellligence in Medicine*, vol. 31, pp. 137-154, 2004.

[47] C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 2, pp. 334-343, 2006.

[48] T. Young and J., Peppard, PE. Skatrud, "Risk factors for obstructive sleep apnea," *The journal of the American Medical Association*, vol. 291, no. 16, pp. 2013-2016, 2004.

[49] A. Otero, P. Felix, J. Presedo, and MR. Alvarez, "Is the average duration of apneas, hypopneas and desaturations useful in the diagnosis of SAHS?," in *7th IEEE International Symposium on Intelligent Signal Processing (WISP)*, Floriana, 2011, pp. 1-6.

[50] R. Sun and L. Bookman, *Computational architectures integrating neural and symbolic processes*. Needham: Kluwer Academic Publishers, 1994.

[51] B. Holmes, W. Wilcox, and S. Lapege, "Identification of enterobacteriaceae by the api 20E system," *J. Clin. Pathol.*, vol. 31, no. 1, pp. 22-30, 1978.

[52] T. Bayes and R. Price, "An essay towards solving a problem in the doctrine of chance," *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370-418, 1763.

[53] Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*.: Morgan Kauffman Publishers, 1988.

[54] LA. Zadeh, "The concept of a linguistic variable and its applications to approximate reasoning," *Inform Sci*, vol. 8, pp. 199-249, 1975.

[55] P. Szczepaniak, *Fuzzy Systems in Medicine*.: Physic-Verlang, 2000.

[56] S. Barro and R. Marín, *Fuzzy Logic in Medicine*.: Physic-Verlang, 2002.

[57] JR. Smith, M. Negin, and AH. Nevis, "Automatic analysis of the electroencephalogram by hybrid computation," *IEEE Transactions on Systems Science and Cybernetics*, pp. 278-283, 1969.

[58] JR. Smith, "Computer aided polysomnography," in *Sleep and Health Risk.*: Elsevier, 1991.

[59] Neurotronics. (2007) Validation Methodology for Pediatric Records. [Online]. http://www.neurotronics.com/index.php?option=com_content&task=view&id=15&Itemid=

[60] G. Klösch et al., "The SIESTA project polygraphic and clinical database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 51-57, 2001.

[61] The SIESTA group. [Online]. http://www.thesiestagroup.com

[62] Royal Philips Electronics. Philips. [Online]. http://www.healthcare.philips.com/pwc_hc/main/homehealth/sleep/somnolyzer/

[63] P. Anderer et al., "An E-Health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24x7 utilizing the Siesta database," *Neuropsychobiology*, vol. 51, pp. 115-133, 2005.

[64] P. Anderer et al., "Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: validation study of the AASM version of the Somnolyzer 24x7," *Neuropsychobiology*, vol. 62, pp. 250-264, 2010.

[65] M. Woertz et al., "Agreement of apnea-hypopnea indexes based on visual and automatic detection," in *ESRS2006*, Innsbruck, 2006.

[66] Grass Technologies. (2012) AURA PSG Ambulatory System. [Online]. http://www.grasstechnologies.com/products/clinsystems/aurapsg2.html

[67] B. Barreiro, G. Badosa, S. Quintana, L. Esteban, and J. Heredia, "Comparación entre el análisis automático y manual de la polisomnografía convencional en el diagnóstico del síndrome de apnea-hipopnea obstructiva del sueño," *Arch. Bronconeumol.*, vol. 39, no. 12, pp. 544-548, 2003.

[68] Compumedics Limited. (2012) Somte PSG. [Online]. http://www.compumedics.com/products.asp?p=39

[69] B. Kemp and M. Roessen, "Polyman: a free(ing) viewer for standard EDF(+) recordings and scorings," in *Sleep-Wake Research in The Netherlands*, SF Ruijgt et al., Eds.: Dutch Society for Sleep-Wake Research, 2007, pp. 71-73.

[70] B. Kemp, AH. Zwinderman, B. Tuk, HAC. Kamphuisen, and JJL. Oberyè, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185-1194, 2000.

[71] R. Golpe, A. Jimenez, and R. Carpizo, "Home sleep studies in the assessment of sleep apnea-hypopnea syndrome," *Chest*, vol. 122, pp. 1156-1161, 2002.

[72] C. Iber et al., "Polysomnography performed in the unattended home versus the attended laboratory setting," *Sleep*, vol. 27, no. 3, pp. 546-540, 2004.

[73] E. Ballester et al., "Evaluation of a portable respiratory recording devide for detecting apnoeas and hypopnoeas in subjects from a general population," *European Respiratory Journal*, vol. 16, pp. 123-127, 200.

[74] PR. Westbrook et al., "Description and validation of the Apnea Risk Evaluation System," *Chest*, vol. 128, no. 4, pp. 2166-2175, 2005.

[75] I. Ayappa, RG. Norman, V. Seelall, and DM. Rapoport, "Validation of a self-applied unattended monitor for sleep disordered breathing," *Journal of Clinical Sleep Medicine*, vol. 4, no. 1, pp. 26-37, 2008.

[76] M. Zucconi, L. Ferini-Strambi, V. Castronovo, A. Oldani, and S. Smirne, "An unattended device for sleep-related breathing disorders: validation study in suspected obstructive sleep apnoea syndrome," *European Respiratory Journal*, vol. 9, pp. 1251-1256, 1996.

[77] JH. Ficker, GH. Wiest, J. Wilpert, FS. Fuchs, and EG. Hahn, "Evaluation of a portable recording device (Somnocheck) for use in patients with suspected obstructive sleep apnoea," *Respiration*, vol. 68, pp. 307-312, 2001.

[78] AC. Oliveira et al., "Diagnosis of obstructive sleep apnea syndrome and its outcomes with home portable monitoring," *Chest*, vol. 135, pp. 330-336, 2009.

[79] Weinmann Medical Technology. (2012, January) Somnocheck effort. [Online]. http://www.weinmann.de/en/homecare/sleep_diagnostics/somnocheck_effort/

[80] CA. Nigro, S. Aimaretti, S. Gonzalez, and E. Rhodius, "Validation of the WristOx 3100 oximeter for the diagnosis of sleep apnea/hypopnea syndrome," *Sleep Breath*, vol. 13, pp. 127-136, 2009.

[81] S. Viner, J. Szalai, and V. Hoffstein, "Are history and physical examination a good screening test for sleep apnea?," *Annals of Internal Medicine*, vol. 115, pp. 356-359, 1991.

[82] W. Flemons, W. Whitelaw, R. Brant, and J. Remmers, "Likelihood ratios for sleep apnea: clinical prediction rule," *American Journal of Respiratory and Critical Care Medicine*, vol. 150, pp. 1279-1285, 1994.

[83] B. Crocer et al., "Estimation of the probability of disturbed breathing during sleep before a sleep study," *American Review of Respiratory Disease*, vol. 142, pp. 14-18, 1990.

[84] J. Rodsutti, M. Hensley, A. Thakkinstain, C. D'Este, and J. Attia, "A clinical decision rule to priorize polysomnography in patients with suspected sleep apnea," *Sleep*, vol. 27, no. 4, pp. 694-699, 2004.

[85] SK. Sharma et al., "Prediction of obstructive sleep apnea in patients presenting to a tertiary care center," *Sleep Breath*, vol. 10, no. 3, pp. 147-154, 2006.

[86] Y. Sweere et al., "The validity of the dutch sleep disorders questionnaire (SQD)," *Journal of Psychosomatic Research*, vol. 45, no. 6, pp. 549-555, 1998.

[87] SD. Kirby et al., "Neural network prediction of obstructive sleep apnea from clinical criteria," *Chest*, vol. 116, pp. 409-415, 1999.

[88] AA. El-Solh et al., "Validity of neural network in sleep apnea," *Sleep*, vol. 22, pp. 105-111, 1999.

[89] J. Victor-Marcos et al., "Radial basis function classifiers to help in the diagnosis of the obstructive sleep apnoea syndrome from nocturnal oximetry," *Medical and Biological Engineering and Computing*, vol. 46, pp. 323-332, 2008.

[90] P. Caseiro, R. Fonseca-Pinto, and A. Andrade, "Screening of obstructive sleep apnea using Hilber-Huang decomposition of oronasal airway pressure recordings," *Medical Engineering and Physics*, vol. 32, no. 6, pp. 561-568, 2010.

[91] NA. Eiseman, MB. Westover, JE. Mietus, RJ. Thomas, and MT. Bianchi, "Classification algorithms for predicting sleepiness and sleep apnea severity," *Journal of Sleep Research*, vol. 21, pp. 101-112, 2012.

[92] P. West and MH. Kryger, "Continuous monitoring of respiratory variables during sleep by microcomputer," *Methods of Information in Medicine*, vol. 22, pp. 198-203, 1983.

[93] JR. Smith, "Computers in sleep research," *CRC Critical Reviews in Bioengineering*, vol. 3, pp. 93-148, 1978.

[94] C. George, T. Millar, and M. Kryger, "Identification and quantification of apneas by computer-based analysis of oxygen saturation," *The American Review of Respiratory Diseases*, vol. 137, pp. 1238-1240, 1988.

[95] H. Rauscher, W. Popp, and H. Zwick, "Computerized detection of respiratory events during sleep from rapid increases in oxyhemoglobin saturation," *Lung*, vol. 169, no. 1, pp. 335-342, 1991.

[96] BH. Taha et al., "Automated detection and classification of sleep-disordered breathing from conventional polysomnography data," *Sleep*, vol. 20, no. 11, pp. 991-1001, 1997.

[97] T. Salmi, T. Telakivi, and M. Partinen, "Evaluation of automatic analysis of SCSB, airflow and oxygen saturation signals in patients with sleep related apneas," *Chest*, vol. 96, pp. 255-261, 1989.

[98] A. Otero, P. Félix, and MR. Álvarez, "Algorithms for the analysis of polysomnographic recordings with customizable criteria," *Expert Systems with Applications*, vol. 38, pp. 10133-10146, 2011.

[99] KM. Al-Ashmouny, AA. Morsy, and SF. Loza, "Sleep apnea detection and classification using fuzzy logic: clinical evaluation," in *27th Annual Conference on IEEE Engineering in Medicine and Biology*, Shanghai, 2005, pp. 6132-6135.

[100] SD. Pittman et al., "Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing," *Sleep*, vol. 27, no. 7, pp. 1394-1403, 2004.

[101] H. Nakano, M. Hayashi, E. Ohshima, N. Nishikata, and T. Shinohara, "Validation of a new system of tracheal sound analysis for the diagnosis of sleep apnea-hypopnea syndrome," *Sleep*, vol. 27, no. 5, pp. 951-957, 2004.

[102] A. Yadollahi, E. Giannouli, and M. Zahra, "Sleep apnea monitoring and diagnosis based on pulse oximetry and tracheal sound signals," *Medical and Biological Engineering and Computing*, vol. 48, pp. 1087-1097, 2010.

[103] C. Guilleminault, SJ. Connolly, R. Winkle, K. Melvin, and A. Tilkian, "Cyclical variation of the heart rate in sleep apnoea syndrome. Mechanisms and usefulness of 24 h electrocardiography as a screening technique," *The Lancet*, vol. 1, pp. 126-131, 1984.

[104] T. Penzel et al., "Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings," *Medical and Biological Engineering and Computing*, vol. 40, pp. 402-407, 2002.

[105] MO. Mendez et al., "Automatic screening of obstructive sleep apnea from the ECG based on empirical mode decomposition and wavelet analysis," *Physiological Measurement*, vol. 31, pp. 273-289, 2010.

[106] A. Figliola, O. Rosso, and E. Serrano, "Detection of delay time between the alterations of cardiac rhythm and periodic breathing," *Physica A: Statistical Mechanics and its Applications*, vol. 327, pp. 174-179, 2003.

[107] C. Maier and H. Dickhaus, "Central sleep apnea detection from ECG-derived respiratory signals," *Methods of Information in Medicine*, vol. 5, pp. 462-466, 2010.

[108] O. Amir et al., "An automated sleep-analysis system operated through a standard hospital monitor," *Journal of Clinical Sleep Medicine*, vol. 6, no. 1, pp. 59-63, 2010.

[109] M. Clabian, C. Nussbaum, and H. Pfützner, "Artificial neural networks for apnea detection," in *EANN*, 1996, pp. 601-608.

[110] M. Clabian and H. Pfützner, "Determination of decisive inputs of a neural network for sleep apnea classification," in *EANN*, 1997, pp. 171-178.

[111] T. Zemen, M. Clabian, and H. Pfützner, "Classification of sleep apnea events by means of radial basis function networks," in *ICSC/IFAC Symposium on Neural Computation (NC98)*, 1998, pp. 351-357.

[112] O. Fontenla-Romero, B. Guijarro-Berdiñas, A. Alonso-Betanzos, and V. Moret-Bonillo, "A new method for sleep apnea classification using wavelets and feedforward neural networks," *Artificial Intelligence in Medicine*, vol. 34, pp. 65-76, 2005.

[113] ME. Tagluk, M. Akin, and N. Sezgin, "Classification of sleep apnea by using wavelet transform and artificial neural networks," *Expert Systems with Applications*, vol. 37, pp. 1600-1607, 2010.

[114] D. Peteiro-Barral, B. Guijarro-Berdiñas, and E. Hernández-Pereira, "Classifying sleep apneas using neural networks and a combination of experts," *Lecture Notes in Artificial Intelligence*, vol. 5988, pp. 271-280, 2010.

[115] J. Oliveira, A. Tome, J. Chunha, LM. Catalao, and J. Azevedo, "Sleep data integration and analysis: an object oriented approach," in *15th Annual Conference of the IEEE Engineering in Medicine and Biology Society*, 1993, pp. 455-456.

[116] P. Guedes-Oliveira, J. Principe, A. Cruz, and A. Tome, "HIDRA: A hierarchical instrument for distributed real-time analysis of biological signals," *IEEE Transactions on Biomedical Engineering*, vol. 24, no. 12, pp. 921-927, 1987.

[117] J. Oliveira and A. Tome, "Sistema automatico para a monitorizaçao de exames neurofisiológicos de sono," *Revista do DETUA*, vol. 2, no. 3, 1998.

[118] A. Fred, J. Filipe, M. Partinen, and T. Paiva, "PSG-Expert - An expert system for the diagnosis of sleep disorders," in *Studies in Health Technology and Informatics*, T. Paiva and T. Penzel, Eds.: IOS Press, 2000, vol. 78, pp. 127-147.

[119] Software Technology Branch, Lyndon B. Johnson Space Centre, "CLIPS Reference Manual, CLIPS Version 6.04," 1997.

[120] E. Hernández-Pereira, Técnicas de inteligencia artificial e ingeniería del software para un sistema inteligente de monitorización de apneas del sueño, 2000.

[121] M. Cabrero-Canosa et al., "An intelligent system for the detection and interpretation of sleep apneas," *Expert Systems with Applications*, vol. 24, pp. 335-349, 2003.

[122] M. Cabrero-Canosa, E. Hernandez-Pereira, and V. Moret-Bonillo, "Intelligent diagnosis of sleep apnea syndrome," *IEEE Engineering in Medicine and Biology Magazine*, vol. 23, no. 2, pp. 72-81, 2004.

[123] C. Iber, S. Ancoli-Israel, A. Chesson, and SF. Quan, "The AASM Manual for the scoring of sleep and associated events: rules, terminology and technical specifications," American Academy of Sleep Medicine, Westchester, IL, 2007.

[124] SensorMedics, "Somnostar LabManager operator's manual and tutorial," 1998.

4. FUZZY SYSTEMS

In this chapter a description on the use of methods for fuzzy analysis of information is given as a fundamental piece within the technological framework established for the development of the proposed system. Hypotheses that have conducted to the use of artificial intelligence techniques being able to handle imprecise information, and specifically of fuzzy logic, have been previously introduced (see Chapter one "*Scope and objectives*", Chapter three "*Handle of imprecise information*" and "*Critical analysis*"), and include avoiding of categorical results, increasing of generalization capabilities by minimization of effects of variability due to noise, subjectivity and criteria variability, improvement of explanation capabilities, or easier knowledge representation.

On the other hand, it is necessary to clarify that even if technological aspects of the developed system go beyond the use of the techniques here described, detailed description of all of them would result totally unfeasible. However because of the relative importance that fuzzy inference processes have within the system, as well as their differential character with respect to current existent approximations for the intelligent diagnosis of SAHS (see Chapter three), it has been considered adequate to carry out a more detailed description of their fundamentals, which is carried out throughout this chapter.

Thus, in the following, theoretical basis of fuzzy logic are explained, showing the possibilities that fuzzy systems allows at the time of handling of uncertainty and carrying out reasonings with imprecise information. Description followed throughout the chapter is conducted from the general perspective, however examples are developed showing its practical applicability in the domain of SAHS diagnosis. Specific

developments using this technology are then included in the final system which is described throughout the subsequent chapter five.

4.1. Introduction to fuzzy logic

If we go out to the street and see a cloudless sky, then we will not think about the possibility of taking out the umbrella. However if we see a clouded sky, we may consider raining as feasible. If, besides, there are very dark clouds and we are within the winter months, we will probably think that there is a high chance of getting wet if we do not have an umbrella, or said with other words, that there is a high possibility of rain. Nevertheless, it also happens that sometimes in the presence of a clear sky, the day unexpectedly became overcast and it finally rained, and vice versa. Most of the times, however, we will be right on the previous predictions.

It seems, according to the preceding example, that exact knowledge is quite unusual in real life, while on the contrary, events are always affected by a certain degree of uncertainty. On the other hand, continuing with the example, let us now to consider the possibility of implementing an automatic reasoning system to deal with the correct weather forecast. With the view of developing a model that mimics human's reasonings, the modeling task should face the necessity of quantifying the terms used above. Accordingly, it is known that people do not express using sentences such as *there is a 79% chance of rains*, but they usually express vague terms such that there is a *high possibility* that it rains. However: what is the meaning of *high possibility*? How can it be quantified using a computational schema?

Fuzzy logic, which is based on the theory of fuzzy sets [**1**], is a theory that allow us managing and processing information, in which prevails the use of inexact, imprecise or subjective terms. Similarly as it does human brain, it is possible to carry out reasonings based on imprecise rules[22] and over incomplete data[23].

[22] Which is related with the concept of uncertainty
[23] Which is related with the concept of imprecision

As it was previously outlined (see Chapter three, "*Handling of imprecise information*"), there are other reasoning models being able to handle imprecision and uncertainty, such as for example, evidential theory of Dempster and Shaffer [**2**], or the model of certainty factors of Shortliffe and Buchannan [**3**]. However fuzzy logic, besides containing previous ones in terms of capabilities of the model[24], it also establishes a natural mechanism of correspondence between imprecision supported by the model and the natural language [**4**]. Such correspondence provides of an optimal framework for the current modeling task, allowing approximation to the clinical language, both at the time of explaining system's results, and by allowing knowledge to be represented by means *fuzzy rules*. The former facilitates the knowledge acquisition task[25] by permitting transferring of expert knowledge, practically without the need of its reformulation, through the use of fuzzy rules as the representation schema.

On the other hand, it is important to make a distinction between the different approximations to the problem of uncertainty handling, as provided by fuzzy logic and the theory of probability. Fuzzy theory talks about the degree of membership of an element to a certain set and not about the probability of occurrence of an event. In this respect, many authors rather prefer to talk about a *theory of possibility* to refer to interpretation of uncertainty under the fuzzy logic perspective.

From a historical point of view, theoretical bases of fuzzy logic were enunciated for the first time by Lofti A. Zadeh [**1**], professor of electrical engineering at University of California in Berkeley. However, it was not until 1973 when Zadeh presented the basic theory of fuzzy controllers [**5**]. The main idea consists in that, differently from classical logic where entities are bivalent, within fuzzy logic entities are characterized with respect to a set by a value of membership $\mu(x)$, which is a real number in the interval [0,1][26]. Such an idea about the existence of certain degrees of veracity and falsehood can be founded yet in the time of Aristotle, who already considered this possibility, or in Plato, who had also considered the existence of certain grades of membership.

[24] Under certain conditions, fuzzy logic schema is equivalent to the preceding models. That is, it can be considered as a superset of the two previous, or these as particular cases of the fuzzy paradigm
[25] The great bottleneck in the development of expert systems
[26] Where in the extreme case 0 indicates total absence of membership whereas 1 indicates total membership

Although initially the work of Zadeh was coldly received (especially in the US), from his work several researchers began to apply fuzzy logic to several processes. In this manner, and mainly after Mamdani applied fuzzy logic to control systems [**6**], several applications were developed. In this respect, without any doubt, the first place where fuzzy systems have obtained a high success was in Japan, where they were importantly applied for the first time in Japanese underground with excellent results. Special mention deserves the creation of LIFE (*Laboratory for International Fuzzy Engineering research)* in 1989, promoted by the Ministry of Economy Trade and Industry (METI) of Japan. In the US and Europe, only when Japan started to inform about the numerous practical applications, fuzzy logic was given importance. Since then northamerican companies such as NASA, Boeing, Ford, Rochwell or Bell commenced to apply fuzzy logic on their projects, and nowadays fuzzy logic has proved to be success in a large variety of applied domains with special relevance in the modeling of control processes [**7**].

4.2. Fundamentals

It has been mentioned in the previous section that fuzzy sets theory extends classical sets theory, allowing an element to be partially included within a set with a certain degree of membership. That is, in classical sets theory the membership function can be understood as function with just two discrete output values, 0 or 1, if the element is not included in the set or it does, respectively. Therefore, the main difference within fuzzy set theory is that the membership function is not discrete anymore but it becomes, on the other hand, continuous in the interval [0, 1].

Formally, given a universe of discourse *U* or universal reference set, it can be defined the *membership function* with respect to the fuzzy subset *A*, of the elements of the universe *U*, as follows:

$$\mu_A(x): U \rightarrow [0,1], \forall x \in U.$$

where A is a *linguistic label* that identifies the fuzzy subset. Then, given a certain element $x_i \epsilon U$, conceptually $\mu_A(x_i) = u$, means that x_i is included with a degree of membership u in the concept represented by A (see Figure 4.1).

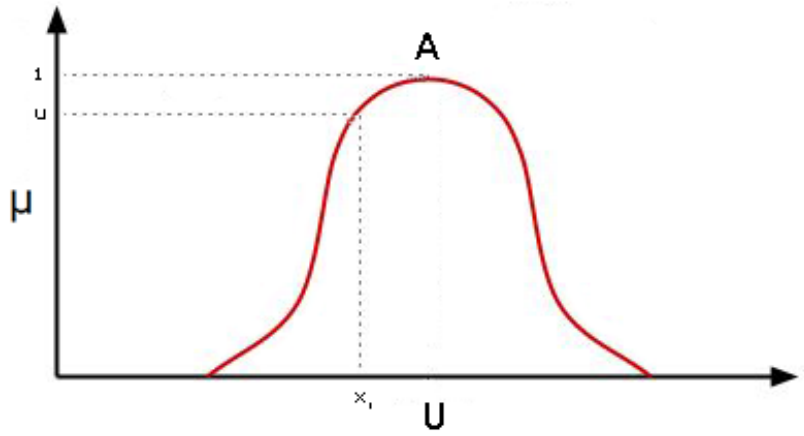


Figure 4.1. How to obtain the degree of membership from a fuzzy set A from a given value x_i

For the definition of these membership functions one normally turns to the use of certain conventionally defined families of standard forms. Ultimately the choice of one family or another, as well as the exact value of their respective parameters depends on its suitability to represent the desired concept in the actual application domain. Some of the most frequently used are functions of type trapezoidal, singleton, triangular, *S* or bell-shaped, which are shown in Figure 4.2.

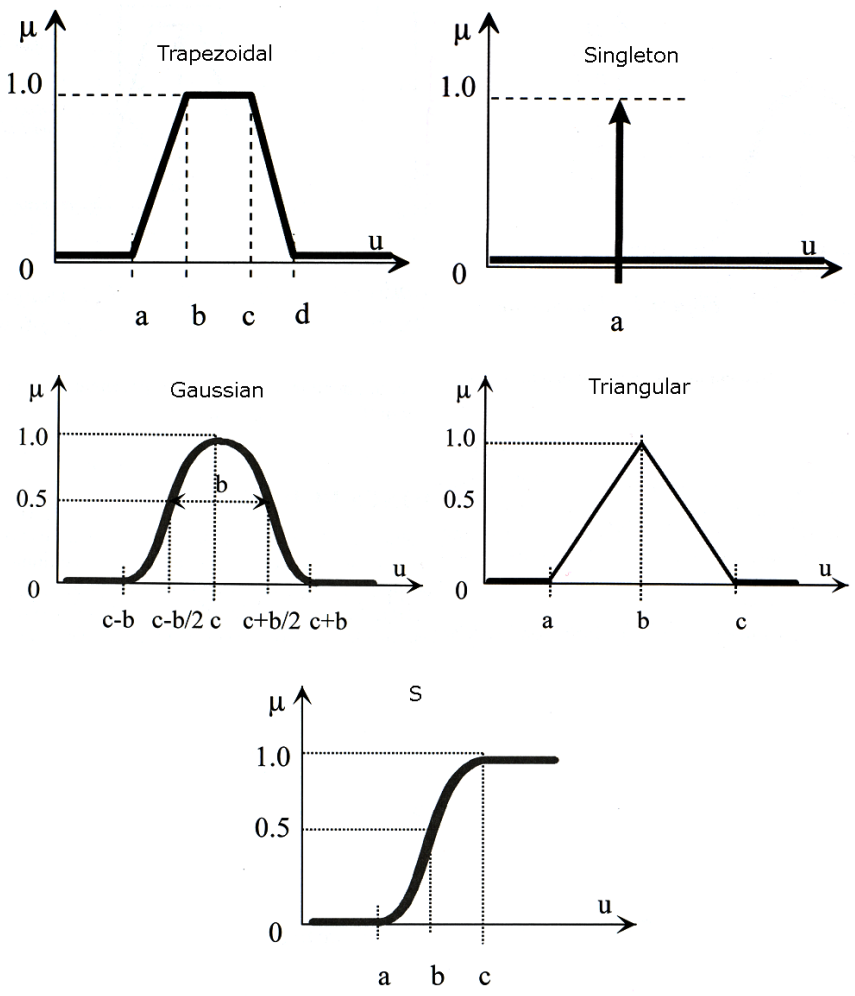


Figure 4.2. Examples of some of the most used families of membership functions: trapezoidal, singleton, Gaussian, triangular and *S*

It is necessary to make a distinction between the *universe of discourse* or *referential U* and the concept of *linguistic label*. The universe of discourse represents the set of possible values which a variable can take, normally numeric. It indeed represents its domain and it is expresses in quantifiable terms. When we talk about linguistic labels, on the other hand, we do it from the qualitative perspective, associating a certain set of values from the referential *U* to a linguistic value. The set of different linguistic values can be interpreted, similarly as in the numeric case, as the set of possible values of a variable, in this case, a *linguistic variable.*

In this manner, within the same universe of discourse, there can be defined several linguistic variables, each one with domain a certain subset of values of the referential, not necessarily disjoints. At the same time within each linguistic variable a set of

linguistic labels can be defined, each one associated with a subset of the domain of the corresponding linguistic variable. To the result of projecting the set of linguistic labels of a linguistic variable over the corresponding referential subset is often referred as *partitioning* of the linguistic variable.

The former is valid both from the fuzzy point of view as well as from the perspective of the ordinary sets (also referred as *crisp*[27]). The only thing that establishes the difference is the concept of set. In this respect, in the classical sets theory, the values pertaining to a certain linguistic label have as their membership function the unity. On the other hand, in the theory of fuzzy sets the membership function is defined as $\mu_A(x): A \rightarrow [0,1], \forall x \in A \subset U$.

For example, one may consider the linguistic variable *age*. It can be established an arbitrary delimitation of the different numeric values that classifies a person within the categories *young*, *adult* and *old*. Under classical sets theory the partition may look like as in Figure 4.3. In contrast, by using fuzzy sets the corresponding partition over the referential *years of a person*, with domain $\mathbb{R}^+$, can be considered as in the following Figure 4.4.

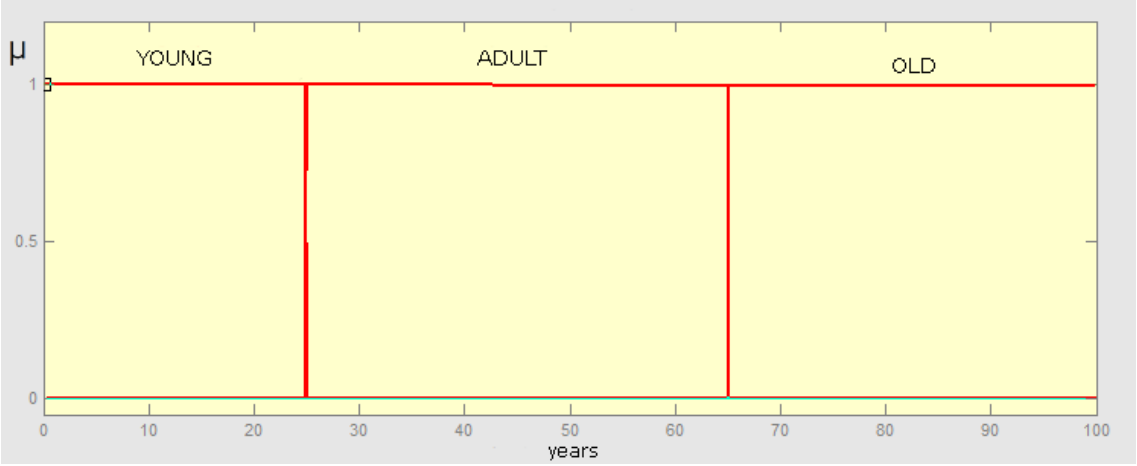


Figure 4.3. Example of a possible crisp classification for the variable *age*

[27] Often with the term *crisp* one refers to everything that derives from the concept of classical set, to differentiate it from that related with the *fuzzy* approximation

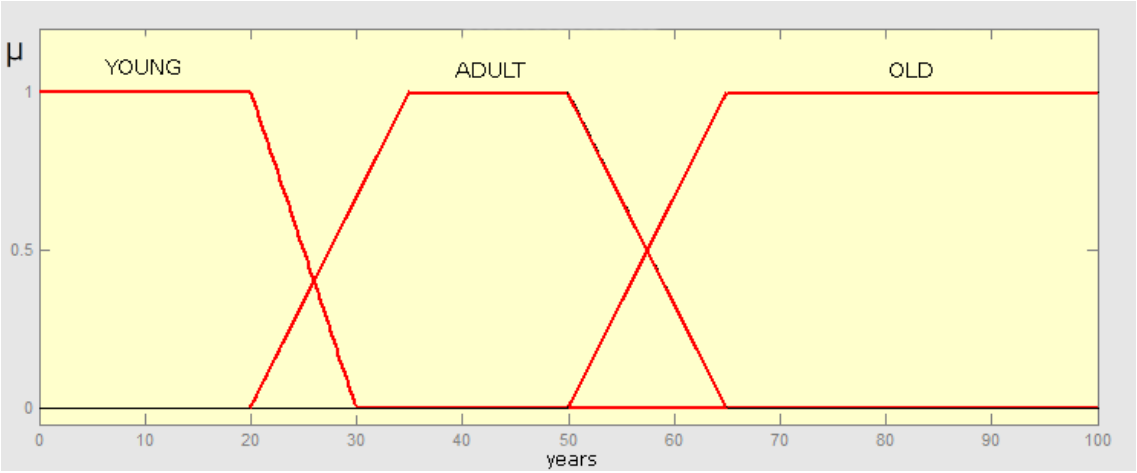


Figure 4.4. Example of a possible fuzzy partition for the variable *age*

Finally let us consider an example showing a fuzzy partition applied to SAHS domain. In order to evaluate the degree of respiratory airflow reduction with respect to the normal respiration, the following possible fuzzy partition over the linguistic variable *respiratory airflow reduction* can be considered:

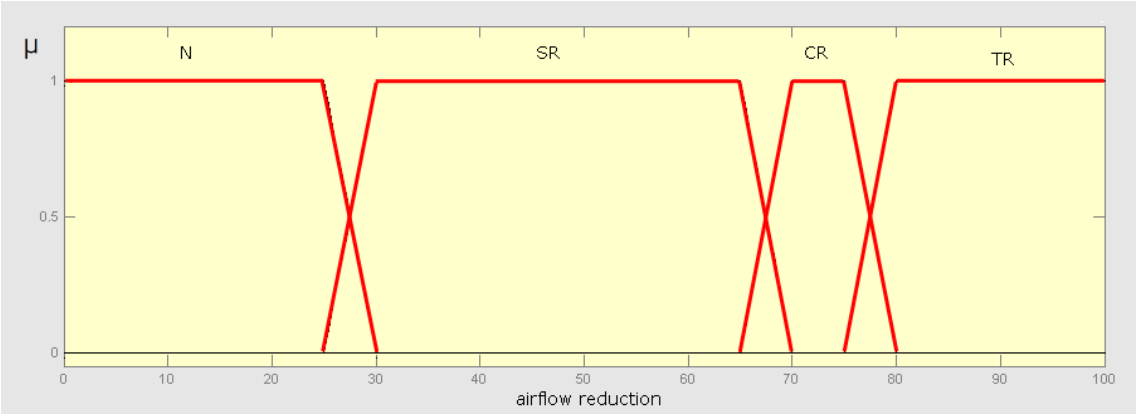


Figure 4.5. Fuzzy partition of variable airflow reduction; N = normal; SR = Slightly Reduced; CR = Clearly Reduced; TR = Totally Reduced

4.2.1. Operations with fuzzy sets

In his paper *Fuzzy Sets* in 1965 [**1**], Zadeh provides the basic definitions over fuzzy sets which are the natural extensions of the corresponding definitions for ordinary sets. In this respect a fuzzy set A defined on the universe U is *empty* if and only if its membership function is zero for all the values $x \epsilon U$:

$$\mu_A(x) = 0, \forall x \epsilon U$$

Two fuzzy sets *A* and *B* are *equal*, written as *A* = *B*, if and only if

$$\mu_A(x) = \mu_B(x), \forall x \epsilon U$$

On the other hand, A is *contained* in B (or, equivalently, $A \subset B$, *A* is a subset of *B*, or *A* is smaller than *B*) if and only if $\mu_A(x) \leq \mu_B(x), \forall x \epsilon U$.

Taking into account the previous definitions, Zadeh proposes also definitions for the basic operators of *complement*, *union* and *intersection*. Behavior of these basic operations is similar to their corresponding equivalents in the classical sets theory. As it has been previously pointed out, fuzzy set theory reduces to classical set theory if uncertainty is set to zero, that is, it only admits values of 0 and 1 for the output of the membership functions. Thus, according to Zadeh, given two fuzzy sets *A* and *B* defined over a referential *U*, such that $A \subset U$ and $B \subset U$, the previous operations are defined as follows:

- *Complement*: $\mu_{\bar{A}}(x) = 1 - \mu_A(x), \forall x \epsilon U$
- *Union*: $\mu_{A \cup B}(x) = \max[\mu_A(x), \mu_B(x)], \forall x \epsilon U$
- *Intersection*: $\mu_{A \cap B}(x) = \min[\mu_A(x), \mu_B(x)], \forall x \epsilon U$

Note that both ∪ and ∩ satisfy the associative and the commutative properties. One may also prove that with the previously defined operations, fuzzy sets satisfy distributive and De Morgan's laws [**1**]. At this point it seems that fuzzy sets with previous definitions of complementation, union and intersection and taking the empty set $\mu_\perp(x) = 0, \forall x \epsilon U$ as the null element and $\mu_T(x) = 1, \forall x \epsilon U$ as the unitary element, have structure of Boolean algebra. However, fuzzy sets do not satisfy neither the law of non-contradiction nor the law of excluded middle, which prevents fuzzy sets to fully verify the required conditions of Boolean algebra [**8**].

Another common operators used among fuzzy sets are the namely *modifiers*. These operators receive such a name because they allow obtaining the result of applying a modifying term in the common language. For example, natural language expressions of *very* and *more or less* can be applied over a fuzzy set linguistically represented by

certain label *A*. For example, if *A = hot*, for a given value *x*, then it can be obtained the resulting degree of membership with respect to the modified fuzzy set *very hot* as follows:

$$\mu_{very\ hot}(x) = (\mu_{hot}(x))^2$$

Analogously:

$$\mu_{more\ or\ less\ hot}(x) = (\mu_{hot}(x))^{1/2}$$

4.2.2. Fuzzy logic

At the time of establishing the correspondence between sets theory and logic, operations described in the previous subsection found their corresponding analogue in the operations of negation, disjunction –logical OR- and conjunction –logical AND- respectively. More specifically fuzzy logic can be defined as a family of logics pertaining to the broader class of *many-valued logics*. Moreover, in practice, most of them usually belong to the so-called *t-norm fuzzy logics* [**9**].

Without the aim to go into excessive details for the purposes of this introductory chapter, it is enough to consider that, generally speaking, in this kind of fuzzy logics operators of conjunction and disjunction, are generalized with the condition of satisfying certain restrictions. The functions that satisfy these restrictions are respectively known as *t-norms* (or unabbreviated, triangular norm) and *t-conorms* (also called S-norms).

Definition. The binary operator $*: [0,1] \times [0,1] \rightarrow [0,1]$, if it satisfies the following properties:

1. Commutativity: $a * b = b * a$;
2. Associativity: $(a * b) * c = a * (b * c)$;
3. Monotonicity: $a \leq b\ y\ c \leq d \Rightarrow a * c \leq b * d$;
4. Identity element the unit: $a * 1 = a$,

where operands $a, b, c, d\ \epsilon\ [0,1]$, then the operator $*$ is said to be a T-norm in [0,1], and is denoted by Δ.

Definition. The binary operator $*: [0,1] \times [0,1] \rightarrow [0,1]$, if it satisfies the following properties:

1. Commutativity: $a * b = b * a$;
2. Associativity: $(a * b) * c = a * (b * c)$;
3. Monotonicity: $a \leq b \; y \; c \leq d \Rightarrow a * c \leq b * d$;
4. Identity element zero: $a * 0 = a$,

where operands $a, b, c, d \; \epsilon \; [0,1]$, then the operator $*$ is said to be a T-Conorm in [0,1], and is denoted by ∇.

T-conorms can also be defined as the dual forms of T-norms -and vice versa- under the order-reversing operation which assigns *1- x* to *x* on [0,1]. Indeed for example given a T-norm, the complementary T-conorm is defined by

$$\nabla(a, b) = 1 - \Delta(1 - a, 1 - b)$$

In practice, in fuzzy logic the concrete form of the previous defined operators depends on the concrete application domain, with the sole condition that the resulting operators must satisfy the previous properties. Some of the most common implementations of T-norms and the corresponding T-conorms are shown in Table 4.1. In Table 4.1 *a* and *b* respectively represent the degrees of membership with respect to the associated fuzzy sets *A* and *B*.

Table 4.1. Examples of some of the most common triangular norms and conorms

T-norm	**T-conorm**
$min(a, b)$	$max(a, b)$
$a \cdot b$	$(a + b - a \cdot b)$
$max(0, a + b - 1)$	$\min(1, a + b)$

Each concrete implementation leads to a specific logic within the family of *t-norm fuzzy logics*. In any case, as it was mentioned above, ultimately the concrete implementation will depend on the concrete application domain. Equivalently to the case of classical logic, triangular norms and conorms obey the DeMorgan's laws that

relate them. It can also be proved that functions $\min(\cdot)$ and $\max(\cdot)$ -the ones proposed by Zadeh- are the most restrictive forms of T-norms and T-conorms respectively. For a more detailed analysis of the different norms and conorms and their implications the reader is referred to [**10**].

4.2.3. Inference in fuzzy logic

As it has been outlined in the previous subsection, fuzzy sets can be reinterpreted as predicates in propositional logic. In fact, in the same manner that an isomorphism between logic and classical sets theory can be defined, it is possible to define an isomorphism between fuzzy logic and theory of fuzzy sets. Reasoning based on natural language is a kind of approximate reasoning, which makes use of propositions and predicates expressing information of imprecise nature. In fuzzy logic, knowledge must be interpreted as a collection of fuzzy constraints which operate over a set of variables. Therefore it is just about incorporating imprecision into classical logic of predicates by means of theory of fuzzy sets. The analogy is represented in the schema of Figure 4.6.

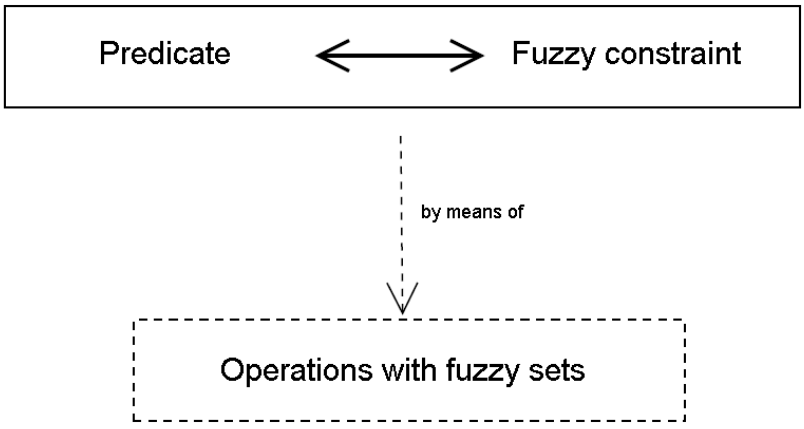


Figure 4.6. Interpretation of logic of predicates through fuzzy logic

Because it is not the objective of this chapter to carry out an exhaustive analysis about reasoning with predicates and its interpretation in fuzzy logic, we will not go into further details. What is intended to remain clear is that fuzzy logic allows us to reinterpret logic of predicates adding mechanisms for handling of uncertainty. In this regard, truthfulness of a predicate is not categorical anymore, but it is evaluated as a function of the degree of membership of each individual clause, and it is subsequently

combined by the corresponding equivalent fuzzy operators. This is carried out by establishing an application from clauses –or sentences- in the natural language, and the numerical context where the degree of membership of each sentence is quantified with respect to its associated fuzzy set. For example, in the context of SAHS, the truthfulness of the sentence "*Desaturation is high or airflow is clearly reduced*" can be evaluated as follows:

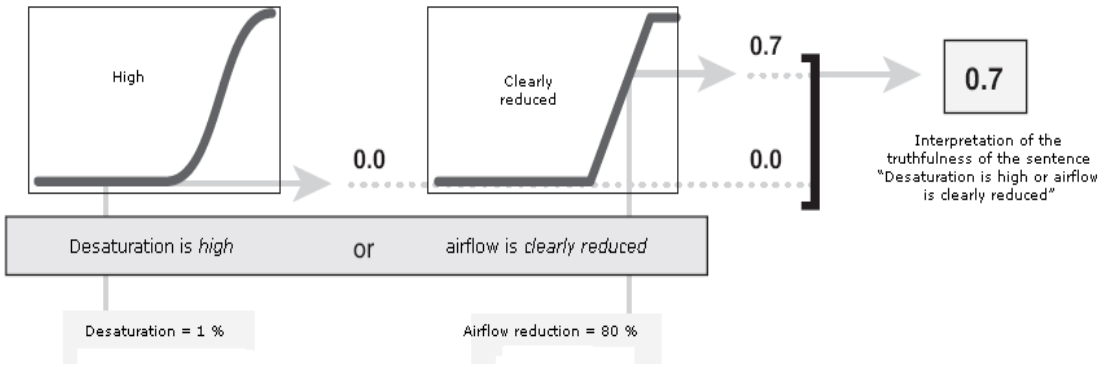


Figure 4.7. Example of the possible interpretation of truthfulness of a proposition using fuzzy logic

Definition. Given two universes of discourse *U* and *V*, a *fuzzy relation* is defined as a fuzzy set in *R* in the space *UxV*, which membership function is denoted as $\mu_R(u, v)$, with $u \epsilon U$ y $v \epsilon V$.

In practice, at the discrete level such relationship can be represented by a table where, for example, to each pair formed by a column element of the referential *U* and a file element of the referential *V*, is assigned a degree of membership in the space *UxV*. In the continuous level the resulting space can be represented through a surface. Indeed it is the bidimensional interpretation of a fuzzy set, but this interpretation can be easily extensible to the *n*-dimensional case.

On the other hand it has to be taken into account that the number of relationships which can be established between two fuzzy sets *A* and *B*, respectively defined over spaces *U* and *V*, is potentially infinite[28].

[28] It is easy to think about the number of possible functions which can be defined between two variables

Definition. Given two fuzzy sets *A* y *B* in *U* y *V* respectively, a *fuzzy implication* of *A* in *B*, which is indicated $A \rightarrow B$, is a concrete type of fuzzy relationship in *UxV* which represents an IF-THEN rule, or from the fuzzy logic point of view, the fuzzy implication operator. Implementation of implication can also vary, but in its more general form it obeys the following formula:

$$\mu_{A \rightarrow B}(u, v) = \mu_A(u) * \mu_B(v)$$

where the concrete implementation of *Star* $(*)$ operator can be any T-norm as those defined in the previous subsection.

Definition. Given two fuzzy relations *R* and *S* defined over *UxV* and *VxW*, with *U*, *V* and *W*, referentials, the *composition* of relations *R* and *S* is defined as a new fuzzy relationship over the space *UxW*, and denoted $R \circ S$.

Interpretations of operator composition to define the concrete formula for the membership function $\mu_{R \circ S}(u, v)$ of the resulting fuzzy relationship are, again, diverse. One of the main used is the so-called *Sup-Star composition*, which is defined as follows:

$$\mu_{R \circ S}(u, v) = sup_{v \epsilon V}[\mu_R(u, v) * \mu_S(v, w)]$$

with $u \epsilon U, v \epsilon V, w \epsilon W$.

Likewise *Sup-Star composition* leads the concrete choice of the corresponding T-norm to be interpretable. In this respect, for example, Zadeh [**5**] uses the operator min $(\cdot)$ which leads to the *Sup-Min composition* (also known as *Max-Min composition*); on the other hand, if the operator product is used, the resulting formula is known as *Sup-Prod composition* or *Max-Prod composition*.

In any case, once the previous definitions have been stated, all the necessary is available in order to reinterpret classical inference at the fuzzy level. In fact, classical reasoning method of *Modus Ponens* establishes that:

(Knowledge): If *A* then *B*

(Fact): *A*

(Consequence): *B*

In the fuzzy plane, classical *Modus Ponens* is reinterpreted as *Generalized Modus Ponens (GMP)* in which the following inferential schema is established:

(Knowledge): If *x is A* then *y is B*

(Fact): *x is A'*

(Consequence): *y is B'*

In the former schema the fuzzy relationship defined by the implication represents the knowledge, in the sense that it defines how elements *x* defined in the referential *U*, and in relation with the fuzzy set *A*, are related with their corresponding projections *y* in the referential *V* and in relation with the fuzzy set *B*. In other worlds, it formalizes how the membership with respect to the input fuzzy set *A* defines a correspondence with the fuzzy set *B* at the output. On the other hand the fact presented at the input of the inferential process is a certain fuzzy set *A'*, which even if similar to *A*[29], it does not necessarily has to match exactly the represented model –*A*- of the given knowledge. That said, given *A'*, that is, if *x is A'*, then GMP provides of an approximate reasoning mechanism such that for the elements *y* of the referential *V*, a new fuzzy set *B'* can be defined, more or less similar to *B*, as much as *A'* is similar to *A*.

Let us take an example of daily life: if we look at someone's old photo, we can still recognize him/her. In fact recognition will be easier as the more recent the photo is, just because its similarity with the current physical look will be higher. In fact we are certainly able to recognize the person although the photos show different scenes. In other words, the fact allowing us to recognize the person is that the physical aspect in the two photos is similar. Hence, and going back over GMP, the idea is that if *A implies B*, then *something similar to A* should imply *something similar to B.*

[29] A possible way to quantify how similar are *A* and *A'* might be by calculating their fuzzy intersection

Formally, if *R* is a fuzzy relation from *U* to *V*, and *A'* is a fuzzy set in *U*, then the fuzzy set *B'* in *V* induced by *A'* is calculated by means of the composition of *R* and *A'*, that is:

$$B' = A' \circ R$$

Once again, the eventual concrete formula will exclusively depend on the chosen operators to implement the T-norms and the T-conorms implicit in the composition formula.

4.3. Fuzzy inference systems

In general terms fuzzy inference can be considered as the mapping process from the input of our inference system, to obtain the corresponding output through approximate reasoning mechanisms using the schema proposed by the GMP. Such process has been outlined in the previous section, and it involves all the concepts described throughout this chapter: membership functions, operations in fuzzy logic and IF-THEN fuzzy rules.

Fuzzy Inference Systems (FIS) are intelligent systems based on rules which as the mechanism for exploration of their knowledge (rulebase) use fuzzy inference. These systems have been successfully used in several fields such as for example automatic control, data classification, decision analysis, expert systems or computer vision. Precisely, because of their multidisciplinary nature, fuzzy inference systems have been associated to a wider range of names including, among others, fuzzy rule-based systems, fuzzy expert systems, fuzzy modeling, fuzzy associative memories, fuzzy controllers, or simply and ambiguously, fuzzy systems. Even though it might be possible to search the nuances for each definition, fundamental basis that govern their operation are the same. In this respect, the above mentioned denomination of *fuzzy inference systems* will be the one used throughout the rest of the document.

A rule by itself, in general, does not help a lot and there are necessary several rules to consider different circumstances of the facts which complement each other. The use of GMP, which is able to produce reasonings based on similarity, provides with interesting consequences it is generalized within fuzzy inference systems when through

the use of several fuzzy rules simultaneously. One of the most important is that presence of a single fact can activate several rules at the same time. The previous can occur since, due to fuzziness of the corresponding fuzzy sets, it is possible that the same event pertains to several fuzzy sets with different degrees of membership. Thus it can partially match with the antecedents of different rules, not being strictly necessary the exact matching with the antecedent in order to active the rule, as it happens in the classical rule-based systems.

On the other hand, and even though there exist several possible implementations under the concept of FIS, often throughout literature two great families of fuzzy inference systems can be differentiated. In this respect distinction general can be made between FIS of type Mamdani and those of Takagi-Sugeno type.

Mamdani FIS receive their name in honor to Ebrahim Mamdani who in 1975 published their studies on the control of a steam engine and a heater, synthesizing a set of control rules obtained from human operators' expertise [**11**]. At the same time Mamdani's work was based on the paper published two years before, in 1973, by Zadeh, about the use of fuzzy algorithms in complex systems and decision processes [**5**]. Basically a Mamdani fuzzy system produces outputs in form of fuzzy sets, allowing expressing rules into a language closer to natural language, as it was previously explained, by means of assigning linguistic labels to the corresponding fuzzy sets. On the other hand, in the context of control systems often final output is required to be numeric rather than linguistic. In this respect, fuzzy sets obtained at the output of a Mamdani type controller are subject to an aggregation procedure after which a concentration process –also known as defuzzification- takes place, hence finally obtaining the required numeric value. The schematic process is depicted in Figure 4.8.

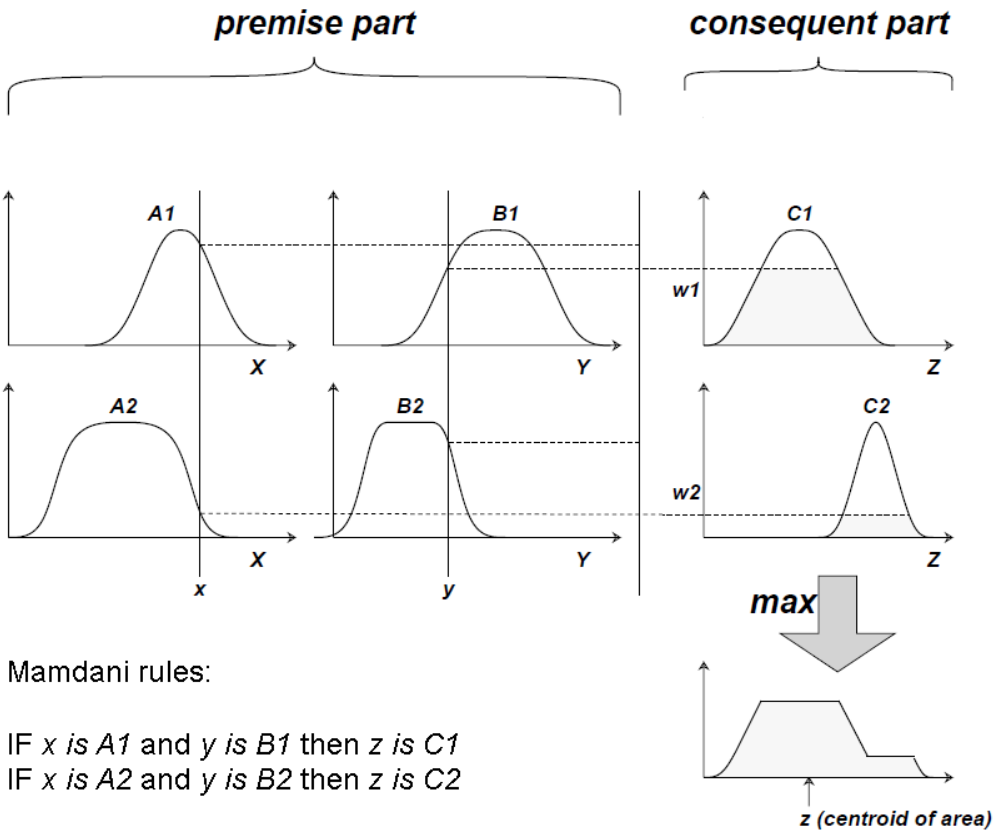


Figure 4.8. Fuzzy rules and general reasoning schema of a Mamdani fuzzy system

On the other hand, in many cases it is more efficient to use a membership function of type *singleton* at the output instead of a distributed fuzzy set. In this manner an increment in the efficiency of the concentration process is produced, because it largely simplifies the required computation with respect to the most general Mamdani method - that searches for the centroid of a bidimensional function. In contrast, in the case of a singleton function type, final value is already numeric and both aggregation and defuzzification can take place simultaneously by using a simple weighted average of a few data points. Indeed, fuzzy systems of Takagi-Sugeno type (or simply Sugeno), introduced in 1985 by Michio Sugeno [**12**], work in this manner, which makes them more adequate for domains where system's efficiency is more important than its capabilities to represent knowledge in a human language manner. In fact, within Sugeno fuzzy systems, output of each rule is directly calculated as the linear combination of the inputs plus a constant term, weighted by the activation weights of each rule, as it is shown in Figure 4.9.

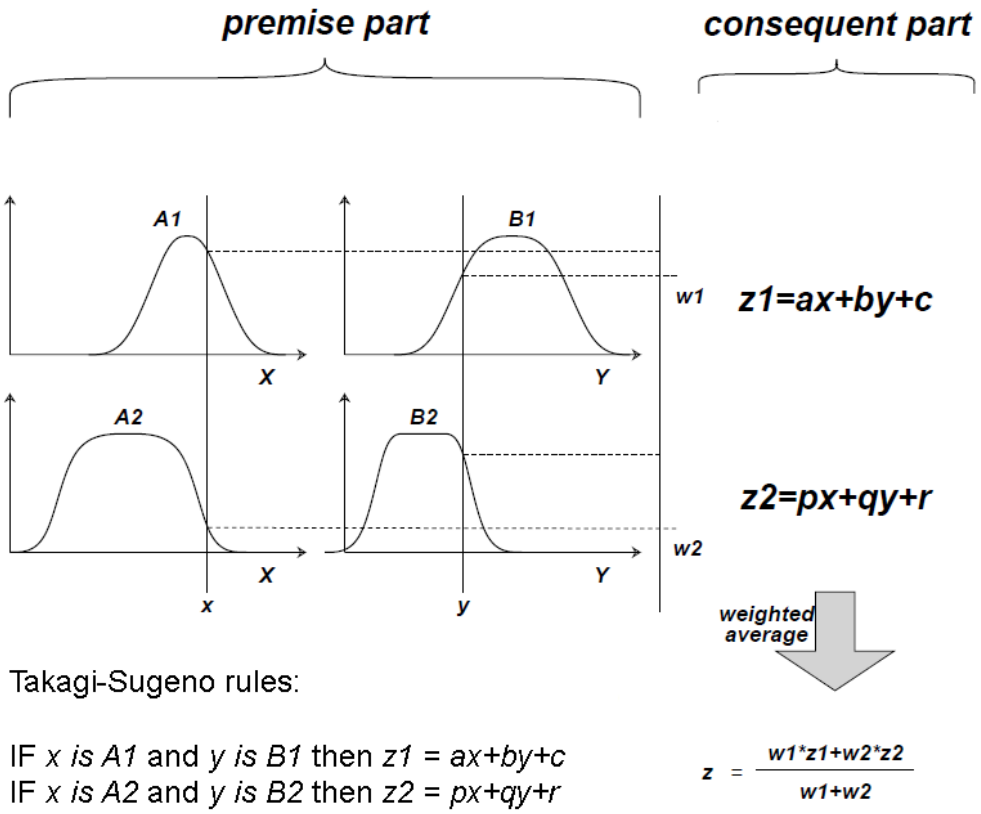


Figure 4.9. Fuzzy rules and general reasoning schema of a Takagi-Sugeno fuzzy system

In order to illustrate more in detail how FIS operate, and how they can be applied to SAHS domain, the general process of a reasoning cycle is shown in the following applied to calculate the possibility of occurrence of an apneic event. In the example the used FIS is of type Mamdani for being these more convenient among reasoning problems which comprise capabilities to explain its behavior, as it happens in the application domain of this doctoral thesis: medical diagnosis of SAHS. FIS of this type are in fact used to control reasoning processes on different parts of the developed system. It has to be remarked, however, that the following example does not necessarily correspond with the concrete implementation followed within the system. The objective here is just to illustrate the working cycle of a Mamdani FIS type with an example. The reader is referred to consult the corresponding sections of Chapter 5 for further details on the concrete implementation of the FIS within the developed system.

That said, for the purposes of this example a FIS can be considered that receives as inputs the reduction in the respiratory airflow and its associated desaturation, obtaining at the output, the degree of membership to the fuzzy variable *apneic event*. Thus, let us suppose that the following inputs are presented to the system: (a) an airflow reduction of 80% accompanied by (b) a 2% desaturation in SaO_2 signal. The reasoning cycle of a Mandami fuzzy system comprises the following steps:

1. ***Input fuzzification:*** Solve all the possible matches with the input variables of the rule antecedents in the knowledge base[30] to obtain the corresponding degrees of membership, between 0 and 1, with respect to the fuzzy sets of the respective partitions (see Figure 4.10).

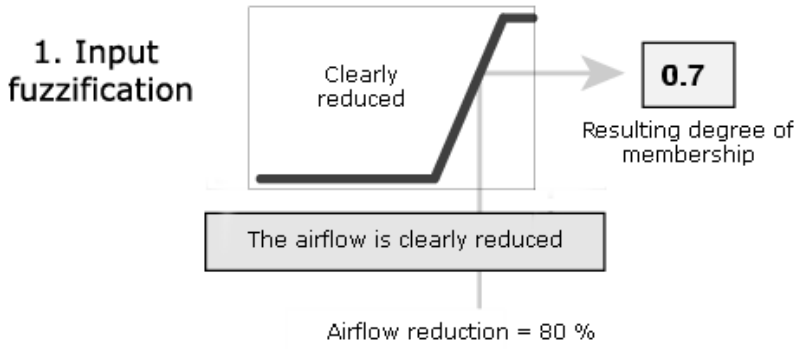


Figure 4.10. Step 1: The figure shows how to fuzzify the numeric input by obtaining the corresponding degree of membership with respect to the fuzzy set *clearly reduced* in the antecedent part of a rule

2. ***Calculate rule activation***: If the antecedent is composed of several elements, then apply the corresponding fuzzy operators to obtain the resulting firing strength of the rule (see Figure 4.11).

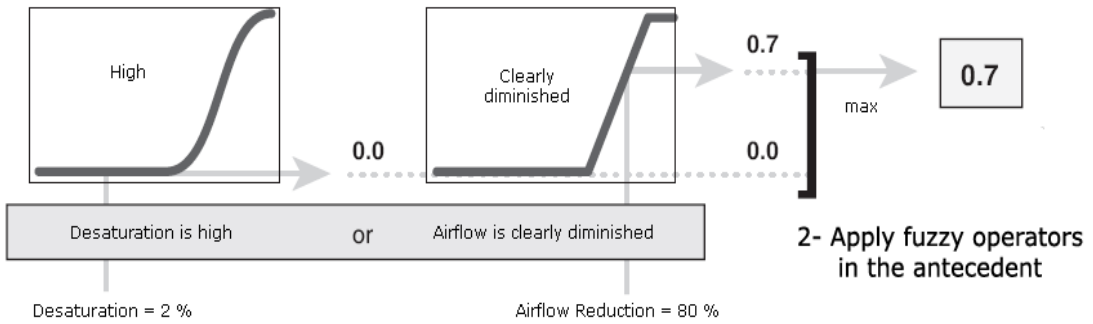


Figure 4.11. Step 2: The figure shows an example in which rule activation is calculated by applying the fuzzy operator on the fuzzy members of the antecedent. In this case OR operator is implemented through using the *max* function

[30] Expressed in the form of fuzzy IF-THEN rules

3. ***Apply implication method:*** Once the degree of activation of each rule's antecedent has been calculated, for those with activation higher than zero, apply the desired implication method to obtain as the output, the degree of membership with respect to the corresponding fuzzy set on the consequent part of the rule (see Figure 4.12).

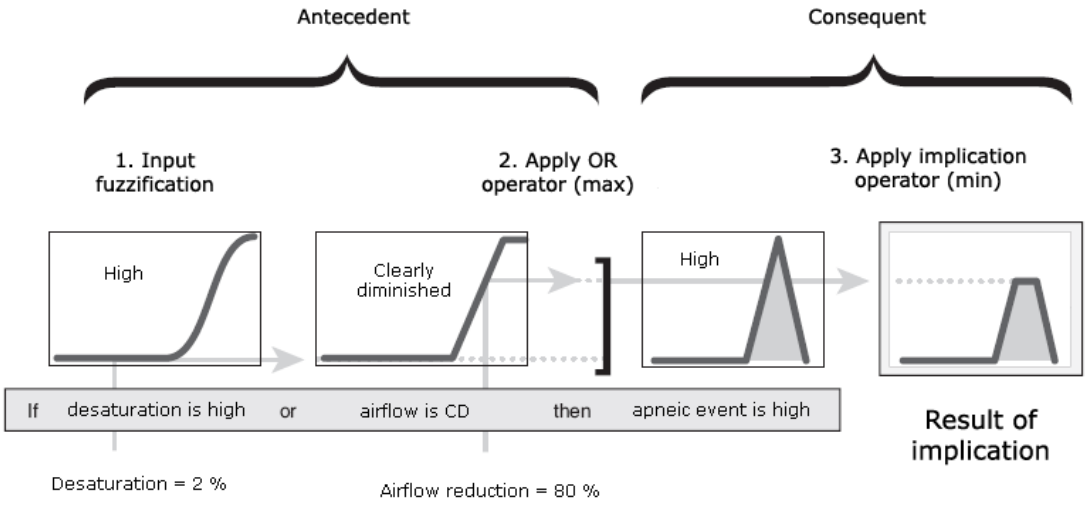


Figure 4.12. Step 3: Figure shows application of implication after rule activation calculation. Here implication is implemented using the *min* operator. Output fuzzy set for the rule is obtained upper bounded according to activation value

4. ***Output aggregation:*** As a consequence of the previous steps and because several rules may be activated with the same input, several membership values are normally obtained with respect to the different fuzzy sets of the partition of output variable. The output fuzzy sets with their corresponding activation are then aggregated to form a unique fuzzy set over the output fuzzy variable (see Figure 4.13).

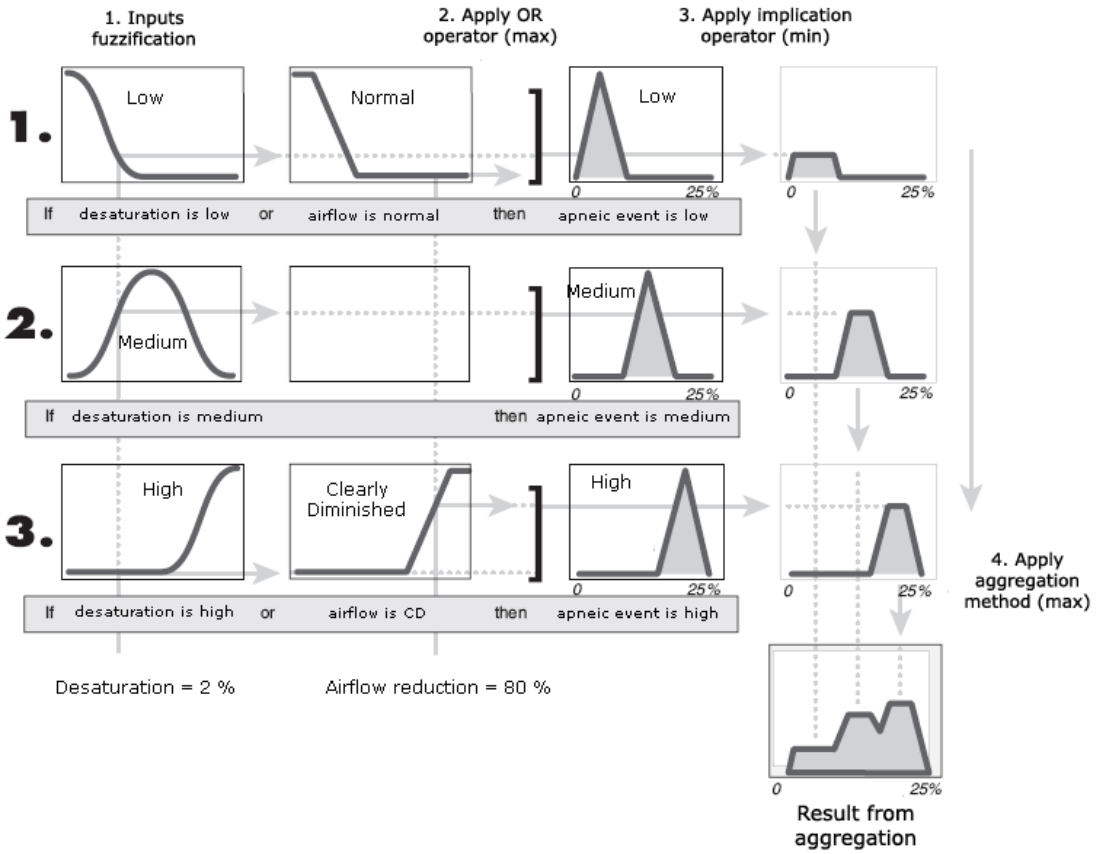


Figure 4.13. Step 4: Figure shows aggregation process after individual outputs for each rule have been calculated. Aggregation method here uses the *max* criterion

5. ***Defuzzification:*** It is the process that transforms the resulting fuzzy set at the output space *V*, into a non-fuzzy value $y \epsilon V$. This operation is normally used among control systems, in which a numeric value is necessary at the output to be used as the input of an actuator mechanism (see Figure 4.14).

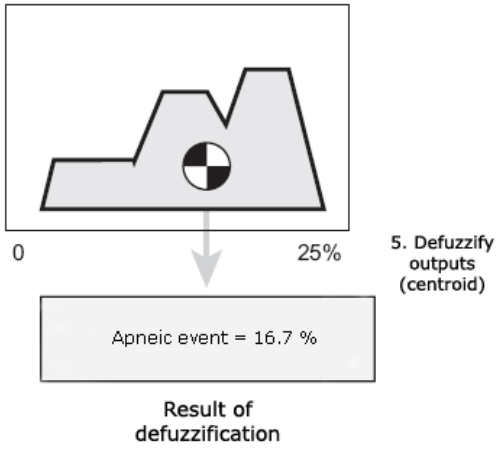


Figure 4.14. Step 5: Output defuzzification by center-of-gravity (centroid) method

4.4. Neuro-fuzzy inference systems

Neuro-Fuzzy Inference Systems (NFIS) come up as a hybrid between FISs and Artificial Neural Networks (ANNs), in an attempt to give response to the problem of FIS modeling. That is, given a certain application domain where certain behavior is wanted to be modeled using a FIS, several questions have to be addressed: How many rules are necessary? In how many fuzzy sets input variables should be partitioned? What kind of fuzzy sets should be used? What are the most appropriate fuzzy operators for the current problem? Which are the correct values for their parameters?…and so on.

Indeed, system modeling task to follow a desired behavior is a complicated job closely related with the problem of knowledge acquisition. Moreover, the problem becomes even harder when for the application domain previous knowledge –normally from an expert- is not available to advice or to guide in the construction of the system. In these cases it is often resorted to machine learning mechanisms that try to optimize configuration of the system automatically.

In this manner, NFIS can be considered as a special case of ANNs, with their consequent learning capabilities, but at the same time, being able to exploit advantages of FISs, including capabilities to handle imprecise information, and their higher expressiveness as compared with ANNs (that behave as black-boxes). In any case, as it was previously outlined, ultimately explanatory capabilities of a FIS are also bounded to the concrete type of FIS used. Indeed, usually within the field of automatic learning there is an inherent compromise between *explanatory capabilities* and *efficacy*, this last usually measured in terms of the final error committed by the system after the training phase [**13**]. That is mainly because a more effective system normally requires of more degrees of freedom, which translates in more complex systems which, in general, are less intelligible from the point of view of human perception.

Advent of neuro-fuzzy systems can be situated starting with the work of Jang in 1993 in which the architecture ANFIS (*Adaptive-Network-Based Fuzzy Inference System*) is described [**14**], although the fuzzy modeling problem had been previously explored by Takagi and Sugeno already in 1985 [**15**]. In the work of Jang, *adaptive networks* are introduced as a superset of all kinds of feed-forward neural networks.

Basically an adaptive network is a structure composed of nodes and directional links through which nodes are connected. Moreover, part or all of the nodes are adaptive, which means each output of these nodes depends on the parameters pertaining to its node. The learning rule specifies how these parameters should be changed to minimize a prescribed error measure. Each node, independently of being adaptive or not, implements a particular function –node function- which depends on the inputs connected to the node and its internal parameters. The function node type can vary from one node to another, and it ultimately depends on the specific general function the networks as a whole is wanted to implement. Figure 4.15 shows the general schema of an adaptive network.

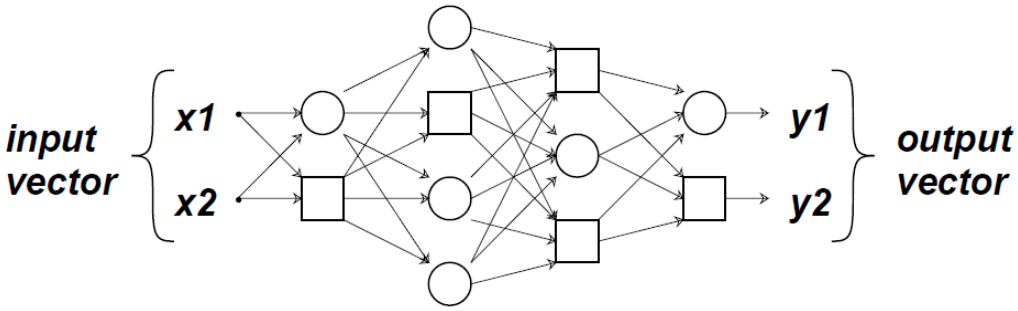


Figure 4.15. General schema of an adaptive network. In the figure squared nodes represent adaptive nodes while circle nodes represent nodes without parameters.

From this general structure it is possible to obtain the equivalent to a concrete FIS, simply by means of choosing the appropriate node functions and establishing the adequate connections between the nodes. In fact Jang shows in the same work how the proposed ANFIS architecture is functionally equivalent to a first-order Sugeno FIS, and also how it is possible to train the architecture for the optimization of its parameters through a hybrid learning rule, which combines adaptation by means of error backpropagation and *Least Squares Estimator (LSE)* [**14**].

In general, the number of potentially possible neuro-fuzzy architectures is very large and, on the other hand, for the same network configuration, there are multiple training variants available: backpropagation of error, genetic algorithms, linear programming, simulated annealing, etc. As a consequence, throughout literature several examples of this kind of systems can be found, among others [**16**] [**17**] [**18**] [**19**] [**20**].

To carry out an exhaustive analysis of the same exceeds the objectives of this chapter as well as that of the doctoral manuscript. However, as an example, in the next subsection the analysis of two approximations for the modeling of neuro-fuzzy systems is performed. The interest on these approximations is to exemplify the use of NFIS for modeling and parameterization of FIS that are used in the development of the proposed system for the diagnosis of SAHS (see Chapter five). Their architectures are discussed in the following subsections. In addition an outline to the modeling of NFIS structures by automatic learning mechanisms is subsequently performed.

4.5. Neuro-fuzzy modeling within the developed system

In this section the introductory perspective to FISs followed along this chapter is left behind and a series of specific developments in the scope of the constructed system to aid in the SAHS diagnosis are described. Besides Mamdani-based FIS, which have been introduced in the preceding section, such realizations refer to neuro-fuzzy modeling processes which have been developed for the implementation of some of the FISs that intervene in the analysis processes of the developed system.

In the following subsections two different neuro-fuzzy architectures are presented to be used, respectively, in regression and classification tasks. Once the architectures have been described, modeling techniques are proposed for both, structure identification and parameter optimization, of general NFIS architectures –including Mamdani FIS. These modeling techniques have been used for the implementation of the FIS developed within the constructed system to aid in the diagnosis of SAHS, which is described throughout subsequent Chapter 5.

4.5.1. An architecture for regression tasks

Here it is presented an architecture to implement a fuzzy system based on adaptive networks. This architecture is based on the well-known structure ANFIS proposed by Jang [**14**], and later generalized by Sun [**21**].

Figure 4.16 shows this architecture for the case of two input variables, x_1 and x_2, and one output variable y. Following Jang's notation, in this kind of networks a circle

denotes a node without parameters; otherwise a square is used. The proposed structure is organized into six functional layers following a feed-forward processing manner.

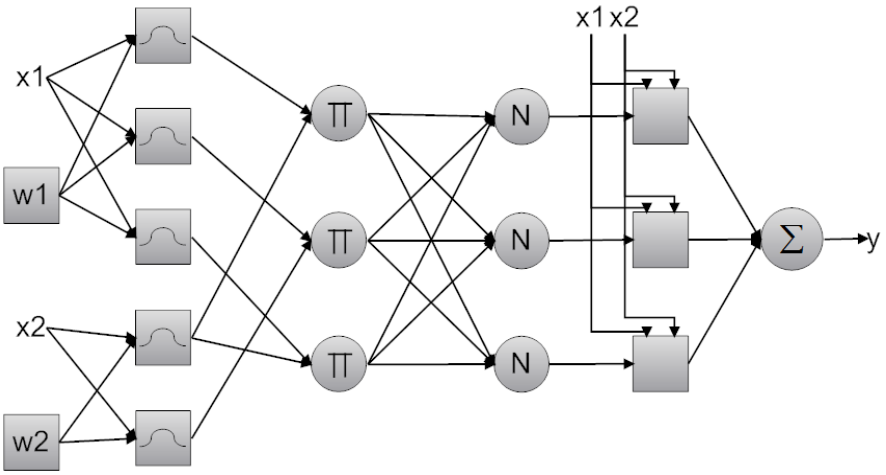


Figure 4.16. Architecture for an adaptive-network-based fuzzy inference system

Each node at the first layer implements a weight of importance w_i associated with the respective input variable i. This allows the implementation of feature selection during the learning adaptation, according to the evolution of the weight assigned to the corresponding input variable. As closer the weight is to 1, the more the influence of the variable on the posterior input fire strength (layer 2). On the contrary as long as the variable is less important on the construction of the output, this weight will tend to approximate to 0.

Every node at layer 2 is a square node associated with a parameterized membership function $\mu_A(x_i)$ where x_i is one of the input variables and A is the linguistic term associated with this node function. Each node connected with an input variable represents a different fuzzy set for the corresponding variable. Thus the resulting degree of membership, taking into account the corresponding weight on layer 1 for the variable i, is:

$$s_i = 1 - w_i(1 - \mu_A(x_i)) \tag{4.1}$$

In this case the input variable x_1 is partitioned in three linguistic terms, whereas in the partition of the input variable x_2 two fuzzy set are used.

Layer 3 combines the degrees of membership from the nodes on layer 2 to which it is connected. This operation can be interpreted as the calculation of the firing strength of a rule. The connections define the premises in the antecedent of the rule, whereas parameters in the nodes of layer 3 implements the conjunction operator. In Figure 4.16 this operation is indicated as general nodes that implement a *t-norm*. The choice of the concrete connective operator depends on the application. More discussion on *t-norms* and their implications can be found on [**10**]. Specifically in the proposed architecture the implemented *t-norm* is the product, as it was proposed in the original publication of ANFIS.

Every node at layer 4 is a circle node that calculates the ratio of the *i*th rule's firing strength with respect the sum of the firing strengths of all rules. Therefore let $\overline{w}_i$ to be the output of the node *i*th at this layer, then $\sum_{i=1}^{R} \overline{w}_i = 1$, being *R* the number of nodes (rules) both at layers 3 and 4.

At layer 5 the linear combination of the input is performed, and the result is multiplied by the corresponding firing strength coming from layer 4. The number of parameters p_k (consequent parameters) per node is *#input + 1*, being the output o_k of the node *k*th of this layer:

$$o_k = \overline{w}_k \sum_{i=1}^{\#input} p_i x_i \tag{4.2}$$

Finally layer 6 takes the summation of the o_k and provides the final output of the system. Note that in this architecture the firing strength of each fuzzy rule is calculated as the conjunction of the membership values in the premise part, the consequence of each rule is a linear combination of the inputs, and the final output is obtained as the weighted average of each rule's consequence. Thus resulting architecture is equivalent to a first-order Takagi-Sugeno FIS [**15**]. Table 4.2 summarizes the number of nodes and parameters per node at each layer.

Table 4.2. Number of nodes and number of parameters per node for architecture of Figure 4.16. V is the number of input variables; L_i is the number of linguistic terms for the variable i; and R is the number of rules.

Layer	1	2	3	4	5	6
#Nodes	$2V$	L_i	R	R	R	1
Parameters per node	1	2	0	0	V+1	0

4.5.2. An architecture for classification tasks

In this subsection an architecture to implement a fuzzy classification system based on adaptive networks is presented. Figure 4.17 shows the architecture for the case of two input variables, x_1 and x_2, and three output classes C_1, C_2 and C_3. The proposed architecture is based on the initial model proposed for classification by Sun and Jang [**22**], to which weighting nodes have been added to layer one. Modification of the connection links between layers two and three was also incorporated. The resulting structure is organized into five functional layers following a feed-forward processing flow.

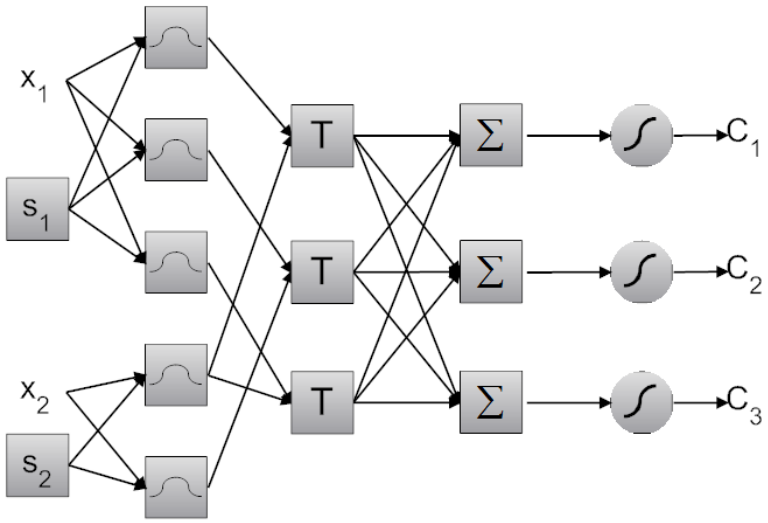


Figure 4.17. Architecture for an adaptive-network-based fuzzy classification system

Each parameter node at layer 1 implements a weight of importance s_i associated with the respective input variable i. Analogously to the structure presented at the previous subsection, the former allows the implementation of feature selection during the learning adaptation, according to the evolution of the weight assigned to the corresponding input variable.

Every node at layer 2 is associated with a parameterized membership function $\mu_A(x_i)$ where x_i is one of the input variables and A is the linguistic term associated with the corresponding node function. Each node connected with an input variable represents a different fuzzy set for the partition of the input variable. Thus the activation in the node k at this layer results:

$$net_k^2 = 1 - s_i(1 - \mu_A(x_i)) \tag{4.3}$$

In this example the input variable x_1 is partitioned in three linguistic terms, whereas in the partition of the input variable x_2 two fuzzy sets are used.

Layer 3 combines the degrees of membership from the nodes on layer 2 to which it is connected. This operation can be interpreted as the calculation of the firing strength of a rule, and the connections between layers 2 and 3 define the premises in the antecedent of each rule. The parameters at the nodes of layer 3 implements the integrator operator (conjunction), and the number of nodes determines the resulting number R of fuzzy rules.

In Figure 4.17 this operation is indicated as general *t-norm* nodes. In the current architecture concrete node function obeys to a parameterized version of Hamacher's t-norm:

$$T_H(a, b, \gamma) = \frac{ab}{\gamma + (1 - \gamma)(a + b - ab)} \tag{4.4}$$

where γ is a non-negative parameter and a and b represent inputs to the operator. On the other hand, due to the associative property of *t-norms* when more than two inputs achieve a layer 3 node, resulting activation can be recursively calculated as:

$$net_r^3 = \underset{i \in r}{T_H}(net_i^2, \underset{j \in r, j \neq i}{\overset{M-1}{T_H}}, \gamma) \tag{4.5}$$

being *M* the total number of nodes at layer 2 connected to the rule $r \in R$. Note also that this *t-norm* is differentiable allowing the architecture to be trained using back-propagation [**23**].

At layer 4 a linear combination for the rule's activations is done regarding each output class *c*. Therefore the total number of nodes performing this weighted summation is equal to the number of output classes *C*. The number of parameters per node is equal to the number of rules because total interconnection exists between nodes of layers 3 and 4:

$$net_c^4 = \sum_{r=1}^{R} net_{cr}^3 w_{cr} \tag{4.6}$$

Finally a sigmoid function is applied at the output layer. Since we are interpreting the output respect to each class as a degree of membership, a logistic function is chosen providing output values between 0 and 1:

$$net_c^5(p) = \mu_c(x_p) = \frac{1}{1 - e^{-net_c^4(p)}} \tag{4.7}$$

where *p* represents the input index pattern.

Table 4.3 summarizes the number of nodes and parameters per node at each layer.

Table 4.3. Number of nodes and number of parameters per node for architecture of Figure 4.17. *V* is the number of input variables; L_i is the number of linguistic terms for the variable *i*; *R* is the number of rules; and C is the number of classes.

Layer	1	2	3	4	5
#Nodes	*V*	$\sum_{i=1}^{V} L_i$	*R*	*C*	C
Parameters per node	1	2	1	*R*	0

4.5.3. Structure identification and parameter optimization

System modeling mainly consists of two parts: structure identification and parameter optimization. On fuzzy modeling, and taking the previously proposed architectures as reference, the former implies finding the appropriate fuzzy partitions on the input space, as well as determining the number of rules, and their antecedent and consequent parts which yield to the concrete connectivity between the different layers. Parameter optimization, on the other hand, consists on finding out best parameter values on each node to minimize model's error according to a defined fitness function, also known as *cost function*.

Leaving aside knowledge-based approaches to manually set up system's structure[31], several methodologies and algorithms have been proposed through the literature to handle the problem of structure identification on fuzzy inference systems from a machine learning approach [**24**] [**25**]. Without the aim to be the optimal, a clustering algorithm is proposed here in order to initialize the structure of the network [**26**]. The algorithm, called *Subtractive Clustering*, is a fuzzy extension of the Mountain Method [**27**] for estimating the number and initial locations of cluster centers from a set of *n*-dimensional points. Under this approach each data point x_i is considered as a potential cluster center and defines a measure of potential

$$P_i = \sum_{j=1}^{n} e^{-\frac{4}{r_a}\left\|x_i - x_j\right\|^2} \tag{4.8}$$

with r_a a positive constant defining a radius of neighborhood. Once the potential of every point has been computed, the point with highest potential is selected as the first cluster center. Subsequently, in function of a decay parameter r_b, and in proportion to the distance to the first cluster, the potential is reduced (*subtracted*) from the rest of data points. The point with the remaining highest potential is then selected as the second cluster center. The algorithm continues until the maximum potential reaches a certain threshold.

[31] Leaved aside since here we are dealing with automatic learning procedures in the lack of knowledge about the ideal structure. However when expert knowledge is available that guides structure identification, this should be preferably used.

After the cluster centers have been determined, each cluster center can be thought as a prototypical data point that exemplifies a characteristic behavior of the system. Hence, each cluster center can be used as the basis of a rule that describes the system behavior [**26**]. Although we use this algorithm in order to initialize the structure, we do not claim for its optimality and the problem of structure identification is out of the scope of this work. On the other hand, it allows an initialization for the structure as estimation of the necessary number of rules. Additionally, compared with other types of structure initialization as grid partition [**22**], the number of rules it returns is much lower, therefore preserving the interpretability requisite related to fuzzy systems.

Once determined the structure of the model parameters of the adaptive nodes have to be optimized using a fitness criterion. This corresponds with the second stage of system modeling, i.e. parameter optimization. As commonly happen in supervised training over neural-based systems, learning to find the optimal parameters for the mapping of the input-output data, is usually based on the minimization/maximization of a cost function that governs the system adaptation. The Mean Squared Error (MSE) has been so far the most popular criterion for system adaptation. However the optimality of second-order statistics depends heavily on the assumption of Gaussianity. Effectively, if the Probability Density Function (PDF) of the errors is not Gaussian distributed, there is information that is not being used to adapt the weights when the squared loss function is minimized. Although such assumptions provide successful engineering solutions to most practical problems, it has become evident that when dealing with nonlinear systems, a criterion that not only considers second-order statistics, but also takes into account higher order statistical behavior is rather desired.

In this respect, Information Theoretic Learning (ITL) has been recently developed extending the concept of mean squared error adaptation to include information criteria [**28**]. ITL preserves the nonparametric nature of learning since the cost function is still directly estimated from the data, but it extracts more information from it, yielding to more accurate solutions than MSE, especially in non-Gaussian and nonlinear signal processing. Inspired by ITL thus two new criteria called respectively *Minimum Error Entropy* (*MEE*) and *Maximum Correntropy Criterion* (*MCC*) are proposed as alternatives to the MSE in order to guide the parameter's adaptation process. A formal definition of MSE and the two ITL-based alternatives can be found in Appendix A. To

go further in the details of ITL-based cost functions will exceed the scope of this manuscript. For more details about the training and performance comparison on using both MSE and ITL-based cost functions the reader is referred to [**28**]. In addition, experimental results using the previous presented architectures can be found in references [**29**] and [**30**], which are also included in Appendix B as relevant publications of the author related with the doctoral thesis.

Independently on the used cost function to be minimized/maximized, parameter's optimization necessarily involves a search across the range of possible values for the parameters. This process is carried out throughout an iterative adjusting process in which the cost function is expected to return a better fitness value at each iteration. Typical algorithms to search in the parameter's space are based on the direction marked by the gradient on the error surface.

This is done by means of the chain rule, and the method is generally referred to as the *back-propagation learning rule* because the gradient vector is calculated in the direction opposite to the flow of the output of each node. In general, given certain cost function *J(E)* defined over the output error *E*, and a certain parameter α to be updated within an adaptive node of the network, then the general updating formula is

$$\Delta\alpha = -\eta \frac{\delta^{+} J(E)}{\delta\alpha} \tag{4.9}$$

in which η is the learning rate and δ^{+} refers to the *order derivative* as defined by Werbos [**31**]. The learning rate can be further expressed as

$$\eta = \frac{\kappa}{\sqrt{\sum_{\alpha}\left(\frac{\delta J(E)}{\delta\alpha}\right)^{2}}} \tag{4.10}$$

where κ is the *step size*, i.e. the length of each transition along the gradient direction in the parameter space. Note, on the other hand, in eq. (4.9) transition on the opposite direction of the gradient is performed, thus assuming the objective is the minimization of the cost function. For further discussion on parameter optimization in adaptive networks the reader is referred to excellent paper by Jang and Sun [**32**].

However, these locally guided techniques often suffer from the problem of local minima. They also require of differentiable functions to be implemented within the adaptive nodes. To avoid the previous problems a different possibility is to use a global optimization approaches that also do not rely on the use of differentiable functions, such as for example, Genetic Algorithms (GAs).

On this context, a global optimization process based on GAs is subsequently presented which is used to accomplish the parameter optimization task on the various FIS implemented within the system. As introduced in Chapter 3 (see subsection "*Knowledge and intelligent systems in medicine*") GAs are search and optimization procedures that are based on the concepts of natural evolution including selection, crossover, mutation and survival of the fitness. GAs work with a population of *N* individuals, with each individual being a candidate solution of the problem. A new generation of solutions is then created from the old generation through genetic operations [**33**].

When using GAs for optimization, and besides the choice of the appropriate fitness function, there are two important factors affecting the performance: the representation schema used for the individuals and the concrete implementation of the genetic operators.

The representation schema describes the parameter structure of an individual, which regarding the structure of a FIS, it includes parameters of the fuzzy membership functions, fuzzy rules, node connectivity, and the encoding method used to convert an individual into a chromosome (i.e. the array of numerical values that represents the resulting FIS in the GA). In the used representation each fuzzy membership function is represented as float (real-coded) vector containing the parameters of the corresponding membership function (see Figure 4.18). The number of parameters varies according to the type of membership function, i.e. if the membership function is of Gaussian type, for example, then two genes are used encoding respectively the mean and the standard deviation that define the shape of the Gaussian, on the other hand, for example, if the membership function is of trapezoidal type, then four parameters are required to represent the trapezoid. A summary of the most popular membership functions and their corresponding parameters has been shown in Figure 4.2. The genes of the several

membership functions are then linked together as shown in Figure 4.18. The number of membership functions has been determined in the structure identification phase and it determines the length of the resulting part of the chromosome. If the NFIS structure comprises membership functions both at the input and at the output (as in the case of Mamdani-type FIS) then this representation is used among the input as well as over the output membership functions.

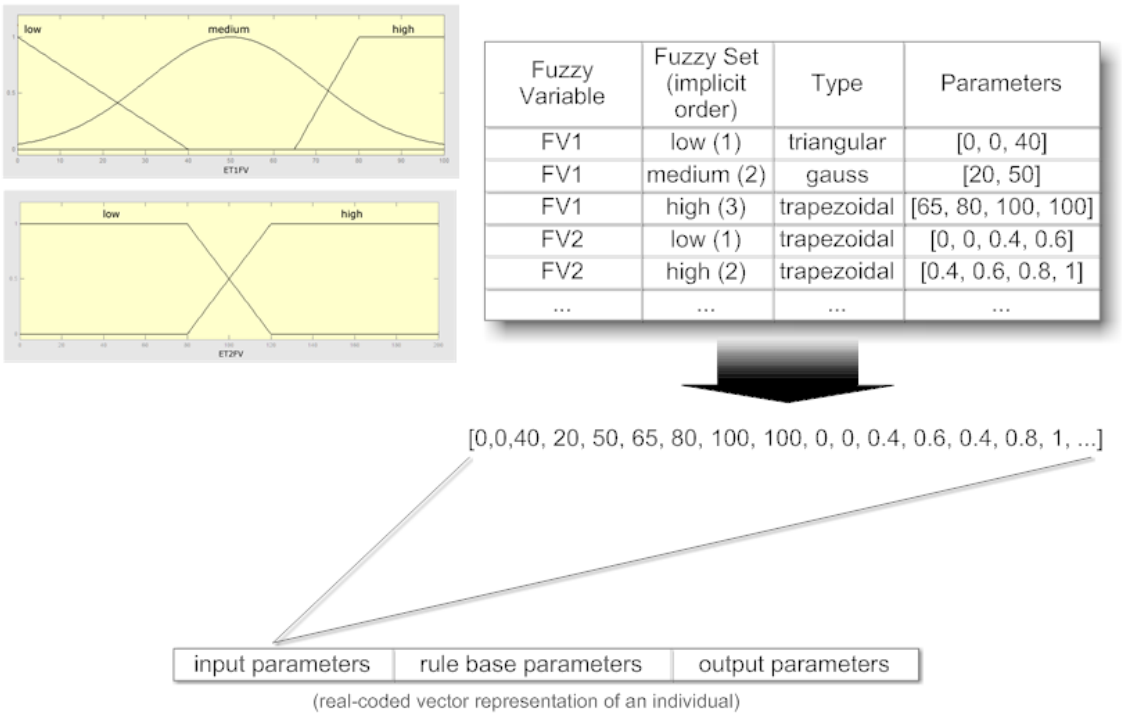


Figure 4.18. Example showing genetic coding of the input fuzzy sets of an individual

The representational schema used for coding the rulebase has also implicit the connectivity between the nodes. This is achieved by indexing the inputs and the outputs variables as well as keeping an implicit order between the different fuzzy sets within each fuzzy variable. To illustrate the process, and without loss of generality, let us consider the following Mamdani rule:

If input1 is MF1 or input2 is not MF3, then output1 is MF2 (weight = 0.5)

Assuming that there are *m* inputs and *n* outputs, then this rule is turned into a structure according to the following logic, where the first *m* vector entries of the rule structure correspond to inputs 1 through *m*:

- The entry in column 1 is the index number for the membership function associated with input 1
- The entry in column 2 is the index number for the membership function associated with input 2, and so on (*m* times)
- The NOT operator is represented by using a negative index. Also, it is possible to not use one of the feature elements in a rule. This situation is handled by assigning a value of "0"
- The next *n* columns work the same way for the outputs
- Column *m*+*n*+1 is the weight associated with that rule and column *m*+*n*+2 specifies the connective used (for example, AND = 1 and OR = 2)

According with the previous logic, the resulting vector associated with the preceding rule is: [1, -3, 2, 0.5, 2].

This coding procedure is repeated for each rule implemented in the system and the different vectors encoding each rule are concatenated. The initial number of rules R_{init} is prefixed at the structure identification phase and it determines the length of the resulting part of the chromosome. On the other hand, dynamical search of the appropriate number of rules is allowed by considering a maximum number of rules $R_{max} \geq R_{init}$ and setting a configurable threshold value on the rule weight, such that rules with weight below the threshold value are not further considered to calculate the final output. In the case of the two previous adaptive networks, that do not implement Mamdani-type fuzzy rules, the representation is similar, omitting fuzzy output partitions from the representation, but rather including the consequent parameters for each Sugeno fuzzy rule in the regression architecture (see eq. 4.2), or the corresponding weight parameters (see eq. 4.6) for the case of the neuro-fuzzy classifier.

With regard to the implementation of the genetic operators, the following are used within the proposed approach for NFIS modeling based on GAs:

Selection operator chooses parents for the next generation based on their scaled values from the fitness scaling function. For this purpose, rank of the raw scores of each

individual is used, being the rank of an individual, its position in the sorted scores of the fitness function. Rank is then scaled, so that for an individual with rank n its scaled value is proportional to $\frac{1}{\sqrt{n}}$. A stochastic selection function is then applied that lays out a line in which each parent corresponds to a section of the line of length proportional to its scaled value. The algorithm moves along the line in steps of equal size, and at each step, the algorithm allocates a parent from the section it lands on [**34**].

Mutation operator specifies how the genetic algorithm makes small random changes among the individuals in the population to create mutation children. Mutation provides genetic diversity and enables the genetic algorithm to search a broader space. A hybrid adaptive feasible function is used that randomly generates directions that are adaptive with respect to the last successful or unsuccessful generation. The feasible region is bounded by constraints and inequality constraints set to fulfill integrity of the parameters. For example, when using triangular fuzzy sets, the following relation must hold $a < b < c$ (see Figure 4.2). In this respect a step length is chosen along each direction so that linear constraints and bounds are satisfied [**35**]. For rulebase parameters, in contrast, the algorithm selects a fraction of the vector entries where each entry has a probability $Rate_{rulebase}$ of being mutated. The algorithm then replaces each selected entry by a random number selected uniformly from the range for that entry.

Crossover operator controls how the genetic algorithm combines two individuals, or parents, previously selected by the selection operator, to form a crossover child for the next generation. Scattered crossover is used in this respect that creates a random binary vector and selects the genes where the vector is a 1 from the first parent, and the genes where the vector is a 0 from the second parent.

The preceding genetic operators have been used with independence of the concrete NFIS architecture. On the other hand, additional free parameters have still to be selected for each case, for example, regarding the number of individuals in the population, the maximum number of iterations, the choice of the fitness function or the proportion of crossover and mutation applied between consecutive generations. With regard to the choice of the fitness function, MSE and ITL-based cost functions (see Appendix A) are used as candidates. Appropriateness of choosing one, another, or a combination of both,

ultimately depends on the concrete application, and the reader was already encouraged to consult more specialized references in this respect such as [**28**]. Likewise, further discussion about NFIS optimization using GAs would exceed the introductory purpose of this chapter. The interested reader is again referred to consult the specific literature available at this respect, such as for example [**36**] or [**37**].

4.6. Summary of this chapter

This chapter goes in depth into one fundamental technological framework over which the clinical decision support system object of this thesis is developed. Just as it was identified throughout the previous chapters, one of the main points of this thesis is that the success of a diagnostic system is strongly linked to its capabilities to both handle imprecise information and to make reasoning processes in environments affected by uncertainty. In this regard, it has been shown already how exact knowledge is quite unlikely in the human being, and also how this inaccuracy becomes apparent in the context of SAHS diagnosis. This, together with the necessity of expressing the results of the system on the basis of approximate linguistic labels, guide us to the use of artificial intelligence techniques being able to manage such kind of information. The fuzzy logic paradigm, for its characteristics expressed all over this chapter, looks especially suitable for accomplishing this task. Development of its theoretical framework and to show how it can be applied to model reasoning processes in the context of SAHS diagnosis justifies the inclusion of this chapter.

The chapter starts by introducing fuzzy logic and its historical perspective. Fuzzy logic has its fundamentals in the theory of fuzzy sets, and its underlying procedures are developed subsequently, concluding that it is possible to establish an isomorphism between the operations in fuzzy logic and their equivalents in the field of classical formal logic.

The chapter continues showing how an inference process can be established in fuzzy terms through the generalization of the classical inference mechanism of the *Modus Ponens*, yielding to the so-called *Generalized Modus Ponens* which is able to produce reasonings with imprecise facts, also establishing a mechanism for uncertainty propagation toward the new inferred facts.

A rule-based system is then introduced which as the method to exploit its knowledge base uses *fuzzy inference*. The reasoning process that would follow such a system is shown through an example applied to the diagnosis of SAHS. The generalization of this kind of systems, by incorporating learning capabilities and adaptation, is achieved through the use of neuro-fuzzy systems, whose particularities are described in the next part of the chapter.

Once the general features of neuro-fuzzy systems have been enunciated, the chapter ends up by detailing concrete approximations for their modeling. Such approximations describe neuro-fuzzy modeling processes that have been developed for the implementation of the different fuzzy inference systems which are integrated in the proposed solution to support SAHS diagnosis. Next chapter is dedicated to the functional description of the developed system. Several signal processing and artificial intelligence techniques are detailed throughout its contents at this respect. However, since fuzzy modeling is not the main objective of the next chapter, it has been considered more convenient to proceed to its description here -thus abstracting underlying modeling details of the different integrating FIS from the contents of Chapter 5.

4.7. References

[1] LA. Zadeh, "Fuzzy Sets," *Information and Control*, vol. 8, no. 3, pp. 338-353, 1965.

[2] G. Shafer, *A mathematical theory of evidence*.: Princeton University Press, 1976.

[3] EH. Shortliffe and B. Buchanan, "A model of inexact reasoning in medicine," *Mathematical Biosciences*, vol. 23, pp. 351-379, 1975.

[4] L. Zadeh, "From computing with numbers to computing with words - from manipulation of measurements to manipulation of perceptions," *IEEE Transactions on Circuits and Systems*, vol. 45, no. 1, pp. 105-119, 1999.

[5] LA. Zadeh, "Outline of a new approach to the analysis of complex systems and decision processes," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, no. 1, pp. 28-44, 1973.

[6] EH. Mamdani, "Application of fuzzy algorithms for the control of simple dynamic plant," *Proceedings of the Institution of Electrical Engineers*, vol. 121, no. 12, pp. 1585-1588, 1974.

[7] A. Piegat, *Fuzzy modeling and control*. Heidelberg; New York, Germany: Physica-Verlag, 2001.

[8] V. Moret-Bonillo, A. Alonso-Betanzos, M. Cabrero-Canosa, B. Guijarro-Berdiñas, and E. Mosqueira-Rey, *Fundamentos de inteligencia artificial*, 2nd ed., University of A Coruña, Ed. A Coruña, 2005.

[9] V. Novák, I. Perfilieva, and J. Mockor, *Mathematical principles of fuzzy logic*. The Netherlands: Kluwer Academic Publishers, 1999.

[10] PP. Bonissone, "Summarizing and propagating uncertain infromation with triangular norms," *International Journal of Approximate Reasoning*, vol. 1, pp. 71-101, 1987.

[11] EH Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, vol. 7, no. 1, pp. 1-13, 1975.

[12] M. Sugeno, *Industrial applications of fuzzy control*.: Elsevier Science Pub. Co., 1985.

[13] D. Nauck and R. Kruse, "Obtaining interpretable fuzzy classification rules from medical data," *Artificial Intelligence in Medicine*, vol. 16, pp. 149-169, 1999.

[14] JSR. Jang, "ANFIS: Adaptive Network-based Fuzzy Inference System," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23, no. 3, pp. 665-685, 1993.

[15] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 15, pp. 116-132, 1985.

[16] CT. Lin and CSG. Lee, "Neural-network-based fuzzy logic control and decision systems," *IEEE Transactions on Computers*, vol. 40, pp. 1320-1336, 1991.

[17] MF. Azeem, M. Hanmandlu, and N. Ahmad, "Generalization of adaptive neuro-fuzzy inference systems," *IEEE Transactions on neural networks*, vol. 11, no. 6, pp. 1332-1346, 2000.

[18] J. Kim and N. Kasabov, "HyFIS: adaptive neuro-fuzzy inference systems and their application to nonlinear dynamical systems," *Neural Networks*, vol. 12, pp. 1301-1319, 1999.

[19] E. Zhou and A. Khotanzad, "Fuzzy classifier design using genetic algorithms," *Pattern Recognition*, vol. 40, pp. 3401-3414, 2007.

[20] D. Nauck, F. Klawon, and R. Kruse, *Foundations of neuro-fuzzy systems*. U.K./Chichester: Wiley, 1997.

[21] CT. Sun, "Rule-based structure identification in an adaptive network-based fuzzy inference system," *IEEE Transactions on Fuzzy Systems*, vol. 2, no. 1, 1994.

[22] CT. Sun and JS. Jang, "A neuro-fuzzy classifier and its applications," in *IEEE International Conference on Fuzzy Systems*, San Francisco, CA, USA, 1993, pp. 44-48.

[23] DE. Rumelhart, GE. Hinton, and RJ. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, DE. Rumelhart and JL. McClelland, Eds.: The MIT Press, 1986, ch. 8, pp. 318-362.

[24] M. Sugeno and GT. Kang, "Structure identification of fuzzy model," *IEEE Transactions on Fuzzy Sets and Systems*, vol. 28, no. 11, pp. 15-33, 1988.

[25] M. Fazle-Azeem, M. Hanmandlu, and N. Ahmad, "Structure identification of generalized Adaptive Neuro-Fuzzy Inference Systems," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 5, pp. 666-681, 2003.

[26] SL. Chiu, "Fuzzy model identification based on cluster estimation," *Journal of Intelligent and Fuzzy Systems*, vol. 2, pp. 267-278, 1994.

[27] RR. Yager and DP. Filev, "Approximate clustering via the mountain method," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24, no. 8, pp. 1279-1284, 1994.

[28] JC. Príncipe, *Information Theoretic Learning*.: Springer, 2010.

[29] D. Alvarez-Estevez, JC. Principe, and V. Moret-Bonillo, "Neuro-fuzzy classification using the correntropy criterion: application to sleep depth estimation," in *10th International Conference on Artificial Intelligence*, Las Vengas, NV, 2010, pp. 9-15.

[30] D. Alvarez-Estevez, JC. Principe, and V. Moret-Bonillo, "Information theoretic fuzzy modeling for regression," in *IEEE World Congress on Computational Intelligence (WCCI2010) - FUZZ-IEEE*, Barcelona, 2010, pp. 1979-1983.

[31] P. Werbos, "Beyond regression: new tools for prediction and analysis in the behavioral sciences," Harvard University, PhD Thesis 1974.

[32] JSR. Jang and CT. Sun, "Neuro-Fuzzy Modeling and Control," *Proceedings of the IEEE*, vol. 83, no. 3, pp. 378-406, 1995.

[33] DE. Goldberg, *Genetic Algorithms in search, optimization and machine learning*. Boston, MA: Kluwer Academic Publishers, 1989.

[34] MathWorks, "Global Optimization Toolbox," Product Documentation 2010.

[35] TG. Kolda, RM. Lewis, and V. Torczon, "Optimization by direct search: new perspectives on some classical and modern methods," *SIAM Review*, vol. 45, no. 3, pp. 385-482, 2003.

[36] O. Cordón, F. Herrera, F. Hoffman, and L. Magdalena, *Genetic fuzzy systems: Evolutionary tuning and learning of fuzzy knowledge bases*. Singapore: World Scientific Printers, 2001.

[37] JSR. Jang, CT. Sun, and E. Mizutani, *Neuro-fuzzy and soft computing*.: Prentice Hall, 1997.

5. DESCRIPTION OF THE SYSTEM

In this chapter functional description of the developed system is performed. Description starts from a software engineering perspective, making reference to the methodology used for its development, analyzing the requirements and making a brief description of the system from the architectonical point of view. In the following sections system's construction is described regarding its design, and explaining and detailing how all its integrating modules work. In this respect the processing algorithms are described including, signal acquisition, artifact detection, analysis of the neurophysiological signals, processing of the respiratory signals, temporal data integration, reasoning mechanisms, detection and classification of the apneic events, and diagnosis generation.

5.1. Development model

The *software process* is defined as a framework for the tasks involved in the construction of high quality software [**1**]. The software process brings the basis for the control and management of the software projects, and it establishes the context in which the technical models are applied, working products are generated, fundamentals are established, quality is garanteed, and changes are appropiately accomplished [**2**].

To solve the problems that come up in a working environment, it is necessary to incorporate a process model or software engineering paradigm. Process models define a set of activities within the framework, a collection of tasks to accomplish each activity, working products generated as a consequence of the tasks, and a set of *umbrella activities* that accompany the whole process [**2**].

There are several types of process models, each one adapts the best to the concrete characteristics of the software to be developed, such as time factors to execute the project, availability of human resources and materials, or the kind of software to be developed. There is not a unique ideal process, moreover, within the same company, several different processes can coexist. Some of the most popular are cited subsequently [**1**] [**2**]:

- **Waterfall model**. It perceives the software process as a sequence of phases or states that start with the specification of requirements from the client, and it continues with the design, implementation and verification, ending up with the maintenance of the finished product. In the sequence, once definition of a state ends, the process continues toward the following state.

- **Evolutionary model**. This approach intertwines activities of specification, development and validation. A first prototype is rapidly developed from abstract specifications. Then the prototype is refined in an evolutionary process ending up with a system that satisfies the necessities of the client.

- **Incremental model**. Developing scaled sequences are applied as long as time advances. Each sequence of developing may involve a software model process, and at the end of such developing, a new *software increment* is produced [**3**]. The first increments are incomplete versions of the final product, but they offer full functionality over part of the initial requirements, thus producing an operational prototype with each increment. This process model is applied in situations in which the initial software requirements, even though well-defined, imply a global developing effort that excludes a purely linear process.

- **Development based on reusability**. This model is based on the use of reusable components. The development process is then focused toward integration of these components more than starting development from the scratch.

- **Spiral model**. Proposed by Boehm [**4**], it is a generator of the process model guided by risks, which is used to conduct intensive systems of concurrent software engineering and with multiple users. It merges, in one hand, the cyclic approach for the incremental growing of the degree of definition and implementation of a system, while reducing the risk. On the other hand, a set of checking points are defined to ensure the compromise with the user in terms of feasible and mutually satisfactory solutions [**5**].

With regard to the system object of this doctoral thesis the evolutionary model has been used. As it has been introduced, evolutionary model is based on the idea of developing of a first initial implementation, expose it to validation results, and refine it through different versions until an adequate system -according to validation and established requirements- is obtained (see Figure 5.1).

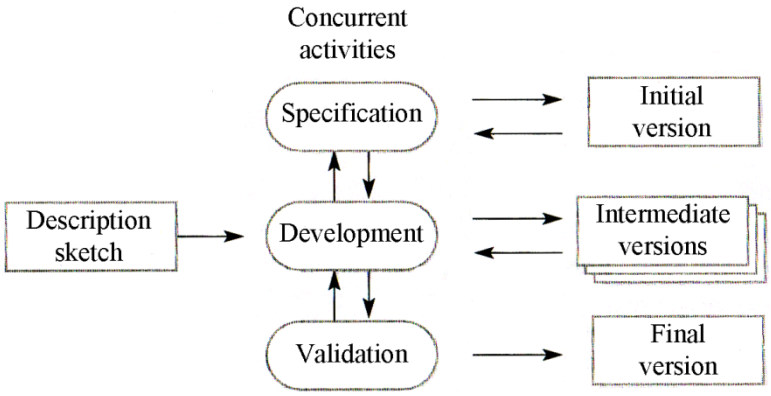


Figure 5.1. Evolutionary model schema

The choice of an evolutionary model is justified in the nature of the project under consideration, which is aimed at the development of a model of intelligent behavior and with a clear research orientation. In this manner, since expert knowledge is of heuristic nature, it is not possible to establish a detailed specification of the software aimed at its emulation. Specifically for the analysis of the PSG for SAHS diagnosis, reasoning processes from the experts involve extraction of relevant information from the signals which is difficult to model. Thus, an evolutionary perspective is considered the most suitable for the development of a system of such characteristics.

5.2. Requirement's specification

One fundamental question that needs to be answered before carrying out any development is if the system *is of utility*. It is very important to set from the beginning what is the problem to solve, who are the potential users of the system, and which is the expected impact of the system in the organization[32]. For this purpose it is necessary to carry out a requirement's definition phase, and to determine the kind of problems to be solved or in which environments are they going to be executed.

Requirement's specification phase in software consists in an abstract description of the services the system is expected to provide and the restrictions under which the system is supposed to work. According to the IEEE definition, software requirements can be classified in two categories: *functional* and *non-functional*. A functional requirement defines a function of a software system or its components. A function is described as a set of inputs, the behavior, and the obtained outputs. Functional requirements may be calculations, technical details, data manipulation and processing, and other specific functionality defining *what* a system is supposed to accomplish. On the other hand functional requirements are supported by non-functional requirements, which impose constraints on the design or implementation, such as performance requirements, security, reliability or usability. Non-functional requirements are also known as *quality requirements* [**2**]. In the following subsections an analysis of both functional and non-functional requirements of our system is carried out.

5.2.1. Functional requirements

In this section there are specified the set of services the user expects to obtain from the system. As it has been commented in the beginning of the chapter, in the context of research and modeling of expert knowledge, the former is not always easy to achieve.

The diagnostic process in the sleep apnea-hypopnea syndrome is a complex task. At the time of determining if a patient suffers from SAHS, it is necessary to have into mind several factors. The ultimate objective, however, is to establish a particularized diagnosis that determines, as close as possible as the human expert does, if the patient

[32] A system working correctly but not adapted to the way the users carry out their tasks will not have any value since it will not be used

suffers from the apneic syndrome, as well as to assess its associated severity and its concrete type. For that purpose, following the classical procedure described throughout Chapter 2, as well as the necessities and objectives identified during the introductory chapter, functional requirements of our system can be established. Each functional requirement can be described as a set of inputs, a behavior and outputs:

- **Construction of the hypnogram**. An analysis of the neurophysiological activity should be performed in order to characterize patient's sleep macrostructure. The previous provides of fundamental information to assess the sleep pattern of the patient and evaluate his/her sleep quality. It also serves as contextual framework to interpret the respiratory events. Input to this function involves the set of neurophysiological signals including EEG, EMG and EOG. As the output the hypnogram of the patient is obtained.

- **Detection of micro-arousal events**. EEG arousals constitute one of the main indicators of disrupted sleep. Their appearance break up the normal sleep cycle and it therefore results in restless sleep. Appearance of micro arousals during sleep can be associated to a number of circumstances, but in apnea patients they are usually related with an apneic origin [**6**]. As a consequence, its detection and quantification results also of interest for the detection of the apneic event. Input to this function includes EEG and EMG signals. As an output the number and the temporal location of the arousal events is obtained.

- **Detection of apneic event intervals among the respiratory signals.** According to AASM's definition, apnea refers to a total respiratory absence, whereas in the case of periods of partial reduction the respiratory event is named hypopnea [**7**]. These reductions are mainly localized in the airflow signal; however, occurrence of the apneic event may also be reflected in the signals of thoracoabdominal respiratory movements. This situation suggests that the analysis for detection of apneic intervals should take into account this set of signals (airflow, and thoracoabdominal movements). After this analysis localization of apneic intervals among these signals is obtained as output.

- **Analysis of behavior of the oxygen saturation signal**. The system should accomplish the analysis of the SaO_2 signal since, physiologically, the apneic event is associated with a drop in the oxygen concentration levels [**8**]. It is also a desirable requirement that the system was able to count the number of produced desaturations and resaturations, and to classify desaturations according to their associated reduction percentage. Characterization of SaO_2 signal is fundamental at the time of determining the SAHS severity in the patient [**7**]. Input to this function is the SaO_2 signal; output includes detection and quantification of desaturation and resaturation intervals.

- **Apneic event interpretation**. During sleeping the different respiratory signals are subject to both amplitude and frequency changes due to sleep phase changes, or because of contextual events such as a change in the sleeping position. In this regard, correct interpretation of the information obtained by the processing of the respiratory signals and the characterization of SaO_2, should be performed in the context of the sleep structure and the remaining contextual information. Input to this function includes both neurophysiological and respiratory information as well as additional contextual information. As output apnea and hypopnea events occurring in the PSG are obtained, while false positives by context are discarded.

- **Classification of detected apneic events.** For each individual detected event the system should be able to determine its class. This allows syndrome classification to be performed. Such a classification has to be done in the basis of information provided by signals of abdominal and thoracic respiratory movements, which bring evidence about presence or absence of respiratory effort during the occurrence of an apneic event. Therefore, output to this function comprises the classification of each apneic event previously detected as obstructive, central or mixed.

- **Calculation of significant numerical parameters.** Throughout the whole analysis process, several parameters and indexes result of interest at the time of evaluating the presence of SAHS in the patient, and if it is the case, the corresponding associated severity. A list of AASM recommended

parameters to be reported for polysomnography can be found in [**7**] and are used as a guideline to conform the output to this function. Input involves quantitative information from all the detected events.

- **Issuing of final diagnosis and explanation of the results.** As the final product of the integration of all the information, the system has to offer a particularized diagnosis concluding about the presence of SAHS in the patient, its type and its severity. It is important that the system was able to provide an explanation of its results, with a correct reasoning about them, so that allowing the clinician to adequately evaluate the diagnosis.

5.2.2. Non-functional requirements

As previously introduced non-functional requirements should establish restrictions in the product under development, in the developing process itself and also with regard to additional specific restrictions the product may have. A good definition of non-functional requirement is provided by Thayer [**9**]: *it is a software requirement that describes not what the software will do, but how it will do it*. A typical example of non-functional requirements is performance. Non-functional requirements are sometimes difficult to be objectively verified and therefore they are often evaluated subjectively. They are also usually associated with the concept of *software usability*.

In our case, success of the system depends on several factors, not just in the capability to correctly detect and classify the apneic events and to obtain a correct diagnosis. The following non-functional requirements should be considered during the development of the system:

- **Ease of use.** It has to be taken into account that it is about the development of a tool to help the clinician, for which managing of the resulting system should not overpass technical capabilities of the final user. The opposite situation might imply the rejection of the system.

- **Performance.** One of the main objectives of the software is to reduce the diagnostic time necessary for each patient. In this respect analysis time per recording should be minimized as much as possible and, in any case, it should never be higher than analysis time for manual revision from the part of the clinician.

- **Natural interaction.** System-clinician interaction has to be as natural as possible, without introducing unnecessary complexity to the analysis process of the PSG. It is also important to carry out the interaction in a language as close as possible to the own language of the clinician. It is the objective to develop a tool that is used by the clinician, not by the engineer.

- **Tidy presentation of the results.** System's results should be presented in a systematic and organized manner. When data volume is high then presentation of an integrated version is essential. In this compact report, the most relevant information is summarized allowing the user to breakdown the different items of interest at each time.

- **Flexibility of the system**. Since physiological signals that constitute the input to the system come from a hardware acquisition device, it would be a desirable requirement that the system was able to operate over data coming from the maximum possible number of recording devices. Therefore, the analysis could be performed independently of particularities of the acquisition device. Flexibility must be also understood at the time of presenting results of the analysis. In this respect we search for a system avoiding categorical results as much as possible, but providing of weighted level of confidence over each possible hypothesis.

- **Extensibility and modifiability**. System's design should be modular so that the different components can be easily exchanged. This requirement is important in a context of research and within an evolutionary life cycle. In this regard incremental design of the system is facilitated as well as the possibility to perform continuous improvement over the analysis algorithms.

5.3. System's architecture

In the introduction of this chapter it has been justified the use of an evolutionary developing model, mainly because of the inherent difficulties to the design of a model of intelligent behavior. The key to the success of this development methodology lies on the use of techniques allowing modification of the system, thus changes can be incorporated and tested as soon as possible [**1**] [**2**].

The above requirement has to be taken into account during the development stage, and it is reflected in the architecture of the system, which is highly modular. In this respect it is feasible the identification and the modification of each one of the tasks that the system carries out, as well as the incorporation of new ones. An architectonical schema of the proposed system including the main modules is depicted in Figure 5.2.

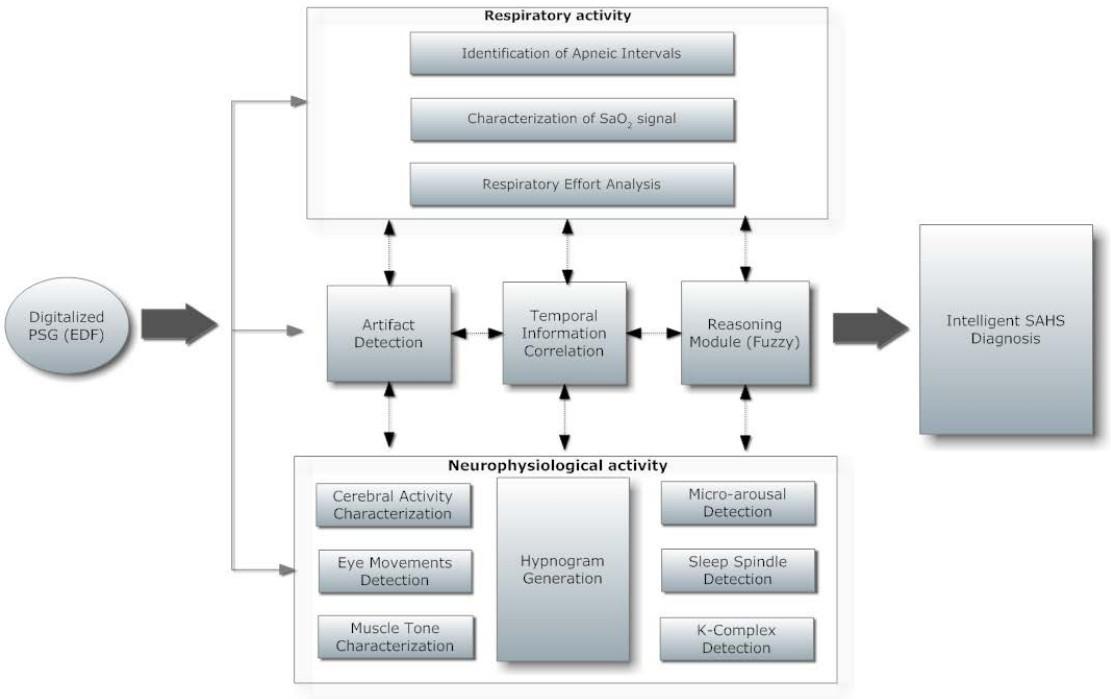


Figure 5.2. Architecture of the system

As it can be seen from Figure 5.2 functionality of the system is organized into several interconnected modules. Each module is in charge of specific functions within the analysis process. Input to the system is given from a digitalized polysomnographic recording containing raw physiological signals from the patient. Currently, data format

used by the system is EDF which is the de-factor standard for PSG digital recordings [**10**]. Data is then fed into the system for its analysis throughout the different processing modules and a SAHS diagnosis is obtained as the output.

There are two well differentiated groups of modules: (i) those specialized in the analysis of the respiratory activity and (ii) those specialized in the processing of neurophysiological activity. Within each major group integrating submodules are in charge of different specific subtasks. Respiratory analysis is structured into three submodules for the identification of apneic intervals, characterization of SaO_2 signal and analysis of the respiratory effort. On the other hand neurophysiological analysis is organized into seven modules: Three are responsible for cerebral activity characterization, eye movements' detection and muscle tone analysis. This information is then fed to the module in charge of obtaining patient's hypnogram. Other three additional modules deal with the detection of transient events including micro-arousals, sleep spindles and K-complexes. These two last (sleep spindles and K-complexes) are also used for the generation of the hypnogram.

The whole analysis process is concurrently assisted by functionality of artifact detection, temporal information correlation and reasoning modules. These are organized as general supporting modules, since they do not belong to a specific task but intervene at different time instants supporting remaining modules. For example, temporal correlation processes occur both for relating individual events among the neurophysiological signals to detect micro-arousals and for the correlation of respiratory events for the detection of apneic patterns (APs). Artifact detection and fuzzy reasoning processes also intercede at several stages throughout the analysis cycle.

In the following sections detailed functionality of the proposed system is carried out. To that purpose, and due to interoperability between modules in the architecture, an information flow perspective is preferred to the description of each module separately. In this respect the different analysis tasks are described in the same way the system operates.

Firstly, description of inputs to the system and some considerations regarding the artifact detection approach are discussed. Once the signals have been acquired by the

system, analysis of neurophysiological activity follows. Description starts by detailing developed approaches for the detection of transient events related to these signals. Methods for detection of micro-arousals in the EEG are firstly described. After that, implemented algorithms for the detection of sleep spindles and K-complexes are discussed. Both sleep spindles and K-complexes are then used as inputs for the hypnogram generation task which is subsequently discussed. This task additionally involves analysis procedures for the characterization of cerebral activity, eye movements and muscle tone.

After neurophysiological processing methods have been discussed, analysis continues among respiratory signals. Preprocessing algorithms for the setting-up of the signals are firstly applied. The objective is to detect and repair –if possible- artifacts causing overflow and loss of focus in the signals. After that, analysis of airflow and respiratory movements is performed for detection of possible apneic intervals. Procedure for the characterization of SaO_2 signal is finally described which objective is to localize desaturation and resaturation intervals.

All the previous information coming from the neurophysiological and the respiratory analyses is then integrated in time in order to form diagnostic patterns –also referred as *apneic patterns (APs)*. Apneic patterns represent pieces of information composed of several different events related in time. In this manner we allow interpretation of apneic events by considering all the possible sources of evidence. Reasoning processes are then conducted in order to identify the actual apneic events, i.e. to detect apneas and hypopneas throughout the PSG. Next step in the analysis involves classification of previous detected events according to their corresponding type (obstructive, central or mixed) for which an analysis of the respiratory effort is carried out involving thoracic and abdominal derivations.

Final step in the description comprises evaluation of data output from signal analyses, which is afterwards considered in order to compute significant numerical parameters and indexes to issue the final diagnosis.

5.4. Description of the inputs

As it has been described in the architectonical schema, input information to the system consists of the digitalized signals acquired during PSG examination. However, in addition to the physiological signals, also contextual information constitutes a direct input necessary for the correct interpretation of patient's biosignals. Such contextual information can be classified according to two different categories: *static contextual information* and *dynamical contextual information*.

Static contextual information mainly involves physical and demographic data from the patient. Such information is usually taken from the clinical records or it is acquired by the clinician during an interview with the patient. In most of the cases this information is decisive for the polysomnographic prescription.

Dynamical contextual information refers to that depending on a concrete instant of time in the PSG. It comes from the recording of environment variables during at the time of the PSG, as they are patient's body position during sleep, state of ambient lights or the recording of the acoustic snore signal. Also within this category, for example, might be considered the information coming from the sequence of sleep phases of the patient. However, it has to be taken into account that this dynamical contextual information is precisely derived from the trend of a subset of signals in the polysomnographic recording. In this respect, there is derived or inferred information and, therefore, it does not constitute a direct input to the system but it is taken into account in consecutive stages of interpretation.

Digitalization process of the physiological signals included in the PSG is carried out through a monitoring system. This monitoring system acts as a whole as a transducer that transforms physical measures, by means of the corresponding sensors, and through an analog-to-digital (A/D) conversion process, into the digitalized signals.

The physical monitoring system is often a commercial hardware device. This can represent an inconvenient at the level of non-functional requirements (see subsection "*System's requirements*"), since potentially, each commercial system may use its own digital data format. In this sense it is desirable a system to be flexible, so that

information treatment can be isolated as much as possible from how this is presented to the system. Therefore, the objective is to achieve a system being as independent as possible from the concrete hardware for data acquisition. The use of a standard format for the digital representation of the polysomnographic recordings should help to solve this problem in a great extent. Accordingly many efforts have been made for the elaboration of standard formats for the representation of biomedical signals. In this line the European Data Format (EDF) stands out as the result of an effort for the standardization of a simple and flexible format for the storage and exchange of multi-channel physical and biological signals. EDF is an open and free format and its description and their specifications are published in [**10**] and accessible from the web [**11**]. From 1992 this format has popularized and it has become one of the main formats for acquisition, storage and exchange of digital PSGs, both for commercial equipments as well as for multi-center research projects.

The proposed system assumes input of digitalized data (both signals and contextual information) to be in EDF format. It is also interesting to stress that EDF counts with the support of several toolkits, also open, which allow conversion from several data formats to EDF and vice versa. To cite one, the project BioSig [**12**] for example, offers an open-source library for the processing of biomedical signals which, in addition to other features, provides support for data acquisition. BioSig provides of a common function calling interface that abstracts acquisition from the concrete format in which data have been digitalized. This interface implements support for several digital formats of biomedical signals, among them EDF, allowing conversion from one format to another.

The use of a standard format such as EDF, in combination with available toolkits that facilitate data conversion offers, in part, a solution to the problem of format diversification. However, the problem of proprietary format is still inevitable since new formats continuously appear. Ultimately, the creation of a translator from new unsupported formats to EDF may be necessary in these cases. On the other hand, it is interesting to stress that nowadays the great majority of the proprietary software in the market includes the option to export its format to EDF.

Besides the digitalization format, another important factor that has to bear in mind when dealing with data acquisition is specifications and requirements for signal monitoring. These include the number and types of derivations to be used for each signal, its sampling rate, the number of bits per sample, or its dynamical range. These specifications may also be susceptible of interpretation with regard to the sleep lab's background or the concrete expert that carries out the polysomnography, existing different recommendations and schools of thought. In this respect, and in an effort to promote an standard on these and additional requirements, the AASM has recently (2007) published a manual that collects the recommendations, procedures and specifications at the time of recording the signals involved in the sleep diagnosis through the PSG [**7**]. Specifications picked up in the AASM manual are aimed at updating and substituting the former rules proposed by Allan Rechtschaffen and Anthony Kales in 1968 [**13**]. The main objective is the standardization and the unification of terms and definitions used worldwide for recording of PSG. In practice, however, convergence will still take several years, among other circumstances, because it implies substitution of the hardware monitoring devices at the hospitals that do not fulfill with the proposed specifications, as well as training of current personnel in the new protocols. Because of this, nowadays standardization of the monitoring protocol is not yet a reality.

All the above mentioned greatly determines the development of analysis software, making it dependent on the set of available signals. In this respect, for example, if the montage does not include derivations for the recording of the thoracic or abdominal movements, then it would not be possible to perform the classification of the apneic events; at least, not directly by following the standard recommended procedures.

Subsequently, Table 5.1 shows a description of the set of signals used as the input for the developed system. Consequently, the system works in the basis of such configuration by extracting the corresponding set of signals from the given EDF file. The concrete set of signals as well as their respective configuration lies in the montage of the standard reference that has been used for the validation of the developed system (see Chapter 6, "*Design of validation tests*"). Input acquisition module can deal with variations in the amplitude –dynamical range- or in the sampling rate of the signals. In this respect, scaling and resampling operations can be applied in the case of being

necessary, and allowing the user to modify input configuration of the system. On the other hand, as stated before, if the available set of signals varies, then partial readaptation of the analysis algorithms should be performed to adapt them to the new input set.

Table 5.1. Specification of input signals to the system according to the available PSG database

Respiratory signals			
Signal	**Derivation/cannel**	**Sampling rate**	**Comments**
Airflow	Thermistor	10 Hz	Amplitude normalized to [-1,1]
Oxygen saturation (SaO2)	Finger pulse oximetry	1 Hz	Expressed in saturation percentage (%) [0,100]
Abdominal respiration	Inductive plethysmography	10 Hz	Amplitude normalized to [-1,1]
Thoracic respiration	Inductive plethysmography	10 Hz	Amplitude normalized to [-1,1]
Neurophysiological signals			
Electroencephalogram (EEG)	C4/A1	125 Hz	Amplitude in µV [-125,125]
	C3/A2	125 Hz	Amplitude in µV [-125,125]
Electromyogram (EMG)	Submental	125 Hz	Amplitude in µV [-31.5,31.5]
Electrooculogram (EOG)	Left	50 Hz	Amplitude in µV [-125,125]
	Right	50 Hz	Amplitude in µV [-125,125]
Additional signals			
Body position	Mercury gauge sensor	1 Hz	1 = Supine 2 = Prone 3 = Left lateral 4 = Right lateral
Lights recording	Light sensor secured to the recording garment	1 Hz	On / Off

5.5. Handling of artifacts

Measuring physical variables on a living being has associated several problems. Without the aim to be exhaustive the following can be cited:

- Inaccessibility of variables to be measured
- Data variability
- Interaction between the physiological systems
- Effects of the measuring transducer
- Noise artifacts

In medicine and in biology the term *artifact* refers to any component from a signal being strange or odd to the physical variable that it represents. In this manner there are

considered as artifacts, for example, the white noise generated by the measuring device, electrical (mains) interference, signal coupling, and in general, any unexpected variation in the signal. It has to be taken into account that because of the characteristics of biomedical signals, these normally present a high sensitivity to noise (low signal-to-noise ratio). Especially among the polysomnographic signals, the set of neurophysiological signals (EEG, EOG and EMG) are the most sensitive, mainly because of their bandwidth and their low amplitude –usually in the scale of microvolts, which makes their recording quite a difficult task.

Some of the above mentioned problems are partially overcome by the acquisition devices themselves, which before the A/D conversion usually apply analog filters to get rid of part of the noise; some of them even perform digital filtering once the signal has been digitalized.

Movement is another important source of perturbations in the measurement of physiological variables of a living being, which rarely can be solved by filtering techniques. Movement causes the displacement of the sensors and of the recording device itself generating variations in the recording of the signals by the transducer. On other occasions, bad calibration of the measuring device may generate signal overflow causing incorrect measures on the extreme values. Overflow artifacts can also be caused by other artifacts such as transient interferences or movements of the patient.

In any case, all these situations, since not avoidable, should be detected and corrected as much as possible. However, at the same time, fixing of the detected artifacts should be carefully done since an excessive or incorrect filtering may cause the removing or alteration of relevant information. In general, establishing the limit on what is a spurious value and what is not is not a trivial task.

All these kinds of situations make of physiological signal acquisition and setting up to be delicate processes, and turns artifact detection to be a fundamental task within the analysis cycle.

It is for this reason that in the developed system artifact detection is conceived as a as a general supporting process (see architectonical design at subsection *"System's*

architecture") that has to act on different fronts and with different strategies during the whole analysis process. Each concrete artifact detection strategy has to be applied at the correct time, and depending on the specific kind of artifacts present at the corresponding level. Some of the artifacts handling strategies used throughout the analysis cycle are now cited from a general point of view. The following are described later on in more detail when carrying out the description of the corresponding analysis phases where they are implemented:

- From the point of view of signal conditioning, it is necessary the use of digital filters to reduce noise effects. As it has been pointed out, this task is especially important in the set of neurophysiological signals (EEG, EMG, EOG) although the use of proper filtering is also performed over the set of respiratory signals when necessary.

- Artifact detection due to situations of signal overflow as well as in the case of a loss of focus is treated specifically. In these cases artifact marking and classification is done, and signal reconstruction is performed where possible.

- Intervals with possible spurious values are detected and marked in the case of either (1) a body movement of the patient during sleep or (2) signal coupling from one channel to another.

- From a reasoning perspective, the set of previously detected artifacts are taken into account in conjunction with information coming from contextual signals in order to provide of an integrated interpretation of the set of significant events detected over the biological signals. In this respect, apneic patterns can be interpreted in the context of artifacts occurrence and refine the results of the analysis by discarding false positives.

5.6. Analysis of neurophysiological signals

Although sleep analysis might be considered as a general and independent process, not necessarily bounded to the diagnosis of SAHS, analysis of the sleep structure is a very important task at the time of evaluating the polysomnographic recording. Diagnosis of SAHS is not an exception, and in this regard, any event detected in the recording must be considered in the context of all the signals that integrate the PSG, and not just in the signal (or signals) in which it has been detected.

In the case of SAHS, for example, it is known that the apneic events often are produced with more frequency during deep sleep (N3) or REM sleep [**14**]. On the other hand, a significant reduction in the airflow that might be pointing out to an apneic event, should be considered a false positive or an artifact if the patient is awake (W), and consequently it should be discarded. It is also known that when respiratory activity significantly decreases, it is common that the patient undergoes a brief alertness state that causes the transition from deep sleep to light sleep (N1 or N2). In fact, the more apneic events occurring during sleep, the more fragmented the sleep cycle of the patient [**15**]. These previous examples point out to some of the reasons for which correct interpretation of the detected apneic intervals in the context of the sleep structure, turns to be of extreme importance for a correct diagnosis of the syndrome.

Analysis of sleep structure comprises the set of so-called *neurophysiological* signals that includes EEG, EOG and EMG. Specifically, in the scope of the developed system, such an analysis pursues a dual objective: (i) determination of the characteristic events from the sleep microstructure such as EEG micro-arousals, K-Complexes and sleep spindles, and (ii) construction of the sleep map of the patient or *hypnogram*. The former helps the temporal localization of specific events which, as in the case of EEG arousals, may be useful even for the localization of apneic events. This happens since occurrence of both apneic events and EEG arousals has, in many cases, a cause-effect relationship (the apneic event triggers an EEG arousal) [**16**]. On the other hand, construction of the hypnogram that –among other things- also depends on the detection of microstructure events such as sleep spindles or K-Complexes, serves as a context for the interpretation of the apneic event with regard to the sleep state of the patient. This

allows the scorer to sometimes confirm and sometimes discard –depending on the concrete sleep state- the possible occurrence of the apneic event.

In what follows a set of algorithms are discussed to be used for determination of the above mentioned items of information. Methods are based on the application of several signal processing and artificial intelligence techniques which are subsequently described.

5.6.1. Identification of EEG arousals

As it has been outlined in the previous subsection, in SAHS and as a consequence of hypoventilation associated to the occurrence of the apneic event, a decrease in the oxygen concentration levels in blood is produced. This lack of oxygen usually triggers a response in form of alert that, when produced during sleep, is known with the name of EEG micro-arousal. As the name suggests, the previous term of micro-arousal does not necessarily reflect a total awakening of the subject[33]. It should be rather understood at the level of sleep microstructure, and most of the times, it implies a change from a deeper sleep state in the patient to a lighter sleep phase. Occurrence of EEG arousals, on the other hand, is not always associated with an apneic origin, and it can be produced by other causes. In any case, whether the origin is, it is an indicator of disrupted sleep and therefore its detection is a good marker for the assessment of the sleep quality.

According to the AASM, an electroencephalographic arousal during sleep is defined as an abrupt shift in the EEG frequency including alpha, theta and/or frequencies greater than 16 Hz (but not spindles), that lasts at least 3 seconds and with at least 10 seconds of previous stable sleep [**7**]. Normal sleep architecture is altered by the presence of these events, and the sleep fragmentation they cause is one of the main reasons for the daytime sleepiness associated with a number of sleep disorders. Therefore, localization of these events is important, not just for the diagnosis of SAHS, but for the sleep studies in general. For scoring arousals at least one central derivation of EEG needs to be recorded. Arousal scoring can also incorporate information from the occipital region. In addition, the scoring of arousals during the REM phase requires of a concurrent increase in the submental EMG lasting for at least 1 second [**7**].

[33] Although it happens with a certain frequency

As a growing area of interest, a number of studies have been published in the recent years regarding automatic identification of EEG arousals. For example, in Cho et al. [**17**], a SVM performs a classification over a signal obtained from the C3/A2 derivation. Another method which also uses just one EEG derivation is the one proposed by Gouveia et al. [**18**]. However, these approaches have an intrinsic inconvenience because in accordance with medical standards, detection of EEG arousals during phase REM requires the inclusion of the EMG in the analysis. Other studies have focused more on alternative procedures centered on the use of indirect measurements for the detection – without using the EEG signal- such as heart rate variability [**19**] or peripheral arterial information [**20**] [**21**]. A drawback of these methods is again that they do not follow methodology of the standard clinical process.

Working on multi-channel data is a more complicated problem since one has to deal with the time correlation of the information from each individual signal. The proposal of Agarwal [**22**] is based on the analysis of the alpha and beta frequency bands over two derivations of EEG. Candidate parameters are then selected through a statistical analysis process. Although two EEG derivations are used on this work, the absence of the EMG channel makes it susceptible of the exposed limitations regarding event detection within REM stage. It is important to remark that in the same work it is pointed out to the high degree of discrepancy existent between different experts, as one of the main obstacles preventing the method to obtain better results. The work of De Carli et al. [**23**] refers to a method for the automatic detection of arousals based on wavelet analysis. The algorithm handles two EEG derivations and one from the EMG, and after feature extraction, the weighted average is taken on the overlapping events. More recently, Sugi et al. have reported a method for the detection of arousals in multi-channel data for the case of the sleep apnea syndrome [**24**]. The previous studies base their decision of the presence of the event upon the value of the extracted features, exceeding or not, certain established thresholds. Finally, the work of Shmiel et al. [**25**] should also be mentioned, proposing the use of data mining techniques for the extraction of the implicit patterns on several signals to perform the detection.

In the following the proposed method to detect EEG arousals within our system is presented. In this respect several methods and techniques from the machine learning

field are studied in order to investigate their applicability within the method. The work developed in this area is structured in several parts:

- Firstly, a method to perform automatic detection of EEG arousals in patients with SAHS is presented. This method is at the same time subdivided in three stages:

    1. A first stage of signal analysis in the time and in the frequency domains, in which a series of events are obtained from the raw signals of the patient.

    2. From previous detected events a number of features are extracted. Events are then related in time in order to construct characteristic patterns representing possible evidence of arousal in the recording. Adding features from the different events a total of 42 features define each characteristic pattern.

    3. The set of characteristic patterns previously found constitute a dataset. Several machine learning models acting as classifiers are then trained and compared to determine the best model. The objective is to achieve the best discrimination in order to detect EEG arousals from the dataset.

- After the best machine learning model has been obtained to act as classifier within the global method, a study on feature selection methods is scheduled. The objective is to try to reduce the number of necessary features while maintaining good performance of the classifier. Several feature selection techniques based on both filters and wrappers are studied at this respect.

A method to detect EEG arousals using machine learning models

In accordance with available inputs to our system (see Table 5.1 in section "*Description of the inputs*"), the proposed method works with three signals: both EEG C3/A2 and C4/A1 central derivations, and submental EMG, all sampled at 125 Hz.

Following a similar approach to a physician examining the recording, the method firstly looks for individual events occurring on each separate channel. A first stage of signal analysis, (both amplitude and frequency based) is carried out for this purpose extracting features from the patient's recorded biosignals. Once events are marked separately in each signal, a set of temporal rules are applied in order to group them in terms of *characteristic patterns.* These patterns represent (characterize) a time interval in the PSG where possible arousal events occur. In order to make the final decision on the presence of the arousal, the set of features contained in the pattern are fed into a two step classification phase. A number of machine learning models are compared for this purpose, including Fisher's linear and quadratic discriminators, Support Vector Machines (SVM) and Artificial Neural Networks (ANN).

1) Processing of EEG and EMG signals

In the first phase, the signal processing occurs in the three signals in a similar fashion, taking two moving windows over each channel. These two temporal windows are used to compare consecutive temporal segments representing the past and the future related to the current instant of time under analysis. This process is accomplished for each second of the signals.

In the case of EEGs, for every second and going back from the corresponding sample, a 10-second window is used to represent the prior information, and immediately consecutive to it, a new 3-second window is used to compare the evolution of the signal against the former. As the definition of EEG arousal is based on the localization of an abrupt shift in EEG frequency, a frequency-based comparison between these two windows is performed. The chosen duration of each window is based on the idea that the EEG frequency shift must be 3 seconds -or greater- to be scored as an arousal, and

that a minimum of 10 seconds of intervening sleep is necessary to score a second arousal once the former is detected [**7**].

The frequency comparison of the two windows is made as follows:

1. Each of the two segments is transformed into frequency domain making use of the Fourier Transform

2. A band-pass filter is applied in the range 8-30 Hz, so that it includes frequency bands in the range of *alpha* (8-12 Hz) and *beta* (13-30 Hz).

3. Power Spectral Density (PSD) is calculated over each filtered band according to the following formula:

$$PSD = \frac{1}{N^2}\sum_{n=1}^{N}\left|X_{bp}(n)\right|^2 \tag{5.1}$$

where N is the number of samples, $X_{bp}(n)$ is the filtered signal in the 8-30 Hz band, and $1/N^2$ is a weighting value which allows comparing power between the two windows even though time length of each one differs[34]. Using equation (5.1) comparison of the corresponding power values (their ratio Φ) on the two windows is performed, and a scalar magnitude is obtained which represents the frequency shift within the band. Repeating this procedure throughout the two EEG channels, a new signal $\Phi PSD_{\alpha,\beta}(n)$ (1 Hz resolution) is obtained for each EEG channel. This signal is normalized to the range [-1,1], growing wherever the evidence of a shift in the EEG frequency in the alpha and beta range occurs, decreasing as the evidence disappears (see Figure 5.3). The use of a change in the alpha-beta range as a continuous marker for the sleep depth (and hence good evidence for the occurrence of arousal events) seemed also to be a good approach as suggested by Asyali et al. [**26**]. This supports the use of the 8-30 Hz band as a valid marker for the localization of arousals.

Taking this $\Phi PSD_{\alpha,\beta}(n)$ signal (which is centered in zero) those intervals in the EEG surpassing the zero value are marked. These intervals, which are considered to be

[34] Recall that we are comparing two temporal windows, one of duration 10 seconds and the other with duration 3 seconds

indicative of the possible presence of an EEG arousal, are considered later on when searching for *arousal patterns -characteristic patterns* (see Figure 5.3).

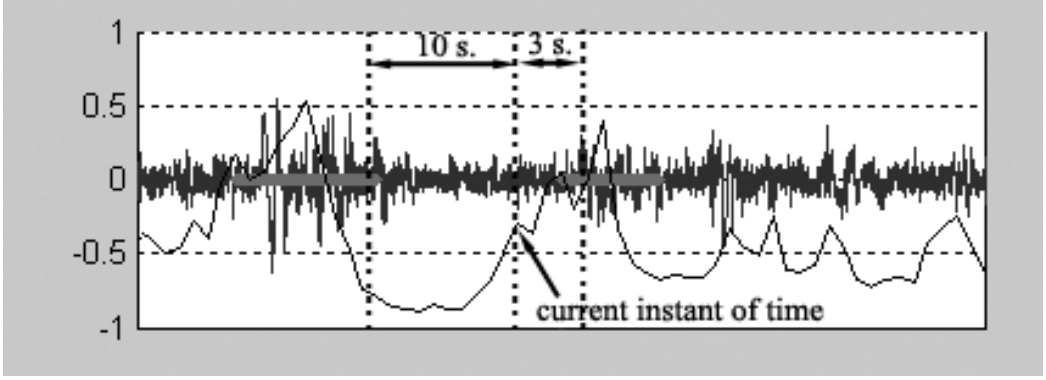


Figure 5.3. Example of a EEG derivation during signal processing. In the figure one minute of EEG signal with amplitude normalized to the interval [-1,1] is shown. Signal $\Phi PSD_{\alpha,\beta}(n)$ is superimposed and events are marked where this signal is above zero

Although the possible events are marked by the above-mentioned procedure, (based on shifts in the band of 8-30 Hz), a similar process is done individually for the delta (0.5-3 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta (13-30 Hz) and sigma (12-14 Hz) bands. These frequency bands are the most relevant ones in the case of sleep studies [**13**]. In this way, an evolution of the different relevant frequencies over time is individually obtained. Thus, once a possible event is detected with the 8-30 Hz marker, information from these individual signals is also available, which is also used for the extraction of relevant features. These features are afterwards used in a subsequent stage to train the classifiers.

In the case of the EMG signal, searching for possible events is performed in a similar way, shifting two moving windows. As it was the case for EEG, processing is done second-by-second. In this case the first moving window, which represents past history of the signal, has duration of 30 seconds, whereas the window representing future information has duration of 3 seconds. This time the values are chosen empirically, since the normative does not specify minimum required duration for the muscular event. In the case of EMG, the objective is to detect increases in the amplitude of the signal. To achieve this, a comparison between the amplitudes of the two windows is performed, in contrast to the frequency-based analysis performed in the case of the EEG.

Similarly, the result of applying this comparison throughout the entire signal is a new signal *a(n)* of 1 Hz resolution. The dynamic range of *a(n)* is also contained in the interval [-1,1], increasing when the amplitude of the EMG signal grows, and decreasing when reduction in the amplitude occurs. Analogously, intervals of EMG in which *a(n)*>0 are marked as the possible events (see Figure 5.4).

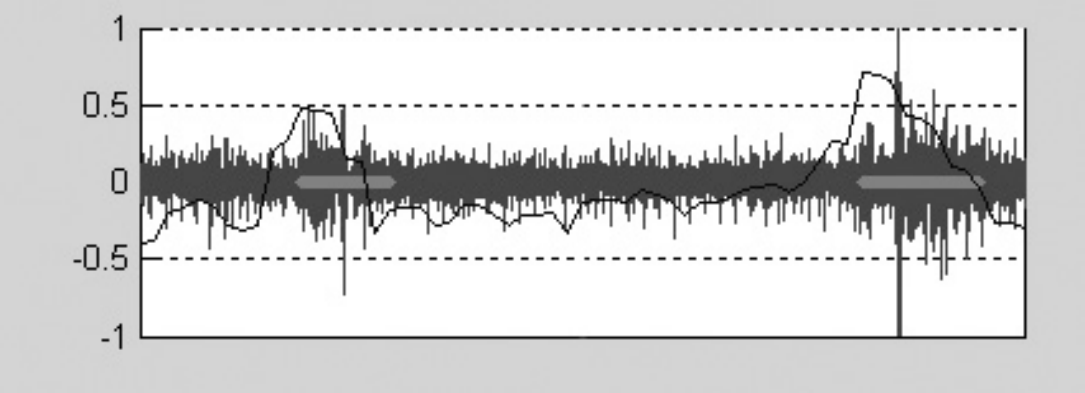


Figure 5.4. Example of EMG signal during signal processing. In the figure one minute of EMG signal with amplitude normalized to the interval [-1,1] is shown. Signal *a(n)* is superimposed and events are marked where this signal is above zero

After the signal processing stage for the marking of individual events[35], features are extracted from the corresponding intervals. These features are subsequently used in order to construct *arousal patterns*. The set of extracted features is summarized in Table 5.2. In Table 5.2 $\Phi PSD_{\alpha}(n)$, $\Phi PSD_{\beta}(n)$, $\Phi PSD_{\theta}(n)$, $\Phi PSD_{\delta}(n)$ and $\Phi PSD_{\sigma}(n)$ reference the ratio of power values between the two moving windows at time *n*, and for each individual frequency range (respectively, *alpha*, *beta*, *gamma*, *delta* y *sigma*).

[35] Indicators of possible evidences for the detection of arousals

Table 5.2. Features extracted from possible events detected in EEGs and EMG used for pattern construction

Source signal	Feature number	Description
(a) Features from the signals		
EEG C3/A2	1	Area of alpha-beta marker $\Phi PSD_{\alpha,\beta}(n)$ over the threshold
	2	Maximum value of $\Phi PSD_{\alpha,\beta}(n)$ during the possible event
	3	Duration of the possible event
	4,5,6,7,8	Area of $\Phi PSD_{\alpha}(n)$, $\Phi PSD_{\beta}(n)$, $\Phi PSD_{\theta}(n)$, $\Phi PSD_{\delta}(n)$ and $\Phi PSD_{\sigma}(n)$ over the threshold
	9,10,11,12,13	Maximum value of $\Phi PSD_{\alpha}(n)$, $\Phi PSD_{\beta}(n)$, $\Phi PSD_{\theta}(n)$, $\Phi PSD_{\delta}(n)$ and $\Phi PSD_{\sigma}(n)$ during the possible event
	14,15,16,17,18	Minimum value of $\Phi PSD_{\alpha}(n)$, $\Phi PSD_{\beta}(n)$, $\Phi PSD_{\theta}(n)$, $\Phi PSD_{\delta}(n)$ and $\Phi PSD_{\sigma}(n)$ during the possible event
EEG C4/A1	19	Area of alpha-beta marker $\Phi PSD_{\alpha,\beta}(n)$ over the threshold
	20	Maximum value of $\Phi PSD_{\alpha,\beta}(n)$ during the possible event
	21	Duration of the possible event
	22,23,24,25,26	Area of $\Phi PSD_{\alpha}(n)$, $\Phi PSD_{\beta}(n)$, $\Phi PSD_{\theta}(n)$, $\Phi PSD_{\delta}(n)$ and $\Phi PSD_{\sigma}(n)$ over the threshold
	27,28,29,30,31	Maximum value of $\Phi PSD_{\alpha}(n)$, $\Phi PSD_{\beta}(n)$, $\Phi PSD_{\theta}(n)$, $\Phi PSD_{\delta}(n)$ and $\Phi PSD_{\sigma}(n)$ during the possible event
	32,33,34,35,36	Minimum value of $\Phi PSD_{\alpha}(n)$, $\Phi PSD_{\beta}(n)$, $\Phi PSD_{\theta}(n)$, $\Phi PSD_{\delta}(n)$ and $\Phi PSD_{\sigma}(n)$ during the possible event
EMG	37	Area of a(n) over the threshold
	38	Maximum value of a(n) during the possible event
	39	Duration of the possible event
(b) Contextual features		
EEG (C3/A2 & C4/A1) and EMG	40,41,42	Binary attributes indicating the existence or not of a marked event in the channels EEG (C3/A2 & C4/A1) and EMG respectively

2) Construction of characteristic patterns

Once events are localized along each single channel, the aim is to associate them, and to determine if they represent a pattern of clinical interest for the detection of an arousal. In this manner, a certain interval *t* in the PSG can be characterized using the information provided by the combined events. Eventually the objective is to see if the evidence carried by this *characterizing pattern*[36] is enough to classify the corresponding

[36] Formed by the aggregation of the individual events from the different signals

interval as representing an actual EEG arousal event. Subsequently it is described a procedure to do so by using a machine learning classification approach.

A classical problem of pattern classification by supervised learning can be represented as a 2-variable term <*v*,*d*>, where *v* is a vector of features which describes the example, and *d* is a label indicating the membership class or desired output. This term *d* is the one that guides the machine learning algorithm to fit the model parameters during the learning process, thus supervising the learning.

However, in this case, one must also consider the time factor. This is not the case of a static problem, but a series of events occurring in a concrete instant of time. Moreover, although the physiological event (arousal) occurs isolated in a more or less precise time location *t*, the diagnostic process implies searching for evidence in all the three biomedical signals –the two EEGs and the EMG. This fact is not envisaged as being problematic for physicians, as experts are used to recognize the arousal through the observation of the independent evidence in the single signals over time. However it complicates the classification problem from the computational perspective. The initial problem of the classification of patterns of examples can be better represented as an 8-variable term <*v1*,*t1*,*v2*,*t2*,*v3*,*t3*,*t*,*d*> where in a certain instant of time *t*, a physiological event takes place classified as *d*, and occurs due to three individual events represented by the three corresponding feature vectors *v1*,*v2*,*v3* –EEG1, EEG2 and EMG-, and localized respectively in times *t1, t2* and *t3* (although certainly, in the environment of the time instant *t).*

One possible solution for this problem comes through the reduction or the projection of the 8-variable problem into the classic 2-variable one. As stated before, a method based on the time association of the events is proposed at this respect. The process is outlined in Figure 5.5. Basically, it consists of grouping together those events happening –*t1*, *t2, t3*- related to a certain time interval *t*. This association can be physiologically interpreted as a period of time in the recording where the occurrence of an arousal event is possible. This group can then be represented by a unique feature vector resulting from the union of the 3 vectors *v1*, *v2* and *v3* included in the group. Supposing *v1*, *v2* and *v3* contain features *v1* = $\{v_{11}\ldots v_{1n}\}$, *v2* = $\{v_{21}\ldots v_{2m}\}$ and *v3* =

$\{v_{31}…v_{3l}\}$ respectively, the resulting vector $vf = \{v_{11}…v_{1n}, v_{21}…v_{2m}, v_{31}…v_{3l}\}$ is thus composed of $n+m+l$ features.

Each group, represented by the corresponding union vector, can be assigned with the label *d* (representing its desired output) by means of the temporal classification of the time interval *t* by the expert. To achieve this, the recording is segmented into classifiable intervals called epochs. Usually in the topic of sleep studies, a commonly used measure of time used for the epoch is 30 seconds. Location of an event unequivocally in one epoch can be made by using its midpoint [**27**].

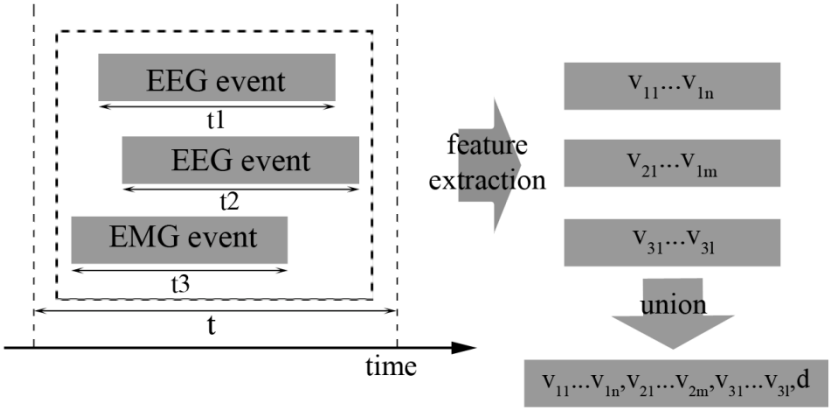


Figure 5.5. In a certain period of time *t*, classified as d by the expert, evidence of the physiological event is seen by the presence of events in the three signals during times *t1*,*t2* and *t3*. By grouping these events, a learning vector is constructed by uniting their respective features

In the case whereby more than one event for the same signal is present in the same epoch, the association process is accomplished by using a set of relatively easy temporal constraints. For example, let us consider following the situation illustrated in Figure 5.6. If an EMG event is present, the event is grouped with those from EEGs which midpoints are the closest to the midpoint of the EMG event. In Figure 5.6, epoch C shows that the events which relate the best are V11C and V21C, because of their proximity to the EMG event (V31C). In a different situation where V12C were the closest to V31C instead of V11C, the latter would be left aside and V12C would be the event integrated in the group. Alternatively, if no EMG event is present over a certain period of time, those events with the highest difference between their respective beta and delta energy shifts are chosen (Figure 5.6, epoch A)

As it can be appreciated in Figure 5.6 (epochs A and B), there are not always events available on the three channels, i.e. three events cannot be included in a group. When this happens, special values indicating the absence[37], are assigned to the corresponding position of the feature vector that represents the *arousal pattern*. When a missing value is present, the value in the corresponding feature is set to zero. Taking into account that signal domain is within the range [-1,1], zero points out to the absence of evidence, either positive (0,1] or negative [-1,0), with respect to the corresponding feature.

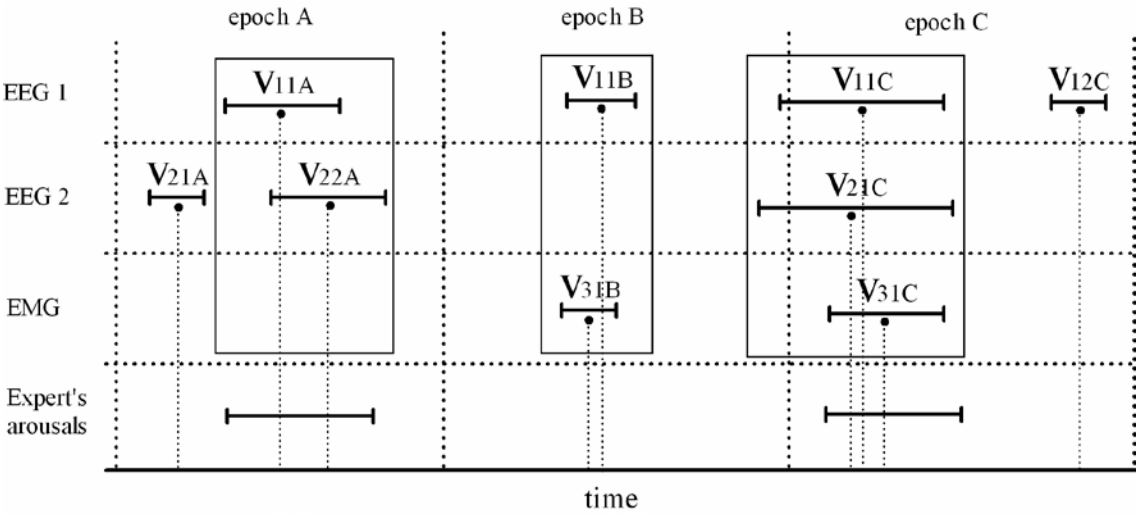


Figure 5.6. Event's association; any possible event detected in the signals is assigned to an epoch based in its midpoint. Then for each epoch, all possible events assigned are selected and grouped based on temporal constraints. Each group is surrounded with a square. The desired output is then constructed based on expert annotations for this epoch. Note: In this example we are assuming that the event V22A has a larger difference between beta and delta frequencies than V21A

In addition, during the process for the construction of the characterizing patterns, besides the features previously extracted and detailed in Table 5.2, three more are added. They are referred as *contextual features* and its description is summarized in Table 5.2(b). These features reflect the presence or the absence of an event in a channel, by properly setting the value of the corresponding binary feature. Therefore one expects the machine learning algorithm to be able to extract information on the missing values, not only by the values on the features themselves, but also using these contextual features.

[37] *Missing values*

Finally there is the possibility whereby no event is detected in any of the single channels; should this occur, no *arousal pattern* is constructed for the corresponding epoch, this epoch being directly discarded as having any presence of an *arousal event*.

It has to be mentioned that the previous set of temporal constraints is actually an *ad-hoc* method. It is not claimed to be a clinical procedure, but it is empirically based on the observation related to the expert's annotation of the events in the recording. The ultimate objective is to relate the output events from the signal processing phase in order to form the so-called characterizing patterns, and to construct a set of learning patterns. More information on the temporal issues related to diagnosis in sleep studies from a knowledge-based approach can be found in [**28**].

As the output of this process, for each epoch of the recording (in which at least one single event is present) a feature vector *v* representing the evidence of an arousal event in the corresponding time interval is obtained, as well as the desired output *d* extracted from an expert's classification of the epoch. The whole set of patterns constructed by this procedure is then used in the subsequent classification phase to compare the precision of different machine learning models analyzed.

3) Classification stage

A comparison of several models for classification with different configurations is investigated at this phase. Here the best classifier for the detection of arousal events is investigated based on the features included in the characteristic patterns constructed as explained in previous sections. Several machine learning models are compared including: (i) Fisher's linear discriminant, (ii) a Quadratic discriminant, (iii) SVMs, and (iv) ANNs. In the following a brief description on the different classifier models is given:

Fisher's linear discriminant [**29**] is based on the idea of splitting the space into classes by using a linear combination of the attributes at the input. These linear discriminants are defined between each pair of classes. The projection maximizes the distance between the means of the two classes while minimizing the variance within each class.

The quadratic discriminant [**30**] is similar but the frontiers between each decision region are now quadratic surfaces. One discriminant per class is calculated as the logarithm of the PDF estimated for the amount of samples available for the corresponding class.

The SVMs [**31**] are based on the maximization of the margin between two data sets for which two parallel hyperplanes are constructed, one on each side of the separating hyperplane. Effective separation is achieved by the hyperplane with the largest distance to the neighboring data points of both classes. In 1995, Cortes and Vapnik [**32**], introduced the concept of *soft margin* to deal with spaces not linearly separable. This introduces a parameter *C* as a compromise term between the margin size and the classification error. Also in 1992, Boser et al. [**33**] suggested the *kernel trick* allowing SVMs to perform non-linear classifications. This introduces two new parameters for the configuration of the SVM: (1) the shape of the kernel function *K*, and (2) the smoothing parameter *S* for the chosen kernel function. Several kinds of kernel functions can be chosen, e.g. linear, sigmoid, polynomial or radial basis function (RBF). The latter was chosen in our case. The idea behind choosing a RBF kernel is that it maps the samples into a higher dimensional space, so unlike linear kernels, it can handle the case when the relation between class labels and attributes is nonlinear. Furthermore, the linear kernel is a special case of RBF [**34**], and the sigmoid behaves like the RBF for certain parameters [**35**]. In comparison with polynomial kernels, RBF kernels have fewer hyperplanes. This influences the complexity of model selection, and in addition, they have shown to have less numerical difficulties when the training set is large [**36**].

Finally, ANNs are mathematical models biologically-inspired on how the biological neurons work. They are basically composed of a set of interconnected nodes. Each connection has a weight which is a measure of the relative importance of this connection. Different models of ANNs are available throughout the literature depending on the architecture (the amount of connections between the neurons or their organization in layers), on the process to adjust their weights, or on the propagation of the information from the inputs to the outputs [**37**]. In our case, a feed-forward network with one hidden layer trained with a scaled conjugate gradient backpropagation algorithm [**38**] is used.

Although both linear and quadratic discriminants are nonparametric models, SVMs and ANNs on the contrary are parameter-dependent. Therefore prior to comparing the four types of classifiers, a configuration procedure is performed to determine the best parameters for both the SVM and the ANN. Optimal configuration for the SVM is tested for combinations of *C* and S, trying exponentially growing sequences: $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$, $S = 2^{-15}, 2^{-13}, \ldots, 2^{3}$. In the case of the ANN, similar exponential increments on the number *H* of neurons in the hidden layer are used: $H = 2^{1}, 2^{2}, \ldots, 2^{6}$. It has to be pointed out that whereas the choice of different parameters *C* and *S* does not necessarily increment the complexity of the SVM model, in the case of an ANN, the number of neurons in the hidden layer directly entails an increment in the complexity of the ANN in terms of training computational complexity. Results on the choice of best model configurations are shown in Chapter 6 "*Design of validation tests: identification of EEG arousals*".

Once configuration parameters are determined, comparison of the four classifier models is performed. The objective is to choose the classifier which best determines, from an *arousal pattern*, if it represents a true EEG arousal or on the contrary, it must be discarded. Nevertheless, when considering the patient's full recording, the final classification accuracy of whole epochs does not uniquely depend on the performance of the classifier. Remember that the method first performs a single-event detection on each signal. Thus, when for a determined epoch, no events are detected in any of the single channels, the method directly assumes that no-arousal event is present on that epoch, i.e. a characteristic pattern has not been constructed, and therefore, the decision is no longer based on the classifier. A schematic representation of the classification stage is depicted in Figure 5.7.

Design of validation tests to select the best classifier can be found in Chapter 6 "*Design of validation tests: identification of EEG arousals*". Corresponding results can be found in Chapter 7 "*Identification of EEG arousals*".

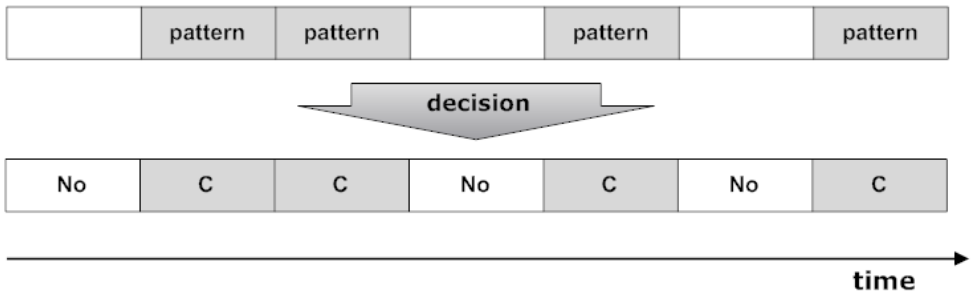


Figure 5.7. Classification stage; in the figure 7 epochs are displayed, of which 4 contain a characteristic pattern after grouping the events from the 3 channels. Epochs not containing a characteristic pattern are directly classified as *non-arousal* (Classification step 1). C means the decision on the classification depends on classifier's decision (Classification step 2)

Study on the reduction of the number of features

In this section the use of different methods of feature selection is suggested to study their influence for the detection of micro-arousals from datasets obtained in the previous sections. In this respect, in the basis of the results of the preceding comparative study (see Chapter 7 "*Detection of EEG arousals*"), it is suspected the possible presence of redundancy in the dataset, which may hinder the correct learning of the models, especially of the SVMs. In addition, by taking a look to the set of extracted features in Table 5.2, existence of redundancy is even expected since some of them, although in a different manner, are in some way reflecting single physiological evidence. For example, given an increment in a certain frequency band, both the area contained under the signal and its maximum value, are influenced by –they are proportional to- the amount of increment produced. In fact, this *a posteriori* analysis over the set of selected features will allow us to answer to the following kind of questions: *Which is the best set of features for the detection of micro-arousals?* For example*: Is it better to use the area below an increment as a relevant feature, or rather to use the maximum value of the increment?*

Feature selection methods allow us to reduce the number of features, identifying the most relevant ones, while preserving a good performance on the classifier [**39**] [**40**]. Thus, on this section it is proposed the use of different feature selection methods to study their influence on the identification of arousals in sleep studies. Each one of the tested feature selection methods ranks the features according to different relevance criteria (information gain, accuracy, etc.) and, therefore, different rankings are expected

to be obtained. Moreover, the study does not only compare the effectiveness of the individual methods in the domain, but it also investigates their combination. Very few studies have been conducted on the combination of different rankings. For example in [**41**], the authors combine the results for three filters. However, this combination study is more restricted, using only the intersection with those features selected by the three methods being labeled with *very high relevant*. Here, besides the individual methods, both intersection and union are tried, and the combination of features is explored adding one by one to the candidate subset, and checking whether classification accuracy is improved after incorporating new features. Additionally, in the previous study only one classifier -a decision tree- is used after the whole subset of candidate subsets is constructed. In the present approach, in contrast, two very well-known classification methods, ANNs and SVMs, are included, accordingly with the previous experimentation carried out for the detection of EEG arousals. The use of two different classification methods is also aimed to demonstrate that the improvement in results depend much more of the subset of features than on the classifier used.

1) Feature Selection

Knowledge Discovery in Databases (KDD) is a field in the computing sciences evolving to provide automatic analysis solutions in order to extract the potentially useful information from data. This information is not typically retrievable by standard techniques but is approached through the use of AI techniques. In particular, dimensionality reduction methods may significantly support these processes by means of finding a sub-dimensioned representation of the problem but preserving the maximum possible information about original data.

Data dimensionality may negatively influence the experimental results in machine learning problems. The higher the dimensionality of the input to the learning system, the higher the amount of examples needed to obtain a good model. In many real problems, there are not enough samples, so the learnt models can be over-fitted. Moreover, data usually contain noise and redundant information which may hinder and slow down the learning process. Reducing the dimensionality of the input space implies a decrement in the number of system’s parameters, thus decreasing the complexity of the model and its execution time. Generalization capabilities of the model increases as well. For that

reason, reducing the number of inputs or features can benefit in a variety of application domains in which machine learning algorithms are adequate, such as stock market analysis or medical diagnosis.

There are two main techniques to obtain feature reduction: feature extraction, which aim is to find a new set of *r* dimensions that are a combination of the *n* original ones, and feature selection, where a subset of *r* relevant features is selected from a set *n*, which remaining features are ignored. In the medical diagnosis field, it is important that features remain meaningful for the physician; therefore, feature selection is preferable.

At the same time, feature selection methods can be grouped into two main categories: *filters* and *wrappers*. Filter methods carry out the selection as a pre-processing step without using an induction algorithm, so only intrinsic characteristics of the training data are employed to select the relevant features. On the other hand, wrappers use an induction algorithm to evaluate each candidate subset of features. Filter approach is faster than the wrapper approach, and results in a better generalization because it acts independently of the induction algorithm. Besides, wrappers are very time-consuming because they demand to train an induction algorithm several times, and therefore, for some data sets with a large number of instances they are intractable. However, wrappers usually turn to better performance results than filters, although they may obtain good results with the inherent induction algorithm and may perform poorly with an alternative algorithm [**42**].

In the case of using wrappers another decision variable is the election of the searching strategy. The main reason to employ a search strategy falls in the fact that, normally, the exhaustive search across all the space of possible subsets of features is usually unviable. Generally, distinction is done between forward and backward searching. Starting from the empty set of features and adding features on each iteration to the subset being evaluated in the first case or, starting from the full set and deleting features in each step of the searching in the second case. As previously indicated, the exhaustive search across all the space of possible subsets of features is usually unviable. Therefore, it is common the use of heuristics to guide the searching process.

In this case, both search techniques are to be used in order to perform a comparative study. Next subsections briefly describe these methods.

1.1) Filter methods

There are many filter algorithms described in the literature. Therefore, different filter methods are selected based on their use of different measures (entropy, distance, etc.), so that they are expected to lead to significantly different rankings of features. Furthermore, filters can return a ranked list of input features, or a subset of significant features. In the present study, all the filters chosen are rankers. Filters returning a subset of features are not considered in order to facilitate a fair comparison. Entropy Minimization Discretization (EMD) [**43**] is used in order to discretize numeric attributes when required. The filters employed are:

-*RelieF*. The original RELIEF algorithm [**44**] estimates the quality of attributes according to how well their values distinguish between instances that are near to each other. For this purpose, given a randomly selected instance, x_s, RELIEF searches for its two nearest neighbors: one from the same class, called nearest hit H, and the other from a different class, called nearest miss M. It then updates the quality estimate for all the features, depending on the values for x_s, M, and H.

-*InfoGain* [**45**]. It evaluates the worth of an attribute by measuring the information gain with respect to the class:

$$\text{InfoGain(Class, Attribute)} = H(\text{Class}) - H(\text{Class} \mid \text{Attribute}), \quad (5.2)$$

i.e. it measures the change in the information entropy H achieved by learning the state of the random variable *Attribute*.

There are different filter methods based on entropy, such as GainRatio [**46**] and Symmetrical Uncertainty (SU) [**47**], which is defined as the ratio between the information gain and the entropy of two features, x and y:

$$SU(x,y) = 2\frac{IG(x/y)}{H(x)+H(y)} \tag{5.3}$$

These methods were also tested, but better performance was achieved by InfoGain, so only results related to the latter are included here.

-*OneR* [**48**]. It produces simple rules based on one attribute only. It operates by generating a separate rule for each individual feature. Each feature is discretized into bins and within each bin the percentage in which each class appears is calculated. Each bin is assigned to the class that has the highest percentage within that bin. After forming the rules, the single feature with the smallest error rate during training is selected. The class for any test case is assigned based on the bin into which its value for the selected feature falls [**49**].

1.2) Wrapper Methods

Wrappers require an induction algorithm to determine the quality of each candidate feature subset, i.e., instead of using subset sufficiency, entropy, or another explicitly defined evaluation function, a kind of “black box” function is used to evaluate the features. Wrappers need, in addition, of a search strategy to explore all the possible feature subsets. An exhaustive search is not adequate for most of the problems because it demands high computational resources. Therefore, sub-optimal strategies are adopted. Well-known strategies are sequential forward selection and sequential backward selection (SFS and SBS, respectively). SFS starts with an empty set of features and adds features one by one, while SBS begins with a full set and removes features one by one. Features are added or removed on the basis of improvements in the evaluation function. In the present study, both strategies are considered using a greedy Hill-climbing algorithm with backtracking. Backtracking is allowed after five iterations, if the current subset does not improve the best accuracy achieved until that moment. Besides, as stated before, two inductions algorithms are selected: ANN and SVM. Combining these search strategies with both induction algorithms, four different wrappers are built, namely: ANN-backward, ANN-forward, SVM-backward and SVM-forward. Therefore, a different ranking of features is derived for each one of the previous methods.

2) Experimentation

Experimentation for studying the application of the feature selection techniques described above is done in the following manner:

(i) A data set is constructed based on the previous gathered features for the detection of arousals in PSGs. The list of features is summarized in Table 5.2.

(ii) The resulting dataset is further split in two new datasets: one dataset used for training (TR), while the remaining patterns are used as a separate external test set (TS). Both datasets consist of a balanced number of patterns belonging to each class. Expert's annotations on the scoring of arousals following the AASM criteria are used in order to establish the desired output for each pattern contained in the datasets.

(iii) Over the TR dataset, several feature selection methods based on both wrappers and filters are applied in order to discard the irrelevant features. As a result, several candidate subsets –subsets of the most relevant features selected by the method- are constructed based on the measures obtained by each selection method. The process to construct these candidate subsets, which is based on the TR data set, is explained in Chapter 6 "*Design of the validation tests: Feature selection on the detection of EEG arousals*".

(iv) Once the candidate subsets are constructed, their respective predictive power is measured. These tests are carried out by using machine learning models in form of classifiers. Specifically, an ANN and a SVM, accordingly to previous experiments on arousal detection, are used. It has to be remarked that it is not the objective of this study to perform a comparative study over the different machine learning models; however the use of two models is aimed at confirming that the predictive power does not depend on the subsequent model used for classification. For each candidate subset, the classifiers are first trained using the TR dataset. Later, the test set is used to check its generalization capabilities. It is important to note that each classifier is also trained and tested using the whole set of 42 features in order to obtain a reference value of the performance. In this manner it can be compared the benefit of using a reduced set of features with regard to the use of the full set of features.

More details about the construction of datasets and the concrete classifiers' parameterization are given in Chapter 6 "*Design of the validation tests: Feature selection on the detection of EEG arousals*". Results of the comparison can be found in Chapter 7 "*Feature selection on the detection of EEG arousals*".

5.6.2. Sleep Spindles

Sleep spindles are one of the hallmarks of human N2 sleep stage and are also one of the few transient EEG events which are uniquely related to sleep [**50**]. Although the term sleep spindle was introduced by Loomis et al. [**51**], actually, it was described for the first time by Berger [**52**]. Generally speaking, this kind of event is characterized by a group of rhythmic waves which progressively increase their amplitude, then gradually decreasing. They normally appear linked to low voltage background EEG, superimposed to delta activity, or temporally locked to a vertex sharp wave or to a K-Complex.

The interest in sleep spindles has been enhanced by recent neurophysiology discoveries that point out to variations in membrane potentials in the thalamocortical network, that oscillate in the frequency range of spindles at an intermediate level of hyperpolarization and in the frequency range of delta at a higher level of hyperpolarization. These findings found a close relationship between changes at a neuronal level in the thalamocortical network and at the macroscopic EEG level, with a reciprocal relationship between sleep spindles and slow waves [**53**] [**54**].

Sleep spindles have been shown to have intra-cycle variations in form of U-shape within the first four sleep cycles [**55**]. With regard to inter-cycle evolution, visual scoring [**56**] and spectral analysis [**57**] [**58**] have univocally shown that spindles increase over consecutive sleep cycles. It has also been widely reported the reciprocal relation of sleep spindle density with delta activity [**50**] [**59**] [**60**]. The great intra-individual variability makes it difficult to put forward any general interpretation of the effect of ageing on sleep spindles, however several changes in spindle parameters have been reported, generally indicating a decrease in spindle amplitude and density with age, together with slight increase in frequency of the oscillations [**61**] [**62**]. In relation to

sleep disordered breathing and sleep apnea, it has been reported a decrease in sleep spindle index in patients when compared to controls [**63**] [**64**].

From a more quantitative point of view, standards for characterization of sleep define the sleep spindle as "a train of distinct waves with frequency 11-16 Hz (most commonly 12-14 Hz) with a duration $\geq$ 0.5 seconds, usually maximal in amplitude using central derivations". Specifically this definition refers to visual classification of these transient events in the PSG [**7**]. However, many times literature refers to sigma activity as the method to characterize cerebral activity within such frequency band. In this respect, there is a debate regarding if changes in powers values in a certain frequency band actually corresponds to the changes observed through visual inspection of sleep spindles. In fact, spectral analysis based on Fourier Transform, by itself, it is not capable of making a distinction between EEG background activity and phasic activity in form of transient sleep spindle event.

In any case, research regarding automatic methods for the detection of sleep spindles is a prominent area of increasing interest in the last years. That is because of the interest that detection of this kind of events causes, not only for the characterization of N2 sleep stage and the consequent construction of patient's hypnogram, but for the progress of the neurophysiology research in general.

As a consequence, several developments on automatic methods for sleep spindle detection have been reported. One of the first developments in this regard is the work of Campbell et al. [**65**] in which comparison of two phase-locked loop spindles detectors is made based on previous developments by Broughton et al. [**66**] and Kumar [**67**]. Respective 65% and 72% of true positive detections are reported. Later on, Declerck et al. [**68**] reported better performance (around 90%) of software over hardware methods. In the 90s Jobert et al. [**69**] applied matched filtering obtaining performance of 80%. Another standing out approach at that time is the one from Dijk et al. [**54**] based on power spectra analysis of the EEG. Spindle detection using artificial neural networks can be found in Huupponen et al. [**70**], Shimada et al. [**71**] or in Ventouras et al. [**72**]. In the framework of spectral analysis techniques the wavelet transform [**73**] [**74**] and the matching pursuit [**75**] have also been proposed. Olbrich and Acherman [**76**], on their part, carried out an analysis of oscillatory patterns by fitting autoregressive models.

More recently, the work of Huupponen et al. [**77**] examines four sleep spindle detection methods obtaining best performance of 70% sensitivity and 98.6% specificity during EEG analysis in sleep stage 2. In addition, Ray et al. [**78**] describes a method that performs a subject-specific adjusting of amplitude threshold, and reports 99% sensitivity and 88% specificity over stage 2 from a group of 10 healthy young subjects.

Algorithm for the detection of sleep spindles

The approximation for the detection of sleep spindles has its fundamental in the frequency analysis by using the Short-Time Fourier Transform (STFT). Main advantage on the use of this technique resides in its compromise between temporal and frequency resolution, so that it allows temporal localization of dominant frequencies in the signal. For that purpose, the signal is subdivided into temporal intervals of finite duration in which the classical Fourier transform is applied, enabling in this manner analysis of spectral frequencies for the corresponding time interval. There is, on the counter part, a loss in the frequency resolution due to the reduction of the analysis period. In any case, in practical applications it is usually possible to find an adequate compromise between duration of the temporal window and the necessary frequency resolution. In this regard, for example in the case of sleep spindles, the frequency band of interest is set around 12-14 Hz, and therefore any temporal window of duration higher than 0.1 seconds should be enough to capture the oscillation in this band.

The detection algorithm implemented in the system acts over both C4/A1 and C3/A2 EEG central derivations, which are the ones recommended by AASM [**7**]. Basically, and by means of the use of STFT, the algorithm searches for increments in the spindle frequency band which, accordingly with formal definition of these events, should present duration ≥ 0.5 s. For that purpose, in a first step the developed procedure uses a moving hamming window of 2 seconds duration which is shifted throughout the signal with an overlapping of 1.8 seconds, that is, each time step the window is shifted 0.2 seconds. Within each temporal window the Fourier Transform is then computed in order to calculate average power in the band between 12-15 Hz -same formula as in equation (5.1). Justification on both duration and time shifting of the used temporal window, is related, as it has been commented above, with an adequate comprise between temporal and frequency resolution. In this regard, a time step of 0.2 seconds

allows us to detect alterations in the spectral band over intervals with less duration than the minimum spindle event.

Once average power in the sigma band (PSDσ) has been established throughout the signal, the algorithm performs a second analysis over obtained values in order to establish a baseline (averaging of 10 previous seconds of average power) to which compare instant power values of spindle activity. The former allows characterization of intervals with increments or decrements with respect to normal sigma activity in the signal. This process is repeated for each instant power value previously calculated, after which those intervals exceeding 2 times baseline value and with duration ≥ 0.5 seconds are marked as possible spindle events.

An example of the previous process is illustrated in Figure 5.8, in which an interval of EEG signal is shown containing periods of spindle activity (see Figure 5.8, A). Evolution in time of computed sigma power (blue) and corresponding baseline (red) are shown at the bottom (see Figure 5.8, B), where peaks correspond to periods of high spindle activity. Subsequent Figure 5.9 plots a zoom over central peak in Figure 5.8. It can be shown, by counting the number of oscillations, that it effectively corresponds to a sleep spindle event.

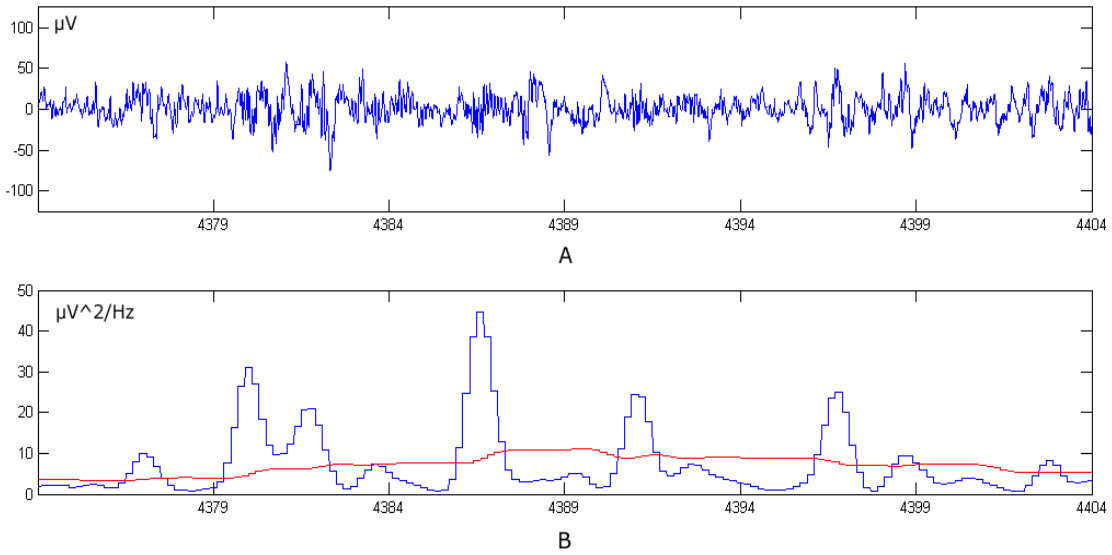


Figure 5.8. Sleep spindle detection: A) EEG channel showing periods of spindle activity; B) Spindle power (blue) and corresponding baseline (red). Peaks correspond to intervals in which there is an increase within the spindle band

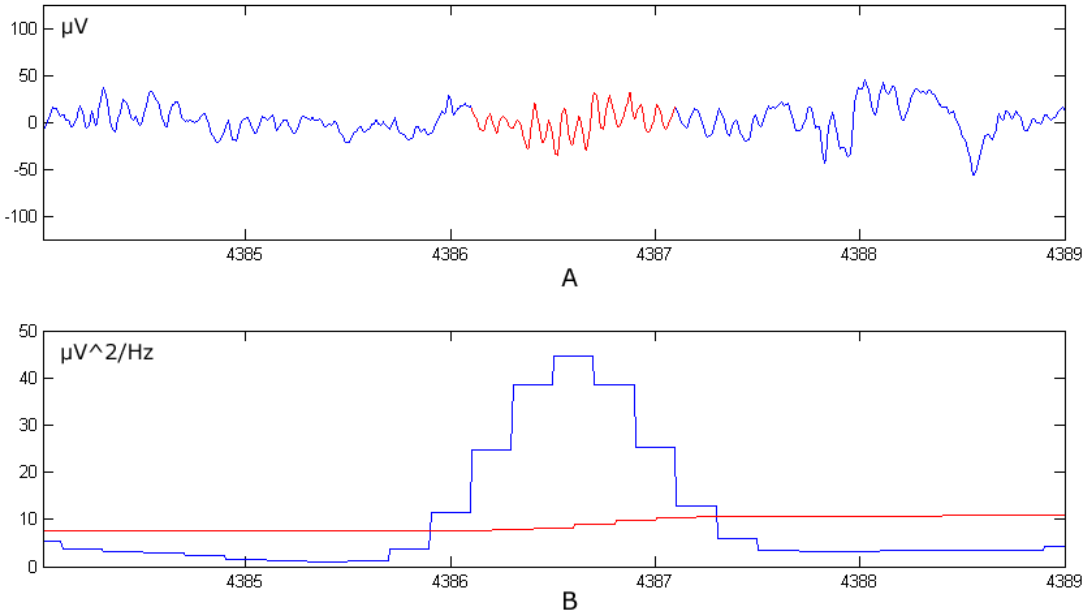


Figure 5.9. Sleep spindle detection: A) Zoom over central peak of Figure 5.8, detected sleep spindle event is marked in red; B) Sigma power (blue) and corresponding baseline (red) within the interval

However, the use of the above described procedure causes that within the set of possible events detected, there may be several false positives that it is necessary to discard. These false positives are mainly caused by (1) interferences produced by increments in the contiguous frequency bands such as alpha (8-12 Hz) and EMG artifacts (> 16 Hz) that can cause harmonics in the spindle band, and (2) events occurring during non-sleep periods such as clear stable intervals of alpha or beta activity.

Thus, in order to discard these false positives a dual strategy is followed, which tries to give response to the two previous commented situations. To solve false positives due to interference of contiguous bands, a detection of transient arousal intervals is performed by using the same approximation as for the detection of spindle intervals, that is, by means of the described STFT method, but now marking those intervals within frequencies of alphas (8-12 Hz) and betas (≥ 16 Hz). Then spindle events that have been detected in presence of these arousal intervals are finally discarded. On the other hand, detection of stable non-sleep periods is performed by computing power percentages within each band with respect to full spectra of each temporal window. That is:

$$P_{8\text{-}12} = AvgPow_{8\text{-}12} / AvgPow_{0\text{-}62.5}$$

$$P_{12\text{-}15} = AvgPow_{12\text{-}15}/AvgPow_{0\text{-}62.5}$$

$$P_{16\text{-}62.5} = AvgPow_{16\text{-}62.5}/AvgPow_{0\text{-}62.5}$$

and considering as false positives those intervals which satisfy:

$$P_{12\text{-}15} < P_{8\text{-}12} \text{ v } P_{12\text{-}15} < P_{16\text{-}62.5}$$

Examples on artifact detection are shown in Figure 5.10 and Figure 5.11. In Figure 5.10 a peak in the sigma band is detected during an arousal interval (excited EEG containing EMG artifact). In Figure 5.11 another false positive is detected, in this case, because of intrusions contiguous frequency bands. In both the two previous situations the false positive is detected and discarded.

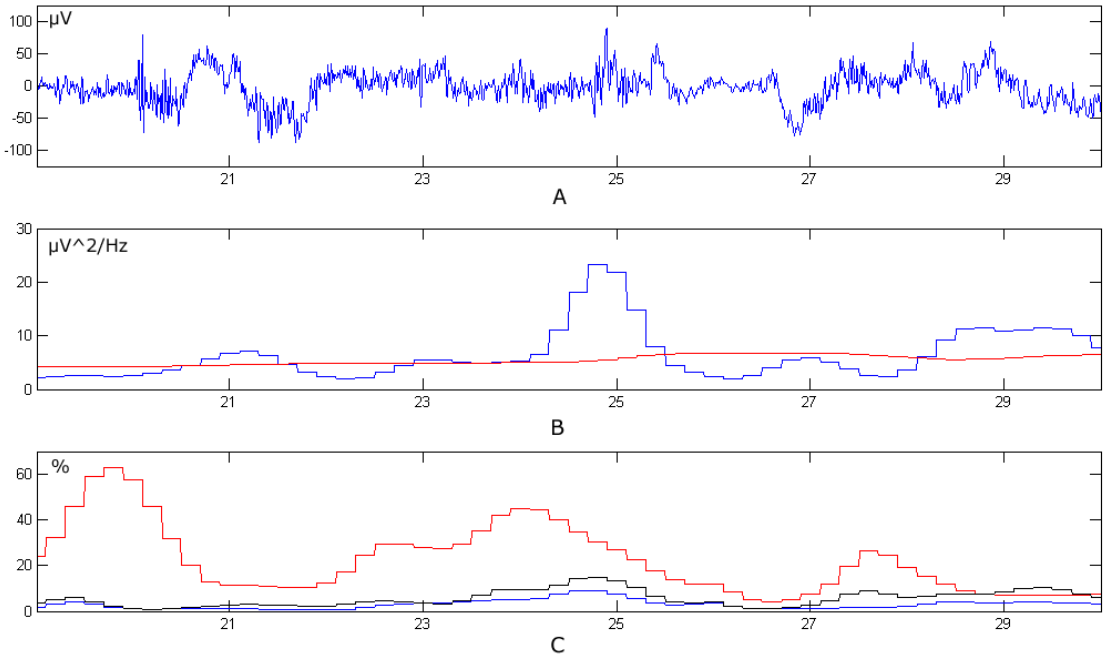


Figure 5.10. Sleep Spindle artifact detection: A) A period of EEG signal affected by continuous EMG artifact; B) Spindle power (blue) and corresponding baseline (red). Note that a peak is produced but in this case there is no spindle event present; C) Power percentages corresponding to $P_{16\text{-}62.5}$ (red), $P_{8\text{-}12}$ (black) and $P_{12\text{-}15}$ (blue). Note that EMG (16-62.5 Hz) is dominant and therefore peak in spindle frequency is considered a false positive

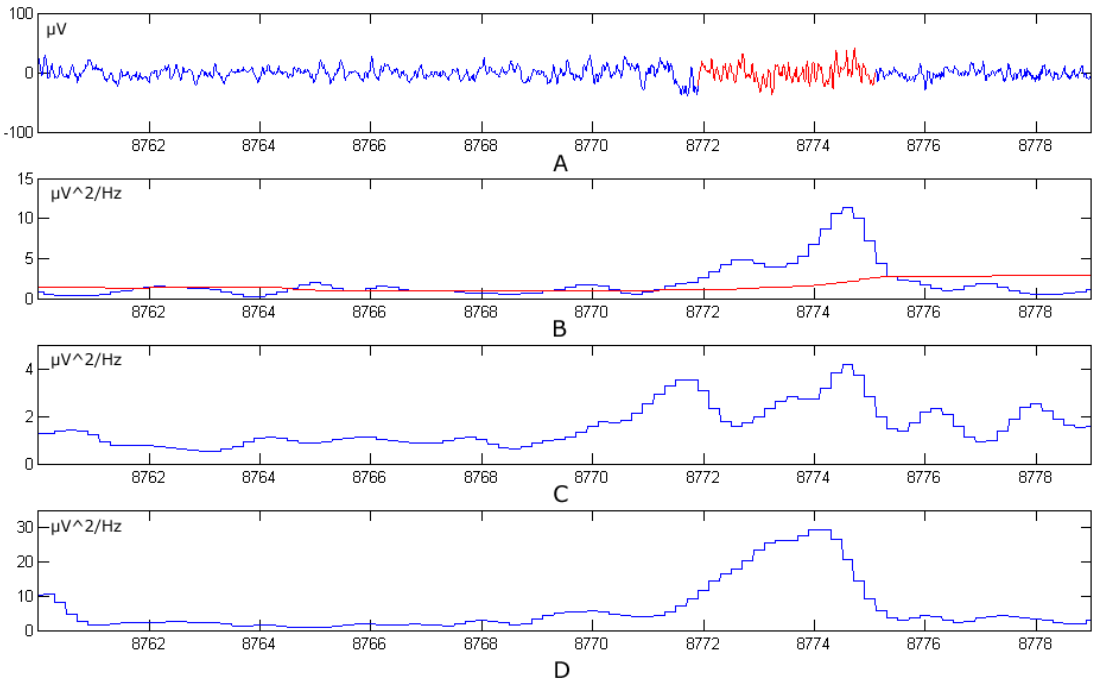


Figure 5.11. Sleep Spindle artifact detection: A) EEG signal in which a possible sleep spindle event is marked in red; B) Spindle power (blue) and corresponding baseline (red) showing corresponding increase in the sigma band; C) EMG power band showing two peaks at the same time; D) Alpha power band which also presents a peak in the time interval corresponding to peak in sigma band (B)

5.6.3. K-Complexes

With regard to the current standard sleep scoring procedures [**13**] [**7**], K-Complexes (KCs) are, together with occurrence of sleep spindles, one of the fundamental transient events significant of human stage 2. However, differently from sleep spindles which are uniquely related to the sleep process, in the case of KCs there has been a lot of discussion regarding its duality: on one hand being one regular building-stone of NREM sleep EEG, and on the other hand being a reactive element elicited by sensory stimuli. Indeed, yet from the very first time the KC was described by the Loomis group [**79**], it was described as a characteristic large potential change occurring as a result of tone stimulation during *C* state of sleep (analogous to current N2 state). Nevertheless, at the same time, Loomis also reported in the same paper that although it can be evoked by external stimulus, it can also appear spontaneously.

Accordingly to the above mentioned characteristic, there is usually a distinction between *evoked KCs*, as those related to a known sensory stimulus, and *spontaneous KCs*. It has to be noted, however, that some studies point out to the possibility that, in

fact, spontaneous KC results from response to unnoticed internal –interoceptive-stimulus [**80**] [**81**]. In any case, the paradoxical character of KC, generated debates on its function: an elementary arousal-response in NREM sleep [**80**] or a sleep-protective component [**82**].

Recent investigation from Amzica and Steriade [**83**] has also opened a third interpretation, considering KC as an emergent phenomenon, result of sleep-specific slow-voltage oscillations in the thalamocortical circuitry. Thus, being a purely passive phenomenon reflection of a particular sleep microstate. Moreover, investigations proving its close relation to the slow wave sleep [**84**] seem to no longer support the arousal hypothesis.

Independently of the controversy regarding its generating mechanism or function, it is a fact that from the visual scoring perspective, in the EEG recording the KC can appear isolated or accompanied by sleep spindles (K-spindle) –most of the times- but also preceding EEG arousals (K-alpha) or even delta waves (K-delta). There is also an abundance of studies demonstrating autonomic and muscle activity conjoining KCs [**85**].

Although with a possible common origin and similar shape [**85**], for sleep scoring purposes, it is important to differentiate KCs from vertex sharp waves (see Chapter 2, "*Structural analysis of sleep*"). Vertex sharp waves are more typically seen in N1 and usually present lower amplitude than KCs, which are more related to N2 and N3. In N3 however they are buried in the high-amplitude background EEG and it may be impossible to distinguish them from delta activity at a glance. This is probably the reason why isolated presence of KCs is a good marker for visually staging of N2.

In this respect, recalling the AASM definition, a KC is defined as a well-delineated negative sharp wave immediately followed by a positive component standing out from the background EEG, with a total duration ≥0.5 seconds, usually maximal in amplitude when recorded using frontal derivations. In practice however, as previously outlined, sometimes KCs are difficult to be scored due to its similarity with vertex sharp waves, its camouflage within delta bursts and, in general, because of alterations produced by background EEG activity.

First attempts to set out automatic detection of KCs can be found in the work of Bremer et al. [**86**] in the decade of 70s. They derived values for the maximum peak-to-peak amplitude and timing parameters, including total duration and time between zero crossings for the initial short positive wave and the longer later negative wave. They set a minimum amplitude criterion of 100 μV and a minimum interval between successive KCs of 2 seconds. Their method proved to be quite comparable to human scorers in detection KCs in stage 2, but had a tendency to overestimate KC rates in other stages. Nonetheless their work provided the foundation for later efforts to establish automatic scoring systems. Da Rosa et al. [**87**] proposed a model of sleep phasic events which consists of feedback loops that are driven by white noise (simulating tonic delta and sigma activity) and by isolated random impulses, simulating vertex waves or KCs, depending on the background tonic activity. The detector was tested on real EEG signals and was able to detect KCs and vertex waves quite reliably in spite of their variable shapes. Bankman et al. [**88**] proposed an approach based on feature selection and the use of neural networks. Respective contribution of the features and that of the neural network were demonstrated by comparing results to those obtained with raw data presented to neural networks and features presented to Fisher's linear discriminant. Other relevant works on methods for automatic KC detection can be found in Richard et at. [**89**], Jansen et al. [**90**] and Jober et al. [**69**].

Algorithm for the detection of K-complexes

The approximation here proposed is quite simple and it basically consists in combination of amplitude analysis of the EEG derivations complemented with information extracted from spectral features. In this respect possible occurrence of a KC is considered when an abrupt shift in the corresponding EEG amplitude is detected and this amplitude correlates with concurrent increment in delta frequency band (0.5–4 Hz). Once a phasic amplitude event has been marked then additional checking is performed in order to confirm or discard the event as an actual KC.

Hence, in a first step a moving window of 2 seconds is shifted throughout the EEG signal with a time step of 0.5 seconds (1.5 seconds overlapping). Within each temporal window amplitude is computed as the difference between the maximum and the minimum signal values. Once instant amplitude has been computed along the EEG by

the above windowing method, corresponding signal baseline is obtained at each time step by averaging instant amplitude values in the 30 seconds preceding interval. As a result, EEG instant amplitude and baseline representations are obtained at each time step. In a second step, similar STFT processing as used for detection of sleep spindles (see subsection "*Sleep Spindles*") is used to detect frequency shifts in delta band. Based on these two signals, possible KCs are marked by thresholding when both, instant amplitude value and average power in the delta band, are at least twice as their corresponding baselines. Additionally the resulting event has to have duration of at least 0.5 seconds, whereas maximum event duration is established to be 3 seconds (see Figure 5.12).

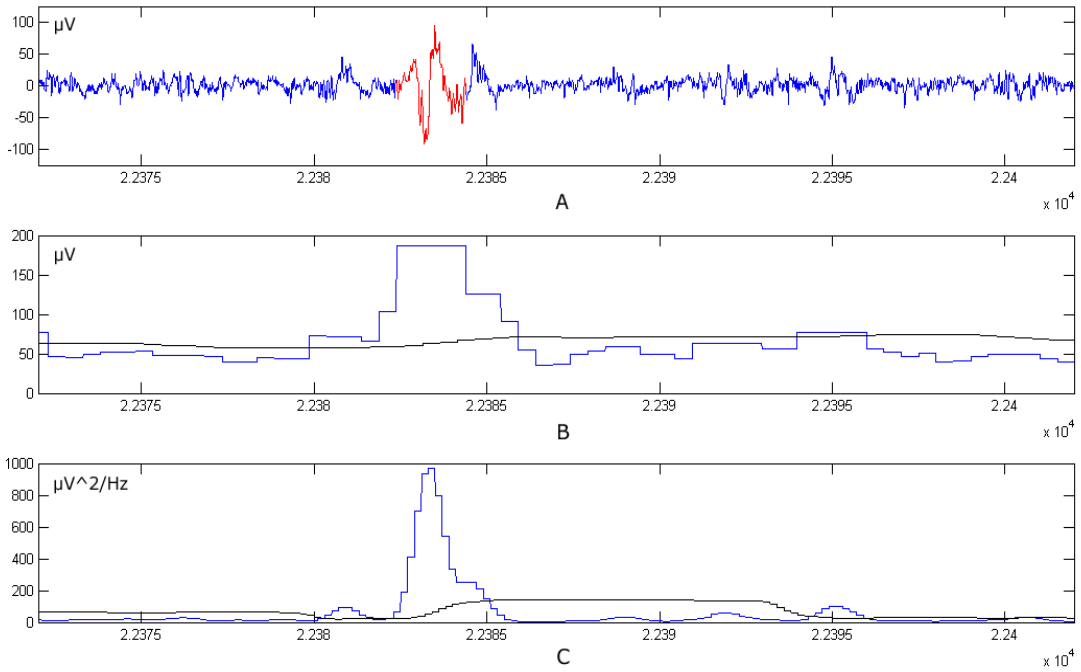


Figure 5.12. K-Complex detection: A) EEG channel showing a detected K-Complex in red; B) EEG instant amplitude (blue) and corresponding baseline (black); C) Delta power (blue) and corresponding baseline (black)

Artifact detection strategy is developed in order to discard false positives from the previous set of possible KCs. Many false positives occur in the presence of EMG artifact or arousal that can produce transient increments in EEG amplitude. These intervals obviously do not correspond with occurrence of a KC. For this purpose analogous STFT processing is used to mark transient increments of alpha (8-12 Hz) and beta (≥ 16 Hz) bands and resulting events are then used to discard false KC positives.

Abrupt shifts in submental EMG amplitude are also evidence of arousal (see subsection "*Identification of EEG arousals*"). In this regard, EMG signal amplitude is computed in the same manner as previously explained for the EEG, and resulting events are also used to discard false positives. An example of false positive discarding is illustrated in Figure 5.13 and Figure 5.14.

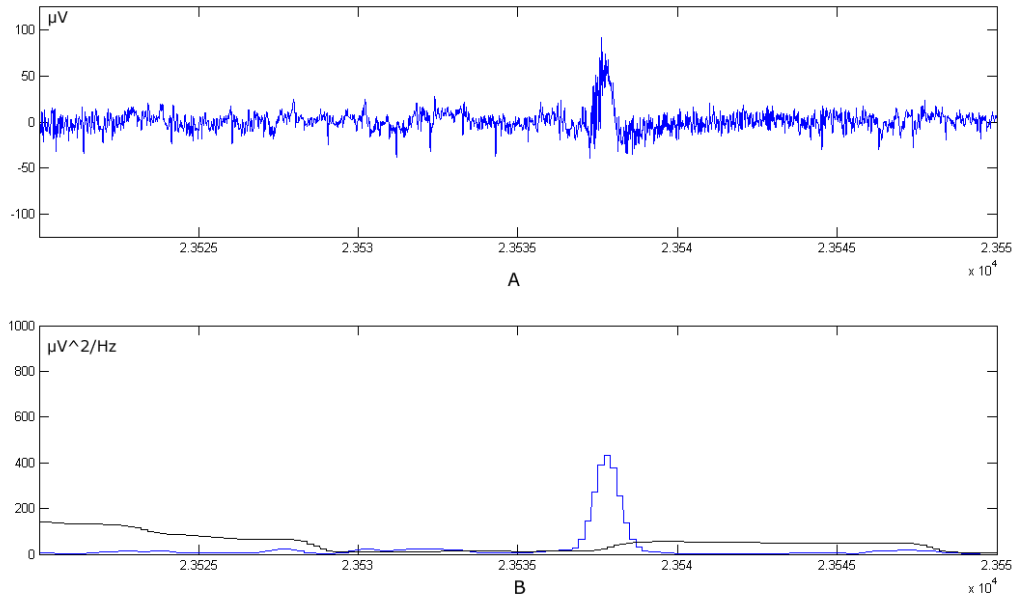


Figure 5.13. K-Complex artifact rejection: A) EEG channel showing an abrupt shift in the EEG amplitude that does not correspond with a K-Complex. B) Delta instant average power (blue) and corresponding baseline (black) showing a peak which corresponds to the transient amplitude shift marking a possible K-Complex

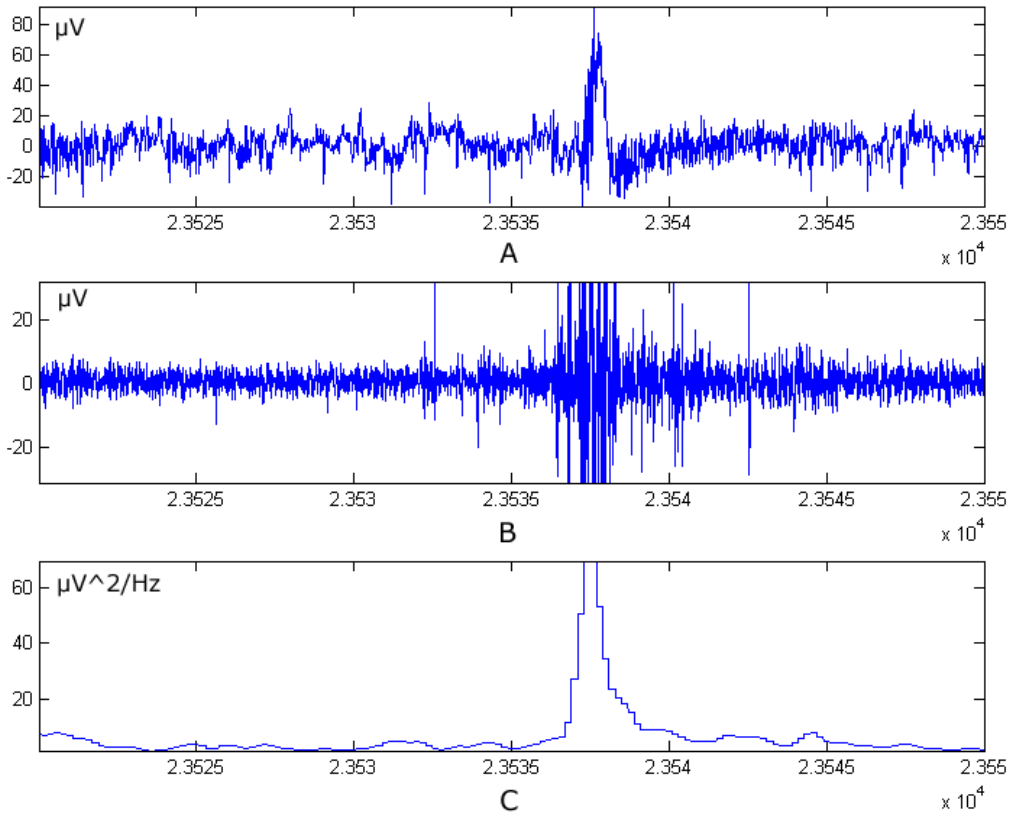


Figure 5.14. K-Complex artifact rejection: A) Same EEG interval as in Figure 5.13; B) Submental EMG channel in which EMG artifact can be shown; C) EMG power reflecting the same EMG artifact in the time-frequency domain

It is important to remark that although KCs may be followed by EEG arousal, and still being considered a KC –the so-called "K-alpha", here discarding is performed only when both the supposed KC and the arousal events occur at the same time; in other words, when the two events overlap. Hence, occurrence of the K-alpha pattern is allowed if certain separation between the KC and the alpha event exists.

5.6.4. Hypnogram generation

One of the most important tasks within the analysis of the PSG is the characterization of the patient's sleep macrostructure. This leads to the construction of the hypnogram in which the voluminous chart recordings of electrical activities recorded trough the PSG are summarized in a simple graph aimed at showing evolution of the different states of sleep throughout the night.

The procedure for obtaining the hypnogram was initially proposed in 1968 by Rechtschaffen and Kales (R&K) [**13**]. As it has been mentioned in subsection "*Structural analysis of sleep*" of Chapter 2, R&K method establishes a set of rules to assign to a time interval in the PSG a label representing certain state of sleep: Wakefulness (W), stages 1 to 4 (S1, S2, S3 and S4), and Rapid Eye Movement (REM) phase. Therefore the sleep recording is segmented into these classifiable intervals, called epochs, being its length arbitrary established to 30 seconds. Sleep states are then assigned based on the trend of the signals within the epoch, which following the R&K method involves monitoring of brain activity by EEG, characterization of muscle tone by EMG, and localization of eye movements by EOG.

R&K method has been the gold standard to the scoring of sleep macrostructure for more than 40 years, being only recently modified by the AASM (see Chapter 2, "*Structural analysis of sleep*"). As it has also been stated, sleep staging is in general a tedious task entailing too much time and effort from the physician, requiring around 5 hours per patient examination. Automatic sleep scoring should help to reduce the time needed by the physician to construct the hypnogram and, accordingly, attempts to develop automatic sleep scoring are almost as old as the R&K rules. Literature provides with several examples of approximations involving different techniques: pattern recognition [**91**] [**92**], evidential theory [**93**], probabilistic models [**94**], stochastic modeling of physiological feedback structures [**95**], artificial neural networks [**96**] [**97**] or wavelets [**98**]. A detailed review of the literature can also be found in [**99**], in which it is also stated that, besides the number of efforts in this field, automation of hypnogram generation is still an open area of research interest.

On the other hand, in the recent years several criticisms have been associated with such a method of sleep characterization [**100**] [**101**] [**102**]. Major drawbacks are associated with its low temporal resolution –one label for 30 seconds- and the unnatural classification of sleep based on fixed-duration discrete epochs. Effectively, evolution of biological processes rather occurs in a continuous manner in which a soft transition takes place between the different considered states. Moreover, restriction of the analysis to a small number of sleep states and the use of fairly large epochs lengths, seems to obey more to practical criteria to avoid the manual scoring of an entire night's sleep to be a prohibitively time consuming task. However at the cost of an increase in the intra-

state variability and information loss, which nowadays is summarized under the broad level of microstructure of sleep, including microarousals, sleep spindles, k-complexes and any other activity with latency shorter that the half-minute of the epoch-based staging methodology [**103**] [**104**].

On this context, there is an interest on exploring different approaches that could overcome such limitations. Specifically, it is interesting to develop computer approximations on the goal of achieving a more continuous characterization of sleep. Thereby avoiding the limitations of sleep staging previously outlined. For example, a guideline to this effect has been proposed as part of an European Community concerted action toward a methodology for the analysis of the sleep/wakefulness continuum [**105**].

In the line through the achievement of a continuous marker describing the sleep structure some approximations can be found in the literature. In this respect a first interesting step is to obtain a continuous sleep depth estimator. Some examples can be found in the contributions of Asyali et al. [**26**], Choi et al. [**106**], Swarnkar et al. [**107**], or Saastamoinen et al. [**108**]. However, the previous approaches suffer from the lack of information on the intra-sleep periods. A realization that could also account for a continuous characterization of NREM intra-states is desirable. Flexer et al. [**109**] developed a continuous probabilistic sleep stager (considering three states: wakefulness, deep sleep and REM) based on a single EEG signal. Nevertheless a problem with probabilistic approximations is that they assign one minus the probability of an event to the complement of the event, i.e. there is no possibility of differentiating between uncertainty about an event and the probability of its complement.

In the system object of this doctoral thesis an alternative approximation is proposed to the problem of the continuous sleep staging by using the fuzzy logic paradigm. As it has been stated throughout Chapter 4, fuzzy logic allows us to quantify a decision in terms of a fuzzy degree of membership which avoids binary decisions based on categorical classifications. It also allows us to deal with uncertainty and imprecision, common aspects of medical diagnostic domains (see Chapter 3, "*Critical analysis*"). Moreover, given the membership $\mu_H(x)$ of a certain hypothesis *H*, the membership of the complementary hypothesis ($\neg H$) should not necessarily be 1- $\mu_H(x)$, and therefore the problems of the probabilistic approaches can be overcome. In fact, the use of fuzzy

logic for the classification of sleep stages is not new [**110**] [**111**] [**112**] [**113**]. However none of these realizations presented a solution in continuum, but they limit to classify the discrete epoch in fuzzy terms.

In addition, the fuzzy paradigm presents an important property to allow the continuous representation of biological processes: soft transitions between class memberships, i.e. fuzzy classifiers yield similar outputs for similar input patterns. A property that it does exploit the work of Heiss et al. [**114**], which however contains some drawbacks with respect to our approach, mainly: (1) it is focused for its use on infants and, (2) it uses the architecture ANFIS [**115**] in order to implement the fuzzy classifier and to represent the knowledge of the domain. Indeed, to fulfill the requirement of a system being able to explain its results, there is an additional property a system for medical decision support should satisfy: the system should not behave as a black-box, i.e. it should be possible to check why and how a certain recommendation is given [**116**]. In this respect, the ANFIS structure can be considered more as a special kind of neural network since the use of Sugeno-like rules considerably reduces its explanation capabilities. In the proposed approach a Mamdani FIS is used instead, which offers better understandability because of its knowledge representation schema (see Chapter 4, "*Fuzzy inference systems*").

Thus, the objective with the proposed approach is dual: (i) first a method for the automatic classification of sleep macrostructure and generation of the hypnogram is proposed, and (ii) at the same time overcoming the limitations of the epoch-based staging methods is attempted by using fuzzy logic to allow smooth transitions between the different sleep states in continuum.

Specifically the proposed method works with four different sleep stages, directly related with R&K and AASM states: wakefulness (W), drowsy sleep (DS), deep sleep (DEEP) and REM sleep. Note in the current approach DS includes classical S1 and S2 sleep stages.

The general approach can be organized in three main sequentially related processing steps. The first step is in charge of performing *parameters extraction* in which features over the biological signals in the PSG are obtained. Then, after relevant parameters have been extracted at the first phase, information is fed into the second

processing step where a *reasoning process* occurs obtaining as output, a degree of membership with respect to each considered state of sleep, as stated before: W, DS, DEEP and REM. In order to achieve such an output, the reasoning module is organized into several sub-modules, each one regarding to a different sleep state, therefore obtaining a respective degree of membership, i.e. μ_W, $\mu_{DS,}$ μ_{DEEP} and μ_{REM}. Note that microstructure events (K-complexes and sleep spindles) detected by the procedures described in the preceding subsections are also used as input at the reasoning stage.

The whole previous process is accomplished in a second-by-second granularity, thus with higher resolution in comparison to epoch-based procedures. This, together with the properties derived from the use of a fuzzy reasoning paradigm, allows us obtaining a new representation of the hypnogram in which current evolution of the different sleep states is individually characterized.

Eventually (third step) this representation is used in order to go back over the classical hypnogram representation, thus showing how the continuous representation preserves the information contained in the discrete hypnogram. To do so, some post-processings are applied over the continuous hypnogram.

The method and each one of its processing steps are subsequently described into more detail.

A method for continuous characterization of the sleep macrostructure

1) Step 1: Parameter extraction

As previously mentioned, according to input information and standard recommendations for sleep analysis, the method proposed works with five signals: both EEG C3/A2 and C4/A1 central derivations, submental EMG and EOG signals from left (EOG_L) and right (EOG_R) eye electrodes. Depending on the type of the signal different parameters are extracted during the analysis.

1.1) Detection of the eye movements

In the case of the EOG the interest is to characterize the eye movements. In order to achieve such a task, an overlapping moving window of 3 seconds is shifted second-by-second throughout both EOG_L and EOG_R channels, and computing the amplitude of the corresponding signal interval within the window. Amplitude is calculated as the difference between the maximum and the minimum values of the signal inside the window. Thereby a value for the amplitude of the signal is obtained for the current sample. By repeating this process throughout the recording two amplitude signals –one for each derivation- are obtained. Finally a new signal A_{EOG} (see Figure 5.15) is constructed by averaging the two previous amplitude signals obtained for each channel (left and right), thus obtaining a single parameter to represent the EOG amplitude independently of the channel. It can be shown in Figure 5.15 that the amplitude of A_{EOG} signal increases in presence of EOG movements while it is almost flat for a relaxed EOG.

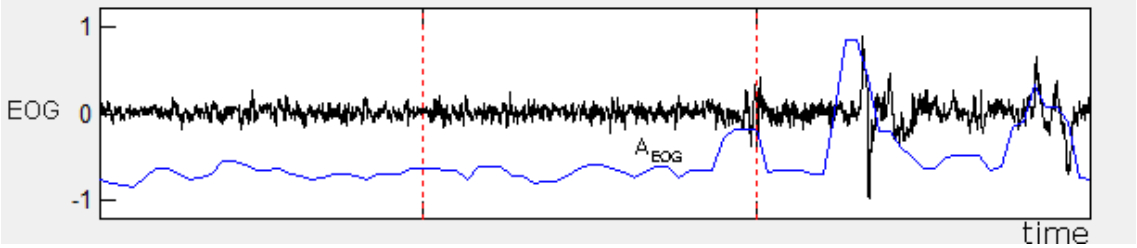


Figure 5.15. In the figure 3-epochs of 30 seconds are shown. Signal amplitudes are normalized in [-1,1]. A_{EOG} signal is superimposed in blue

1.2) Characterization of muscle tone

In the case of EMG, to distinguish between presence and absence of muscle tone, a similar amplitude-based analysis is performed. Using a window of 3 seconds and moving it second-by-second throughout EMG, a new A_{EMG} signal is obtained. Differently from the amplitude computation in the EOG, this time each *i*-th sample of A_{EMG} is calculated as the mean of the absolute value of the EMG samples included in the window (see Figure 5.16). The main reason to compute the amplitude in this manner is that it better supports the higher frequency nature of the EMG signal.

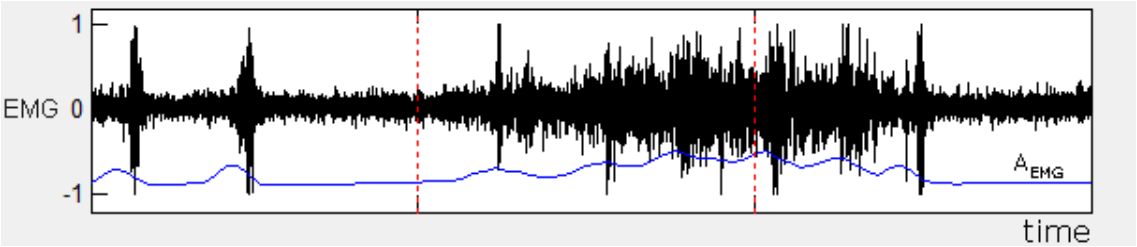


Figure 5.16. In the figure 3-epochs of 30 seconds are shown. Signal amplitudes are normalized in [-1,1]. A_{EMG} signal is superimposed in blue

1.3) Processing of electroencephalographic activity

Regarding the EEG, the different sleep stages are characterized by the different proportion of characteristic frequencies in the most representative bands: alpha (α, 8-12 Hz), beta (β, 13-30 Hz), theta (θ, 4-7 Hz) and delta (δ, 0.5-3 Hz) (see Chapter 2, "*Structural analysis of sleep*"). Short-Time Fourier Transform (STFT) method (3-second window, 2-second overlapping) is used to compute spectra on every analysis window of the EEG. Then, within each window PSD is estimated for each band using a band-pass filter, and integrating the corresponding squared spectrum on the filtered window. Four measures of PSD are obtained for each window through the previous process: PSD_α, PSD_β, PSD_θ and PSD_δ. Let PSD_T to be $PSD_\alpha+PSD_\beta+PSD_\theta+PSD_\delta$, then the relative power proportion (rPSD) for each band x is calculated as $rPSDx=PSDx/PSD_T$ where $x = \{\alpha, \beta, \theta, \delta\}$.

Finally, and similar to the case of EOG, since two derivations of EEG –C3/A2, C4/A1- are available, final values for the parameters are calculated by averaging the respective rPSD values over the two derivations. Figure 5.17 shows the resulting four parameter signals ($rPSD_\alpha(i)$, $rPSD_\beta(i)$, $rPSD_\theta(i)$ and $rPSD_\delta(i)$) characterizing the activity of the EEG.

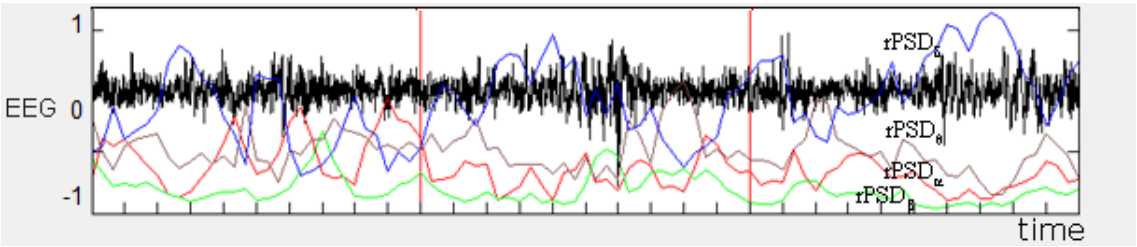


Figure 5.17. In the figure 3-epochs of 30 seconds are shown. Signal amplitudes are normalized in [-1,1]. $rPSD_\alpha$, $rPSD_\beta$, $rPSD_\theta$ and $rPSD_\delta$ signals are superimposed, respectively, with red, green, brown and blue colors

Note that in addition to the previous parameters extracted through the above described analysis of the EOG, EMG and EEG signals, microstructure events (K-complexes and sleep spindles) are also considered as sleep descriptors to be used in the subsequent fuzzy reasoning stage. Algorithms for the detection of these transient events have been described already in the preceding subsections of this chapter (see Chapter 5, "*Sleep Spindles*" and "*K-Complexes*").

Summarizing, as the output for the parameter extraction phase it is obtained, for each *i*-th second of recording, a set of six values characterizing eye movements ($A_{EOG}(i)$), muscle activity ($A_{EMG}(i)$) and EEG activity ($rPSD_{\alpha}(i)$, $rPSD_{\beta}(i)$, $rPSD_{\theta}(i)$ and $rPSD_{\delta}(i)$). Quantification of the presence of sleep spindles and K-complexes complete the set of extracted sleep descriptors (thus, eight in total). This information is then fed into the subsequent fuzzy reasoning modules to obtain the respective fuzzy memberships for each state (W, DS, DEEP and REM).

2) Step 2: Fuzzy Reasoning Process

The second step of the analysis is divided into four submodules, each one being the responsible to accomplish the analysis regarding to one of the four considered sleep stages (wakefulness, light sleep, drowsy sleep and REM).

Each submodule is implemented in the form of a Fuzzy Inference System (FIS) of type Mamdani (see Chapter 4 "*Fuzzy Inference Systems*"). This allows us to fulfill the requirement that knowledge can be accessible and extracted in form of human-like decision rules (*fuzzy rules*). Actually, overall knowledge –extracted from medical expertise- is structured into four independent sets of rules (or knowledge bases), each one involving a particular sleep stage. Thus, the output of each submodule consists of a value μ in the real interval [0, 1], which represents the degree of membership for the current instant of time under analysis, i.e. μ_W, μ_{DS}, μ_{DEEP} and μ_{REM}.

Input vector to these submodules is composed of the parameter information extracted on the previous analysis steps regarding EOG, EMG and EEG activity. In

order to mimic humans' procedure to capture dynamics of the signals and promote smooth transitions, parameters of the signals are averaged in the environment of the current sample under analysis. Specifically, input for instant *i* is calculated by averaging input parameters on the interval [*i*-15, *i*+15]. Trapezoidal fuzzy sets are used for the partition of the input variables. For the A_{EOG} signal 3 fuzzy sets (*low*, *medium* and *high*) are used. Similarly 3 fuzzy sets (*relaxed, medium* and *tense*) are established for the parameter $A_{EMG.}$ In the case of the EEG, each of the corresponding frequency bands (α, β, θ, δ) resulted in a variable partitioned again in 3 sets namely *low*, *medium* and *high*. Output variables were partitioned by defining 5 fuzzy sets uniformly distributed along the interval [0, 1] (*very low*, *low*, *medium, high*, and *very high*). All the fuzzy sets are partially superimposed in order to explode smoothing transitions and improve generalization capabilities of the FIS. Examples of implemented fuzzy rules are:

IF: (1) A_{EMG} is *tense*
AND: (2) A_{EOG} is *high*
THEN: (1) W is *very high*

IF: (1) A_{EMG} is *tense*
OR: (2) A_{EOG} is *high*
OR: (2) $rPSD_{\alpha}$ is *high*
OR: (3) $rPSD_{\beta}$ is *high*
OR: (4) $rPSD_{\delta}$ is NOT *high*
THEN: (1) DEEP is *low*

Optimal parameterization of both the fuzzy sets and the rulebase is performed after initial knowledge structuration by automatic learning mechanisms as described in subsection "*Structure identification and parameter optimization*" of Chapter 4. For the configuration of the FIS the *minimum* was chosen as the T-norm operator for the conjunction and for the implication. On the other hand, the *maximum* was chosen as the S-norm operator for the disjunction and for the aggregation. Defuzzification is performed by using the center-of-gravity method.

Once all the seconds of the recording have been analyzed, a continuous evolution of the degrees of membership for the different sleep stages is obtained. This output can be observed in Figure 5.18, in which evolutions are represented for a full PSG recording. Note that this representation provides more information than the discrete hypnogram since (i) it is provided in a higher sample rate -for each second- compared to 30s epochs

of R&K/AASM, (ii) the natural continuous evolution of biological processes is maintained and (iii) the information regarding each sleep state is also kept individually available.

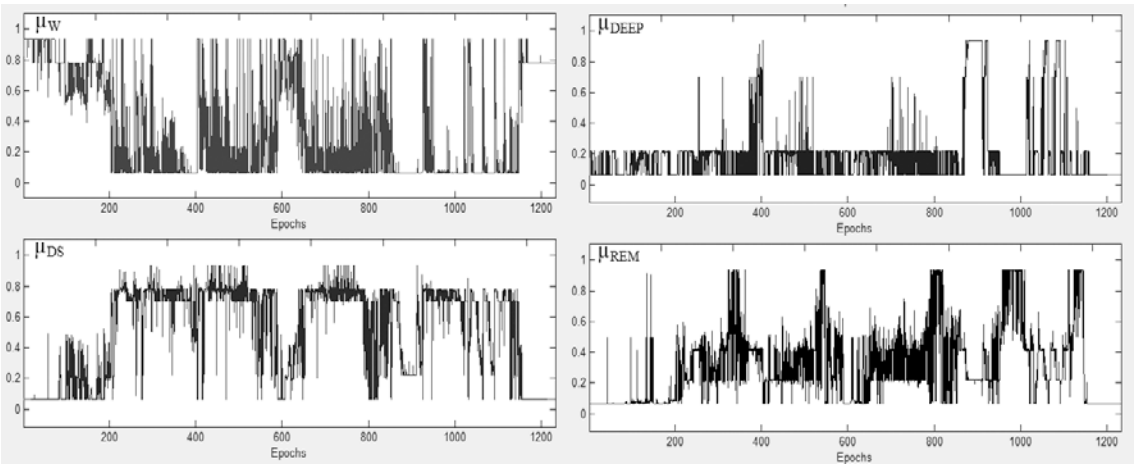


Figure 5.18. Evolution of sleep states throughout a full PSG recording estimated by their respective degrees of membership

3) Step 3: Hypnogram generation

The proposed representation based on the evolution of the corresponding degrees of membership is used here to generate the classical hypnogram. The interest in going back to the epoch-based hypnogram from the continuous representation is diverse: (1) it can be a way to show how this new proposed representation preserves all the hypnogram information, in fact being a superset of the information contained in the epoch-based hypnogram, (2) it shows how this fuzzy representation can be used as an alternative method within available literature regarding the problem of the automatic hypnogram generation in sleep. On the other hand (3), the only way to asses on the validity of (1) and (2) is to perform a validation process against experts' manually generated hypnograms, only possible though discretization of the continuous representation and going back to the epoch-based hypnogram.

Thus taking the continuous representation some post-processings are performed:

1) An average of the second-by-second output of each subsystem within each epoch is performed to be used as the resulting degree of membership for the corresponding epoch. The epoch is finally assigned to a discrete stage (W, DS,

DEEP, REM) by taking the corresponding maximum averaged degree of membership.

2) Previous processing could lead to noisy isolated epochs that break up the normal evolution of sleep. Therefore a second step searches for unusual phase transitions such as direct transitions from W to DEEP or DEEP to REM, and then assigns the most possible one according to the normal sleep evolution [**117**] [**94**] [**118**].

3) In cases where the degree of membership regarding DS and REM is similar, final assignation is performed taking into account the presence of sleep spindles and/or K-complexes within the corresponding epoch.

4) Additionally, in regions where similar degrees of membership to various states are achieved, final labeling is decided based on the trend marked by the immediately previous and subsequent epochs.

Figure 5.19 shows the result of applying such post-processing to the outputs presented at Figure 5.18. Resulting hypnogram can now be used to perform a validation process against human expert's classification. Design of validation tests and results are carried our respectively in the corresponding sections of Chapter 6 and Chapter 7.

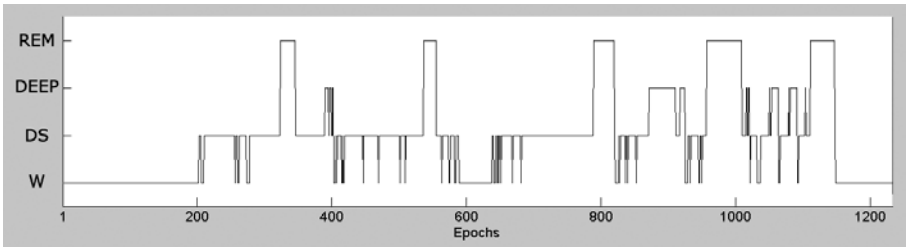


Figure 5.19. Discrete hypnogram obtained after post-processings

Next, Figure 5.20 provides of a better comparison between epoch-based and continuous hypnogram representations from the resulting method. In the figure eight epochs are shown comprising the subsequent channels (from bottom to top): (1) right EOG derivation; (2) left EOG derivation; (3) submental EMG; (4) C3/A2 EEG; (5) C4/A1 EEG; (6) full epoch-based obtained hypnogram, blue mark showing

corresponding time interval that is being displayed; (7) continuous evolution of the individual fuzzy degrees of membership. Epochs are delimited through red vertical lines. Linguistic labels are used on the top of the figure showing the resulting classification of the epochs according to discrete hypnogram. On the other hand it can also be shown how fuzzy degrees of membership evolve within the epoch. Color code is as follows: blue for W, green for DS, yellow for DEEP and red for REM.

It can be shown by taking a look to the signals, how the continuous representation provides of more information on the actual evolution of the signals: overall belief on each state can be thus quantified and intra-state evolution of the sleep states can be evaluated. It can be appreciated for example how although the fifth epoch has been scored as DEEP, a rise tend of μ_{DS} starts almost from the beginning of the epoch while μ_{DEEP} commences to decrease. It can also be shown how although the third epoch has been scored as DS, confidence with respect to a DEEP classification achieves similar levels. This situation may represent the typical case where subjective interpretation plays a role in the final classification. Isolated epoch of DS (sixth) within a cycle of DEEP might seem strange at a first sight; however it can be shown from the figure how slow waves decrease while being a concentration of sleep spindles (shown in pink over EEG channels). Within the seventh epoch slow EEG activity recovers and therefore μ_{DEEP} rises again while μ_{DS} starts to decrease again.

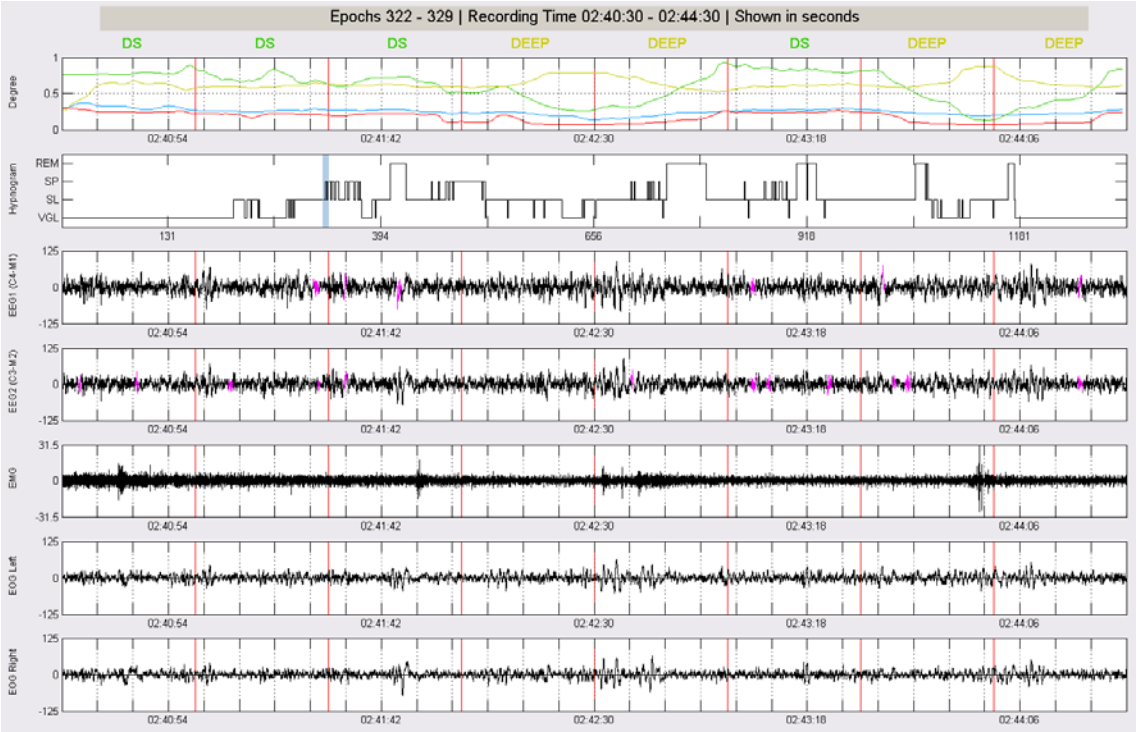


Figure 5.20. PSG interval showing both epoch-based classification scoring and continuous evolution of the individual degrees of membership

5.7. Analysis of respiratory signals

Analysis of the respiratory signals comprises in one hand analysis of airflow, abdominal and thoracic excursions, for the detection and quantification of respiratory pauses. These pauses are characterized by intervals of amplitude reduction with respect to the normal respiration. On the other hand, analysis of SaO_2 signal is performed in order to detect and quantify desaturation and resaturation intervals, which are indicative of the presence of apneic events. Main objective is localization of individual apneic evidences across the respiratory signals to be afterwards correlated in time forming diagnostic patterns in subsequent stages.

In the following the corresponding developed algorithms for the processing of the respiratory signals are described. However, previously to their analysis, a preprocessing stage is accomplished with the objective of getting rid of possible artifacts that may be hampering correct characterization of the respiratory activity.

5.7.1. Preprocessing of respiratory signals

At this stage a first preprocessing analysis is carried out to search for artifacts in the respiratory signals. The objective is the correct setting-up of the signals checking for spurious values which may obscure posterior analyses. More specifically, here two kinds of artifacts are intended to be detected: those caused by bad calibration of the sensor (*overflow artifacts*) and those caused by signal interruption (*loss of focus*) which may be due to a number of circumstances –for example a displacement of the sensor.

Normally *overflow artifacts* derive in a signal that in its digital representation appears saturated within its dynamic range. An example can be seen in Figure 5.21 which shows a saturated interval in the thoracic respiration. Overflow artifacts often occur in respiratory airflow or in the respiratory movement signals (either abdominal or thoracic).

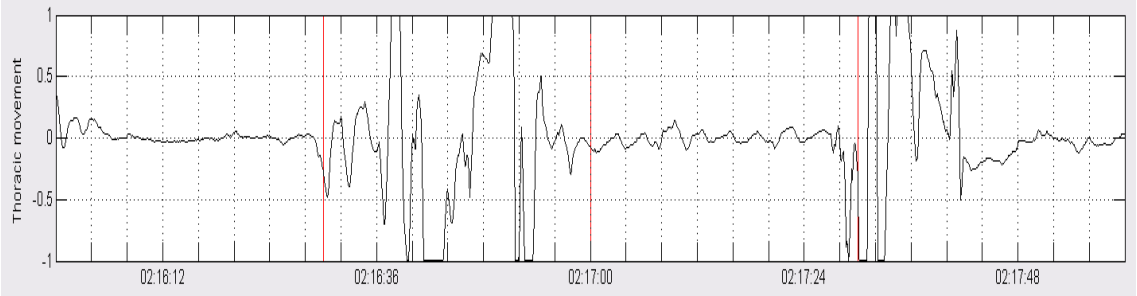


Figure 5.21. Example of an overflow occurring in the thoracic respiratory signal. Physical signal exceeds the digital representation range probably because of a bad calibration of the sensor

On the other hand, when a *loss of focus* occurs, trend of the signal suddenly interrupts. A clear example can be seen in Figure 5.22 regarding the saturation signal. Although according to the digital representation a signal overflow may also be considered –in fact although not shown in the figure, the signal actually achieves zero value- it is easy to visually differentiate it from an overflow artifact since an abrupt shift can be perceived in the natural evolution of the signal.

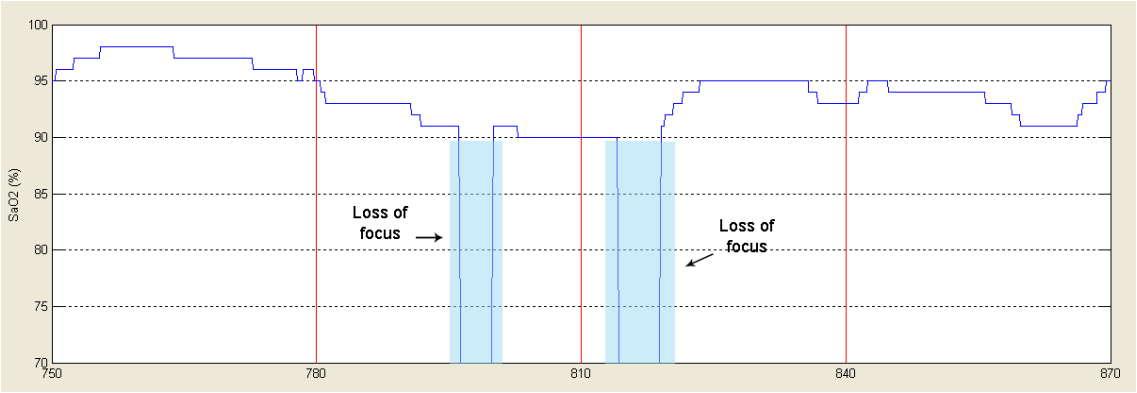


Figure 5.22. Artifacts present in the SaO_2 signal, probably because of a loss of focus in the sensor

In any case, detection of both situations should be performed previously to the application of further analysis algorithms. In this respect, the followed approach is dual:

- On one hand, independently of the concrete artifact type, artifacts detection is performed by marking their starting and ending points. Thus a list of artifact intervals found in each signal is obtained. Additionally a classification of artifact's severity is performed according to its duration in the context of the corresponding signal.

- On the other hand, where possible, a reconstruction of the signal process is carried out. The main reason to attempt signal reconstruction is the minimization of posterior filtering application by subsequent analysis procedures. This is especially important in the case of a loss of focus since application of filter algorithms can introduce fictitious frequencies in the range of the affected signal.

For the detection of overflow artifacts the system checks if the signal under consideration reaches the limits of its dynamic range. To achieve this signal period is calculated and if the overflow interval exceeds period of the signal then the artifact is marked. In this respect, since signal period may vary along the recording, calculation is recomputed every 2 minutes. For that purpose, each 2 minutes signal period is approximated by taking the inverse of the maximal frequency component in the Fourier Transform of the interval.

Detection of loss of focus in the saturation signal is performed by differentiation, marking as artifacts those signal intervals in which its derivative exceeds certain

threshold value. Such threshold value represents the maximum desaturation/resaturation speed which is considered biologically possible. A maximum rate of 40% per second is considered in the algorithm. Values surpassing such a value are considered abrupt shifts in the signal caused by non natural biological transitions and are thus marked as artifacts.

Once an artifact has been detected, classification is based on two values: Δ_{min} y $\Delta_{critical}$. The former represents the minimal time an interval has to contain spurious values in order to be considered a true artifact. Use of Δ_{min} is justified in the case of overflow artifacts since it can be the case of a signal where current value achieves the limits of its dynamic range without being necessary an artifact. In this respect, as previously mentioned, only overflow intervals with duration higher than the signal period are considered as artifacts. On the other hand, artifact periods of time duration higher than Δ_{min} are classified into two types: *Weak Artifacts (WA)* and *Heavy Artifacts (HA)*. From Δ_{min} and while interval duration is less than $\Delta_{critical}$ the artifacts is classified as WA. Durations higher of $\Delta_{critical}$ are marked as HA. Figure 5.23 shows an example of such a classification in the saturation signal.

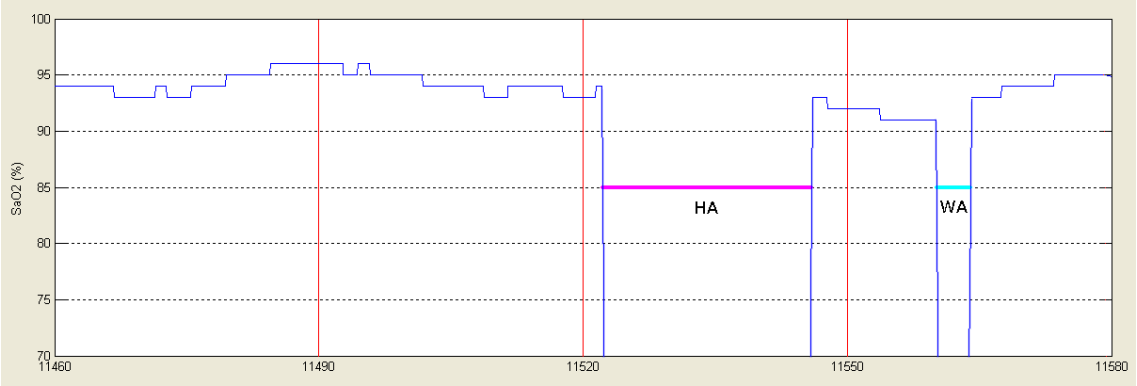


Figure 5.23. In the figure two different artifacts have been detected and classified according to their duration; HA = Heavy Artifact; WA = Weak Artifact

The use of two different labels in order to classify the artifacts is justified in terms of the posterior treatment of this information by subsequent analysis stages. For example, at the time of explaining the results of the system, the presence of a WA is considered as evidence that reasoning processes carried out over facts might be affected by the presence of an artifact, thus analyzed data may not completely be due to physiological events. In this manner, the user is warned over this fact in the

corresponding explanation of the event in the final report. Occurrence of an HA, on the other hand, is considered serious enough it may invalidate reasoning analysis results. If an apneic event is detected in the presence of an artifact of this type, a significant loss of information is considered as a consequence of the artifact, and the detected event is eventually discarded.

Signal reconstruction is independently performed in the case in which a loss of focus is localized in the SaO_2 signal. This unconditional reconstruction is accomplished in order to minimize collateral effects on posterior analysis phases. Application of digital filters, for example, may introduce artificial values as a consequence of the abrupt shifts in the natural signal evolution. In these cases linear interpolation between immediately adjacent samples free of artifact is performed. Resulting signal is assumed to follow regular signal evolution as if the artifact was not present. Figure 5.24 shows resulting signal after reconstruction from the artifacts shown in Figure 5.23. Signal trend previous to reconstruction as well as information about classification of the artifact is maintained.

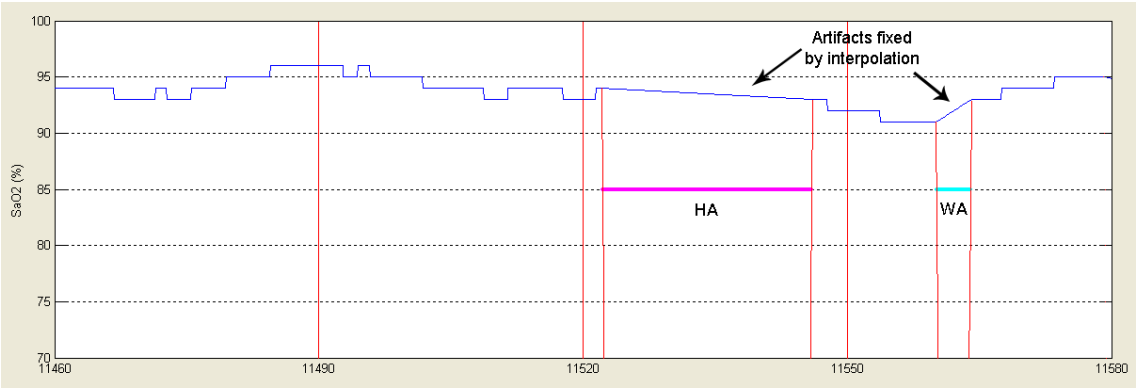


Figure 5.24. Same interval of SaO_2 signal than in Figure 5.23. Signals has been reconstructed by interpolation of extreme values

Summarizing, as the output of the preprocessing of respiratory signals, a list of artifacts detected on each signal is obtained. These artifacts may be produced by overflow in signal's dynamical range or by abrupt shifts caused by a loss of focus in the sensors. Detected artifacts are localized in time and are classified as *weak* or *heavy* according to its relevance for the posterior analyses. Additionally, in the case of loss of focus in the SaO_2 signal, reconstruction of the signal is performed by interpolating previous and posterior signal samples to the artifact interval.

5.7.2. Identification of apneic intervals

Detection of apneic intervals involves analysis of the airflow signal and the movement signals from the thoracic and abdominal excursions. The kind of processing applied to these signals is the same because of their similarities: all are sinusoidal signals of which period represents the respiratory rhythm; additionally because of inputs configuration (see Table 5.1 in subsection "*Description of the inputs*"), the amplitude is always normalized to [-1, 1]. Fundamental difference resides in the derivation used to record the respiratory rhythm, as it has been explained in Chapter 2, and therefore each derivation may offer different information for the interpretation of the apneic event[38].

The signal that constitutes the focus of reference for the localization of apneic events within the classical clinical approach is the airflow [**7**]. However signals of respiratory movements also produce a measure of the amount of respiration since each inspiratory/expiratory cycle causes an increment/decrement in the thoracoabdominal section. In this respect, even though apnea definition requires the detection of a significant reduction localized over the airflow signal, it may be the case that an error in the sensor overlooks this event, however being this reflected both in the saturation and respiratory movement signals. Furthermore, some experts consider that a significant reduction on the amplitude of respiratory movements may be sufficient for the detection of a hypopnea [**119**]. In general, several combinations of situations are possible, and it is the hypothesis of this doctoral thesis that taking all into account will ultimately cause the better characterization of the apneic event. It is for this reason that the system searches for apneic evidences over the three previous mentioned signals: airflow, abdominal and thoracic movements. Relative importance of evidence detected on each signal –from now on, referred as *apneic intervals*- contributing to the overall belief about the existence of a concrete type of apneic event, is later refined at the reasoning level (see subsection "*Detection of apneic events*"). In any case, at this signal processing level, algorithms developed for the detection of apneic intervals –amplitude reductions- perform in the same manner over these three signals.

[38] For example, in the obese patient the apneic event is preferably localized using the thoracic than with the abdominal derivation, since the excess of abdominal mass may obscure the recording of a feasible signal

Taking into account classical definitions regarding apnea and hypopnea events, their minimal duration is established to 10 seconds [**120**]. However it may be interesting to detect significant amplitude reductions of shorter duration. In this manner the detection criterion is relaxed, introducing a first source of imprecision in the system. *It is, in fact, an event of duration 10.1 seconds valid while other with 9.9 seconds is not?* In other words, given a system aimed at emulating intelligent behavior, it seems fair to consider the following question: *Is a human being able, at a glance, to quantify such a difference? What about when he/she has been examining the PSG for three hours?* In this regard, the tests carried out by Otero et al. on visual estimation of 50% of a line in 257 participants show that error committed is 10.9 ± 4.35 (mean ± standard deviation). In the same study when estimating the 10% proportion the achieved error was 4.09 ± 2.25 [**121**].

Taking into account the above mentioned, detection algorithm searching for apneic intervals in the three considered signals works as follows:

1. Processing of the three signals is carried out through shifting a temporal moving window of length Δ_{event}[39] seconds throughout them, so that each sample *i* is associated to the temporal window that starts in the *i*-th sample. Maximum signal amplitude within the window $\Omega_{event}(i)$ is then calculated as the difference between the maximum and minimum signals values contained in the corresponding window. Thus, value Δ_{event} defines minimal duration of the possible amplitude reduction to be considered an apneic interval.

2. The computed Ω_{event} amplitude is compared with respect to the baseline value, calculated over the immediately previous signal interval of duration $\Delta_{baseline}$ seconds. Duration of this comparison interval, usually established to 2 minutes, is used to determine the amplitude value of *normal respiration* over each respiratory signal. It is important to remark that within the baseline interval only periods of stable respiratory amplitude are considered, i.e. previous periods marked as possible apneic intervals do not count to

[39] Configurable length can be established

compute the amplitude value for the baseline breathing. Resulting amplitude reference $\Omega_{baseline}$ is then calculated by averaging the $\Omega_{event}(k)$ amplitudes associated with the corresponding temporal windows of the *k*-th samples within the baseline period. Thus, $\Omega_{baseline}(i)$ represents the amplitude value of normal respiration for the current sample under analysis *i* that has associated amplitude $\Omega_{event}(i)$.

3. Comparing amplitude associated to the current sample with its corresponding baseline value, leads to a new measure $\propto$ representing the associated reduction:

$$\propto(i) = \begin{cases} 1 - \dfrac{\Omega_{\text{event}}(\text{i})}{\Omega_{\text{baseline}}(\text{i})}, & if\ \Omega_{\text{baseline}}(\text{i}) > \Omega_{\text{event}}(\text{i}) \\ -1 - \dfrac{\Omega_{\text{baseline}}(\text{i})}{\Omega_{\text{event}}(\text{i})}, & if\ \Omega_{\text{baseline}}(\text{i}) < \Omega_{\text{event}}(\text{i}) \\ 0, & if\ \Omega_{\text{baseline}}(\text{i}) = \Omega_{\text{event}}(\text{i}) \end{cases} \tag{5.4}$$

4. By repeating this process throughout all the samples of the three signals, a new reduction signal is obtained (see Figure 5.25) that increases in the intervals where the signal amplitude decreases with respect to baseline respiration, and it decreases when respiration recovers.

5. From the previous obtained signal, those periods in which the associated reduction is higher than α_{min} and has duration more than Δ_{min} seconds are considered as possible apneic intervals. Duration Δ_{min} marks minimal duration of the detectable apneic event by the system, thus $\Delta_{min} >= \Delta_{event.}$ Both α_{min} and Δ_{min} are configurable values and default values have been empirically established to 10% and 8 seconds respectively. Note that according to formula (5.4) and the dynamical range of the raw signals (contained in [-1,1], see Table 5.1) a ±10% amplitude change corresponds to a difference of ±0.1 in the signal α(i). The process is illustrated in Figure 5.25.

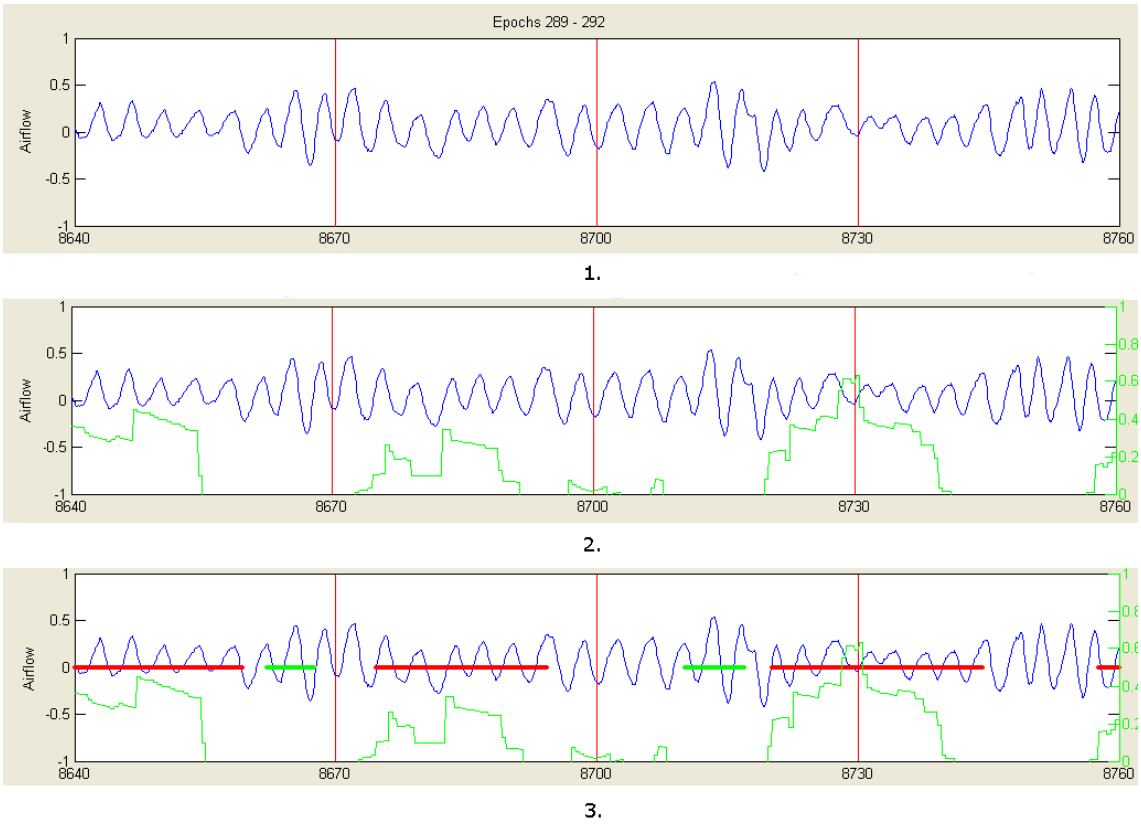


Figure 5.25. Detection of apneic intervals: (1) original airflow signal; (2) computed reduction signal added in green; (3) apneic intervals marked in red, intervals marked in green represent respiration recovering periods

As the last subtask within this module, a post-processing is performed over the previously detected apneic intervals. The post-processing has as its objective tuning the results of the detection algorithm. For example, it may be the case where the proximity of two events in the same respiratory signal causes that, with the above mentioned mechanism, only one apneic interval is detected, however, when actually two different events are present (see Figure 5.26). In this respect an analysis of the suspicious intervals is performed assessing the necessity of splitting the initial event into two or more individual events. The procedure is further detailed next:

1. For those apneic intervals with duration more than Δ_{max} seconds, associated average reduction $\alpha_{average}$ is calculated (see Figure 5.26).

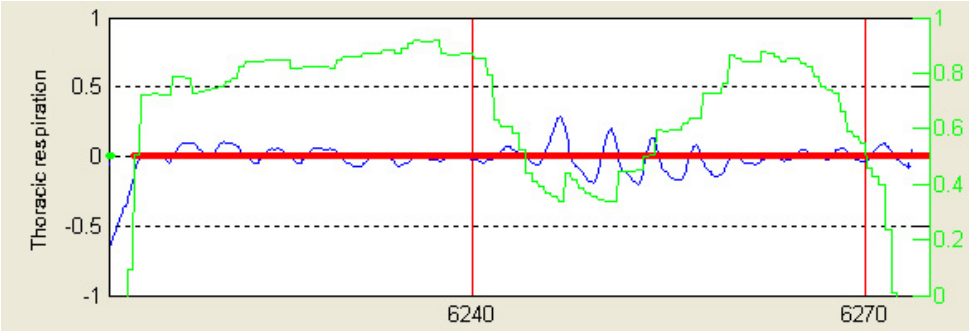


Figure 5.26. Original apneic interval after initial processing algorithm

2. The suspicious event is segmented according to periods in which its associated reduction signal exceeds the $\alpha_{average}$ cut (see Figure 5.27).

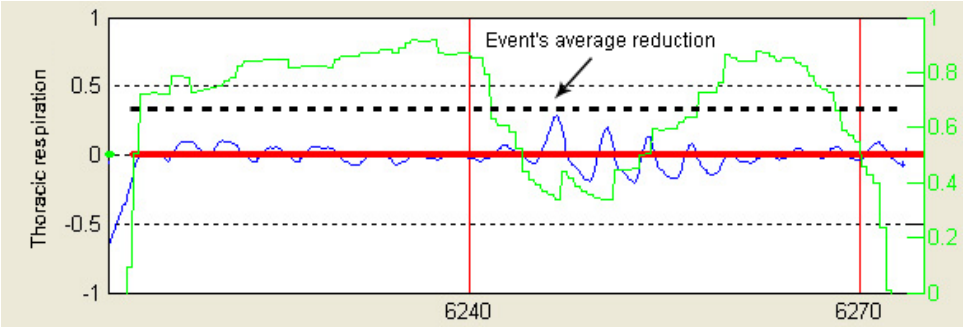


Figure 5.27. Same event as in Figure 5.26, average associated reduction is shown with a black dotted line

3. Segmented intervals of duration more than Δ_{min} are finally considered as individual apneic intervals and are definitively separated from the original event (see Figure 5.28).

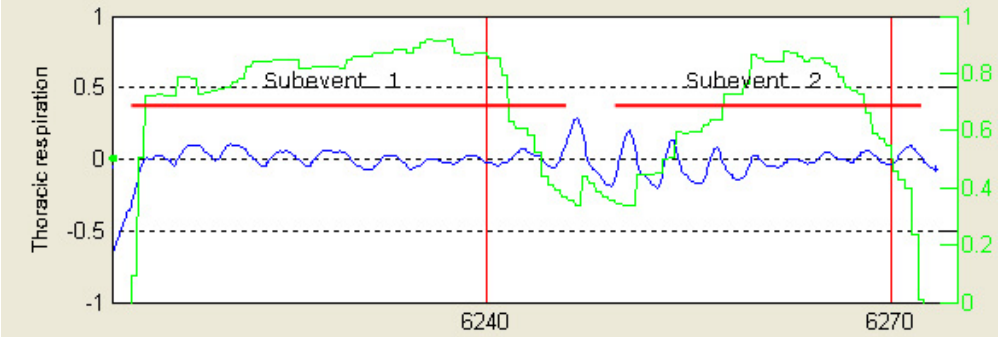


Figure 5.28. Same event as in Figure 5.26, finally original event is split and two different apneic intervals are obtained

5.7.3. Characterization of oxygen saturation signal

In parallel to the analysis of airflow and signals of respiratory movements, analysis and characterization of the oxygen saturation signal (SaO_2) is performed.

It is known that associated to each apneic event a desaturation/resaturation pattern appears reflected in the SaO_2 signal [**8**] [**122**]. During the occurrence of an apneic event, and as a consequence of the airflow reduction during inspiration, an oxygen concentration reduction in the arterial blood oxyhemoglobin is produced. Thus, it is said that an oxygen *desaturation* is taking place. When airflow recovers, oxygen concentration is quickly restored leading to the consequent *resaturation* in the SaO_2 signal.

Criteria that define both events in the context of an apneic event can vary from one expert to another. Normally, after hypoxia observed in the respiratory signals, desaturation in SaO_2 signal is required to start within an interval of 20-40 seconds after starting of the causing apneic event. Duration of the desaturation should be more than 10 seconds and a relative decrease in the saturation levels with respect to normal respiration about 2-4% is required. After reaching its lowest saturation level, this value is maintained for some seconds –*desaturation plateau*- depending on the duration of the apneic event, after which oxygen concentration raises until recovering preceding levels of stable breathing. Because of the hyperventilatory compensation produced as a consequence of the fall in the oxygen levels in blood, the resaturation event often is produced in a more abrupt manner when compared to desaturation. In fact, resaturation events rarely last more than 10 seconds [**122**].

It can be shown, again, that according to previous definitions the detection process itself constitutes a source of imprecision, because of the diversity in the criteria and the necessity to estimate durations, latencies or reductions in the signals from the visual inspection. This imprecision is later on treated by subsequent reasoning mechanisms in the context of fuzzy logic (see subsection "*Detection of apneic events*").

In any case, at this signal processing stage, the algorithm that characterizes the SaO_2 signal to extract relevant information works as follows (see Figure 5.29):

1. Firstly, signal preprocessing is performed by using a means filter in order to remove little disturbances that may interfere in subsequent analysis stages. In this regard convolution of the signal with the filter is made using a 5 seconds temporal window.

2. Once the signal has been filtered, if the sampling rate is higher than 1 Hz then a downsampling is applied up to work with one sample per second, after which a new temporal window of 5 samples –thus 5 seconds- is shifted throughout the resulting signal, sample by sample. This window is centered on the current sample and the difference between the corresponding second previous and second subsequent samples is computed. That is, let *n* to be the current sample, difference calculation is carried out according to the following formula:

$$\Delta x[n] = x[n+2] - x[n-2] \tag{5.5}$$

   where *x[n]* is the value of the saturation signal which corresponds to sample *n*. From this –new- processed signal and by means of a new differentiation, the points representing patterns of the type *start of fall*, *end of fall*, *start of rise* and *end of rise* are marked.

3. Finally, the SaO_2 signal is segmented according to the previous detected patterns. In this respect a *possible desaturation* results from the chaining of one or more *start of fall* patterns, followed by one or more *end of fall* patterns, which mark a complete fall. Analogously, a *possible resaturation* results from the chaining of one or more *start of rise* patterns followed by one or more *end of rise* patterns, which represents a complete rise.

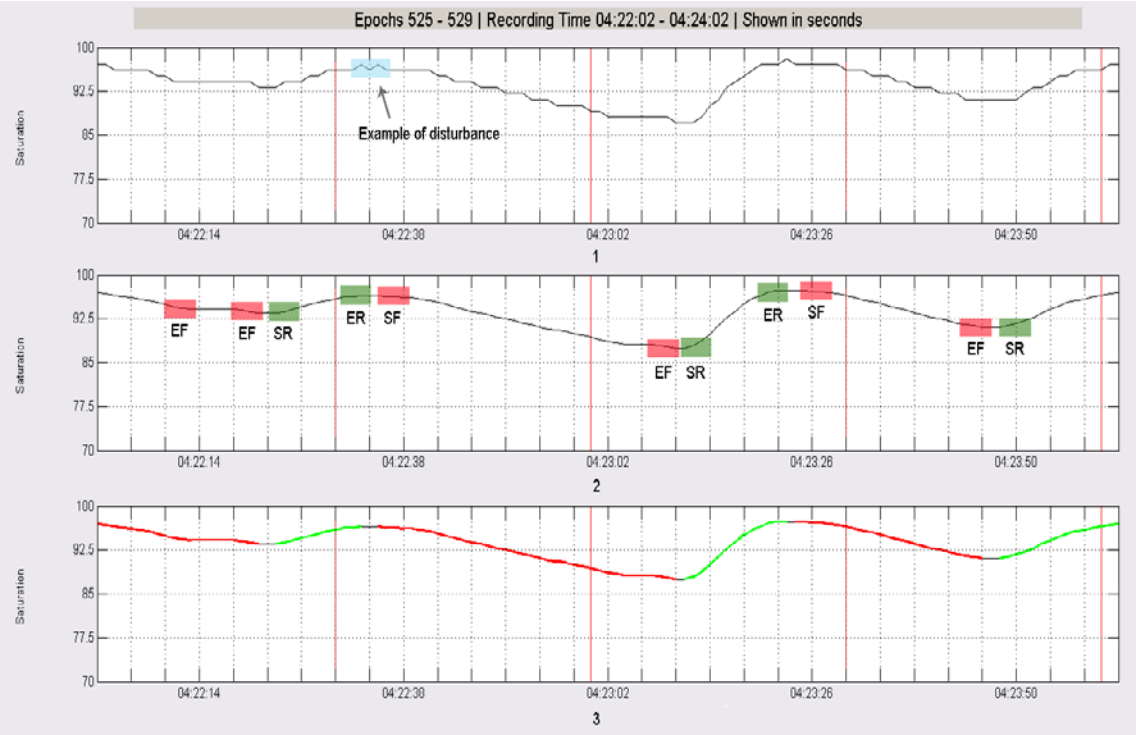


Figure 5.29. Analysis of SaO_2 signal: (1) original signal, a disruption on the signal is highlighted which may interfere in the correct detection of end of rise and subsequent start of fall; (2) smoothed signal where patterns *start of fall (SF)*, *end of fall (EF)*, *start of rise (SR)* and *end of rise (ER)* have been detected; (3) finally signal segmented according to *possible desaturation* (red) and *resaturation* events (green)

A note over the previously described procedure is necessary: during the third step it has been mentioned that segmentation of the signal is done in terms of *possible* desaturations/resaturations. The use of the term "*possible*" is important since here only intervals of sustained increase or decrease in the signal are being detected. Their confirmation or not, as significant enough to be respectively considered actual desaturations or resaturations is carried out later at the reasoning level (see subsection "*Detection of apneic events*").

One more postprocessing algorithm takes place before moving out to the next analysis step that refines the results of the previous segmentation. Sometimes it happens that subsequent chaining of several *end of fall* patterns with no intermediate resaturations produces abnormally long desaturation intervals. In these cases resulting possible desaturations usually present an intermediate plateau that connects two or more subsequent falling intervals. When this plateau is excessively long, it does not have sense anymore to consider that the flat period pertains to the physiological event; a more plausible explanation rather relies on the occurrence of two independent

desaturation events in which no significant resaturation is observed between them. This situation is common in patients where latency between subsequent apneic events is very short. This can be, for example, because of the delay between airflow reduction and the oxygen saturation drop in arterial blood, which among these patients, it sometimes causes the superimposition of two consecutive desaturations without recovering the preceding respiration levels, i.e. no resaturation event is produced. To avoid this situation, when suspiciously long falling saturation intervals are detected (>30 seconds), further analysis is performed trying to localize intermediate plateau subintervals. In the case a flat stable subinterval (> 15 seconds) is detected within the suspicious event, then the plateau subinterval is removed, and two independent falling intervals are finally computed (see Figure 5.30).

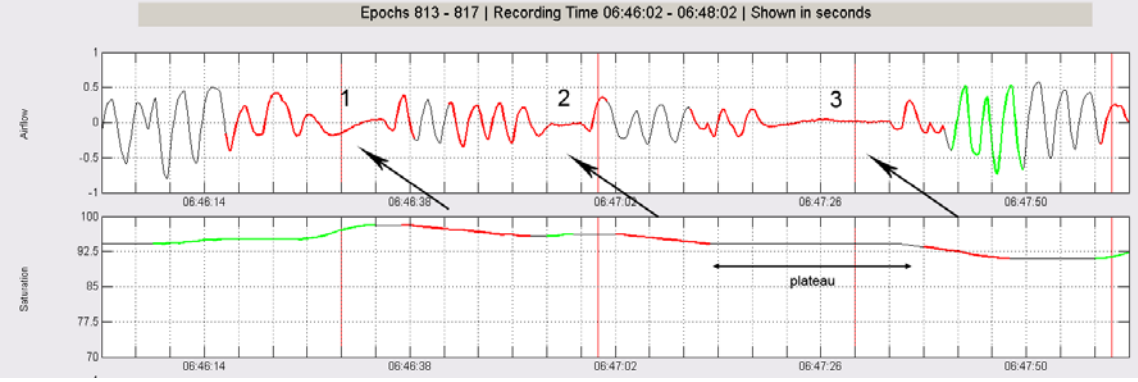


Figure 5.30. In the figure three consecutive apneic events can be observed in a short period of time. Events can be shown over the airflow derivation. Consequent desaturations are marked on the saturation channel (correspondence is indicated by arrows). Little resaturation interval is observed between the first and second events. However, despite two individual events can be observed on the airflow, no resaturation at all is observed between the second and the third apneic event, Postprocessing removes the plateau interval in the SaO_2 signal between these events and two desaturation events are eventually differentiated

5.8. Integration and event characterization

Integration and characterization of all the information generated during the previous analysis phases goes firstly through the relation of the different detected events in time for the construction of relevant diagnostic patterns. Temporal constraint rules are applied at this respect to characterize PSG intervals in the recording in which there is evidence pointing out to the possible existence of an apneic event.

Resulting *apneic patterns* are then evaluated by means of a reasoning process in which respiratory, neurophysiological and contextual information is taken into account to classify the pattern as either apneic or as false positive. In this respect, a fuzzy inference system is used to obtain the corresponding degrees of membership with respect to the categories of apnea, hypopnea and false positive.

The set of apneic patterns that have been considered true positives are then classified according to its origin, i.e. central, obstructive or mixed, by similar fuzzy classification techniques. Previous to classification, additional analysis of thoracoabdominal derivations is necessary to evaluate presence or absence of respiratory effort.

These three outlined processes are described in more detail throughout the following subsections.

5.8.1. Building apneic patterns: temporal event correlation

This analysis stage has as its objective the relation in time of the different events individually detected over each signal. The purpose is the detection of significant patterns –referred as *apneic patterns (APs)*- that characterize temporal intervals in the PSG that evidence the possible existence of an actual apneic event –with independency of its concrete type, i.e. an apnea or a hypopnea.

Time correlation of the individual events is carried out by applying a series of temporal constraints that determine how the different events should be related for the resulting group of events to be considered as a relevant diagnostic pattern.

A first level of temporal constraints involves the events located in the respiratory signals: airflow, oxygen saturation in arterial blood, thoracic and abdominal breathing movements. Each one of the patterns resulting from the correlation of the respiratory events defines a reasoning unit in our system which is subsequently interpreted taking into account the context of the hypnogram and the remaining neurophysiological events, as well as other contextual events such as body position, state of the ambient lights or presence of artifacts.

Hypothesis. *Existence of an apneic pattern is considered when the underlying cause has physiological significance.*

The previous hypothesis will guide the correlation process. In this regard, as it has been commented already, when a respiratory pause occurs as a consequence of an apneic event, the pause is reflected in the sinusoidal signals that monitors the respiratory cycles as an amplitude reduction interval –apneic interval- with respect to normal breathing. On the other hand, as a consequence of the drop in the respiratory flow, oxygen saturation levels in the arterial blood decay producing a desaturation. Desaturations can normally be seen with a certain lag with respect to the occurrence of the causing apneic intervals. Once the apneic event ends, amplitude recovering can be seen usually with presence of compensatory hyperventilation –amplitude recovering slightly higher than previous baseline breathing- as well as resaturation in the oxygen levels, again, with a certain delay with respect to end of the corresponding apneic intervals.

Having the previous sequence of physiological events into mind, the temporal correlation process tries to establish the corresponding relations among the respiratory events. During this process, desaturation events detected in the oxygen saturation signal are taken as the reference from which trying to establish the corresponding cause-effect relationships with the apneic intervals.

In this respect, for each possible desaturation previously detected in the SaO_2 signal:

1. A searching interval is defined beginning at the start of the desaturation and going back in time. This temporal window has duration of 30 seconds, defining a time interval for the localization of apneic intervals in the signals of airflow and thoracoabdominal respiratory movements (see Figure 5.31).

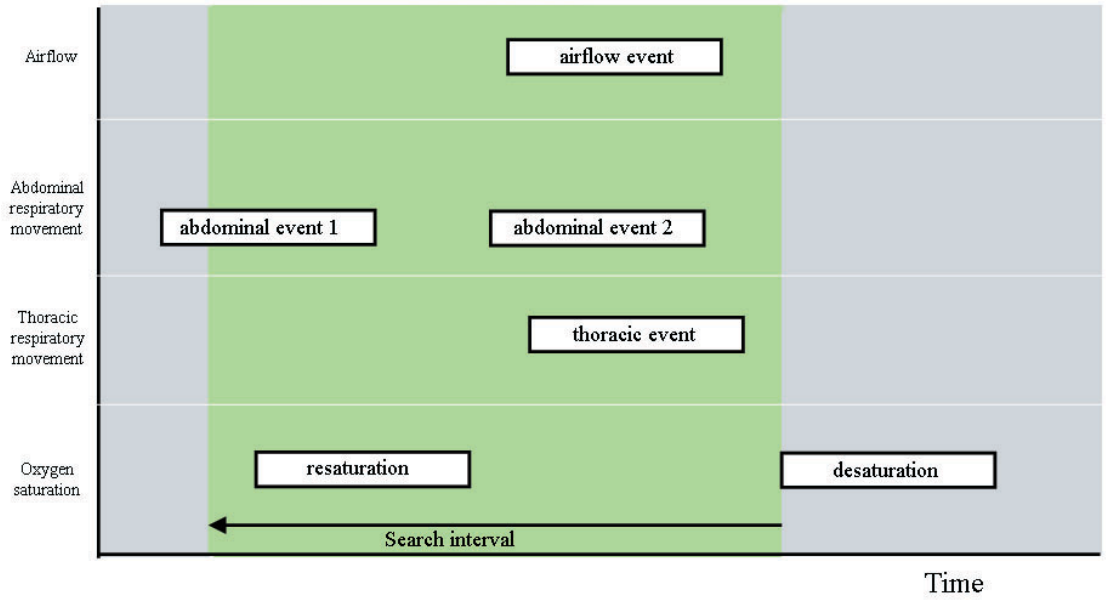


Figure 5.31. Definition of the search interval in the apneic pattern correlation

2. Within the searching interval, there are considered for correlation, those events of which their ending points are within the interval or their starting points are before the start of the fall in the saturation channel. Events exceeding the ending point of the saturation fall are not considered, even if they partially overlap the searching interval (see Figure 5.32).

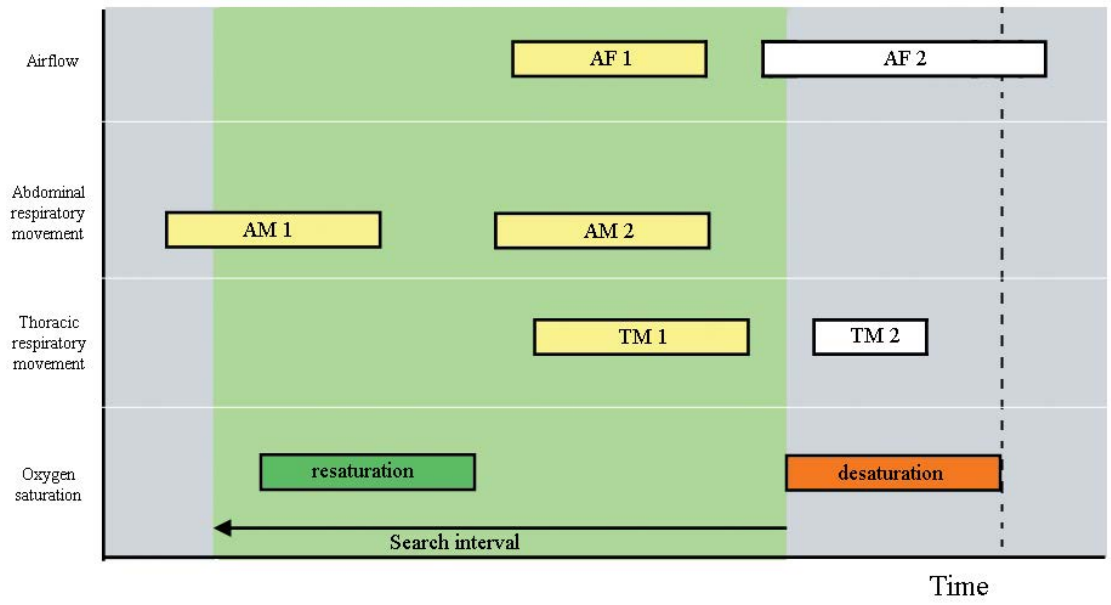


Figure 5.32. Correlation of apneic patterns: event AF2 does not correlate because its end overpasses ending point of the saturation fall. Event TM2 also does not correlate because it starts after starting point of desaturation. AF = Airflow; AM = Abdominal Movement; TM = Thoracic Movement

3. If a rising interval exists in the saturation signal within the searching interval –sign of a possible resaturation- those events ending before the starting of the rise are not considered as well (see Figure 5.33).

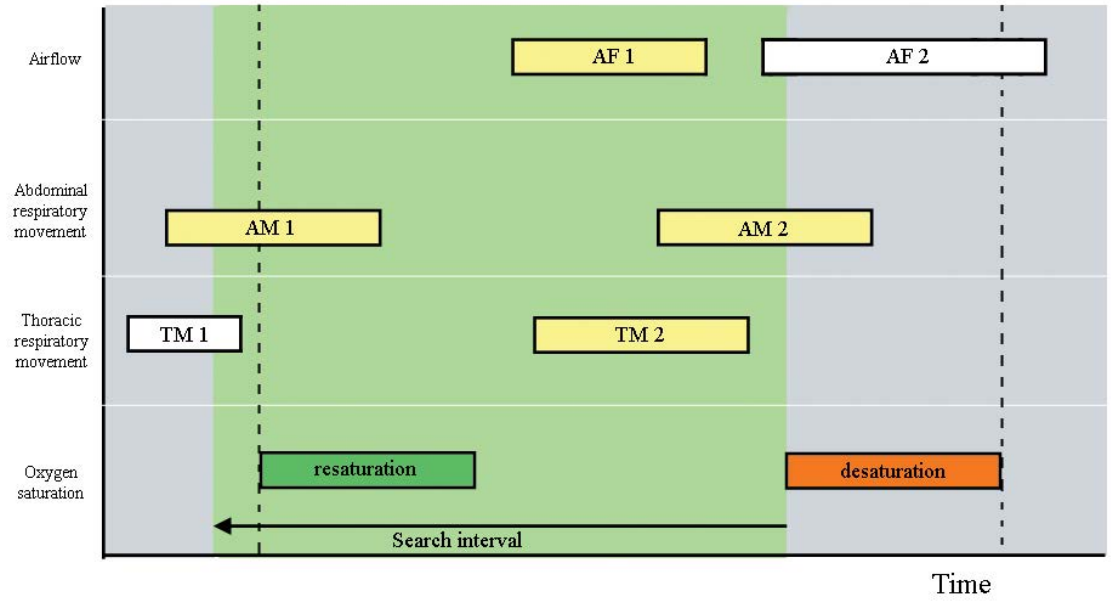


Figure 5.33. Correlation of apneic patterns: event TM1 does not correlate because it ends before starting of resaturation pattern. AF = Airflow; AM = Abdominal Movement; TM = Thoracic Movement

4. For the remaining events within the searching interval, and for each one of the three signals of airflow, abdominal movement, and thoracic movement:

    a. Commencing from the beginning of desaturation and going back in time, the algorithm searches the closest event. This event is related with the possible desaturation.

    b. If more than one event exists in the current channel, the subsequent in order is merged with the previous one if (1) temporal difference between respective ending and starting points is less than Δ_{union} seconds, and (2) a respiratory recovering[40] has not been detected between them (see Figure 5.34, Figure 5.35 and Figure 5.36).

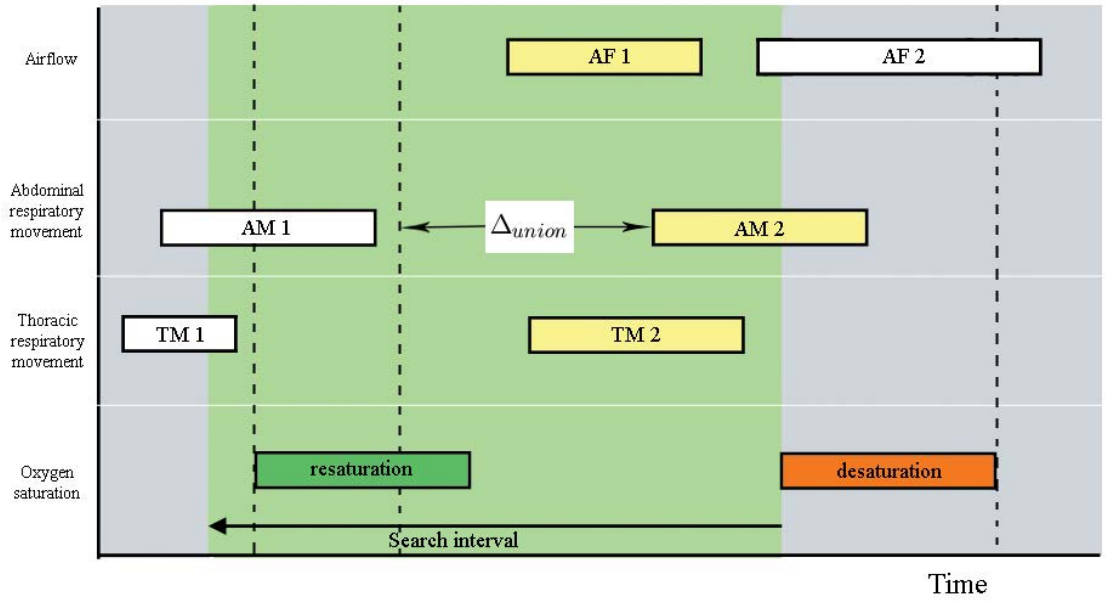


Figure 5.34. Correlation of apneic patterns: Same situation as in Figure 5.33, where event AM1 is discarded for correlation since its distance to AM2 is higher than $\Delta_{union.}$ AF = Airflow; AM = Abdominal Movement; TM = Thoracic Movement

[40] Respiratory signal interval which is found within amplitude levels of normal respiration. In cases of compensatory hyperventilation produced by ending of a previous apneic event, this recovering may even exceed these levels.

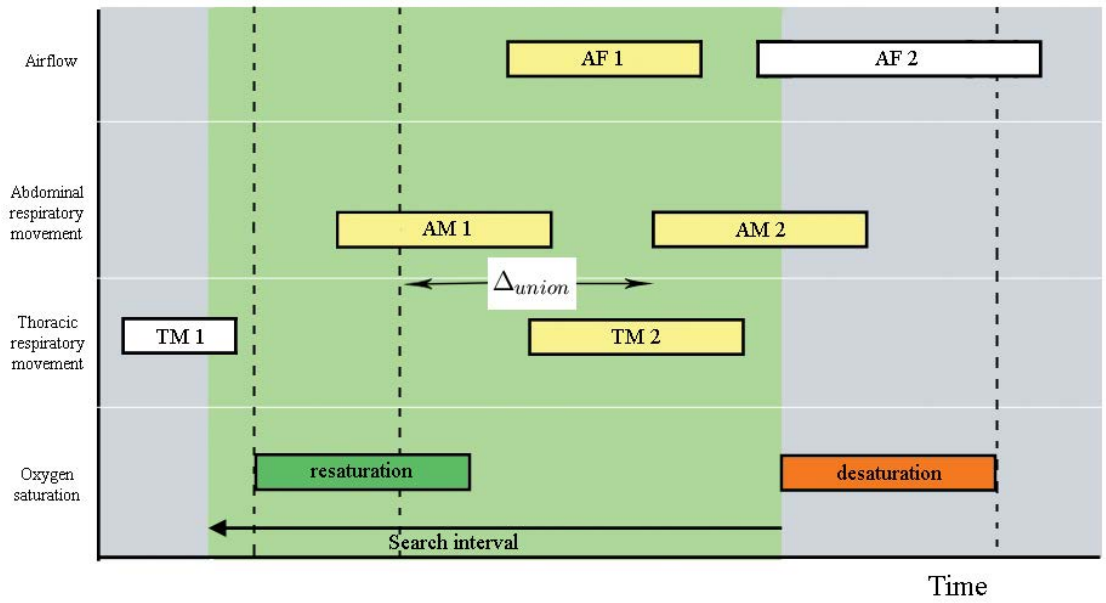


Figure 5.35. Correlation of apneic patterns: Now distance between events AM1 and AM2 is less than $\Delta_{union.}$ AF = Airflow; AM = Abdominal Movement; TM = Thoracic Movement

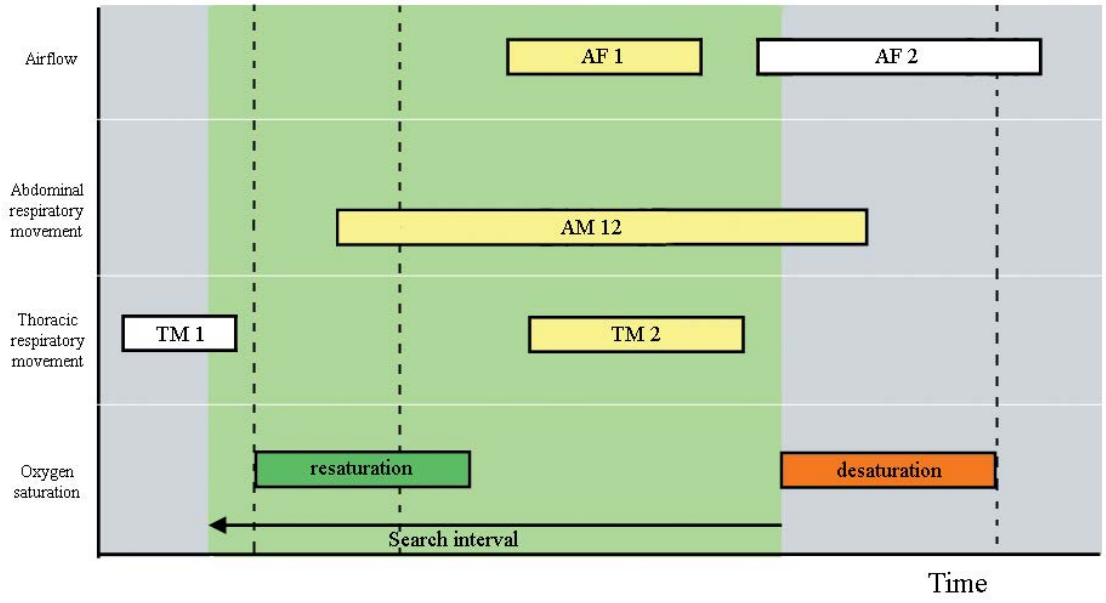


Figure 5.36. Correlation of apneic patterns: Same example as in Figure 5.35. Events AM1 and AM2 are merged to form a new event AM12. Events marked in yellow are those that finally form the apneic pattern together associated with the desaturation. AF = Airflow; AM = Abdominal Movement; TM = Thoracic Movement

After this temporal correlation process, for each rise/fall pattern detected in the SaO_2 signal, an apneic pattern is obtained representing the occurrence of a possible apneic event. The pattern at this stage is composed of the events in the saturation signal plus zero-to-one apneic intervals, for each one of the signals of airflow, abdominal movements, and thoracic movements (see Figure 5.35). In the case in which an apneic interval is not detected on any of these three signals, it is considered that the signal does not show relevant evidence of apneic event in the corresponding channel.

In addition to the previous described procedure, an apneic pattern can also be formed without presence of significant fall in the saturation. This situation can be caused by presence of artifacts in the oxygen saturation channel, or because of the chaining of several apneic events that, because of their proximity, do not reach to trigger two differenced desaturations. In this case, to become a significant diagnostic pattern, there must be apneic intervals over all the three respiratory signals (airflow and thoracoabdominal movements), such that middle point of each of them is contained within time interval of the remaining two (see Figure 5.37).

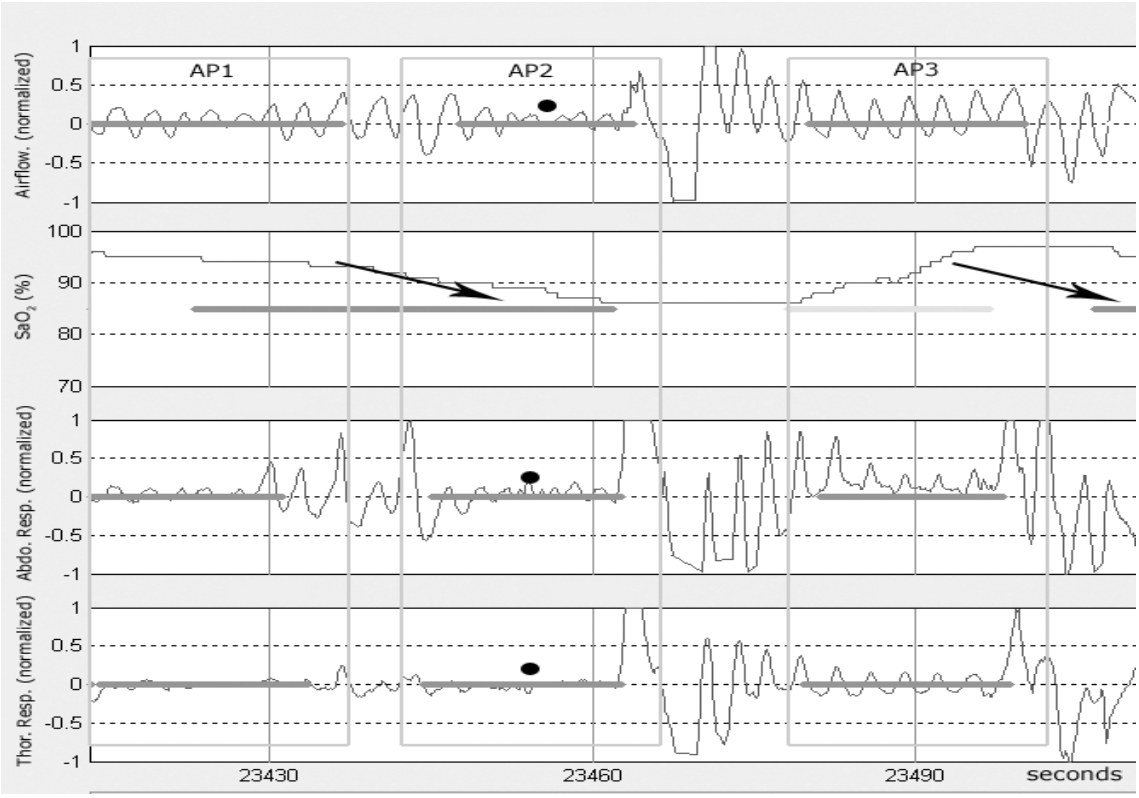


Figure 5.37. AP1 and AP3 are associated with desaturations indicated by arrows. AP2 has no associated desaturation given its proximity to AP1. Nonetheless, AP2 must be classified as an apneic event. The black circles indicate the mid-points of the apneic intervals included in AP2. AP = Apneic Pattern

Apneic patterns constructed on the basis of the previously described procedures only represent respiratory evidence of the presence of an apneic event. However, as it has been continuously sustained throughout the text, correct interpretation requires additional information coming from remaining PSG signals. This information serves as interpretative context and it includes neurophysiological activity and additional contextual information such as body position, presence of detected artifacts or ambient lights. For the integration of this contextual information within the apneic pattern, the patterns previously generated –that involved only respiratory activity- are now completed with the following information:

- *Body position*. The apneic pattern is characterized according to patient's body position during its occurrence. If the sleeping position changes during the occurrence of the apneic interval, the change is also reflected within the apneic pattern.

- *Lights' state*. Information on lights' state (ON or OFF) is added during the occurrence of the pattern.

- *Sleep state*. The pattern is contextualized in function of the sleep state(s) corresponding to the epochs that overlaps with the apneic pattern.

- *Detected artifacts*. If it is the case, each one of the detected artifacts on the different events is integrated in the apneic pattern.

- *EEG Arousals*. An EEG arousal is associated with a respiratory event if the arousal begins less than 5.0 seconds after the end of the event (i.e. 0-4.9 seconds) [**119**].

Eventually, all the different apneic patterns containing the previously described information are then evaluated in order to confirm or discard them, respectively, as true apneic events or as false positives. And then, if they are confirmed, their class, i.e. apnea or hypopnea, and its nature, i.e. obstructive, mixed or central, is determined. The process for the evaluation of the apneic patterns is described throughout the next subsections.

5.8.2. Detection of apneic events

Once all the individual events have been correlated in time to form apneic patterns, analysis is produced in order to confirm or discard them as true apneic events or false positives. In other words, the task is to infer if, which up to this time were *possible* apneic events, can be then confirmed as apneas or hypopneas, or in the contrary, to be discarded as apneic events. It is at this stage in which the approximate reasoning itself takes place through the use of a FIS designed accordingly to the procedures defined in Chapter 4.

Input information to this stage is constituted by all the apneic patterns detected at the previous step. Recall these patterns have resulted from the correlation of the individual events detected among the different respiratory signals, together with the contextual information coming from signals related with the sleep function, patient's body position, light's state and the presence of artifacts. The individual events integrating the apneic pattern have associated quantitative and qualitative information according to respective detection algorithms described throughout the previous sections. This information is summarized in Table 5.3. All this information is of imprecise nature since there is not an exact threshold value that determines the precise –i.e. categorical- evaluation of each one of the evidences. In contrast, it rather exists a range of more or less accepted values that point out to different hypotheses. Moreover, as it was mentioned already, effects of human subjectivity also contribute to this fact. Therefore, particularities of events associated to each pattern should be consequently evaluated from the general perspective, on the basis of approximate judgments and similarity criteria, as it does the human expert.

Table 5.3. Quantitative and qualitative information of the individual events integrating the apneic pattern; *values are determined according to user's preferences

Data Source	Information item	Possible values
Airflow	*Reduction percentage*	[0 – 100] ∈ ℝ
	Duration	[0 – max. duration*]∈ ℝ
Abdominal respiration	*Reduction percentage*	[0 – 100] ∈ ℝ
	Duration	[0 – max. duration*]∈ ℝ
Thoracic respiration	*Reduction percentage*	[0 – 100] ∈ ℝ
	Duration	[0 – max. duration*]∈ ℝ
SaO_2	*Fall reduction percentage*	[0 – 100] ∈ ℝ
	Fall duration	[0 – max. duration*]∈ ℝ
	Rise increase percentage	[0 – 100] ∈ ℝ
	Rise duration	[0 – max. duration*]∈ ℝ
Respiratory artifacts	*Type*	1 – Signal overflow 2- Loss of focus (weak) 3 – Loss of focus (heavy)
	Duration	[0 – max. duration*]∈ ℝ
	Location	1 – Prior to event 2 – Within the event
	Channel	Derivation in which the artifact has been detected
Sleep Stages	*μ(W)*	[0 – 1] ∈ ℝ
	μ(DS)	[0 – 1] ∈ ℝ
	μ(DEEP)	[0 – 1] ∈ ℝ
	μ(REM)	[0 – 1] ∈ ℝ
EEG arousal	*Presence*	YES/NO
	Distance	[0 – max. distance*]∈ ℝ
	Duration	[0 – max. duration*]∈ ℝ
Body position	*Position*	0 – right 1 - left 2 – supine 3 – prone 4 – change (movement)
Lights state	*State*	ON / OFF
EMG	*μ(relaxed)*	[0 – 1] ∈ ℝ
	μ(tense)	[0 – 1] ∈ ℝ

To do so, a fuzzy inference schema is used to obtain, for each apneic pattern, its corresponding degrees of membership, simultaneously, with respect to three different categories: apnea, hypopnea or false positive, in function of its respective similarity according to the expert knowledge expressed in the form of fuzzy linguistic statements (fuzzy rules).

In this respect, at this phase each apneic pattern is analyzed as illustrated in Figure 5.38; knowledge is structured in several knowledge bases composed of fuzzy rules that evaluate the information at different levels of abstraction. This ensures a better structuring of the knowledge. The reasoning process takes place in three steps illustrated in the same figure.

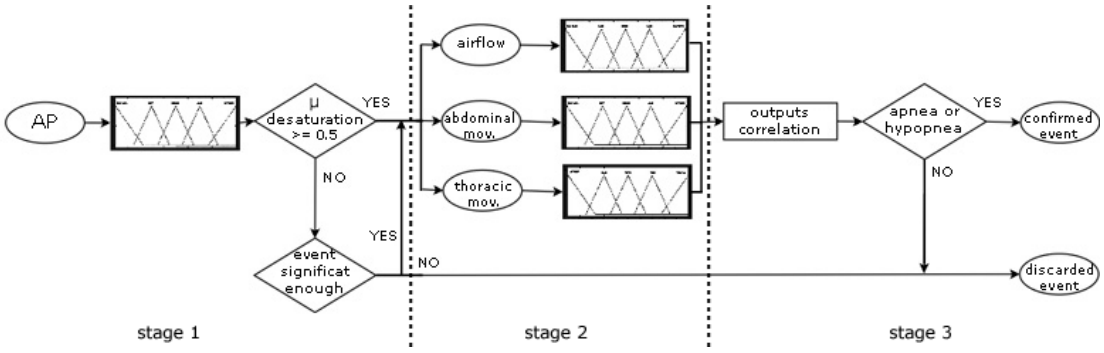
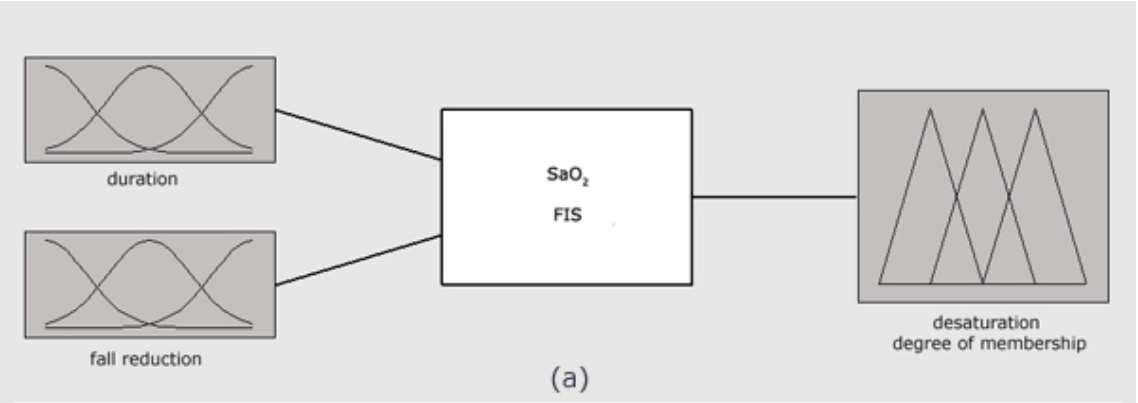


Figure 5.38. Fuzzy reasoning schema split into three stages. AP = apneic pattern; mov = movement

The first stage represents an initial level of abstraction in which the significance of falls detected in the SaO_2 signal are analyzed in order to determine whether they can be considered actual desaturations. For this purpose, a Mamdani-type FIS is used in which two input variables are established representing fall duration ("*duration*") and reduction ("*fall reduction*") that are partitioned using trapezoidal fuzzy sets. The output variable is partitioned into five fuzzy triangular sets so as to establish the degree of membership with regard to the desaturation category ("*event desaturation*"); defuzzification is done using the centroid method. Figure 5.39 illustrates the FIS that intervenes in this stage and the input and output variable partitions for this FIS.



(a)

Input variable *duration* (units in seconds)		
Membership Function Name	**Type**	**Vector Definition**
Very Short	Trapezoidal	[0 0 3 4]
Short	Trapezoidal	[3 4 5 6]
Normal	Trapezoidal	[5 6.5 35 60]
Large	Trapezoidal	[35 45 ∞ ∞]

(b)

Input variable *fall reduction* (units in percentages)		
Membership Function Name	**Type**	**Vector Definition**
Soft	Trapezoidal	[0 0 1 2]
Normal	Trapezoidal	[1.8 2 4 5]
Strong	Trapezoidal	[4 6 8 10]
Very Strong	Trapezoidal	[9 12 100 100]

(c)

Output variable *event desaturation* (units in degree of membership)		
Membership Function Name	**Type**	**Vector Definition**
Very soft	Triangular	[0 0 0.23]
Soft	Triangular	[1.8 2 4 5]
Normal	Triangular	[0.35 0.5 0.65]
Strong	Triangular	[0.55 0.7 0.85]
Very Strong	Triangular	[0.77 1 1]

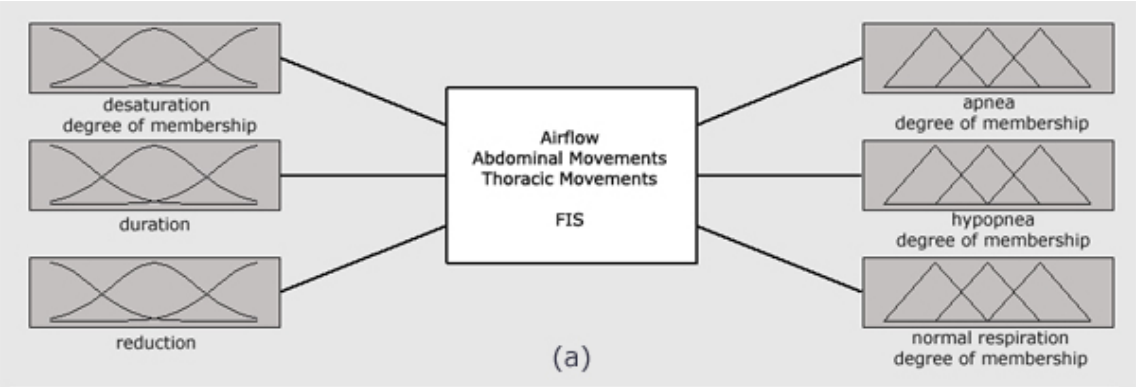
(d)

Figure 5.39. (a) FIS for the first reasoning stage; (b) Partition for the duration input variable; (c) Partition for the fall reduction input variable; (d) Partition for the event desaturation output variable

Since, physiologically, an apneic event should be associated with a desaturation event, in a first instance only apneic patterns in which possible associated desaturations have membership values of $\mu_{desaturation}(AP) \geq 0.5$ (calculated empirically as the threshold value) are considered for the next level of reasoning. It is possible, as previously mentioned, that two apneic events occurring within a short space of time produce a single significant desaturation [**119**] (in other words, there is no perceptible resaturation); it is also possible that the measuring device may fail to reflect a desaturation associated with an apneic event in the SaO_2 signal. Bearing in mind these

circumstances, the apneic patterns that have not been associated with any significant desaturation in the correlation phase ($\mu_{desaturation}(AP) < 0.5$) are also included in the second reasoning phase, but only if they show evidence of including apneic intervals of enough significance. The requirements established for the evaluation of patterns with no clear desaturations, and include them in the second reasoning phase, have been established in the previous section (see "*Building apneic patterns: temporal event correlation*").

In the second stage, each apneic interval included in the apneic pattern—airflow and thoracic and abdominal movement signals—is evaluated concurrently. The aim is to obtain an individual measure of the evidence of apneic event for each respiratory signal separately. For example, it is possible that, by uniquely considering the apneic interval regarding the airflow signal, the evidence was not significant enough to consider the apneic pattern to be a real apneic event. On the other hand, the respiratory movement signals might reflect significant amplitude reductions indicating that the apneic pattern should be classified as a hypopnea [**7**]. At this second stage, for each of the three signals, inputs are the duration of each existing apneic interval ("*duration*"), the associated reduction ("*reduction*") and the degree of membership obtained at the first phase and referring to the desaturation category ("*desaturation degree of membership*"). The output for the second reasoning stage consists of three fuzzy variables: apnea, hypopnea and false positive (i.e. normal respiration). The output variables are partitioned on the basis of fuzzy triangular sets. Figure 5.40 depicts the FIS for the three respiratory signals and the respective partitions. The desaturation input value is the value that has been obtained at the output for the first stage, and hence it is the same for each apneic interval.



Input variable *duration* (units in seconds)		
Membership Function Name	**Type**	**Vector Definition**
Short	Trapezoidal	[0 0 7 9]
Normal	Trapezoidal	[8 10 120 130]
Large	Trapezoidal	[120 130∞ ∞]

(b)

Input variable *reduction* (units in percentages)		
Membership Function Name	**Type**	**Vector Definition**
Normal	Trapezoidal	[0 0 25 30]
Slightly Reduced	Trapezoidal	[25 30 65 70]
Clearly Reduced	Trapezoidal	[65 70 75 80]
Totally Reduced	Trapezoidal	[75 80 100 100]

(c)

Output variables *apnea*, *hypopnea* and *normal respiration* (units in degree of membership)		
Membership Function Name	**Type**	**Vector Definition**
Very soft	Triangular	[0 0 0.23]
Soft	Triangular	[0.15 0.3 0.45]
Normal	Triangular	[0.35 0.5 0.65]
Strong	Triangular	[0.55 0.7 0.85]
Very Strong	Triangular	[0.77 1 1]

(d)

Figure 5.40. (a) FIS for the second reasoning stage involving airflow and thoracic and abdominal respiratory movements; (b) Partition for the duration input variable; (c) Partition for the reduction input variable; (d) Partition for the event desaturation output variable

Finally, it is necessary to correlate the individual outputs obtained in the second stage for each of the apneic intervals in the apneic pattern. Weights are assigned to each of the respiratory signals and the memberships obtained are combined as outputs to the second stage, according to the established weights. In this regard weighted fuzzy logic techniques [**123**] are used, with fuzzy AND/OR operators interpreted as follows:

Let (∇, Δ) be a pair of T-norms and T-conorms that satisfy DeMorgan's laws, let $w = max\{w_1, w_2, \ldots, w_n\}$, with $\sum_{i=1}^{n} w_i = 1$, where each w_i is a weight associated with each operand, and let $\mu(a_1), \mu(a_2), \ldots, \mu(a_n)$ be the degrees of membership to which the AND operator or the OR operator are applied. The resulting degrees of membership are calculated according to the following formulas:

$$\mu_{AND}(a) = ((\frac{w_1}{w} \times \mu(a_1)) \Delta \ldots \Delta (\frac{w_n}{w} \times \mu(a_n)) \tag{5.6}$$

$$\mu_{OR}(a) = ((\frac{w_1}{w} \times \mu(a_1)) \nabla \ldots \nabla (\frac{w_n}{w} \times \mu(a_n)) \tag{5.7}$$

Using these definitions, the process basically consists in performing a *fuzzy OR* of the degrees of membership for the output variables *apnea* and *hypopnea*, and a *fuzzy AND* for the degrees of membership regarding the *normal respiration* output variable, among the individual outputs of the three signals obtained at the second stage. The use of the AND operator in order to evaluate the possibility of discarding the pattern as apneic event tends to favor a positive identification of the apneic event. The former increases the sensitivity in the detection, but it also augments the number of false positives.

The concrete applied weights for airflow (0.4), abdominal movement (0.3) and thoracic movement (0.3), are estimated empirically as the most suitable. The idea of assigning greater weight to airflow is motivated because, in the traditional diagnostic process, this signal represents the main reference for apneic event identification [**124**].

By applying the above described procedure it is obtained, for each apneic pattern (AP), a single measure of its degree of membership with respect to the *apnea*, *hypopnea* and *normal respiration* output variables: $\mu_{apnea}(AP_i)$, $\mu_{hypopnea}(AP_i)$ and $\mu_{normal\ respiration}(AP_i)$.

At this point, artifact information from the respiratory signals is taken into account in the reasoning. In this process two different types of artifacts that may affect the apneic pattern on any of its associated signals are distinguished: (1) artifacts taking place in the event itself, that is, artifacts which its temporal occurrence is within an apneic interval, and (2) artifacts that, even though they do not temporally overlap with a event, on the other hand they may influence its detection. An example of the second situation is when the artifact occurs in the temporal interval used to calculate normal respiration baseline that is used to mark the apneic interval (see subsection "*Identification of apneic intervals*").

The apneic pattern has to be interpreted also with regard to remaining contextual information, that is, patient's sleep position, ambient lights, sleep stages and the presence of EEG arousals.

In this respect, a first procedure is used to discard false positives among apneic patterns that, although according to respiratory activity, they may present the characteristics of an apneic pattern (i.e. $\max [\mu_{apnea}(AP_i), \mu_{hypopnea}(AP_i)] >= \mu_{normal\ respiration}(AP_i)$), they should be discarded according to the contextual evidence. Thus, independently of its associated degree of membership, an apneic pattern is declared a false positive if any of the following situations occur:

- *Body position change*. Change in body position is considered to produce non reliable respiratory signals because of the artifact in the sensors caused by the movement. Hence, when change in sleeping position is detected during occurrence of an apneic pattern, the apparent respiratory reduction in the signals is associated with patient's movement rather than with a real respiratory absence.

- *Artifacts (HA)*. When loss of focus or overflow is detected within the respiratory channels of apneic pattern and the artifact has been classified as HA (see subsection "*Preprocessing of respiratory signals*"), then the corresponding apneic interval is discarded since apparent respiratory reduction can be attributed to the occurrence of the artifact.

- *Patient is awake*. By definition, apneic events must occur while the patient is sleeping. In this respect information on patient's hypnogram is used to assess his/her sleeping state and, if the apneic pattern is localized during stable periods of wakefulness, then the apneic pattern is considered as a false positive.

- *Lights are "ON"*. Another source of evidence to discard false positives comes from the recording of lights. If lights channel is on, either because ambient light is on, or because the sleep technician has annotated the interval as *non valid for scoring*, then all possible apneic events detected on the corresponding interval are automatically discarded.

In addition, besides of false positive discarding, contextual information is also used for adjusting final degrees of membership of the remaining apneic patterns –i.e. those that were not previously discarded. Depending on the corresponding evidence, this adjusting can increase or decrease the degree of membership associated to the different hypotheses (apnea, hypopnea or normal respiration/false positive). The former may only produce a mere adjusting on the final beliefs, but it may also cause a new discarding of false positives, or even the final confirmation of apneic patterns that were initially considered as false negatives. Note that although based on clinical evidence, exact adjusting values have been empirically determined. Specifically:

- EEG arousal is usually triggered as a consequence of the lack of oxygen produced by an apneic event. According to medical criteria [**119**] an arousal is associated with a respiratory event if the arousal begins less than 5 seconds after the end of the event (i.e. 0-4.9 seconds). Consequently, since EEG arousal associated with apneic pattern is indirect evidence of the presence of apneic event, the following rule is applied:

$$\mu_{apnea}(AP_i) = \max(1, \mu_{apnea}(AP_i) + \mu_{apnea}(AP_i) \times 0.2)$$
$$\mu_{hypopnea}(AP_i) = \max(1, \mu_{hypopnea}(AP_i) + \mu_{hypopnea}(AP_i) \times 0.2)$$
$$\mu_{normal\ respiration}(AP_i) = \min(0, \mu_{normal\ respiration}(AP_i) - \mu_{normal\ respiration}(AP_i) \times 0.2)$$

- Muscle relaxation is a symptom of deep sleep stage which favors appearance of apneic events. This situation is common of phase REM. Sleep in supine position can also be a triggering event favoring occlusion of upper airways, increasing likelihood of appearance of apneic events [**14**] [**125**]. Accordingly, final degrees of membership to account for this evidence are slightly modified:

$$\mu_{apnea}(AP_i) = \max(1, \mu_{apnea}(AP_i) + \mu_{apnea}(AP_i) \text{ x } 0.1)$$
$$\mu_{hypopnea}(\text{AP}_i) = \max(1, \mu_{hypopnea}(AP_i) + \mu_{hypopnea}(AP_i) \text{ x } 0.1)$$
$$\mu_{\text{normal respiration}}(AP_i) = \min(0, \mu_{\text{normal respiration}}(AP_i) - \mu_{\text{normal respiration}}(AP_i) \text{ x } 0.1)$$

- Transition from light sleep to deep sleep may cause slight reduction of in the respiratory amplitude. However, this reduction is related to the normal sleep process and therefore it should not be considered evidence of hypopnea [**126**]. Thus, the following adjusting is applied in the case this situation occurs:

$$\mu_{hypopnea}(AP_i) = \min(0, \mu_{hypopnea}(AP_i) - \mu_{hypopnea}(AP_i) \text{ x } 0.5)$$

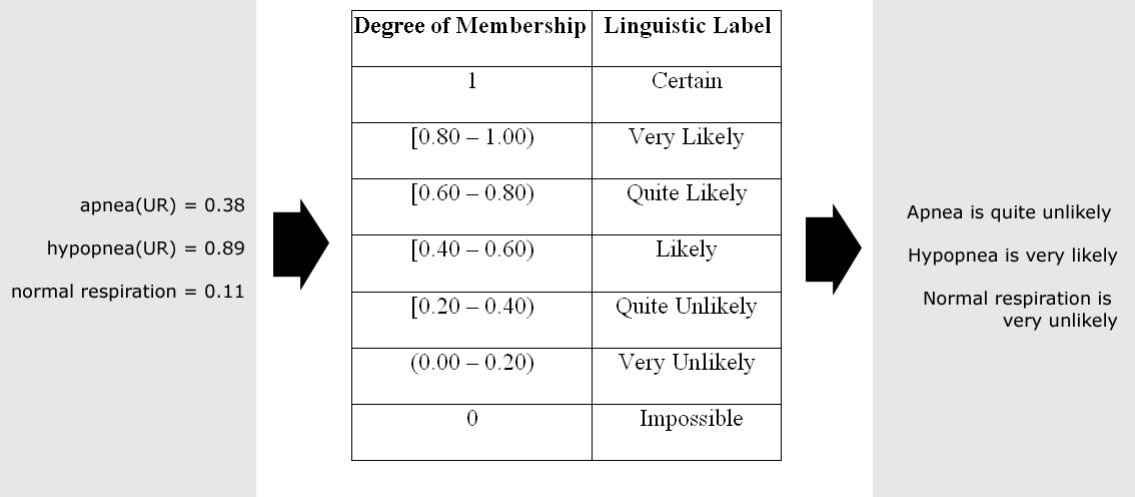
Once the reasoning and contextual adjusting processes have ended, as the output of the apneic detection phase, a set of apneic patterns finally confirmed as apneic events (either *apnea* or *hypopnea*) and temporally localized over the PSG are obtained. Let $\mu_{apnea}(AP_i)$, $\mu_{hypopnea}(AP_i)$ and $\mu_{\text{normal respiration}}(AP_i)$ to be the final degrees of membership in the apnea, hypopnea and normal respiration fuzzy sets of the apneic pattern_i, $i=1..n$, where n is the total number of apneic patterns. The process that determines whether apneic pattern_i is finally confirmed or otherwise is discarded as apneic event is the following:

If : $\{\max [\mu_{apnea}(AP_i), \mu_{hypopnea}(AP_i)] >= \mu_{\text{normal respiration}}(AP_i)\}$

Then: AP_i is confirmed as an apneic event

Else: AP_i is considered to be a false apneic event and is therefore ruled out.

Finally, and with independency of the final confirmation of the apneic event, for each apneic pattern, numerical values representing the corresponding degrees of membership are characterized in terms of linguistic labels (see Figure 5.41).



Degree of Membership	Linguistic Label
1	Certain
[0.80 – 1.00)	Very Likely
[0.60 – 0.80)	Quite Likely
[0.40 – 0.60)	Likely
[0.20 – 0.40)	Quite Unlikely
(0.00 – 0.20)	Very Unlikely
0	Impossible

Figure 5.41. Outputs in the form of linguistic labels for the differing degrees of membership

For example, given AP_i, if the system establishes that $\mu_{apnea}(AP_i) = 0.78$ and $\mu_{hypopnea}(AP_i) = 0.56$ on concluding the reasoning phase, then the output perceived by the user is that the event is *quite likely* to be an apnea but it also is *likely* to be a hypopnea. Note that with the preceding procedure, categorical classifications of the event are avoided and, at the same time, outputs are characterized using expressions in natural language of the clinician.

5.8.3. Classification of apneic events

In the section "*Apneic event classification*" of Chapter 2, classification of the apneic event has been structured, independently of being an apnea or hypopnea, according to its origin or nature, that is, as central, mixed or obstructive. A correct classification of the nature of the apneic event is necessary insofar as treatment of the disease may depend on the dominant type of apneic events present in the patient diagnosed with SAHS.

Classification phase in the developed system has as its objective to perform such a classification from the apneic events previously confirmed at the detection phase.

Clinical approximation to carry out classification from inductive plethysmography focus on analysis of respiratory movements, both thoracic and abdominal, which allows inferring the presence or absence of respiratory effort associated to the apneic event. For that purpose, in the constructed system, analysis is structured in two stages:

- First, recorded signals of respiratory movements are analyzed using signal processing techniques. The objective is to extract relevant information to associate to each apneic event previously confirmed, features gathering data which is subsequently evaluated at the reasoning stage.

- Second, reasoning stage carries out an analysis of previous extracted features by using a FIS which performs a fuzzy classification of the corresponding apneic event with respect to three possible outputs: *obstructive, mixed* and *central*.

More specifically, first classification processing stage proceeds in the following form. For each apneic event confirmed at the detection phase:

1. Derivations of thoracic and abdominal respiratory movements are analyzed in the context of the corresponding apneic pattern, firstly by filtering high frequencies (low pass filter, 2 Hz) in order to remove possible high frequency noise present in the signals.

2. Resulting filtered signal intervals are processed quantifying amplitude of internal breathing cycles. Let φ_{cycle_i} to be amplitude of the *i*-th respiratory cycle within the interval. Amplitude is measured as difference between maximum and minimum values of the respiratory cycle.

3. Taking as reference the starting point of the apneic event, a reference interval is taken going Δ_{ref} seconds back in time. Then the same processing is performed over this reference interval as in points 1 and 2. That constitutes a baseline period to compare amplitude of respiratory cycles during the apneic event, with respect to

average of respiratory cycle's amplitude within this reference interval. Let φ_{ref} to be such average amplitude.

4. Then, the two respiratory channels are segmented in the time interval of the apneic event, by comparing amplitude of the respiratory cycles with their corresponding reference values. Segmentation is performed in accordance with the following criterion:

$$\varphi_{cycle\_i} < \varphi_{ref} * \alpha \Rightarrow absence\ of\ movement_i$$

$$\varphi_{cycle\_i} \geq \varphi_{ref} * \alpha \Rightarrow presence\ of\ movement_i$$

That is, if the amplitude reduction shown by the corresponding *i*-th cycle with respect to average reference amplitude is higher than certain threshold α, then the respiratory cycle is labeled as *absence of movement*. On the other hand if the reduction does not overpass the value α, then the respiratory cycle is labeled as *presence of movement*. Empirically value of α has been set to 0.15. The process is illustrated in Figure 5.42 and Figure 5.43.

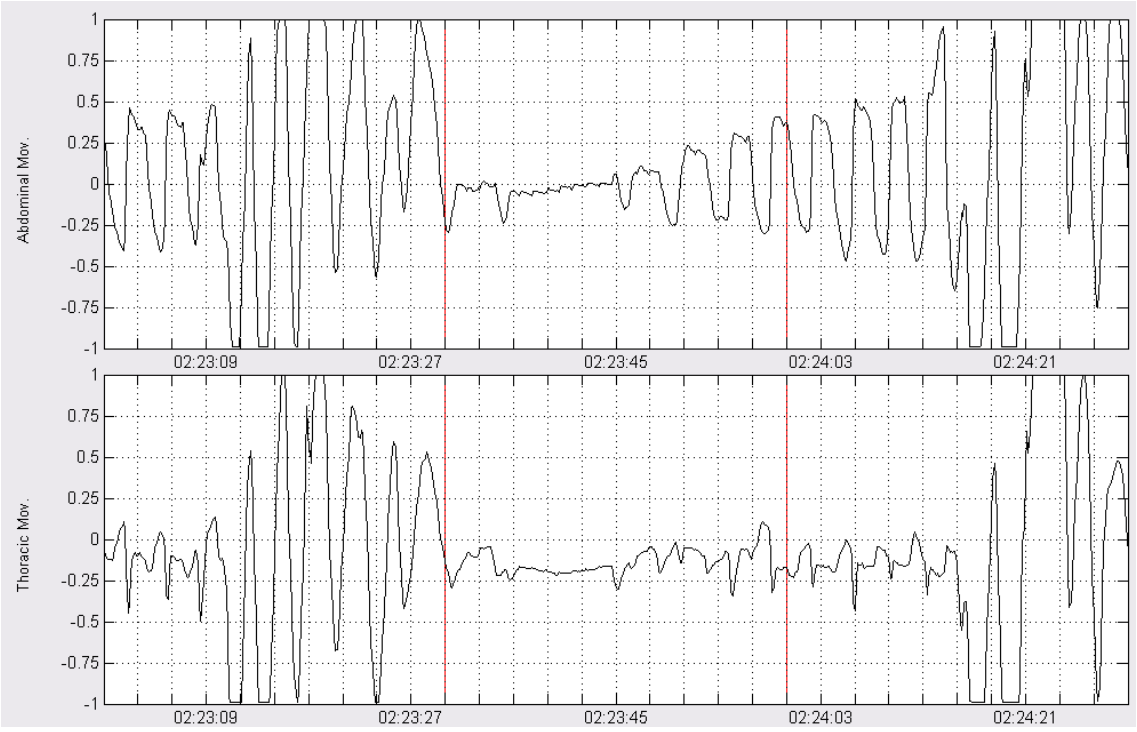


Figure 5.42. Classification processing: apneic event shown over abdominal and thoracic respiratory movements

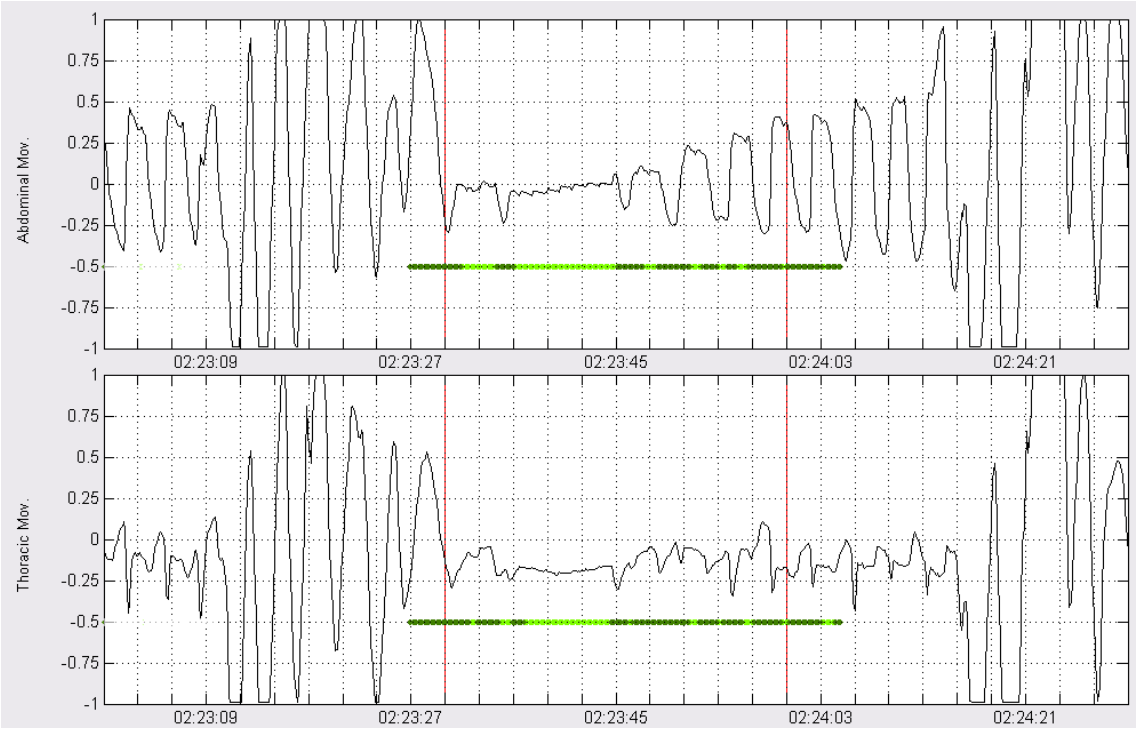


Figure 5.43. Classification processing: segmentation has been made according to periods of presence of movement (dark green) and absence of movement (light green)

5. Once the signals have been segmented on intervals of presence/absence of movement, the whole analysis interval is subdivided into 3 subsegments of equal length, respectively representing starting, center and ending of the event. For each one of these subintervals the percentage of absence and presence of movement is quantified. Additionally, for each one of the corresponding subintervals in the two signals, linear correlation is computed (see Figure 5.44). Linear correlation between the different sections of the event complements with information on phase of the two signals previous quantification on the amount of registered movement (absence/presence). As it has been described in Chapter 2, central events are characterized by thoracic movements in phase with those from the abdominal respiration –positive correlation. On the other hand, during occurrence of an event of an obstructive origin, sinusoidal waves from thoracoabdominal derivations tend to present certain phase lag, which results in negative correlation.

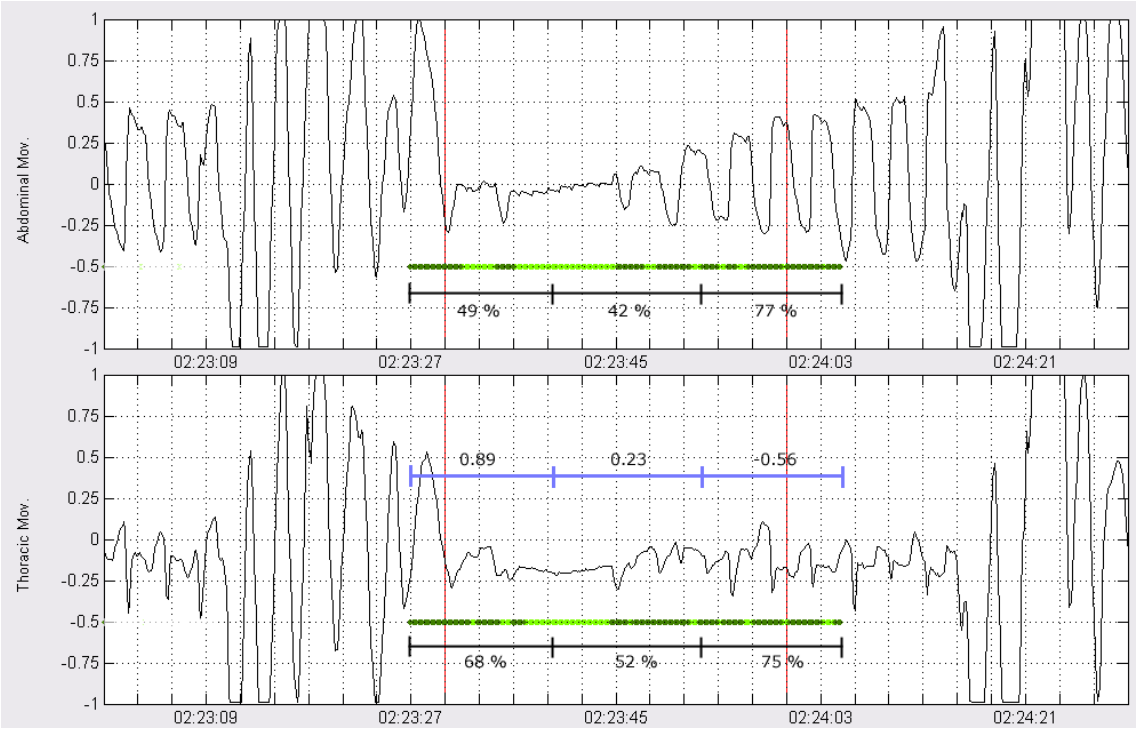


Figure 5.44. Classification processing: each derivation is divided into three intervals and amount of presence of movement is computed. Percentages are shown below marks of previous segmentation. Linear correlation between the two signals is also computed for each subinterval. Resulting correlation is shown in the middle of the figure.

Once processing of thoracoabdominal signals has concluded, resulting extracted features are fed into a FIS to perform classification of the event. A total of six features are used for classification: average of presence of movement over the two signals for each of the three subintervals, and liner correlation obtained for each one of them (see Figure 5.44).

The FIS used at this reasoning stage has been optimized following learning procedures and neuro-fuzzy architecture for classification described in Chapter 4 (see section "*Neuro-fuzzy modeling within the developed system*"). Unlike the FIS used at the detection stage, here the inferential process is carried out in just one step. Therefore the number of input variables is six, in correspondence with the number of input features. At the output, a fuzzy classification of the corresponding apneic pattern is obtained in basis of three fuzzy output variables: obstructive, central and mixed. A different degree of membership results to each one of them, and the maximum degree is taken as the default final classification. As it has been described in previous subsection, linguistic classification is also performed over the obtained numerical degrees of membership to express results in natural language.

5.9. Diagnostic generation

Diagnostic generation basically consists in compilation of all the relevant information produced during the analysis to compute significant numerical parameters and indexes that are useful for the physician to issue a diagnosis. Resulting information objects are then organized in form of report, where information is synthesized according to several categories and levels of detail. The clinician can therefore access on demand to the different items and request explanations about system's results.

A list of AASM recommended parameters to be reported for polysomnography can be found in [**7**] and are used here as guideline. With regard to sleep scoring data, besides of the hypnogram, the following parameters are included in the report:

1. Total lights out clock time: Total time spent with lights OFF (hr:min)
2. Lights on clock time: Total time spent with lights ON (hr:min)
3. Total sleep time (TST): Total time in any sleep state different from awake in min
4. Sleep latency (SL): Time with lights OFF to the first epoch of any sleep in min
5. Sleep start (ST): Start of the recording to the first epoch of any sleep in min
6. Stage REM latency: sleep onset to first epoch of stage REM in min. Sleep onset is considered as the first epoch of any sleep
7. Stage DEEP latency: sleep onset to first epoch of stage DEEP in min
8. Wake after sleep onset (WASO): stage W during (1) minus (5) in min. Note that WASO includes all wake activity, including wake out of bed
9. Percent sleep efficiency: (TST / (1)) x 100
10. Time in each stage (min) and corresponding percentage with respect to TST

Total number of arousal events is counted and the Arousal Index (ArI) is computed as:

$$ArI = \frac{(Total\ number\ of\ arousals\ x\ 60)}{TST}$$

Main objective of diagnosis is to determine whether the patient suffers from SAHS and, if this is the case, to indicate the type. In order to do this, the Apnea-Hypopnea Index (AHI) is computed as:

$$AHI = \frac{(Total\ number\ of\ apneic\ events\ x\ 60)}{TST}$$

ApI (apnea index) and HI (hypopnea index) are also computed analogously but considering only the number of apneas and hypopneas respectively. On the basis of the value of AHI, linguistic classification of the severity of the syndrome is performed, according to Table 5.4, into four different categories. Nevertheless, it has to be taken into account that besides clinical recommendations [**127**], there is still a discussion on the appropriate distinction between the different categories of SAHS severity based on thresholding the number of apneic events per hour of sleep [**128**]. Therefore, although based on common extended segmentation values of AHI over the literature, the resulting categories used in Table 5.4 should be considered just as a guideline for the clinician. Final severity consideration for the diagnosis should in fact account for additional symptomatology of the patient [**129**].

Table 5.4. Classification of SAHS severity

AHI	**Linguistic label**
$0 \leq AHI < 10$	Not significant
$10 \leq AHI < 20$	Mild
$20 \leq AHI < 30$	Moderate
$AHI \geq 30$	Severe

Once the existence of SAHS has been confirmed, syndrome classification is performed based on recalculation of AHI taking into account each kind of apneic event, i.e. obstructive, mixed or central. The syndrome is then classified in terms of the greatest value given by the relative AHI. Prevalence of *Mixed* events does not represent a specific diagnosis from the point of view of a clinical diagnosis [**129**]. For this reason, when calculating AHIs with respect to SAHS classification, *mixed* events and *obstructive* events are computed together.

If the number of central events is reduced in comparison with the number of obstructive events (i.e. less than 20% of obstructive events), then it can be concluded that the detected central events are not a consequence of the existence of central sleep apnea-hypopnea syndrome (CSAHS), and therefore the syndrome is classified as obstructive (OSAHS). On the other hand, if the number of central events is greater than the number of obstructive events (i.e. more than 50% of obstructive events) then it is indicated that the patient suffers from CSAHS [**130**] [**131**]. In any case, a minimum occurrence of at least 10 events per hour of sleep (AHI $\geq$ 10) for both, central and obstructive events, should occur for a patient to be either scored respectively with CSAHS or OSAHS. Finally, in the case where the number of central events is between 20%-50% of obstructive events, the patient is diagnosed as with "*mixed* OSAHS". Clarification is needed at this point, since as it has been commented already, according to clinical guidelines, the category mixed OSAHS is not recognized as a proper syndrome type [**129**]. On the other hand, the use of the term "*mixed*" preceding OSAHS output is aimed at pointing out to the significant number of central events detected, thus the clinician can perceive that although the patient has been diagnosed with OSAHS, there is a significant number of central events which may suggest further investigation about the pathological origin.

Syndrome classification according to patient's sleeping position is also performed. Information regarding the position of the patient is recommended in order to do not overestimate the severity of SAHS in patients with apneic events dependent on sleeping position (see Chapter 2 "*Contextual interpretation of apneic events*"). In this regard, once the existence of SAHS has been confirmed, in the case of being classified as obstructive, then it can be determined whether or not the syndrome is related to position by using the following criteria [**132**]:

$$OSAHS_{POSITIONAL} \Leftrightarrow \frac{AHI^{NS}}{AHI^{S}} \leq 0.5$$

$$OSAHS_{NON\_POSITIONAL} \Leftrightarrow 0.5 < \frac{AHI^{NS}}{AHI^{S}} \leq 2.0$$

$$OSAHS_{NON-SUPINE\_POSITIONAL} \Leftrightarrow \frac{AHI^{NS}}{AHI^{S}} > 2.0$$

Additional reported statistics comprise distribution of the different apneic events together with their classification with respect to the sleep states and the different sleeping positions, distribution of sleeping positions with respect to the sleep phases, distribution of EEG arousals with respect to the hypnogram and the of desaturation patterns classified according to their associated percentage of desaturation (< 2%, 2-3%, 3-4% and >4%).

Based on all the previous information, diagnostic module provides of a summary explanation of its output, with varying degrees of detail on the basis of the obtained results. In a first level, global syndrome characterization comprises final confirmation or discarding of the syndrome, and if it is the case, classification of its severity and type (obstructive, mixed obstructive, central, positional or non-positional) as described above. On a second explanation level, information on the identification and classification of each individual apneic event is provided. In this respect, each event is justified in terms of overall trust as a genuine apnea or hypopnea, as well as to the possibility of being considered a false positive; event classification is also provided in terms of the different degrees of membership achieved with regard to classification categories: obstructive, mixed and central. Finally, a third level comprises related statistics including hypnogram (both, continuous or classical epoch-based), sleep parameters (see above), transient EEG detected events and several other statistics of interest. Next section shows how this is information is available through the main user interfaces of the application.

A final note is intended with respect to the advantage of using fuzzy rules to implement system knowledge in relation with explanatory possibilities of the system. Indeed, as it has been mentioned throughout the text, the use of fuzzy logic allows us to express system knowledge using linguistic expressions in the form of IF-THEN fuzzy rules. This is because of the association of linguistic labels to each fuzzy set representing a category of data (see Chapter 4, "*Fuzzy inference systems*"). The former permits knowledge to be easily interpreted from the medical personnel. As an example, fuzzy rules implemented in relation with evaluation of a respiratory reduction in the thoracoabdominal channels are shown in Figure 5.45.

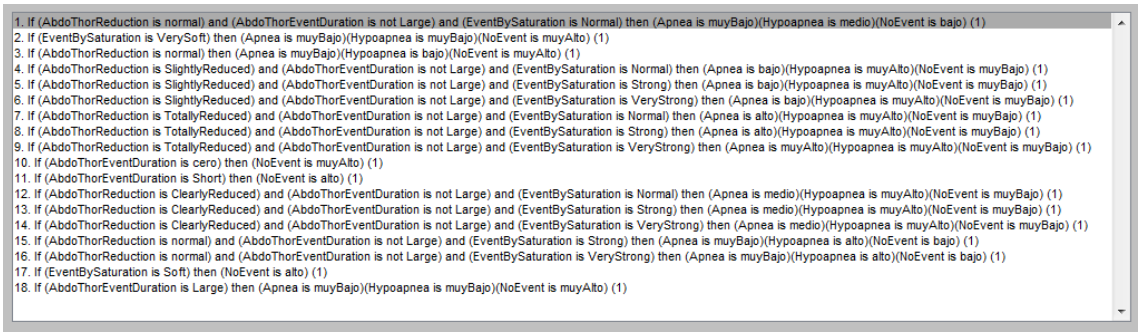


Figure 5.45. Example of linguistic IF-THEN rules implementing knowledge about the interpretation of a respiratory reduction (apneic interval) detected through thoracoabdominal channels

Visual evaluation of the rules from the clinician is then available which enables follow up of the inferential process. For example, in following Figure 5.46 system output tracking is performed for input values of thoracoabdominal related amplitude reduction of 77% with respect to baseline, duration of 28 seconds and associated desaturation degree of membership of 0.68 (see subsection "*Detection of apneic events*" of current chapter). In the figure, individual contribution of each rule to the output (apnea, hypopnea, false positive) can be evaluated.

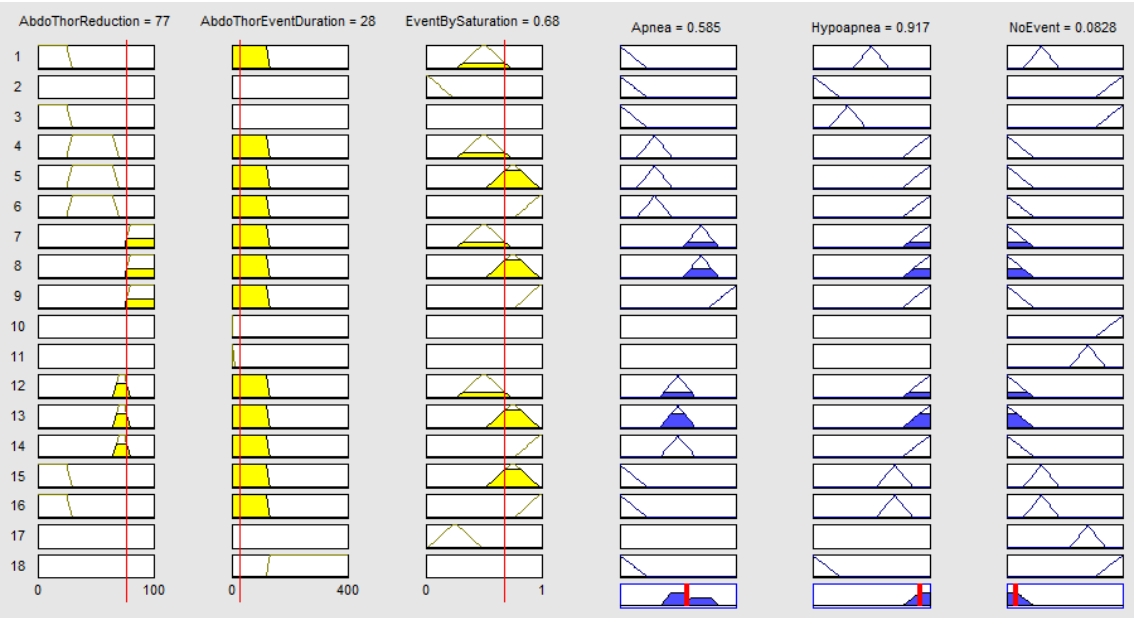


Figure 5.46. Tracking of the reasoning process of the system for the evaluation of an apneic event located in the thoracoabdominal respiratory channels. Associated amplitude reduction of the event is 77% with respect to baseline breathing, event duration is of 28 seconds and the degree of membership of its associated desaturation is 0.68. Obtained output possibilities according to the implemented rules in Figure 5.45 are apnea (0.58), hypopnea (0.92) and false positive (0.08)

The former allows the system to explain the results of its reasoning processes by tracking the set of activated linguistic rules according to a given input. The clinician can therefore evaluate the output of the system and decide on its validity. In addition, besides its intuitive configuration –and in case of high disagreement- the expert may ultimately insert new rules or adapt the existent ones in order to match his/her preferences.

5.10. Main user interfaces

Application's main user interface looks as in Figure 5.47.

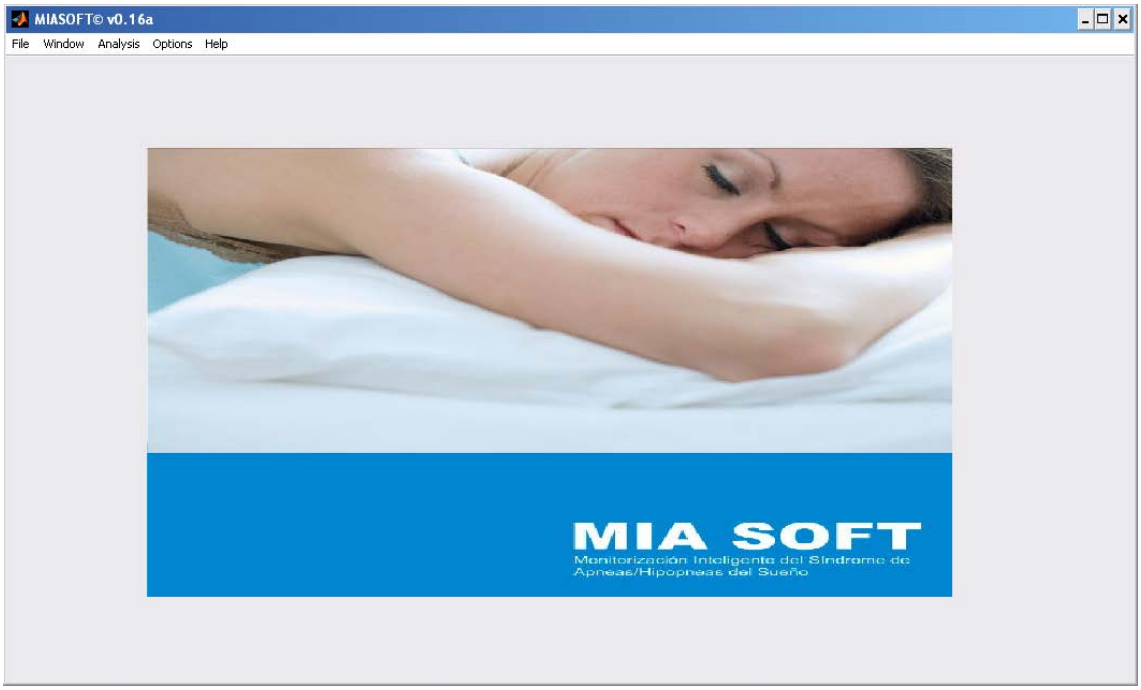


Figure 5.47. Systems main interface

Menu bar is scheduled according to several menus allowing the user to navigate through the different options:

Menu File: Controls different options regarding file input and output to the system. It includes the possibility to save an existing session, so that user configuration and analysis data can be saved and reloaded when desired, without the necessity to repeat the analysis.

Menu Window: Manages the active windows allowing selection for display and hiding of the different interfaces in a concrete instant of time.

Menu Analysis: Displays analysis options (see Figure 5.53 below).

Menu Options: User configuration options (Locale)

Menu Help: Help and copyright info.

Thus, once a PSG recording has been made at the hospital, resulting file can be imported to the system. The option "*File->Import EDF…*" leads the user to import a new PSG for analysis in the system through the acquisition submodule. Next screen allows the user to select the signals to be imported, as well as some configuration options, such as adaptation of sampling rate or dynamical range (normalization) of the input signals (see Figure 5.48).

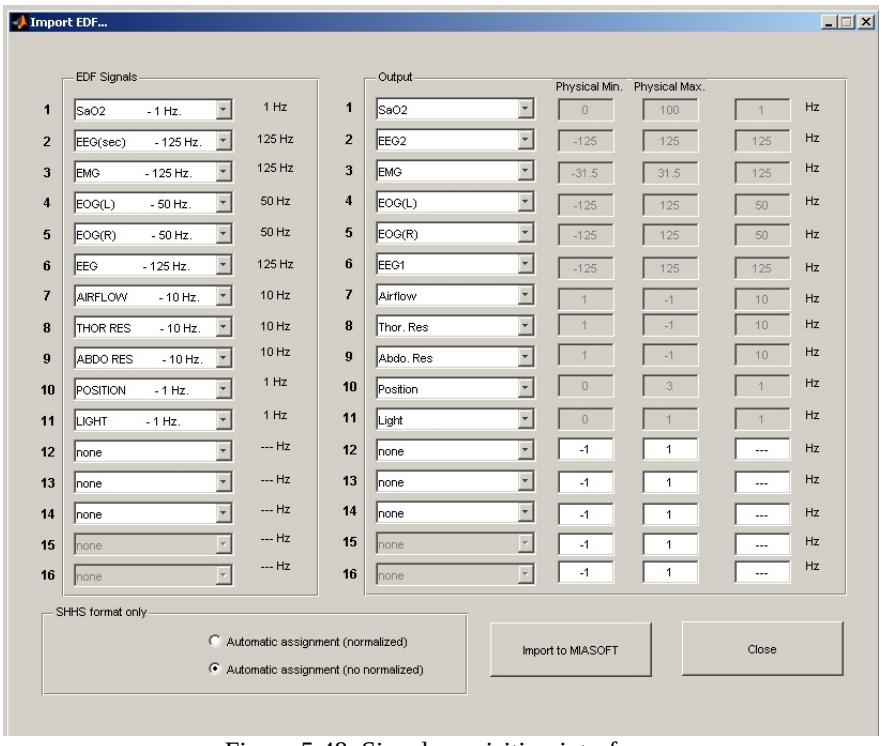


Figure 5.48. Signal acquisition interface

Once the desired configuration has been made, the signals are imported from the EDF file into the system for representation and analysis (see Figure 5.49).

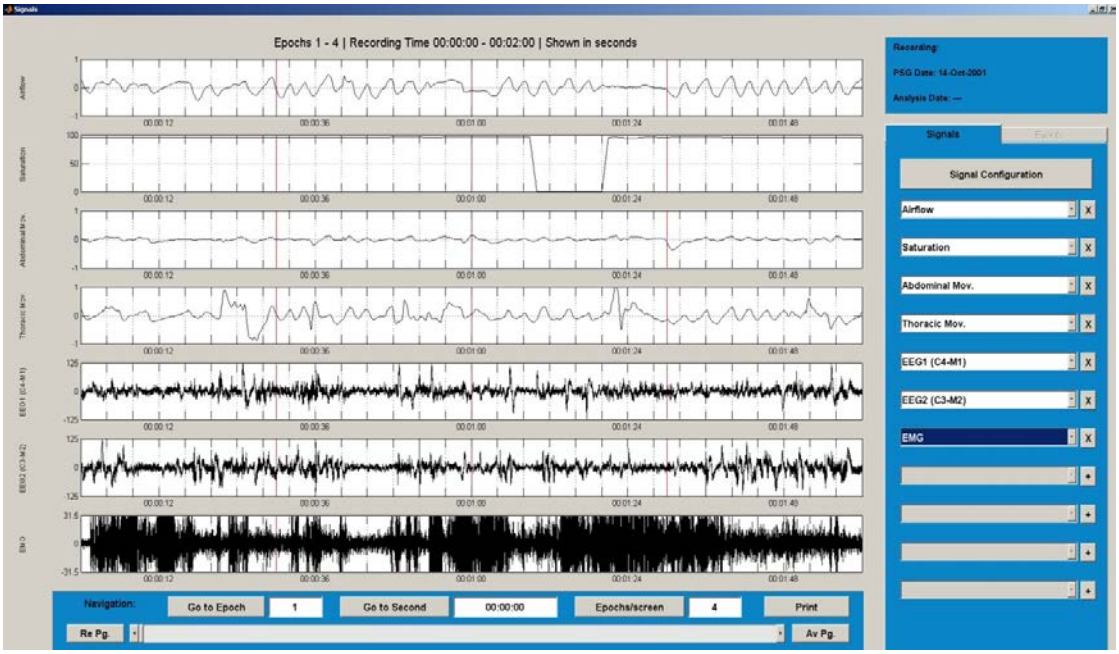


Figure 5.49. System's signal visualization interface

In Figure 5.49, on the right side of the screen there is a series of popup menus to select which signals the user wants to display and their order. The interface shows the corresponding signal intervals according to the selected time configuration display. In the navigation bar, below, the user can move along the full recording, forwards and backwards in time, or directly to go to a specific position by specifying the desired time or epoch. Signal's visualization options can be accessed by pressing "Signal Configuration" button (see Figure 5.50).

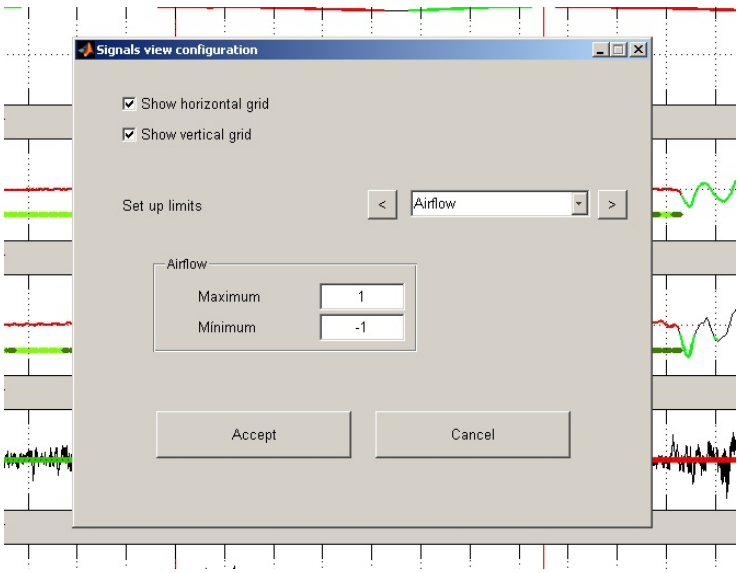


Figure 5.50. Signal visualization options window

Additional configuration options are accessible pressing the right button over signal's visualization interface:

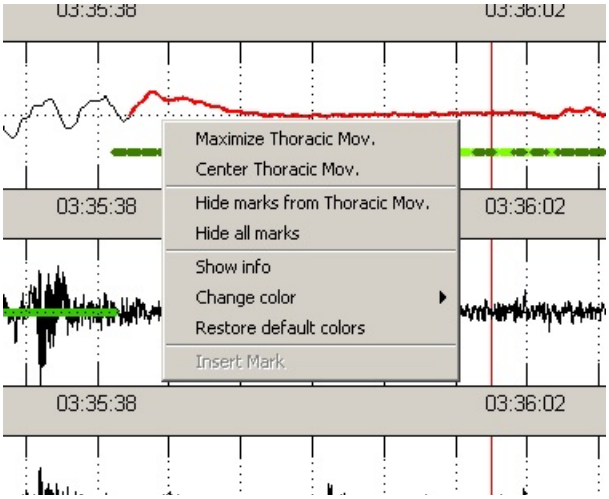


Figure 5.51. Visualization options accessible through the right button

Before the analysis starts, the configuration analysis window (*Analysis->Analysis configuration*) allows visualization and modification of the different analysis parameters (see Figure 5.52). For example, minimum required reductions and durations can be configured for different events.

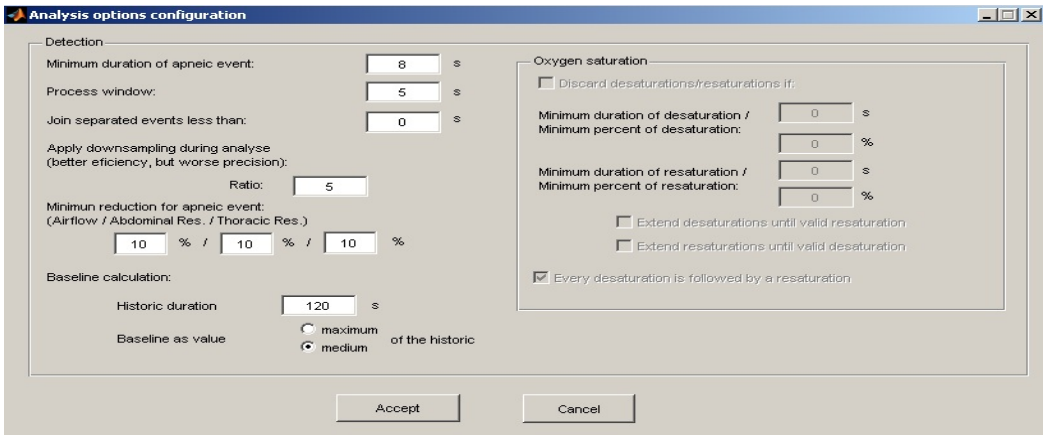


Figure 5.52. Analysis configuration window

After the user has selected the appropriate configuration, the analysis can be started through the "*Analysis*" menu. The program allows separate analysis of the signals involving the sleep function for the construction of the hypnogram (Sleep Analysis option), signals involving the respiratory activity (Respiratory Function Analysis option) or the possibility of making the full analysis involving chaining of the two

previous processes. Note that the respiratory function analysis depends on the sleep analysis, so if the user wants to analyze respiratory signals, sleep analysis has to be firstly performed.

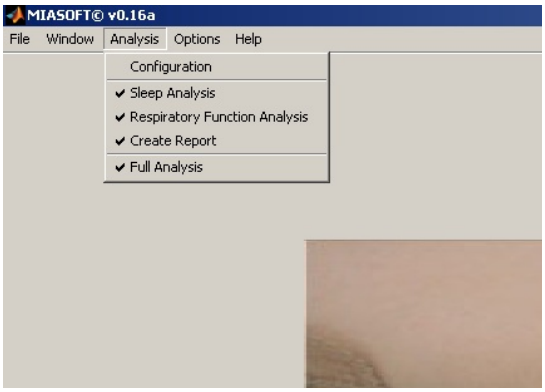


Figure 5.53. Menu showing main analysis options

Once the analysis has finished, the system allows visualization of the several significant events detected, for example, EEG arousals, oxygen desaturations and resaturations, respiratory flow reductions, increases or decreases in muscle activity, alterations in the respiratory movements, etc.

The system enables navigation through all these detected events on the "Event's navigation panel" (see Figure 5.54, on the right). By means of this panel it is possible to select the different types of events available for navigation by selecting the corresponding type of event on the popup menu. A list of detected events of the selected type is shown, and navigation is possible going forwards or backwards from the current event, or just by directly selecting the number of the desired event. Current selected event is marked by superimposition of a blue transparent box in the visualization interface. In case the event is composed by the aggregation of several individual events all the integrating events are marked in blue. At the same time summary information on event's classification is shown on the bottom right section. Classification information varies depending on the selected type of event. Resulting degrees of membership regarding current event's classification are shown both numerically and linguistically. Bar's plot graphically represents the assigned belief (degree of membership) on the corresponding linguistic scale (see Figure 5.54).

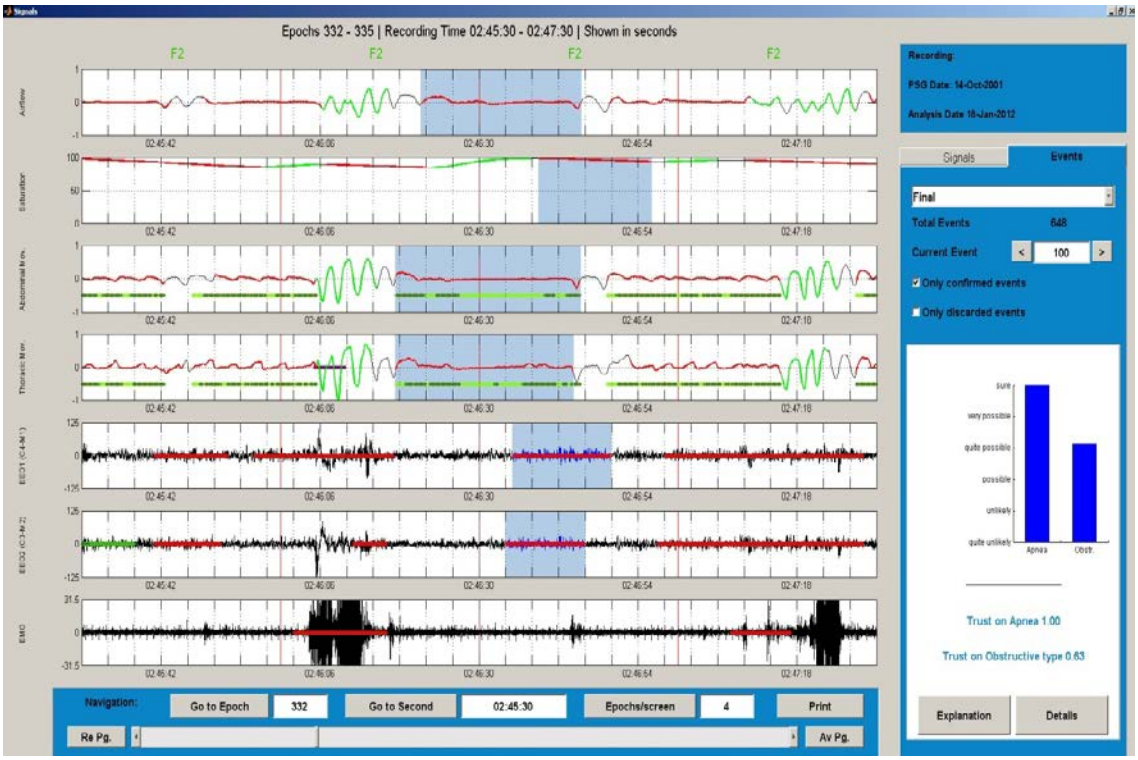


Figure 5.54. Signals visualization interface in which an event is selected for visualization. Integrating events composing the full diagnostic pattern are marked in blue. Summary information on event's characterization can be shown on the right bottom area of the window

Below the summary information, two additional buttons are available: "*Explanation*" and "*Details*". By pressing "*explanation*" button the user access to a text-like explanation of the selected event. This information shows the different evidences and conclusions from the reasoning process, which have been followed to classify the event into the different categories of relevance for the diagnosis (see Figure 5.55). Note that linguistic labels are used in order to describe the different evidences and conclusions on the possible categories for the event. This allows the clinician to evaluate the possibility of an alternative classification.

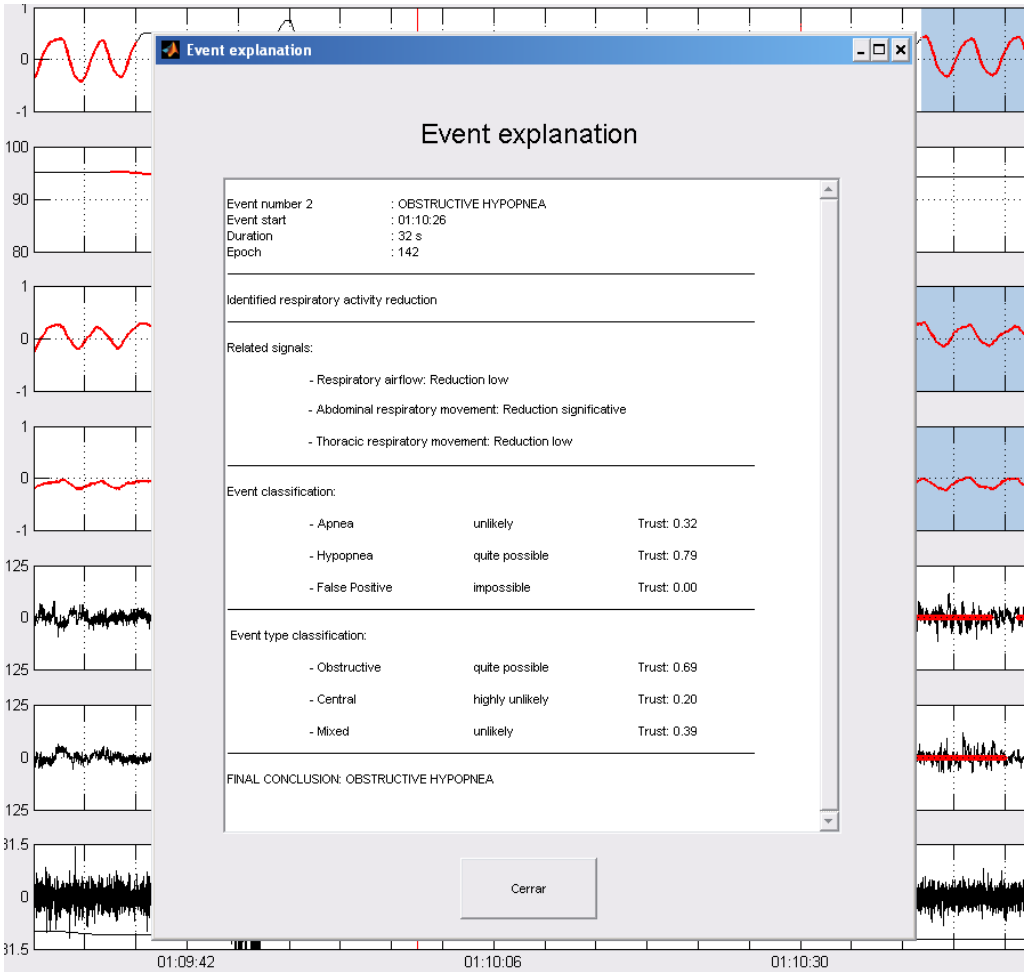


Figure 5.55. Event's explanation window

On the other hand, by pressing "*Details*" button one can access advanced information with more details and quantification of relevant information (see Figure 5.56).

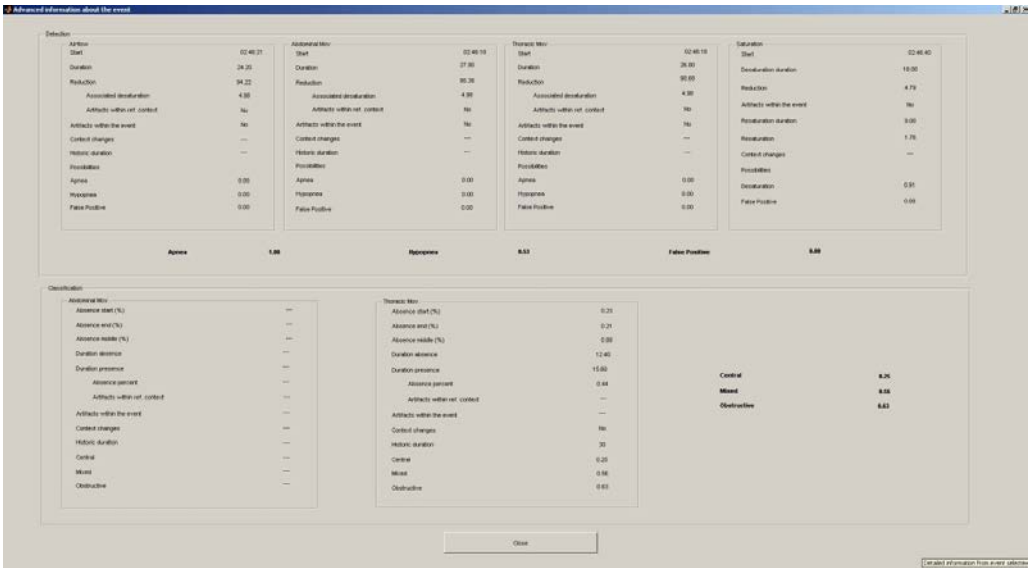


Figure 5.56. Advanced event's information interface

If taking into account all the quantitative and qualitative information provided by the system, the clinician decides to modify system's classification, it can be done as illustrated in Figure 5.57. By dragging the mouse over the corresponding recording interval (red box) the clinician can insert a new event or modify classification of a pre-existing event detected by the system (blue box).

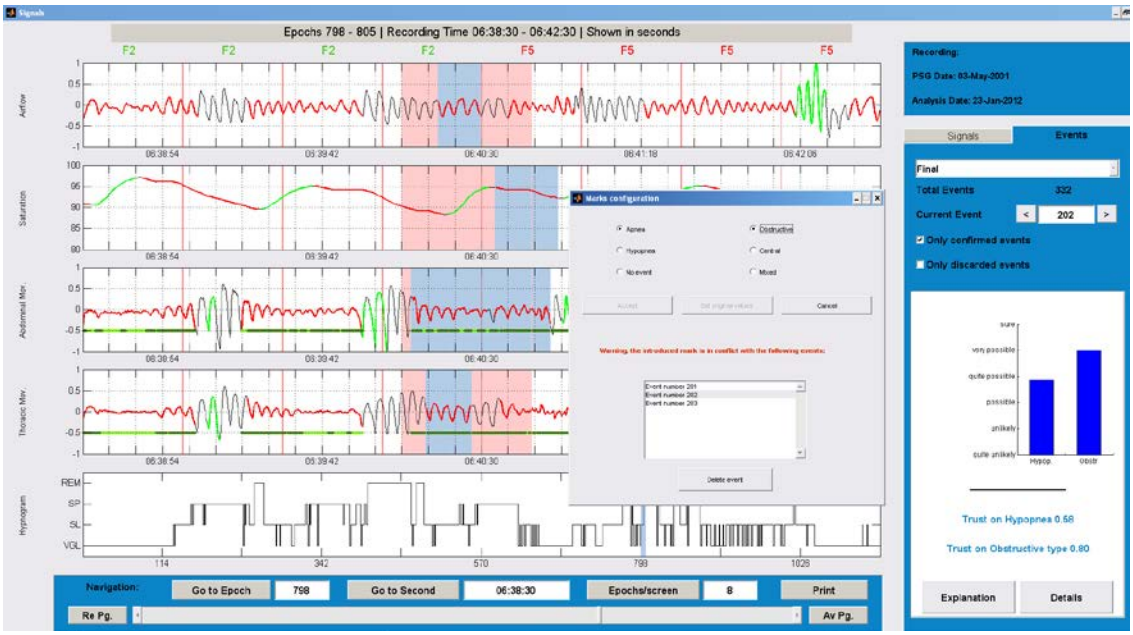


Figure 5.57. Modification of system's classification

In addition, once the analysis of the signals has finalized, the system provides of a report containing summary information and statistics on all the detected events as well as diagnosis information. This option is accessible through the menu "*Window->Final report*". Within this form it is possible to navigate through the different compounding panels that structure the information into four categories.

1. The tab "*report*" shows a general summary regarding patient's diagnosis. Main diagnostic indexes can be observed as well as the total number of events classified according to their type (see Figure 5.58).

Figure 5.58. Final report window: summary diagnosis tab

2. The tab "*advanced statistics*" shows more detailed statistics about event's classification. It includes, for example, statistics on the different sleep stages distribution, classification of apneic patterns regarding patient's sleeping position and sleep stage, indexes of positional SAHS or arousals distribution. The panel also contains information on significant sleep statistics such as total sleep time, lights on/off clock time, sleep efficiency, sleep latency, REM stage latency or waking after sleep onset (see Figure 5.59).

Figure 5.59. Final report window: Advanced statistics

3. Event's report tab shows a list with all the detected events and its localization within the PSG recording. The user can click on any event in the list and access to its classification results and to an explanation on the reasoning process. The button "*Go to event*" can also be clicked to go directly over the visualization interface and show the signal trends involved in the classification of the event (see Figure 5.60).

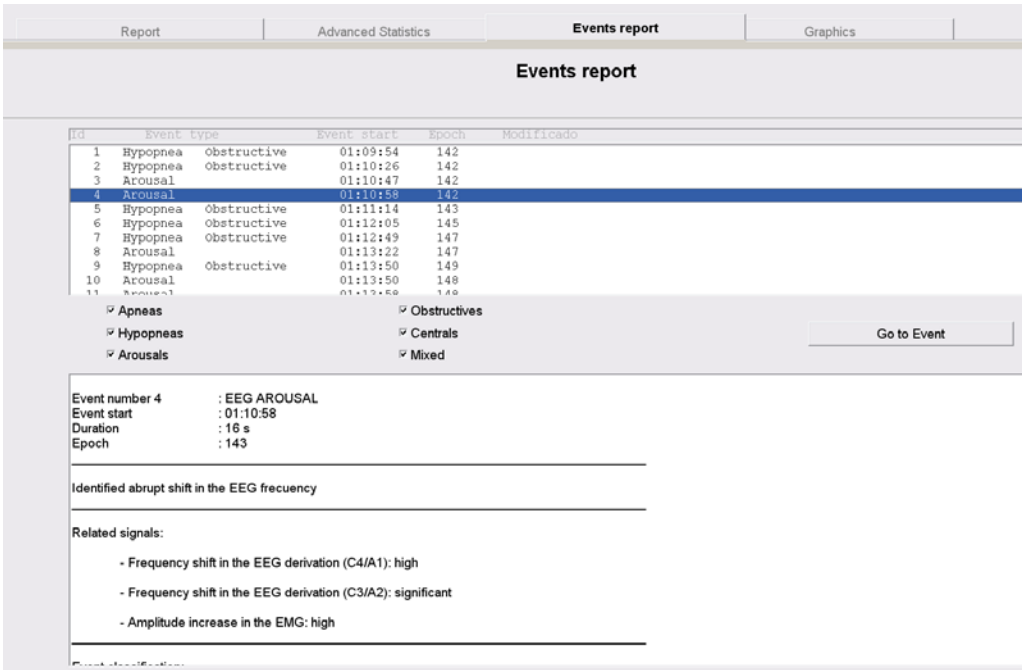


Figure 5.60. Final report window: Event's report

4. In the tab "*Graphics*" the user can access to a series of graphics of interest for the diagnosis, including patient's hypnogram, body position throughout the recording, oxygen saturation evolution and time distribution of the different apneic events and EEG arousals (see Figure 5.61).

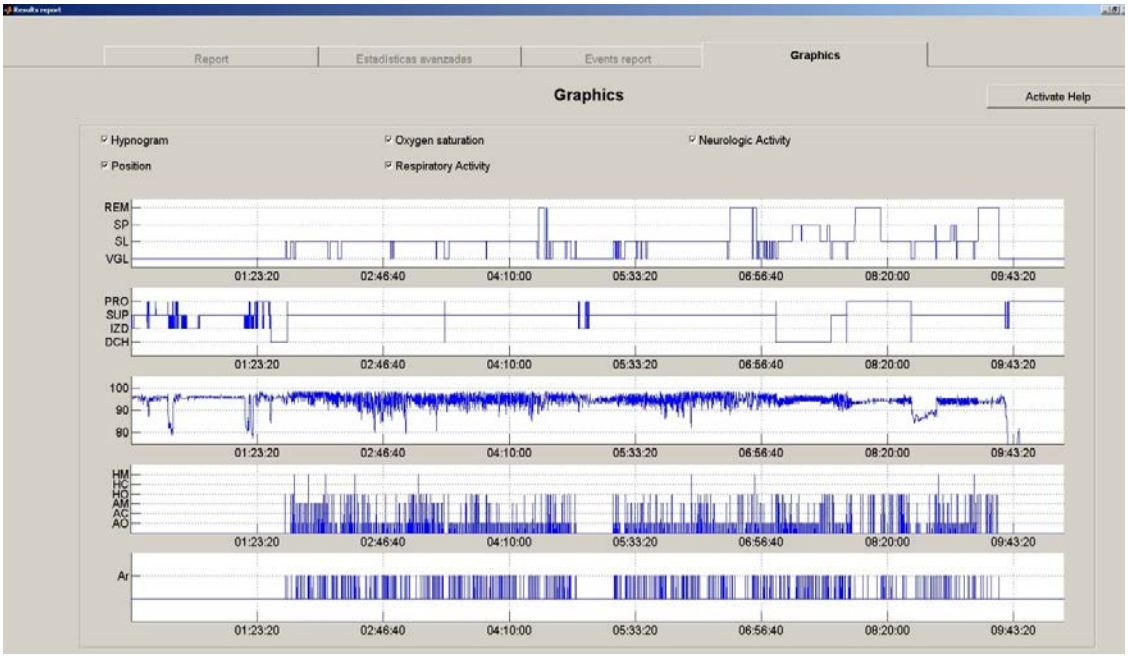


Figure 5.61. Final report window: Graphics

5.11. Summary of this chapter

The contents of this chapter constitute the main part of the manuscript in which description of the proposed system is performed. Description includes the establishment of a developing process, analysis of the corresponding requirements, presentation of systems' architecture and functional description of its integrating parts.

Before carrying out the building of any system, it is necessary to perform an analysis assessing its adequateness as a solution for the problem for which its construction has been scheduled. Having that in mind, throughout the previous chapters there have been enunciated the shortages of the current existent systems. In this respect the proposed system is aimed at improving the PSG analysis through a comprehensive approach that introduces artificial intelligence techniques in the analysis task. Moreover

the economic viability of the system has been justified already, since automatic analysis of the PSG involves a great saving in terms of time and effort for the clinical personnel.

The next step within the engineering method requires the establishment of a developing process that takes into account user necessities. Given the nature of the domain, an evolutionary developing model is proposed. Within this model, but also as a mandatory step over almost all the engineering processes, it is necessary to set up the corresponding requirements of the system. These have been divided into *functional requirements* and *non-functional requirements*. Within the *functional requirements* are gathered all those requirements related to the tasks the system can perform: loading of digital polysomnographic recordings, neurophysiological signals analysis, hypnogram generation, respiratory signals analysis, detection and classification of apneic events, calculation of significant parameters or elaboration of the final reports. On the other hand, within the *non-functional requirements* usability requirements are mostly included, such as ease of use, user-friendly interaction with the clinical user, structured presentation of the results or system's flexibility.

Architecture of the system is subsequently described. Such architecture is highly modular attending to the necessities of the developing process, and it is guided by the previously established requirements. Main functionalities on each one of the integrating modules are also outlined at this point.

The chapter continues going in depth into the functional description of each one of the analysis phases. Following a methodological perspective, the description starts by introducing inputs to the system. In this part the procedure for the signal acquisition is analyzed for which a standard EDF file is read which contains the digitalized signals. An outline of the artifact detection strategy is given which is performed in a decentralized form, as a function of each signal type and over different processing levels. Therefore in the proposed system this task is spread out among its different compounding parts, and concrete mechanisms for artifact detection are detailed throughout the individual description of the different modules that integrate the system.

Functional description of the system continues detailing the processes for the analysis of the neurophysiological signals related with patient's sleep. In this respect the

algorithms for the detection of *EEG arousals* are described first, in which the applicability of several machine learning models is studied. The objective is to choose the best model to perform an analysis of the extracted features from the EEG, and its posterior classification for the detection of the arousal. This study is followed by another one trying to reduce the number of necessary features for the detection of the EEG arousal. Several methods for feature selection are compared at this respect. Afterwards, description of the algorithms used for the detection of additional transient events at the microstructure level is performed. Main interest is in the detection of *sleep spindles* and *K-complexes*.

As the final stage in the analysis of the neurophysiological activity, it takes place the construction of the sleep map of the patient, the so-called hypnogram. To that end an analysis of the dominant frequencies in the EEG is performed. Besides, the method for the characterization of muscle tone in the EMG and the localization of ocular movements in the EOG is described. All this information, together with the transient events previously detected –sleep spindles and K-complexes- is used as the input to a fuzzy inference system that, as its output, characterizes the sleep state of the patient in a continuous manner. After applying some post-processings over the continuous output, the discrete hypnogram of the patient is obtained as well, according to four different sleep states: awake, drowsy sleep (phases N1 and N2), deep sleep (N3) and REM sleep.

The analysis of the respiratory signals takes place subsequently for the identification of *apneic intervals*. This analysis involves the signals of airflow and respiratory movements in the thoracoabdominal channels. The apneic intervals are localized as intervals of reduction with respect to the normal amplitude of the signal. Such intervals are then detected and quantified. Likewise, processing of the oxygen saturation signal comprises the localization of descending intervals (*possible desaturations*) and ascending intervals (*possible resaturations*) in that signal. Similarly, these intervals are marked and quantified. Preceding these analyses, however, a preprocessing of the respiratory signals is performed with specific artifact rejection consisting in the localization of signal overflow intervals or with a focus loss. Specific algorithms used for the detection and the processing of this kind of artifacts are also described in this section.

Once all the respiratory signals have been analyzed and the patient's hypnogram has been obtained, a temporal correlation stage among the detected events occurs. Its objective is the characterization of *apneic patterns*. An apneic pattern is defined as a group of individually detected events in the respiratory signals that, correctly related, determine the possible existence of an actual apneic event in the corresponding time interval. In order to decide on its veracity and, in case of confirmation, on the concrete type of apneic event, the detected apneic patterns are introduced into a new fuzzy inference system. As the output, and for each identified pattern, a degree of membership is obtained that indicates the possibility for the pattern to be classified as an *apnea*, a *hypopnea*, or to be discarded as a *false positive*. Additionally, a linguistic label is assigned to each numeric value of membership which describes it in terms of natural language. This process is described in the section regarding the detection of apneic events.

Finally, over those events which have been confirmed as truly apneic, their classification as obstructive, central or mixed is performed. Classification is achieved by means of an analysis of the thoracic and abdominal movements in the interval where the event has been marked. Similarly classification is carried out by using a fuzzy inference system modeled for this specific purpose.

As the final step in the analysis, all the previous generated data is taken into account to elaborate the pertinent reports. In this regard the system calculates the all the relevant parameters in order to evaluate patient's severity and issue the proper diagnosis. Last part of the chapter explains output parameters and describes main interfaces of the user application.

5.12. References

[1] I. Sommerville, *Ingeniería del Software*, 6th ed.: Pearson Educación, 2002.

[2] RS. Pressman, *Ingeniería del software: un enfoque práctico*, 6th ed.: McGraw-Hill, 2006.

[3] P. Rook, "Software Development Process Models," in *Software Reliabitlity Handbook*, Paul Rook, Ed.: Elsevier Science Publishers, 1990, ch. Apendix B, pp. 413-440.

[4] B. Boehm, "A spiral model for software development enhancement," *Computer*, vol. 3, no. 7, pp. 61-72, 1988.

[5] B. Boehm, "The spiral model as a tool for evolutionary software acquisition," *CrossTalk*, vol. 14, no. 5, pp. 4-11, 2001.

[6] RJ. Thomas, "Arousals in sleep-disordered breathing: patterns and implications," *Sleep*, vol. 26, no. 8, pp. 1042-1047, 2003.

[7] C. Iber, S. Ancoli-Israel, A. Chesson, and SF. Quan, "The AASM Manual for the scoring of sleep and associated events: rules, terminology and technical specifications," American Academy of Sleep Medicine, Westchester, IL, 2007.

[8] C. George, T. Millar, and M. Kryger, "Identification and quantification of apneas by computer-based analysis of oxygen saturation," *The American Review of Respiratory Diseases*, vol. 137, pp. 1238-1240, 1988.

[9] M. Dorfman and R. Thyler, *Software requirements engineering*., 1990.

[10] B. Kemp, A. Värri, AC. Rosa, KD. Nielson, and J. Grade, "A simple format for exchange of digitalized polygraphic recordings," *Electroencephalography and Clinical Neurophysiology*, vol. 82, pp. 391-393, 1992.

[11] European Data Format. [Online]. http://www.edfplus.info

[12] A. Schlögl, C. Brunner, R. Scherer, and A. Glatz, "BioSig - an open source software library for BCI research," in *Towards Brain-Computer Interfacing*, G. Dornhege et al., Eds.: MIT Press, 2007, pp. 347-358.

[13] A. Rechtschaffen and A. Kales, "A manual of standardised terminology techniques and scoring system for sleep stages of human subjects," U.S. Department of Health, Education and Welfare Public Health Service – NIH/NIND, 1968.

[14] R.D. Cartwright, F. Diaz, and S. Lloyd, "The effects of sleep posture and sleep stage on apnea frequency," *Sleep*, vol. 14, no. 4, pp. 351-353, 1991.

[15] S. Redline et al., "The effects of age, sex, ethnicity, and sleep-disordered breathing on sleep architecture," *Archives of Internal Medicine*, vol. 164, pp. 406-418, 204.

[16] DJ. Pitson and JR. Stradling, "Autonomic markers of arousal during sleep in patients undergoing investigation for obstructive sleep apnoea, their relationship to EEG arousals, respiratory events and subjective sleepiness," *Journal of Sleep Research*, vol. 7, pp. 53-59, 1998.

[17] SP. Cho, J. Lee, HD. Park, and KJ. Lee, "Detection of arousal in patients with respiratory sleep disorders using a single channel EEG," in *27th IEEE Engineering in Medicine and Biology Annual Conference*, 2005, pp. 2733-2735.

[18] P. Gouveia, R. Oliveira, and R. Rosa, "Sleep Apnea related micro arousal detection with EEG analysis," in *7th Portuguese Conference on Biomedical Engineering*, 2003.

[19] S. Tesler et al., "Progressive detrended fluctuation analysis and other numerical methods applied on sleep ECG for sleep stage recognition and arousal detection," *Journal of Sleep Research*, vol. 13, 2004.

[20] G. Pillar et al., "An automatic ambulatory device for detection of AASM defined arousals from sleep: The WP100," *Sleep Medicine*, vol. 4, no. 3, pp. 207-212, 2003.

[21] G. Pillar et al., "Autonomic arousal index: an automated detection based on peripheral arterial tonometry," *Sleep*, vol. 25, no. 5, pp. 541-547, 2002.

[22] A. Agarwal, "Automatic detection of micro-arousals," in *27th IEEE Engineering in Medicine and Biology Annual Conference* , China, 2005, pp. 1158-1161.

[23] F. De Carli, L. Nobili, P. Gelcich, and F. Ferrillo, "A method for the automatic detection of arousals during sleep," *Sleep*, vol. 22, no. 5, pp. 561-572, 1999.

[24] T. Sugi, F. Kawana, and M. Nakamura, "Automatic EEG arousal detection for sleep apnea syndrome," *Biomedical Signal Processing and Control*, vol. 4, no. 4, pp. 329-337, 2009.

[25] O. Shmiel, T. Shmiel, Y. Dagan, and M. Teicher, "Data mining techniques for detection of sleep arousals," *Journal of Neuroscience Methods*, vol. 179, no. 2, pp. 331-337, 2009.

[26] MH. Asyali, RB. Berry, MCK. Khoo, and A. Altinok, "Determining a continuous marker for sleep depth," *Computers in Biology and Medicine*, vol. 37, pp. 1600-1609, 2007.

[27] D. Alvarez-Estevez and V. Moret-Bonillo, "Fuzzy reasoning used to detect apneic events in the Sleep Apnea-Hypopnea Syndrome," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7778-7785, 2009.

[28] Á. Fernández-Leal, V. Moret-Bonillo, and E. Mosqueira-Rey, "Causal temporal constraint networks for representing temporal knowledge," *Expert Systems with Applications*, vol. 36, no. 1, pp. 27-42, 2009.

[29] RA. Fisher, "The use of multiple measurement in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179-188, 1936.

[30] D. Michie, DJ. Spiegelhalter, and CC. Taylor, *Machine Learning: neural and statistical classification*, Ellis Horwood, Ed. Chichester, UK, 1994.

[31] N. Cristianini and J. Shawe-Taylor, *An introduction to Support Vector Machines*. Cambridge, UK: Cambridge University Press, 2000.

[32] C. Cortés and VN. Vapnik, "Support-Vector Machines," *Machine Learning*, vol. 20, 1995.

[33] BE. Boser, IM. Guyon, and VN. Vapnik, "A training algorithm for optimal margin classifiers," in *5th Annual ACM Workshop on COLT*, Pittsburgh, PA, 1992, pp. 144-152.

[34] SS. Keerthi and CJ. Lin, "Asymptotic behaviours of support vector machines with Gaussian kernel," *Neural Computation*, vol. 15, no. 7, pp. 1667-1689, 2003.

[35] HT. Lin and CJ. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," Department of Computer Science, National Taiwan University, 2003.

[36] VN. Vapnik, *The nature of statistical learning theory*. New York, NY: Springer-Verlang, 1995.

[37] JC. Príncipe, N. Euliano, and C. Lefebvre, *Neural Systems: fundamentals through simulations*.: John Wiley, 2000.

[38] MF. Moller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, pp. 525-533, 1993.

[39] E. Alpaydin, *Introduction to Machine Learning*.: MIT Press, 2004.

[40] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[41] JS. Aguilar-Ruiz, F. Azuaje, and JC. Riquelme, "Data mining approaches to diffuse large B-Cell lymphoma gene expression data interpretation," in *Lecture Notes in Computer Science*. Zaragoza: Springer-Verlang, 2004, pp. 279-288.

[42] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273-324, 1997.

[43] U. Fayyad and K. Irani, "Multi-internal discretization of continuous-valued attributes for classification learning," in *13th International Joint-Conference on Artificial Intelligence*, 1993, pp. 1022-1029.

[44] K. Kenji and L. Rendell, "A practical approach to feature selection," in *9th International Workshop on Machine Learning*, 1992, pp. 249-256.

[45] T. Mitchell, *Machine Learning*.: McGraw-Hill Companies Inc., 1997.

[46] J. Quinlan, *C4.5: Programs for machine learning*.: Morgan Kaufmann Publishers, 1993.

[47] M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *17th International Conference on Machine Learning*, San Francisco, CA, 2000, pp. 359-366.

[48] R. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63-90, 1993.

[49] A. Shabtai, RY. Moskovitch, and C. Glezer, "Detection of malicious code by applying machine learning classifiers on static features: a state-of-the-art survey," *Information Security Technical Report*, vol. 14, no. 1, 2009.

[50] L. De Gennaro and M. Ferrara, "Sleep spindles: an overview," *Sleep Medicine Reviews*, vol. 7, no. 5, pp. 423-440, 2003.

[51] AL. Loomis, EN. Harvey, and G. Hobart, "Potential rhythms of the cerebral cortex during sleep," *Science*, vol. 81, pp. 597-598, 1935.

[52] H. Berger, "Über das Elektroencephalogram des Menschen. Sechste Mitteilung," *Arch Psychiatr Nervenkr*, vol. 99, pp. 555-574, 1933.

[53] M. Steriade, "Coherent oscillations and short-term plasticity in corticothlamic networks," *Trends Neurosci*, vol. 22, pp. 337-345, 1999.

[54] DJ. Dijk, B. Hayes, and CA. Czeisler, "Dynamics of electroencephalographic sleep spindles and slow wave activity in men: effect of sleep deprivation," *Brain Res*, vol. 626, pp. 190-199, 1993.

[55] SL. Himanen, J. Virkkala, H. Huhtala, and J. Hasa, "Spindle frequencies in sleep EEG show U-shape within first four NREM sleep episodes," *Journal of Sleep Research*, vol. 11, pp. 35-42, 2002.

[56] WR. Jankel and E. Niedermeyer, "Sleep spindles," *Journal of Clinical Neurophysiology*, vol. 2, pp. 1-35, 1985.

[57] E. Werth, P. Achermann, DJ. Dijk, and AA. Borbely, "Spindle frequency activity in the sleep EEG: individual differences and topographic distribution," *Electroencephalography and Clinical Neurophysiology*, vol. 103, pp. 535-542, 1997.

[58] M. Jober, E. Poiseau, P. Jahnig, H. Schulz, and S. Kubicki, "Topographical analysis of sleep spindle density," *Neuropsychobiology*, vol. 26, pp. 210-217, 1992.

[59] DJ. Dijk, B. Hayes, and DA. Czeisler, "Dynamics of electroencephalographic sleep spindles and slow wave activity in men: effect of sleep deprivation," *Brain Research*, vol. 626, pp. 190-199, 1993.

[60] S. Uchida, T. Maloney, DJ. March, R. Azari, and I. Feinberg, "Sigma (12-15 Hz) and Delta (0.3-3 Hz) EEG oscillate reciprocally within NREM sleep," *Brain Research Bulletin*, vol. 27, pp. 93-96, 1991.

[61] JC. Principe and JR. Smith, "Sleep spindle characteristics as a function of age," *Sleep*, vol. 5, no. 1, pp. 73-84, 1982.

[62] A. Nicolas, D. Petit, S. Rompre, and J. Montplaisir, "Sleep spindle characteristics in healthy subjects of different age groups," *Clinical Neurophysiology*, vol. 112, pp. 521-527, 2001.

[63] SL. Himanen, J. Virkkala, E. Huupponen, and J. Hasan, "Spindle frequency remains slow in sleep apnea patients throughout the night," *Sleep Medicine*, vol. 4, pp. 361-366, 2003.

[64] B. Ondze, F. Espa, Y. Dauvilliers, M. Billiard, and A. Besset, "Sleep architecture, slow wave activity and sleep spindles in mild sleep disordered breathing," *Clinical Neurophysiology*, vol. 114, pp. 867-874, 2003.

[65] K. Campbell, A. Kumar, and W. Hofman, "Human and automatic validation of a phase locked loop spindle detection system," *Electroencephalography and Clinical Neurophysiology*, vol. 48, pp. 602-605, 1980.

[66] R. Broughton, T. Healey, J. Maru, D. Green, and B. Pagurek, "A phase locked loop device for automatic detection of sleep spindles and stage 2," *Electroencephalography and Clinical Neurophysiology*, vol. 44, no. 5, pp. 677-680, 1978.

[67] A. Kumar, "The complex demodulation method for detection of alpha waves and sleep spindles of the human eeg in rea-time," in *International Conference on Advances of Signal Processing Techniques*, Lausanne, 1975.

[68] AC. Declerck, WLJ. Martens, and W. Wauquier, "Sleep spindle detection and its clinical relevance," *European Neurology*, vol. 25, pp. 56-60, 1986.

[69] M. Jobert, E. Poiseau, P. Jahnig, H. Schulz, and S. Kubicki, "Pattern recognition by matched filtering: an analysis of sleep spindle and K-complex density under the influence of lormatezapam and zopiclone," *Neuropsychobiology*, vol. 26, pp. 100-107, 1992.

[70] E. Huupponen et al., "Autoassociative MLP in sleep spindle detection," *Journal of Medical Systems*, vol. 24, pp. 183-193, 2000.

[71] T. Shimada, T. Shiina, and Y. Saito, "Detection of characteristic waves of sleep EEG by neural network analysis," *IEEE Transactions on Biomedical Engineering*, vol. 47, pp. 369-379, 2000.

[72] EM. Ventouras et al., "Sleep spindle detection using artificial neural networks trained with filtered time-domain EEG: A feasibility study," *Computer Methods and Programs in Biomedicine*, vol. 78, pp. 191-207, 2005.

[73] SV. Shonwald, GJL. Gerhardt, EL. Santa-Helena, and MLF. Chaves, "Characteristics of human EEG sleep spindles assessed by Gabor transform," *Physica*, vol. 327, pp. 180-184, 2003.

[74] A. Akin and T. Akgul, "Detection of sleep spindles by discrete wavelet transform," in *24th Annual IEEE Northeast Conference on Bioengineering* , Hershey, PA, USA, 1998, pp. 15-17.

[75] J. Zygierewicz et al., "High resolution study of sleep spindles," *Clinical Neurophysiology*, vol. 110, pp. 2136-2147, 1999.

[76] E. Olbrich and P. Achermann, "Analysis of oscillatory patterns in the human sleep EEG using a novel detection algorithm," *Journal of Sleep Research*, vol. 14, pp. 337-346, 2005.

[77] E. Huupponen et al., "Development and comparison of four sleep spindle detection methods," *Artificial Intelligence in Medicine*, vol. 40, pp. 157-170, 2007.

[78] LB. Ray, SM. Fogel, CT. Smith, and KR. Peters, "Validating an automated sleep spindle detection algorithm using an individualized approach," *Journal of Sleep Research*, vol. 19, pp. 374-378, 2010.

[79] AL. Loomis, EN. Harvey, and G. Hobart, "Distribution of disturbance patterns in the human electroencephalogram, with special reference to sleep," *Journal of Neurophysiology*, vol. 13, pp. 231-256, 1938.

[80] M. Roth, J. Shaw, and J. Green, "The form, voltage distribution and physiological significance of the K-complex," *Electroencephalography and Clinical Neurophysiology*, vol. 8, pp. 385-402, 1956.

[81] Y. Niiyama, N. Satoh, O. Kutsuzawa, and Y. Hishikawa, "Electrophysiological evidence suggesting that sensory stimuli of unknown origin induce spontaneous K-complexes," *Electroencephalography and Clinical Neurophysiology*, vol. 98, pp. 394-400, 1996.

[82] P. Halász, I. Pál, and P. Rajna, "K-Complex formation in the EEG in sleep. A survey and new examinations," *Acta Physiologica Hungarica*, vol. 65, pp. 3-35, 1985.

[83] F. Amzica and M. Steriade, "The functional significance of K-complexes," *Sleep Medicine Reviews*, vol. 6, pp. 139-149, 2002.

[84] F. Amzica and M. Steriade, "The K-complex: its slow rhythmicity and relation to delta waves," *Neurology*, vol. 49, pp. 952-959, 1997.

[85] IM. Colrain, "The K-Complex: a 7-decade history," *Sleep*, vol. 28, no. 2, pp. 255-273, 2005.

[86] G. Bremer, JR. Smith, and I. Karacan, "Automatic detection of the K-complex in sleep electroencephalograms," *IEEE Transactions on Biomedical Engineering*, vol. 17, pp. 14-23, 1970.

[87] AC. Da Rosa, B. Kemp, T. Paiva, FH. Lopes da Silva, and HAC. Kamphuisen, "A model-based detector of vetex waves and K complexes in sleep electroencephalogram," *Electroencephalography and clinical neurophysiology*, vol. 78, pp. 71-79, 1991.

[88] IN. Bankman, VG. Sigillito, RA. Wise, and PL. Smith, "Feature-based detection of the K-complex wave in the human electroencephalogram using neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 39, pp. 1305-1310, 1992.

[89] C. Richard and R. Lengelle, "Joint time and time-frequency optimal detection of K-Complexes in sleep EEG," *Computer and Biomedical Research*, vol. 31, pp. 209-229, 1998.

[90] BH. Jansen and PR. Desai, "K-complex detection using multi-layer perceptrons and recurrent networks," *International Journal of Biomedical Computing*, vol. 37, pp. 249-257, 1994.

[91] PA. Estévez et al., "Polysomnographic pattern recognition for automated classification of sleep-waking states in infants," *Medical and Biological Engineering and Computing*, vol. 40, pp. 105-113, 2002.

[92] CA. Holzmann et al., "Expert-system classification of sleep/waking states in infants," *Medical and Biological Engineering and Computing*, vol. 37, pp. 466-476, 1999.

[93] JC. Principe, SK. Gala, and TG. Chang, "Sleep staging automaton based on the theory of evidence," *IEEE Transactions on Biomedical Engineering*, vol. 36, pp. 503-509, 1989.

[94] Y. Kurihara, K. Watanabe, K. Kobayashi, and H. Tanaka, "Observer based on body movement information in sleeping and estimation of sleep stage appearance probability," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 3, pp. 688-695, 2008.

[95] B. Kemp, EW. Gröneveld, AJMW. Janssen, and JM. Franzen, "A model-based monitor of human sleep stages," *Biological Cybernetics*, vol. 57, pp. 365-378, 1987.

[96] J. Pardey, S. Roberts, L. Tarassenko, and J. Stradling, "A new approach to the analysis of the human sleep/wakefulness continuum," *Journal of Sleep Research*, vol. 5, pp. 201-210, 1996.

[97] S. Roberts and L. Tarassenko, "New method of automated sleep quantification," *Medical and Biological Engineering and Computing*, vol. 30, pp. 509-517, 1992.

[98] L. Fraiwan et al., "Classification of sleep stages using multi-wavelet time frequency entropy and LDA," *Methods of Information in Medicine*, vol. 49, pp. 230-237, 2010.

[99] T. Penzel and R. Conradt, "Computer based sleep recording and analysis," *Sleep Medicine Reviews*, vol. 4, no. 2, pp. 131-148, 2000.

[100] SL. Himanen and J. Hasan, "Limitations of Rechtschaffen and Kales," *Sleep Medicine Reviews*, vol. 4, no. 2, pp. 149-167, 2000.

[101] H. Schulz, "Rethinking sleep analysis," *Journal of Clinical Sleep Medicine*, vol. 4, pp. 99-103, 2008.

[102] J. Hasan, "Past and future of computer-assisted sleep analysis and drowsiness assessment," *Journal of Clinical Neurophysiology*, vol. 13, pp. 295-313, 1996.

[103] F. De Carli et al., "Quantitative analysis of sleep EEG microstructure in the time-frequency domain," *Brain Research Bulletin*, vol. 63, pp. 399-405, 2004.

[104] P. Halasz, "Hierarchy of micro-arousals and the microstructure of sleep," *Clinical Neurophysiology* , vol. 28, pp. 461-475, 1998.

[105] B. Kemp, "A proposal for computer-based sleep/wake analysis," *Journal of Sleep Research*, vol. 2, pp. 179-185, 1993.

[106] BH. Choi et al., "Non-constraining sleep/wake monitoring system using bed actigraphy," *Medical and Biological Engineering and Computing*, vol. 45, pp. 107-114, 2007.

[107] V. Swarnkar, U. Abeyratne, and C. Hukins, "Objective measure of sleepiness and sleep latency via bispectrum analysis of EEG," *Medical and Biological Engineering and Computing*, vol. 48, pp. 1203-1213, 2010.

[108] A. Saastamoinen, E. Huupponen, A. Värri, J. Hasan, and SL. Himanen, "Computer program for automated sleep depth estimation," *Computer Methods and Programs in Biomedicine*, vol. 82, pp. 58-66, 2006.

[109] A. Flexer, G. Gruber, and G. Dorffner, "A reliable probabilistic sleep stager based on a single EEG signal," *Artificial Intelligence in Medicine*, vol. 33, pp. 199-207, 2005.

[110] HG. Jo, JY. Park, CK. Lee, SK. An, and SK. Yoo, "Genetic fuzzy classifier for sleep stage identification," *Computers in Biology and Medicine*, vol. 40, pp. 629-634, 20101.

[111] N. Khasawneh, MAK. Jaradat, L. Fraiwan, and M. Al-Fandi, "Adaptive neuro-fuzzy inference system for automatic sleep multistage level scoring employing EEG, EOG and EMG extracted features," *Applied Artificial Intelligence*, vol. 25, no. 3, pp. 163-179, 2011.

[112] P. Piñeiro et al., "Sleep stage classification using fuzzy sets and machine learning techniques," *Neurocomputing*, vol. 58, pp. 1137-1143, 2004.

[113] AC. De Rosa and JM. Lima, "Fuzzy classification of microstructural dynamics of human sleep," in *IEEE SMC*, Beiging, China, 1996, pp. 1108-1113.

[114] JE. Heiss et al., "Classification of sleep stages in infants: a neuro fuzzy approach," *IEEE Engineering in Medicine and Biology Magazine*, vol. 21, pp. 147-151, 2002.

[115] JSR. Jang, "ANFIS: Adaptive Network-based Fuzzy Inference System," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23, no. 3, pp. 665-685, 1993.

[116] D. Nauck and R. Kruse, "Obtaining interpretable fuzzy classification rules from medical data," *Artificial Intelligence in Medicine*, vol. 16, pp. 149-169, 1999.

[117] M. Cabrero-Canosa, E. Hernandez-Pereira, and V. Moret-Bonillo, "Intelligent diagnosis of sleep apnea syndrome," *IEEE Engineering in Medicine and Biology Magazine*, vol. 23, no. 2, pp. 72-81, 2004.

[118] H. Merica and JM. Gaillard, "Statistical description and evaluation of the interrelationships of standard sleep variables for normal subjects," *Sleep*, vol. 8, no. 3, pp. 261-273, 1985.

[119] Sleep Heart Health Study SRC, "Sleep Heart Health Study. Reading center manual of operations," Case Western Reserve University, Tech Report VMLA-039-02, 2002.

[120] C. Guilleminault, F. Eldridge, and W. Dement, "Imsomnia with sleep apnea: a new syndrome," *Science*, vol. 181, pp. 856-858, 1973.

[121] A. Otero, P. Félix, and MR. Álvarez, "Algorithms for the analysis of polysomnographic recordings with customizable criteria," *Expert Systems with Applications*, vol. 38, pp. 10133-10146, 2011.

[122] H. Rauscher, W. Popp, and H. Zwick, "Computerized detection of respiratory events during sleep from rapid increases in oxyhemoglobin saturation," *Lung*, vol. 169, no. 1, pp. 335-342, 1991.

[123] X. Luo, C. Zhang, and J. Cai, "The weighting issue in fuzzy logic," *Informatica: An International Journal of Computing and Informatics*, vol. 21, no. 2, pp. 255-262, 197.

[124] C. Guilleminault, A. Tilkian, and WC Dement, "The sleep apnea syndromes," *Annual Review of Medicine*, vol. 27, pp. 465-484, 1976.

[125] R. Wietske et al., "The role of sleep position in obstructive sleep apnea syndrome," *Eur Arch Otorhinolaryngol*, vol. 263, pp. 946-950, 2006.

[126] M. Cabrero-Canosa et al., "An intelligent system for the detection and interpretation of sleep apneas," *Expert Systems with Applications*, vol. 24, pp. 335-349, 2003.

[127] JE. Lawrence et al., "Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults," *Journal of Clinical Sleep Medicine*, vol. 5, no. 3, pp. 263-276, 2009.

[128] American Academy of Sleep Medicine Task Force, "Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research," *Sleep*, vol. 22, no. 5, pp. 667-689, 1999.

[129] American Academy of Sleep Medicine, *International classification of sleep disorders: Diagnostic and coding manual*, 2nd ed. Westchester, IL: American Academy of Sleep Medicine, 2005.

[130] DJ. Eckert, AS. Jordan, P. Merchia, and A. Malhotra, "Central Sleep Apnea: pathophysiology and treatment," *Chest*, vol. 131, pp. 595-607, 2007.

[131] TI. Morgenthaler, V. Kagramanov, V. Hanak, and PA. Decker, "Complex Sleep Apnea Syndrome: Is it a unique clinical syndrome?," *Sleep*, vol. 29, no. 9, pp. 1203-1209, 2006.

[132] DA. Pevernagie and JW. Shepard, "Relations between sleep stage, posture and effective nasal CPAP levels in OSA," *Sleep*, vol. 15, no. 2, pp. 162-167, 1992.

6. VALIDATION

Validation of intelligent systems can comprise a number of different methods which can be classified in qualitative methods and quantitative methods [**1**]. Qualitative methods use subjective techniques to compare the performance, while quantitative methods are based on the use of statistic measurements. Although both approximations are not incompatible, here the discussion is centered over quantitative validation, since objective measurements allow better comparison of the results. In this respect, quantitative measurement of intelligent systems can use statistical measures of general use, easy to interpret and to compare. It can also include graphical representation methods which permit figuring out how the system under validation is behaving.

Independently of the used method, validation of intelligent systems is normally based on the comparison of the results they achieved versus the so-called *gold standard* or *standard reference*. Such standard represents the ideal behavior to which the system should converge. In this respect two different types of validation can be distinguished:

- *Validation against the problem*, when the comparison is made against a reference which is known to be correct. That is, it is known with total certainty that the desired output is the correct for the actual problem. Unfortunately, the previous procedure cannot be always carried out for the validation of intelligent systems in medicine.

- *Validation oriented to the result* or *against the expert*, when reference is based on the use of interpretations from human experts as the validation criterion. In this case the objective is to achieve behavior of the system to be similar to that of a human expert. As compared to validation against the problem, validation against the expert may not represent such an objective

validation, as long as the standard reference can be affected by subjectivity of the judgments. For example, it can be possible that two experts of the same level conclude different solutions to the same problem.

For the validation of the proposed system, due to the nature of the standard reference, validation against the expert is the only possibility. Therefore, one has to take care at the time of interpreting the results in this kind of validation. As previously said, one has to take into account that expert's opinion is subjective. It is known to vary even when they are confronted with themselves in the same case but in different occasions. Sources influencing appearance of subjectivity among expert's opinions are several and they have their origin regarding different causes:

- Medicine is basically a non deterministic science but based on heuristics and affected by imprecise information.
- Several factors like tension and tiredness can influence the interpretation process carried out by the clinician.
- It is possible to make equivalent decisions and equally valid, although not identical, to solve the same problem.

That said, as with regard to the different quantitative methods available for validation, several statistical approximations can be used to carry out the validation process: hypothesis tests, analysis of variance (ANOVA), confidence intervals, etc. In addition to these statistical tests, additional agreement measures and graphical techniques can be used, allowing quantification and characterization of the error, and providing wide knowledge about the general performance of the system.

The use of one particular measure or another depends, to a great extent, on the specific kind of agreement to be measured and the nature of the data. In this respect, for example, one first categorization can be established regarding if data has can be categorized into nominal classes or, on the contrary, it has a purely numerical nature. For example, with regard to the SAHS diagnosis, much of the data involve categories, for example, sleep stages (W, DS, DEEP, REM) or the different types of apneic events (apnea/hypopnea, obstructive/mixed/central). If the interest is in assessing the agreement over classification of the different sleep stages, quantitative measures should

be used being able to deal with categorical data. On the other hand, if for example the interest is in the comparison of the final AHI values obtained between the system and the standard reference on a set of PSG recordings, then quantitative measures must be used being able to work with the respective numerical distributions.

An additional classification of the validation measures can be established, for example, with regard to the perspective in which the object of the comparison is interpreted with respect to the general reference. Under this perspective, quantitative techniques can be divided in three groups: *pair-wise measures*, *group measures* and *agreement ratios*. *Pair-wise measures* are intended to evaluate the degree of agreement and/or association between the results of two experts (including an intelligent system, human experts or a standard reference). *Group measures*, on the other hand, are oriented at measuring the divergence in the opinions considering the different experts as a group. In this respect, it can be investigated if the opinion of a certain expert differs, in particular, with the general opinion of the group, or in order to obtain measures that can be generalized to a group of experts within the some population. Therefore, they are specially suited when two or more experts are available for comparison. Finally, *agreement ratios* are aimed at measuring the agreement existent between an expert –or intelligent system- and a standard reference. They differ from pair-wise measures in that the first ones handle the interpretation of a variable as a whole, whereas in the case of agreement ratios, the results are analyzed within the different categories in which the interpretation can be divided. They have the need, therefore, of data that can be structured into categories.

In the following the set of validation measures used for validation of the system are discussed. For that purpose, the different measures are organized in two great groups, depending on whether they operate over categorical or numerical data. After definition of the different validation measures, design of the validation tests is carried out. The objective is the description of the specific validation procedures carried out to assess performance of the system according to the different subtasks of interest.

6.1. Measures involving categorical data

6.1.1. Contingency tables

A general procedure for the calculation of agreement measures, when dealing with categorical data, involves the previous construction of a contingency table. A contingency table –or confusion matrix- is a type of table in matrix format that crosses categorical data of different experts. Table 6.1 shows an example of contingency table that relates the results from two experts.

Table 6.1. Contingency table relating results from experts A and B

		Results expert B				
		1	2	…	*k*	Total
Results expert A	1	n_{11}	n_{12}	…	n_{1k}	$n_{1.}$
	2	n_{21}	n_{22}	…	n_{2k}	$n_{2.}$
	…	…	…	…	…	…
	k	n_{k1}	n_{k2}	…	n_{kk}	$n_{k.}$
	Total	$n_{.1}$	$n_{.2}$	…	$n_{.k}$	$n_{..}=N$

Each cell from the table includes a quantity n_{ij} that represents the number of cases in which expert A selects the category *i*, whereas expert B selects the category *j*. *Absolute marginal frequencies* are situated at the margins of the table, and they are calculated as the sum of the values n_{ij} or *absolute frequencies* from the respective files and columns.

It is also possible to generate contingency tables on the basis of *relative frequencies*, or proportions, instead of using absolute frequencies. The relative frequency from cell *ij* (represented by p_{ij}) is nothing but the number of cases within the cell (n_{ij}) divided by the number of total cases (N):

$$p_{ij} = \frac{n_{ij}}{N}$$

Once the experts have classified all the cases in the set of possible categories, then different agreement measures can be obtained by constructing the corresponding contingency tables confronting classifications of the different experts two-by-two.

6.1.2. Pair-wise measures

Pair-wise measures are based on the two-by-two confrontation of the analysis results by a set of experts. Each expert carries out an evaluation of the cases and performs its classification by assigning a certain semantic label. The set of semantic labels[41] must be exhaustive (it has be one for each case), and in addition, the corresponding labels have to be mutually exclusive (only one semantic label can be assigned to each case).

Once the experts have classified all the cases in the set of possible categories, then agreement measures can be obtained from the resulting contingency table. Within this group of validation measures two of them are used in the validation of the developed system: the *agreement index* and the *kappa index*.

Agreement Index

One of the agreement measures most commonly used is the index or proportion of agreement. Also known as *accuracy*, this measure can be simply calculated as the quotient between the number of agreement observations and the number of total observations. To obtain this measure from the contingency table, one can simply sum the absolute frequencies from the main diagonal, and divide them by the total number of cases. Another possibility is just to sum the relative frequencies or proportions from the main diagonal:

$$Agreement\ Index\ (AgrI) = \frac{\sum_{\substack{i=1,j=1 \\ i=j}}^{k} n_{ij}}{N} = \sum_{\substack{i=1,j=1 \\ i=j}}^{k} p_{ij}$$

[41] Each semantic label represents a different category

The agreement index takes values within the range [0, 1] in which 1 represents the maximum agreement and 0 means absolute disagreement. Its value is not affected by the relative order of the categories. The main advantage of this measure is the simplicity of its interpretation which has made its use widespread over different fields and applications. However, it presents the inconvenient that it does not take into consideration agreements due to chance. This situation of concordances or agreements by chance is frequent when, for example, two classifications are performed on the same data but using different number of categories. In this respect, the classification with less number of categories will tend, in general, to have a lower agreement index, since the probability to agree in the classification of an item is lesser as lesser is the number of categories.

A derived measure from the agreement index is the *classification error*, which is calculated as the complementary of the agreement index (*1 – AgrI*). This measure accounts for the amount of disagreement between two experts, i.e. it measures the proportion of misclassified cases.

Kappa index

Kappa index (κ) is a chance-corrected measure of agreement proposed by Cohen [**2**]. The measure is based on the calculation of two quantities:

- p_0 = proportion of observed agreement

$$p_0 = \frac{\sum_{i=1,j=1}^{k} n_{ij}}{N}$$

- p_c = proportion of agreement occurring by chance

$$p_c = \sum_{\substack{i=1,j=1 \\ i=j}}^{k} \frac{n_{i.}}{N}\frac{n_{.j}}{N}$$

The term p_0 represents the proportion of agreement as the number of total agreements divided by the total number of cases analyzed, whereas p_c is the sum of the products of the marginal proportions. In this manner, 1 - p_c represents the maximum

possible agreement once chance has been removed, and p_0 - p_c represents the obtained agreement once chance has been removed. The kappa index κ is then defined according to the following quotient:

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

When the interpretations of two experts are compared, the resulting agreement is expected to be higher than the one obtained if the marginal proportions were taken as probabilities[42]. For that reason observed agreement (p_0) is corrected with agreement due to chance (p_c).

Kappa index is independent of the number of observations and of the number of present categories, and it is also not influenced by permutations on the order of the categories. It is a symmetrical index, and if the observed agreement is equal to the agreement occurring by chance, then value of kappa is zero. If the observed agreement is higher than that expected due to chance, value of kappa is positive, being its maximum value 1. On the other hand, if the observed agreement is less than the agreement occurring by chance, then the value of kappa is negative, being its minimum value -1.

Subsequently, Table 6.2 shows a linguistic interpretation of the parameter κ according to the criterion of Landis y Koch [**3**]:

Table 6.2. Symbolic interpretation of κ according to Landis and Koch

Value of κ	**Agreement level**
< 0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

[42] In this manner the probability of a certain cell to be chosen is the product of the marginal proportions corresponding to that cell

One has to be cautious, however, at the time of interpreting the kappa index, because only in the case in which the number of events classified in each category by the experts is the same, the value of kappa might achieve the value 1. In other words, the former implies that the marginal distributions should be identical. However, this is not the common case, and therefore it is often useful to determine the maximum possible value of kappa that can be achieved with the given marginal distributions. Thus, if the number of cases classified in each category by the experts differs, the maximum value of kappa can determined by:

$$\kappa_M = \frac{p_{0M} - p_c}{1 - p_c}$$

where p_{0M} is calculated by matching the marginal values of each expert by choosing the minor value of each pair and then summing the resulting values. Therefore, the quotient $\frac{\kappa}{\kappa_M}$ tells us that the obtained value of kappa is $\frac{\kappa}{\kappa_M}$ times larger than the maximum possible value given the circumstances[43].

One more observation about kappa is that, in the way it has been defined, kappa can be applied both to nominal or ordinal classes, however it does not penalize deviations within the disagreements, which can be of interest in the case of nominal scales.

Chi-squared test for homogeneity

This test is used to determine whether frequency counts on the categories of certain categorical variable are identically distributed across different populations. It can be, in fact, applied to assess the distribution over two or more populations, and in this sense, it might be classified into the category of group measures rather than be included as kind of pair-wise measure. Here, however, the chi-squared test will be particularized for the case of two populations, which for the aim of the validation, can be though as the corresponding distributions resulting by classification of the cases made by two different experts.

[43] Due to the number of cases classified for each expert in each category

In this respect, formulation of the problem can be represented according to the following particularized view of the contingency table:

Table 6.3. Particularization of the contingency table for the comparison of the frequency counts of two different experts

		Categories				
		C1	C2	…	*Ck*	Total
Expert's classifications	A	n_{A1}	n_{A2}	…	n_{Ak}	$n_{A.}$
	B	n_{B1}	n_{B2}	…	n_{Bk}	$n_{B.}$
	Total	$n_{.1}$	$n_{.2}$	…	$n_{.k}$	$n_{..}=N$

Using the previous contingency table, then the objective is to determine whether the observed sample frequencies significantly differ from each other. The null hypothesis can be enunciated as $H_0: \frac{n_{A1}}{N} = \frac{n_{B1}}{N} \wedge \frac{n_{A2}}{N} = \frac{n_{B2}}{N} \wedge \ldots \wedge \frac{n_{Ak}}{N} = \frac{n_{Bk}}{N}$, against the alternative hypothesis that at least one of the previous statements is false.

Then, once defined the null hypothesis, the next step is to compute the resulting Chi-squared (χ^2) statistic according to the following formula:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=A,B} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

where E_{ij} is the frequency count for expert j at the category i:

$$E_{ij} = \frac{n_{.i} n_{j.}}{N}$$

and the number of degrees of freedom for the case of two experts is $k - 1$.

Once the value of the statistic and its corresponding degrees of freedom have been determined, then the *p*-value can be obtained to test the validity of the null hypothesis according to the general procedure for hypothesis test as described in Table 6.4.

Table 6.4. General procedure for hypothesis test

1. Define the null hypothesis (H_0)
2. Select a statistical test (test statistic) to be used for evaluation of H_0 validity
3. Choose a level of significance α for the test: probability to refuse H_0 if H_0 is true
4. Calculate the *p*-value (probability to obtain higher discrepancy than that observed when H_0 is true)
5. Compare the obtained *p*-value with the significance level:
    a. If *p*-value $\leq \alpha$ then refuse H_0
    b. If *p*-value $> \alpha$ then accept H_0

6.1.3. Agreement ratios

Calculation of agreement ratios is based on the construction of a 2x2 contingency table for each one of the categories in which is divided the interpretation (see Table 6.5).

Table 6.5. Contingency table (2x2) to calculate agreement ratios. D represents the presence of a category in the interpretation whereas ¬D stands for its absence

		Standard Reference		
		D	¬D	
Expert system	D	a	b	a + b
	¬D	c	d	c + d
		a + c	b + d	a + b + c +d

In this table the results from one source to be compared –a system or another expert- are related with respect to the results of the standard reference for a particular category. In the example, values *a, b, c* y *d* are counters respectively representing the number of:

- ***a*: True Positives (TP)**. They represent the cases in which the system[44] classifies the case within the category *D* and so it does the standard reference.

[44] As it has been stated, here we refer to the developed system but in general it might be any source we want to compare with respect to the standard reference

- ***b*: False Positives (FP) or type-II errors**. Classification of the case by the system into category *D*, while standard reference classifies it as ¬*D*.

- ***c*: False Negatives (FN) o type-I errors**. The inverse case in which the system classifies the case as ¬D, but standard reference points out to a positive event (classified as D).

- ***d*: True Negatives (TN)**. Both, system and standard reference agree by classifying the event as ¬D.

Usually these values are not used for direct comparison since they depend on the number of analyzed cases. In their place, relative measures are derived from them and used for comparison. The most important ones are subsequently described.

Agreement Index

It is equivalent to the agreement index previously described for the general case of contingency tables of size *MxN*. For the particular case of agreement ratios, it represents the proportion of cases in which the expert system has matched with the standard reference for a given category:

$$Agreement\ Index = \frac{TP + TN}{TN + TP + FP + FN}$$

It has to be taken into account that the agreement index does not make any distinction between positive agreements (TP) and negative agreements (TN).

Sensitivity

It is defined as the ratio of true positives and it allows measuring of the capability of the system to correctly classify the positive cases. It can be understood as the probability that the system correctly detects an event as positive given that it actually is a positive.

$$Sensitivity = \frac{TP}{TP + FN}$$

A complementary measure related to sensitivity is the ratio of false negatives, which provides a measure of the possibility of failing in the detection of a case that the standard reference has considered a positive case with regard to a given category. That is, it measures the probability that the system wrongly classifies a case as negative.

$$False\ Negative\ Rate = 1 - Sensitivity = \frac{FN}{FN + TP}$$

Specificity

It is defined as the ratio of true negatives. It allows measurement of system's capability to correctly classify the negative cases:

$$Specificity = \frac{TN}{TN + FP}$$

A measure related to specificity is the ratio of false positives, which is calculated as the number of times the system fails at classifying a case as positive, divided by the number of times the standard reference considers a negative event. That is, it measures the probability that the system wrongly classifies a case as positive.

$$False\ Positive\ Rate = 1 - Specificity = \frac{FP}{FP + TN}$$

Specific agreement

There are some cases in which either the positive or the negative cases cannot be measured, or the categorization of the number of respective cases might be difficult. A typical example, in fact, is when scoring events over time. In these cases, appropriate segmentation of the temporal axis is not always possible. In addition, it can be the case in which it can be of interest to measure the agreement over certain category without taking into account either the number of true positives or true negatives, but just the deviations with respect to type-I and type-II errors. This can be the case for example, where the proportion of true negatives and true negatives is very high with respect to each other.

In these cases, measures of positive or negative agreement can be used which are defined as follows:

$$Positive\ Agreement = \frac{TP}{TP + FP + FN}$$

$$Negative\ Agreement = \frac{TN}{TN + FP + FN}$$

ROC Curves

Receiver Operating Characteristics (ROC), or simply ROC curves, are widely used in validation problems because they offer an intuitive graphical mode to interpret discriminant capabilities of a system. They also provide of a numerical value θ which can be inferred from the graphic and that it globally validates the capabilities of the system within a unique comparable index. Comparison can be, therefore, easily and rapidly performed through the use of the θ parameter.

ROC curves are graphical representations that relate the ratio between true positives (sensitivity) with the ratio of false positives (1 – specificity). The resulting crossing point is named *operation point*. For each one of these points, the corresponding ROC curve starts in the origin and it goes until the right upper corner, passing by the crossing point of the values of sensitivity and false positive ratio (see Figure 6.1). By evaluating the system on several situations ,or under different configurations, various operation points can be obtained which joining all together determine the resulting ROC curve that represents performance of the method.

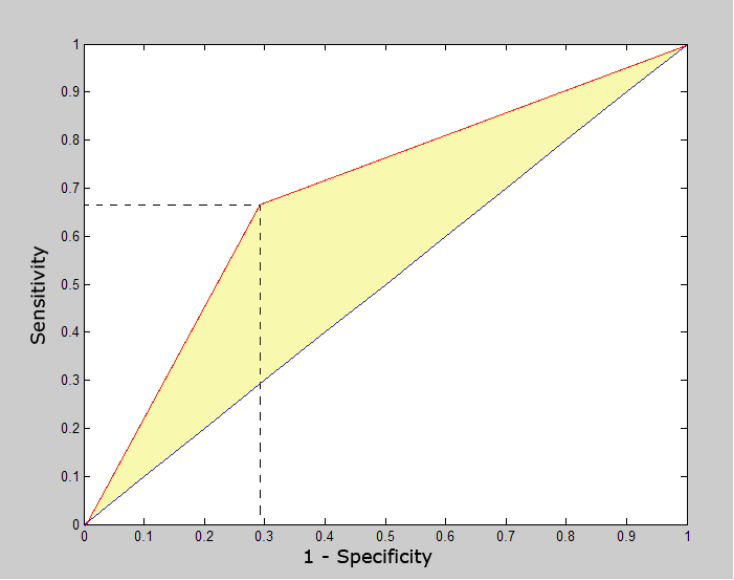


Figure 6.1. Example of one point ROC curve

Basic graphical interpretation is that, the better a system is, the more the curve approximates to the point (x,y) = (0,1). That is, when comparison comprises two or more systems by means of their respective ROC curves, the best system is the one having a curve monotonically superior in all of its points to the rest of the systems [**4**]. However, in the cases in which the curve does not overpass in all of its points to the rest, the comparison does not result so easy. It is for this reason that, often, a supplementary measure is used which allows direct comparison of the quality of the test, the *area under the ROC curve* (AUC). AUC or θ can be interpreted as the probability of a correct classification from the system, whereas the quantity 1 – θ represents the rate of incorrect classifications. Thus, basically, a system with value of θ closer as possible to one is desired, and therefore when comparing two or more systems, the one with the higher value of θ would be the best. Value of AUC can be easily calculated in the case of ROC curves with just one operating point by using the following formula:

$$AUC = \theta = \frac{Sensitivity + Specificity}{2}$$

In the cases where the ROC curve involves two or more operation points, θ can be approximated by trapezoidal integration [**5**]. Although this approximation tends to subestimate the actual value of the area, because of the manner in which the points are connected through straight lines, the computed value asymptotically approximates to the real value as long as the number of operating points increases.

In the subsequent Table 6.6 linguistic interpretation of the value θ is displayed according to [**6**]:

Table 6.6. Symbolic classification of θ representing area under ROC curve

Value of θ	**Linguistic descriptor**
0.90-1.00	Excellent
0.80-0.90	Good
0.70-0.80	Normal
0.60-0.70	Scant
<0.60	Null or almost null

6.2. Measures involving numerical data

When dealing with numerical data, one common question one may want to ask is whether a set of values are consistent between them, in the sense that they can be regarded as being taken from the same population following a certain distribution. Some assumptions can be made, like for example, that data is normally distributed, or that certain moments of the distribution are known, for example, that data distribution is known to have a certain mean μ, or to have a certain variance σ. On the other hand, one may be interested in assessing if there are significant differences over the mean values of a population, or between respective variances from different populations. A wide range of test statistics is available in this respect, depending on the characteristics or information available from the original data distribution, and the object (main descriptors) of the comparison. In the following, some test statistics are described that are used for the analysis of numerical data for the validation of the system. Description here is aimed to be only introductory, thus for detailed comprehension and further details the reader is referred to any reference book on biological statistics such as [**7**].

6.2.1. Pair-wise measures

Pair-wise distribution comparison

One common parametric test statistic used to check whether two random variables *X*, *Y* have equal means (μ_x, μ_y) is the *two-sample Student's t-test*. This test assumes random, independent sampling from the two populations, normal distributions and equal variances (but unknown). The t-test for independent groups is designed to address the question about if the observed differences between means in the populations can be attributed to chance, or it is compelling evidence of a real difference in the populations. For this purpose, it tests the null hypothesis H_0: $\mu_x = \mu_y$ against the alternative that means from the two populations are unequal. The particularization of this test for one variable leads to the *one-sample t-test,* which checks the null hypothesis that data are a random sample from a normal distribution with zero mean and unknown variance, against the alternative that the mean is not zero. In addition, for the purposes of the

validation of the developed system, it is of particular interest the paired version of the test. *Paired t-test* is used when the aim is to compare two different measures from the same target (in our case individuals) from different experts, or from the same expert in different instants of time, and decide if there is a significant difference between the two measures. The main difference with regard to the unpaired version is that, in the first one, the two populations are considered as random selections of individuals, whereas in the paired version, the same set of individuals are used in the experiment. Paired test is interesting, for example, to compare respective AHI values from both the system and the expert on a concrete set or recordings.

However, for the purpose of the validation of the developed system, in general, non-parametric tests are preferred since assumption of normality is not always affordable.

Non-parametric versions of the former test statistics for pair-wise comparisons comprise the *Wilcoxon rank sum test,* for the unpaired comparison of the medians of two random variables, and the *Wilcoxon signed rank test,* in case paired comparison is desired. These tests can be used as alternatives to their respective t-test versions when the population cannot be assumed to be normally distributed, or the data is on the ordinal scale. The null hypothesis in the Wilcoxon test is H_0: $m_1=m_2$ in the first case (unpaired comparison) or H_0: $m_1-m_2 = 0$ in the second case (paired comparison), where m_1 and m_2 are the respective medians from the two populations. The *one sample version of the Wilcoxon test* performs a two-sided signed rank test of the null hypothesis that data are a random sample coming from a continuous, symmetric distribution, with zero median, against the alternative that the distribution does not have zero mean.

In any case, once selected the appropriate test statistic, the procedure to assess the result of the test is analogous to that described above by calculating the corresponding *p*-value (see Table 6.4).

Pearson's linear correlation

Association measures are considered as a special kind of pair-wise measures in which evolution of the values of a variable *X* with respect to the evolution of the values

of a second one *Y* is investigated. They attempt, therefore, to assess the dependence that exists between two numerical variables.

Among these measures, Pearson's product-moment correlation coefficient (*r*) is one of the most extended. It is obtained by dividing the covariance of the two variables by the product of their standard deviations:

$$r = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y}$$

Pearson's *r* is a measure of linear dependence and it is defined in the interval [-1, 1]. The Pearson correlation is 1 in the case of perfect positive linear relationship (increasing regression line), -1 in the case of a perfect negative linear relationship (decreasing regression line), and 0 in the case in which the variables are linear independent. Thus, if the variables are independent, then Pearson's correlation is 0, but the converse is not true because the correlation coefficient only detects linear dependencies between the two variables.

Hypothesis test can be performed in order to check for significant linear correlation by assuming normal data distribution. Under this assumption, sampling distribution of Pearson's correlation coefficient approximately follows Student's t-distribution with *N* - 2 degrees of freedom, where *N* is the number of samples. This also holds approximately, even if the observed values are non-normal, provided sample sizes are not very small [**8**].

Spearman's rank correlation

Spearman's rank correlation coefficient (*rho*) is a non-parametric measure of statistical dependence between two variables *X* and *Y*. It assesses how well the relationship between two variables can be described using a monotonic function. It therefore amplifies Pearson's linear correlation to a wider range of monotonic functions, not necessarily linear, and it also has the property of being less sensitive to the presence of outliers. Similar to Person's *r*, range of *rho* is defined within the interval [-1, 1].

It is defined as the Person's correlation coefficient between ranked variables, that is, it deals with associations between ranges, or categories, of which it is known their order. Therefore, as a difference with Pearson's *r*, Spearman's *rho* is invariant to transformations of *X* and *Y* in which order is maintained. For its calculation, firstly, raw scores X_i and Y_i are converted to ranks x_i and y_i, and then, *rho* is computed from these as follows:

$$rho = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}}$$

One can test for statistical significance of *rho* using the following test statistic:

$$t = rho\sqrt{\frac{N-2}{1-rho^2}}$$

which analogously to Pearson's *r*, is distributed approximately following a Student's *t* distribution with *N* - 2 degrees of freedom under the null hypothesis that *rho* is significantly different from zero.

6.2.2. Group measures

Multiple distribution comparison

Multiple distribution comparison measures generalize comparison of random variables for the case in which comparison involves two or more groups of numerical populations.

For example, if one wants to check if there are significant differences between the means of several populations, one can perform a one-way analysis of variance (*one-way ANOVA*). The statistical null hypothesis is that the means of the numeric variable are the same for the different groups; the alternative hypothesis is that they are not all the same.

The basic idea is to calculate the mean of the observations within each group, and then to compare the variance among these means to the average variance within each group. Under the null hypothesis that all the observations in the different groups have the same mean, the weighted among-group variance should be the same as the within-group variance. As the means get further apart, the variance among the means increases. The test statistic is thus the ratio of the variance among the means, divided by the average variance within the groups (*F*-test).

One-way ANOVA is mathematically equivalent to Student's t-test for the case of just two groups of data. Similarly, ANOVA assumes that data is independent and normally distributed with equal variance within each group. If data do not fit these assumptions, then a non-parametric version of the test can be used, such as for example the *Kruskal-Wallis test*, which generalizes the Wilcoxon rank sum test for comparison of two or more groups of numerical data.

Intraclass Correlation Coefficient

The Intraclass Correlation Coefficient (ICC) is a general measure of agreement in which the measurements used for comparison are assumed to be parametric (continuous and normally distributed). There are several versions of ICC and the concrete definition ultimately depends on the experimental design, and on the conceptual intent of the study. In this regard, for the aim of the validation of the proposed system, specific ICC definition is used in order to measure the agreement of quantitative measurements made by different observers measuring the same quantity.

Let *n* to be the number of targets, and *k* to be the number of experts, under the previous considerations the ICC definition is based on the assumption of the following linear model, which is analyzed using a two-way analysis of variance:

$$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}$$

in which x_{ij} denotes the *i*th rating ($i = 1, \ldots, k$) on the *j*th target ($j = 1,\ldots n$), *a* stands for the rater, *b* stands for the target, *ab* stands for interaction, and *e* stands for the error. Specifically, the ICC(2,1) version of this statistic is used as defined by Shrout and Fleiss

[**9**]. Under this form, ICC accounts for a composite of intra-observer and inter-observer variability, treating the different experts as random effects. ICC(2,1), as defined before, is intended to ask the question whether the different experts involved in the comparison are interchangeable. As a derived measure of correlation, ICC is theoretically defined in the interval [-1, 1], however negative values of ICC rarely occur in practice. In this respect an ICC of 1 is interpreted as total agreement between the raters, whereas an ICC of 0 represents no agreement at all, i.e. the experts are not exchangeable. Confidence intervals and hypothesis test in order to check for statistical significance of the value of ICC can also be obtained. A more precise mathematical definition of such statistic would exceed the objectives of this chapter, and the reader is referenced to the former paper of Shrout and Fleiss for further details.

6.2.3. Model comparison

Throughout the functional description of the system carried out in Chapter 5, some analysis methods have been proposed that require of a comparative analysis of several machine learning models to be done. That implies an additional sort of *internal* validation process –model selection- to be done, in order to determine which model achieves the best performance within the associated method. In this respect, for example for the detection of EEG arousals, four different models have been investigated to act as classifiers in the last phase of the method for EEG arousal identification (see Chapter 5, "*Identification of EEG arousals*"). In order to accomplish this comparative analysis, a general methodological framework is introduced describing specific mechanisms and metrics to carry out model comparison within the field of machine learning.

In this respect, in general, in order to solve a problem in the scope of machine learning, the simplest case normally consists in having a model with a series of customizable parameters –or degrees of freedom- and some data. In this context, experimental design firstly implies to carry on an effective learning or training process to optimize its parameters. Then it should be estimated the real performance of the model. Such estimation is usually measured by calculating the error committed by the model.

In some occasions it would be necessary to perform some previous processing of data. This preprocessing can involve an adequate preparation of the data, reduction of dimensionality or data normalization. The global schema is shown in Figure 6.2.

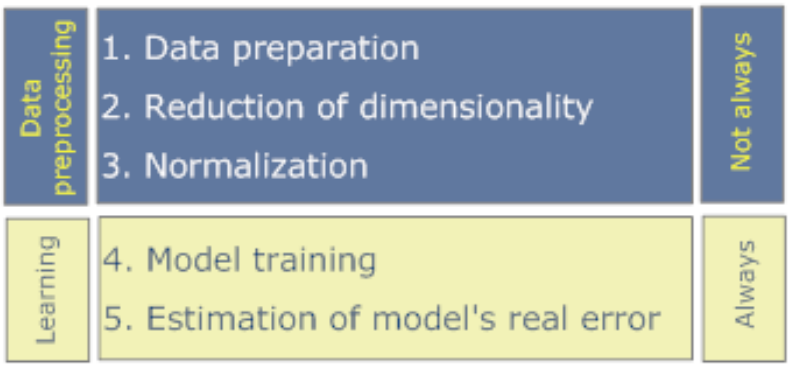


Figure 6.2. Experimental design schema in machine learning

Data preparation may serve, for example, to try to remove cases with incomplete information, or if it is possible, atypical cases –outliers- and noise. It might also be necessary to adapt the inputs to the characteristics of the chosen machine learning model. For example, in the case of time series, it is usual to transform the one dimensional signal into *d*-dimensional patterns using temporal moving windows.

Reduction of dimensionality pursues reducing the effects of the so-called *curse of dimensionality*. In general, the higher the input dimensionality is at the input of the machine learning system, the higher the number of necessary examples to obtain a good model. In many real cases, however, few examples are available which implies a problem when dimensionality is high. Likewise, data can contain a lot of redundant information. Because of this, in many cases it is necessary to reduce dimensionality at the input. Finally, dimensionality reduction at the input space causes a reduction in the number of systems' parameters. This has the advantage of the reduction of the complexity and execution time. It also improves generalization capabilities of the model. Reduction of the number of features within the system has been applied, for example, for the identification of EEG arousals (see Chapter 5, "*Identification of EEG arousals*").

Normalization involves linear rescaling of input variables. It is especially adequate if the different variables have very different values –orders of magnitude. In this respect normalization is done in order to bring data similar values. One possible technique consists of scaling each variable independently, such that for each variable x_i, its mean μ_i and standard deviation σ_i are calculated in the training set. The new variables are then calculated as:

$$\bar{x_i} = \frac{x_i - \mu_i}{\sigma_i}$$

In this manner resulting variables have mean equal to zero and standard deviation equal to one.

Once an adequate preprocessing of data has been made –optional- it is performed the training of the model. Each model –neural networks, SVMs, genetic algorithms, etc. has its particularities in the training process. In each stage optimal parameters are obtained with the training data, normally using part of the whole set of available data. This stage is combined with the estimation of the error where, normally, several trainings of the model are carried out using different training sets.

In order to estimate performance the error of the model has to be evaluated. The error should be estimated using the whole population of which data came from. However, normally only a limited sample of data is available. The simplest solution would be to use the whole dataset to train the model and to estimate the error. Nevertheless such a method carries on some problems since the obtained model will probably overfit data and the obtained error will be too optimistic.

A better estimation of the error can be achieved by using alternative methods based on the idea of always using, at least, two different datasets, one for the training of the model, and the other to be used to estimate the error once the model has been trained. This second testing dataset is independent of the training set. This process is also known as *holdout*. Drawbacks of this method are that when few data is available it is kind of a "luxury" to use an important part of data in the testing set. On the other hand, since only just one experiment is carried out with a training set, the result can be deceptive if the partition is not adequate.

More efficient methods are based on random subsampling or on cross-validation. Random subsampling consists in the realization of *k* experiments using as testing set different subsets of the main dataset. Each testing subset is randomly chosen from the total number of samples –without replacement. Remaining data are used for training.

In a *k*-fold cross-validation the set of patterns are split in *k* disjointed sets of the same length, with the members of each fold randomly chosen from the full set. A total of *k* iterations are done, each one using a different fold as the testing set, while the rest are used as training. In the extreme case of cross-validation for a dataset of *N* samples, *N* experiments are performed in which in each one *N-1* samples are used for training and the remaining simple for testing. This technique is kwon as *leaving one-out*.

The procedure results in *k* error measures $\{e_1^j,...,e_k^j\}$ for each model configuration *j*. The final error for each model is then calculated as the averaged error for each fold $E_j = \frac{1}{k}\sum_{i=1}^{k} e_i^j$, allowing us to obtain a good estimation of the error, and minimizing the effect of the initial conditions.

The choice of the number *k* of folds ultimately depends on the concrete problem. In general, if a great number of subsets is chosen, then the estimated error will tend to be very precise, however the variance of the real error will be high too as well as the computational time because of the high number of experiments. On the other hand, if few subsets are chosen, then computational time is reduced and variance will be low. On the contrary the estimated error will tend to be less precise. A usual choice for the number of folds is $k = 10$.

An additional variant to the design based on two datasets –training and testing- is the experimental design based on three datasets. In this case a third dataset is used as the validation set in situations where one wants to establish the structural configuration of the model, e.g. to determine the best number of layers in a neural network or the smoothing parameter in a SVM.

An important aspect to be taken into account is that of the comparison of the results. It can be necessary, for example, to choose between different configurations of the same learning model –e.g. the number of neurons in the hidden layer of an ANN, or to choose between several machine learning models –e.g. between a SVM, an ANN or a linear discriminant. In this regard the next question should be addressed: *Which is the best model, among all possible, in terms of committed error?*

To ask this question in the case of comparison of two models, it can be assumed that a cross-validation has been already done, using the very same folds for the training and testing on each one of the two models. Then, it has to be determined if the performance of the two models is equivalent, or on the contrary, there are significant differences among them. In such cases statistical methods –as described above- are usually applied which are based on hypothesis testing. Steps to be followed can be structured, again, in accordance with those described above in Table 6.4.

In accordance with previously described methods, typical statistical tests involving comparison of numerical distributions between two populations –in this case of errors- are, for example, the Student's *t*-test or the Wilconxon rank sum test (see subsection "*Pair-wise measures*"). Figure 6.3 illustrates the process for the comparison of two models.

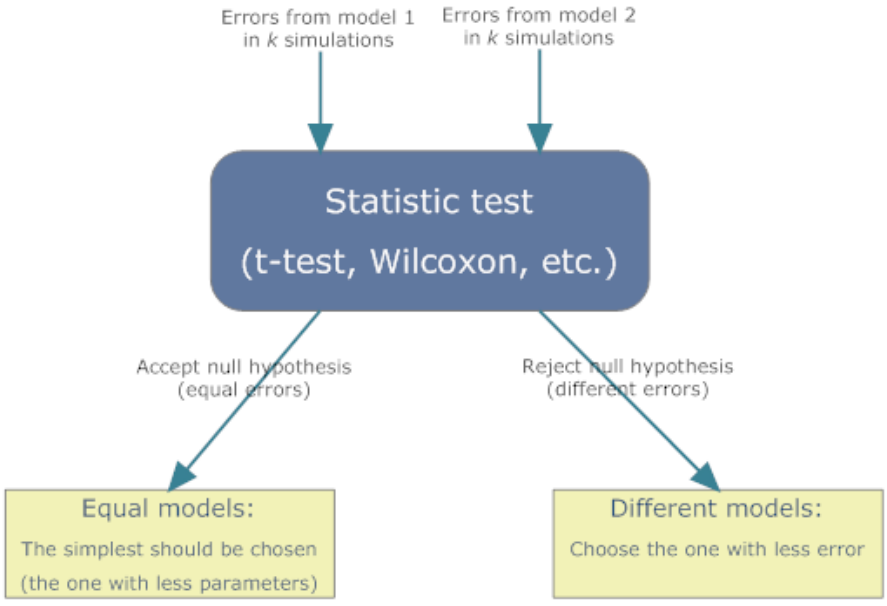


Figure 6.3. Selection between two models in machine learning

On the other hand when there are more than two groups of models, then it is not enough to compare each pair of models separately. The possibility to incorrectly detect a significant difference increases with the number of comparisons. In these cases it is usual to perform an analysis of variance (ANOVA) or a Kruskal-Wallis test to identify if there is a significant difference, respectively, among all the means or all the medians (see subsection "*Group measures*"). In this manner, if the variance test concludes that there are significant differences, then it has to be investigated which the differences are by using a multiple comparison test –e.g. the Tukey method [**10**]. On the other hand if the variance test concludes that there are not significant differences, then it implies that all the models are equivalent, thus the simplest one should be chosen. Figure 6.4 shows the schema for the comparison in which more than two models are involved.

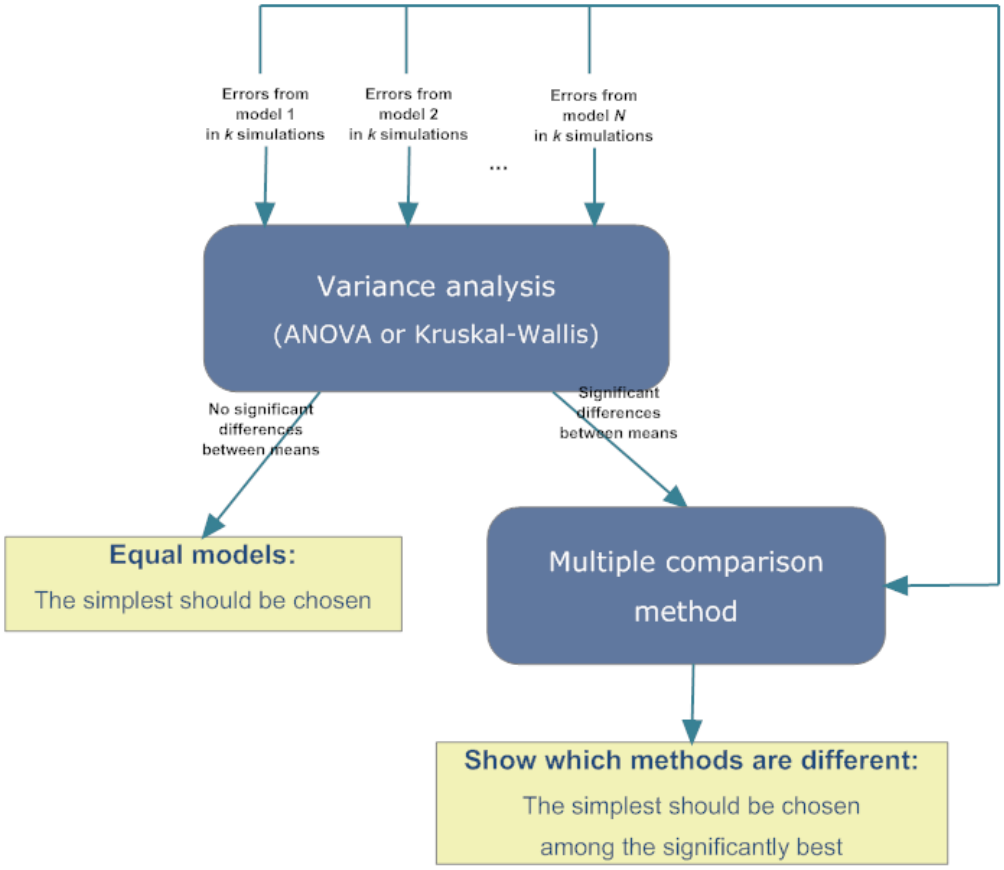


Figure 6.4. Multiple model selection in machine learning

6.3. Design of validation tests

In this section, design of the validation tests is described in order to check system's performance on the SAHS diagnostic task. Tests are organized in such a manner so that the objective is not only to perform validation regarding the capabilities of the system for discriminating the presence of SAHS in the patient with regard to the final diagnosis (SAHS or no SAHS), but also to test performance of the system on the intermediate subtasks that involve the whole diagnostic process, whenever a *standard reference* is available for comparison.

In our case, standard reference for validation is taken from the database resulted from the Sleep Heart Health Study (SHHS). This database, to which access has been granted by the Case Western Reserve University previous agreement for collaboration, emerges from a cohort initiative of several medical centers and universities in the United States, supported by the National Heart Lung & Blood Institute (NHLBI). Its main objective was the study of cardiovascular consequences of sleep-disordered breathing during sleep. The resulting database was then enabled to be used as a resource for subsequent studies related to sleep disorders. Further details about design, motivation and methods of SHHS can be found in [**11**].

SHHS database comprises a set of PSG recordings from real patients in European Data Format (EDF) [**12**]. Each recording is accompanied by respective XML annotation file in which events that resulted from physicians' offline analysis of the PSG are marked. Procedures and criteria followed for the analysis and marking of the events are collected in its Manual of Operations [**13**] that follows recommendations of the AASM for analysis procedures and scoring of events. Thus, validation of the developed system will be based on comparison of its results against the *standard reference* constituted by annotations from the SHHS experts.

Specifically, annotations available through SHHS possess marking over the following types of events:

-**Sleep stages** according to the R&K standard (i.e. W, S1, S2, S3, S4, REM) in the basis of 30-second epochs. Note, however, that SHHS experts do not differentiate between S3 and S4. Same criterion is currently recommended by the recent AASM guidelines [**14**].

-**EEG Arousals**, marked as "*start of arousal*" and "*duration*" in the basis of the ASDA (current AASM) criteria published in [**15**]. The scoring of EEG arousals is independent from the scoring of sleep stages (i.e. an arousal can be scored in an epoch of recording which would be classified as wakefulness by the R&K criteria). An arousal can precede the wake stage or it can be followed by a return to sleep.

-**Apneic events**: hypopneas, obstructive apneas and central apneas marking the start and duration of the apneic event. No attempt is made to distinguish mixed apneas from obstructive apneas. Central hypopneas and increased upper airway resistance events (RERAs) are also not scored by SHHS experts "*because of controversies in the defining these events and the probable need to use invasive monitoring to identify these accurately*" [**13**].

Besides the previous annotations, also desaturations and artifacts in the oxygen saturation channel (start, duration) are included within the annotations. However, these annotations are artificial, i.e. are automatically done by the scoring revision software used by the experts, not by the experts themselves, and therefore they are not reliable. In fact, desaturation events are marked always 30 seconds after the end of the scored apneic event and its duration is established to the same of the apneic event. That is, these default marks neither have validity nor they correspond with the actual start and end of the real desaturation, and therefore they cannot be considered for validation purposes.

In addition to XML annotations, some demographic data is available for some of the recordings. These data does not allow personal identification the patient, but includes information regarding age, sex, height, weight and some clinical conditions.

Hence, taking into account available data from the standard reference, the validation test are structured in the following manner:

(1) Validation of the capabilities for the detection of EEG arousals: two tests are to be carried out first, using the two approximations described in Chapter 5, "*Identification of EEG arousals*": one using the full set of features (see Table 5.2), whereas the other makes use of feature selection methods in order to account for possible redundancy in the first set. Finally the final method that uses a reduced set of features is tested over an independent validation set.

(2) Validation of the sleep staging algorithm, according to the method described in Chapter 5, "*Hypnogram generation*".

(3) Validation of system's capabilities for the identification of apneic events. With a dual purpose: first, to assess its capability in the location of apneic events in the recording (regardless of its concrete type), and in a second place, to determine the discriminative performance for the characterization of the detected events as apneas or hypopneas. For that purpose, output from system's detection phase is used (see Chapter 5, "*Detection of the apneic events*").

(4) Validation of the classification of the apneic events. That is, determination of the concrete class of the apneic event: obstructive, mixed or central. In this respect output from the algorithm described throughout Chapter 5, "*Classification of the apneic events*" is used for validation. Note however that since SHHS annotations do not include "mixed events", only performance regarding classification of obstructive and central events can be tested.

(5) Validation of the final diagnosis of the patient. For that purpose comparisons on the final computed indexes and overall diagnosis are performed against the respective SHHS expert's output.

In order to carry out these tests, a total of 26 patients (mean age ± std.deviation: 68.5±7.7, 8 females) from the SHHS database are randomly chosen without previous knowledge of the clinical history of the participants. Random selection only comprises recordings that were not used for the parameterization and optimization of the integrating modules of the system. In other words, the set of 26 patients forms a completely independent set in order to ensure adequate evaluation of generalization

capabilities of the system. Additional training data is needed for the parameterization of the system and for carrying out comparisons and model selection, especially when learning from examples using machine learning approaches. For this purpose additional PSG recordings from the SHHS database may be used as training sets which is indicated, where applicable, in the design of the corresponding validation task. In any case this *independent validation set* (IVS) of 26 patients is always leaved aside for final validation in order to assess reliable generalization capabilities of the system. In total, the 26 testing recordings included in IVS involve 15540 minutes of sleep.

In the following, more detailed explanation on the design of each one of the previously outlined validation tasks is carried out. Results of the respective validations are given in the corresponding sections of Chapter 7.

6.3.1. Identification of EEG arousals

As previously mentioned, validation tests regarding identification of EEG arousals by the system are suborganized into three subsections: the first one (A) assessing model selection using the full set of 42 features in Table 5.2, the second one (B) in charge of studying different feature selection methods to reduce the number of needed features, and finally (C), validation of the final method using the independent set of 26 recordings.

A) Detection of EEG arousals using machine learning models

In the considered approach to detect EEG arousals, a set of different machine learning models are investigated to act as classifiers. As described in Chapter 5 "*Identification of EEG arousals*", the approach consists of a first signal processing stage after which relevant features are correlated in time to form characteristic arousal patterns. Thus, the aim is to select the best model to act as classifier over the resulting characteristic patterns to detect the arousal. For the purposes of comparison between the different machine learning models, validation strategy is scheduled using an independent set of 20 recordings from SHHS database –different from the IVS- in which data is separated, at the same time, in two differentiated training and testing sets.

According to that, in order to carry out the experimentation the set of 20 PSG recordings is split up into two new sets: 15 of them used as training set (TR), while the remaining 5 are used as a separate testing set (TS). The number of epochs available for classification on each data set is 18090 and 6284, respectively, for the TR and the TS sets. The TR set is used in a training manner in the sense that it is used to determine both the best configurations for SVMs and ANNs, and it is also employed as a training set to compare the different classifiers. On the other hand, the TS set is put aside and used as a separate test set in order to assess generalization capabilities of the different classifiers.

After signal processing and construction of the characterizing patterns (see Chapter 5, "*Identification of EEG arousals*"), the number of characterizing patterns is 15280 for the TR dataset and 4850 for the TS dataset. Note the epochs that do not hold an arousal pattern are directly classified as *non-arousal* at the first classification step. All the patterns are normalized subtracting for each attribute x_i its mean (μ_i) and dividing the result by its standard deviation (σ_i). Thus all the attributes result in a zero mean and standard deviation equal to one.

Due to the fact that usually the majority of epochs during sleep are free of arousal event, even for a patient with high arousal index, the classes both in the TR and TS sets are unbalanced in a proportion of 20% of actual *arousals* versus 80% of *non-arousal* epochs. Therefore, in order to avoid a biased classifier, an under-sampling technique [27] is applied so that the number of examples throughout experimentation is equilibrated within the two sets, resulting in two new balanced TR_{eq} and TS_{eq} sets with 5968 and 1992 patterns respectively.

Two experiments are carried out at this point:

(i) training the classifiers with the TR_{eq} set, we want to test their generalization capabilities over the TS_{eq} set;

(ii) training the classifiers with the TR_{eq} set, we want to test their performance evaluating the entire method. That means we are really using the expanded set *TS+D*,

where *D* denotes the set of epochs not containing a characteristic pattern –which will be directly classified as non-arousal at the first step in the classification, and *TS* denotes the testing set in which classes are unbalanced according to the proportion of arousals present in the recording –which are evaluated at the second classification step;

Therefore, the first experiment pursues the comparison of the different models acting as classifiers over a certain data set, whereas the second determines the performance of the method itself –which comprises two classification steps- using different models acting as classifiers at the second classification step (see Figure 5.7). Comparing (i) and (ii) the consistence on the classification can be measure, moving away from the arousal pattern dataset to a real PSG recording scenario.

Before proceeding with the experimentation, however, the first step is to select the appropriate configurations for the SVMs and ANNs. A 10-fold cross-validation is carried out for this purpose, in conjunction with a grid search on the space of combinations for parameters *S* and *C* for the SVM, and on the number of hidden neurons *H* for the ANN. Using the set TR_{eq} to perform the cross-validation, error measure is taken (the proportion of classes is balanced in this set) as $e_k = 1\text{-}accuracy_k$, where $accuracy_k$ is the proportion of correctly classified patterns in the fold *k*. Through this procedure, best obtained parameters are finally $S = 2^{-5}$ and $C = 2$ for the SVM, achieving an averaged error (*mean*±*standard deviation*) of $E_{SVM} = 1.79\text{x}10^{-1}\pm1.81\text{x}10^{-2}$, whereas the best number of neurons for the ANN is $H = 32$, with an averaged error of $E_{ANN} = 1.71\text{x}10^{-1}\pm1.93\text{x}10^{-2}$.

Once the appropriate configurations have been selected for the SVM and the ANN, comparison is performed on the four models (Linear discriminant, quadratic discriminant, SVM and ANN) to select the best classifier. Results of the comparision are shown in the corresponding section on chapter 7.

B) Feature selection on the detection of EEG arousals

Study of feature selection in the detection of EEG arousals has as its main objective to assess if it is possible to reduce the number of necessary features while keeping adequate detection performance. For *adequate* it has to be understood, at least, to

maintain same detection capabilities as in the case with the whole set of 42 features. Different feature selection methods are studied in this respect and two different classifiers (ANN and SVM) act as evaluators. The use of two classifiers is not aimed at performing a comparative study but at confirming that the predictive power does not depend on the subsequent model used for classification. However as both models have free parameters, a specific configuration has to be chosen for each one, keeping it constant in all the experiments. This has to be done in order to allow comparison of the results among the different number of features. In other words, if the classifier configuration is modified throughout the different subsets of features then the results can be corrupted, influenced by its structure configuration and not by the predictive power of the current set of features.

Thus, in order to allow comparison of the results and fulfill the objectives of the study, two fixed configurations are set for the two classifiers. The ANN is configured within 10 hidden neurons using back-propagation with momentum (learning rate 0.01, momentum 0.9), and the SVM is configured using a radial-basis kernel function with cost parameter C = 1.0. Nevertheless, even with a fixed structure configuration, when dealing with machine learning models a training phase is mandatory to set up internal parameters of the model. Thus for each candidate subset of features, the classifiers are firstly trained using the TR dataset. Later, the TS set is used to check their generalization capabilities. Each classifier is also trained and tested using the whole set of 42 features in order to obtain a reference value of the performance. In this manner it can be evaluated the benefit of using a reduced set of features with respect to the use of the full set of features.

Steps followed for studying the application of the feature selection techniques described above are structured in the following manner:

(i) Based on the previous gathered set of features for the detection of arousals in PSGs (a total of 42 features, see Table 5.2) a subset of 10 patients is randomly chosen from the SHHS database. The use of a subset of 10 patients from the original set of 20 patients is motivated by the high computational cost requirements of wrapper methods. In any case, all the recordings included in this subset are totally independent of those used in IVS.

(ii) The resulting dataset, which consists of 2814 patterns, is split in two new datasets: one dataset composed of 1842 patterns is used for training (TR), while the remaining 942 patterns are used as a separate external test set (TS). Expert's annotations on the scoring of arousals following the AASM criteria [**14**] are used in order to establish the desired output –class- for each pattern contained in the datasets. Both datasets consist of a balanced number of patterns belonging to each class.

(iii) Over the TR dataset, several feature selection methods based on both wrappers and filters are applied in order to discard the irrelevant features. As a result, several candidate subsets –subsets of the most relevant features selected by the method- are constructed based on the measures obtained by each selection method. The process to construct these candidate subsets, which is based on the TR data set, is explained below.

(iv) Once the candidate subsets are constructed, their respective predictive power is measured using both a ANN and a SVM. As stated before, for each candidate subset the classifiers are first trained using the TR dataset. Later, the TS set is used to check their generalization capabilities. Each classifier is also trained and tested using the whole set of 42 features in order to obtain a reference value of the performance. In this manner it can be evaluated the benefit of using a reduced set of features in respect to the use of the full set of features.

That said, the procedure for the construction of candidate subsets (step iii) for the filter methods is done as follows (see Figure 6.5):

- Firstly, over TR dataset, several filter selection methods are used for scoring the different attributes –features-. Each filter method ranks the features in a different way. Therefore, different subsets can be constructed depending on the ranking used and the number of *i*-top positions considered on it.

- Various candidate subsets are explored to compare the adequacy of each method. To do so, each ranking is explored step by step, starting at the top and incorporating features to form a new candidate subset.

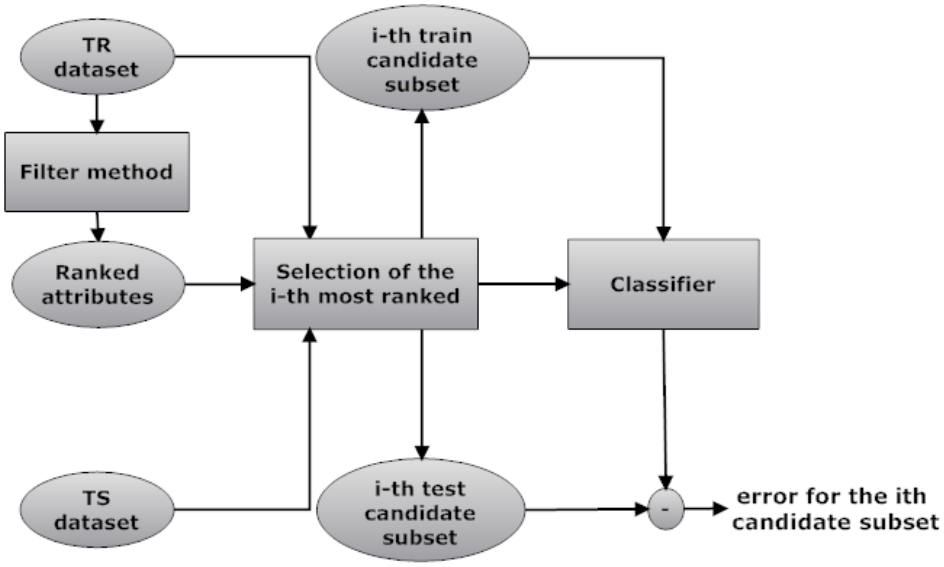


Figure 6.5. Procedure to obtain and evaluate the different candidate subsets using filters

Apart from the three filters described in Chapter 5 "*Identification of EEG arousals*", two more approaches that consist of the combination of the individual rankings provided by each filter are explored. It is expected that a more adequate subset of features, hopefully leading to better results, could be obtained. For that purpose, the intersection and the union of the rankings of the three filters used are also tested. Several different candidate subsets are obtained respectively as follows:

$$\text{subsetI(i)} = \{\cap(\text{OneR(i)}, \text{Relief(i)}, \text{InfoGain(i)})\} \tag{6.1}$$

$$\text{subsetU(i)} = \{\cup(\text{OneR(i)}, \text{Relief(i)}, \text{InfoGain(i)})\} \tag{6.2}$$

where at step *i*, *OneR(i)*, *Relief(i)*, *InfoGain(i)* represent respectively the *i*-first ranked features obtained by the OneR, Relief and InfoGain methods.

On the other hand, in the case of the wrapper methods, by combining both classifiers and the two search strategies, four individual wrapper approaches are obtained. Each wrapper returns a feature subset that is checked by applying the classifier over the test set. The procedure to obtain the candidate subsets, which is slightly different from the case of filters, is illustrated in Figure 6.6, and is as follows:

- In order to avoid the bias of the pattern distribution in the training set (TR), a 10-fold cross-validation (CV) is applied. Therefore a potentially different subset of features is obtained for each fold.

- For those subsets, the potential importance of each feature is measured by counting the number of times the feature appears in the 10 subsets, –so, being the maximum value 10-. The features are then ranked in descending order, but taking into account that several features may have the same importance.

- In this way, each candidate subset is formed by exploring this ranking step by step: starting with the subset of features with the highest frequency of appearance and including, in decreasing order for each step, those features of inferior order of appearance.

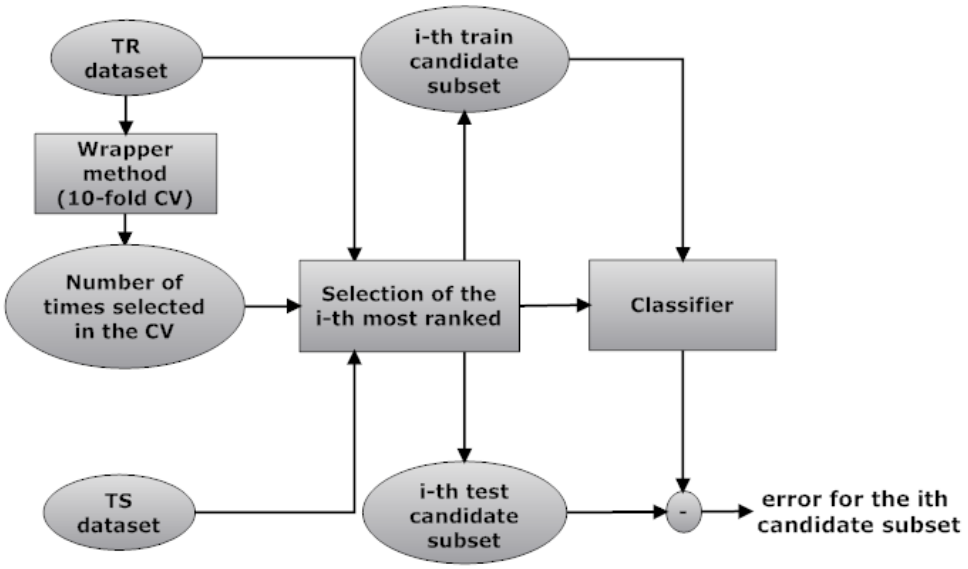


Figure 6.6. Procedure to obtain and evaluate the different candidate subsets using wrappers

Apart from these four methods, and analogously to the case of the filters, the combination of the rankings by the union and the intersection of their features is investigated (see eq. 6.1 and 6.2). Results of the analysis for both filters and wrappers are given in the next chapter in section "*Feature selection on the detection of EEG arousals*".

C) Final validation of EEG arousals detection using the IVS set

The objective of this validation test is to check the performance of the final resulting method for EEG arousal identification, after feature selection methods were applied in order to reduce the number of initial features. For that purpose the independent validation set (IVS) of 26 recordings is used containing a total of 31080 epochs for validation.

Specifically, from the results of the two previous sections (i) a classifier method between the four tested models in A) might have been selected as candidate, and (ii) the best set of predicting features might have been determined in B). Thus, in order to assess final results of the method, final training process is scheduled using the selected machine learning classifier over the best subset of predicting features.

Final validation results are displayed over each individual recording and agreement ratios are calculated. The resulting agreement index, sensitivity, specificity and AUC values are then computed overall and for each individual recording. Results of the validation as well as the corresponding statistical analysis can be found in the corresponding section of Chapter 7.

6.3.2. Sleep Staging

In order to validate the proposed approach annotations made by expert polysomnographic scorers are taken as reference for the validation process. As it has been outlined in Chapter 5 "*Hypnogram generation*", it is important to remark that since the experts follow the R&K procedure (30 seconds labeling), then the epoch-based output (non continuous) from the system, is used in order to allow the validation process. Recall this epoch-based output is the result of the post-processings described in the same section after the continuous hypnogram has been obtained.

Note that for the calculation of agreement ratios, in this case the number of possible categories is four (W, DS, DEEP and REM), so that given a certain category *C*, then a negative case (*not C*) is considered when the related classification involves any of the remaining three categories. In other words, if for example the system scores an epoch as

W and the standard reference classifies it either as DS, DEEP or REM, then a false positive is scored for category W and a false negative is computed for DS, DEEP and REM. That said, agreement indexes, sensitivities, specificities and AUC values can be easily calculated for each category, by constructing the corresponding 2x2 confusion matrix as explained above in section "*Agreement ratios*".

Availability of four possible categories for each classifiable element in this case makes of interest the construction of a contingency table as described in section "*Pair-wise measures*". From this table analysis kappa can be performed to obtain resulting agreement not affected by chance. Percent of positive agreement can also be investigated here since, according to the procedure described above for calculation of agreement ratios, the high number of true negatives for each category might be obscuring actual misclassifications of the sleep stages. Study of overall deviating scorings can be computed as well from the contingency table to determine which sleep stages are involved in the misclassifications.

Results of the sleep staging validation and the corresponding data analysis can be found in section "*Hypnogram generation*" of Chapter 7.

6.3.3. Apneic events detection

Validation here is focused on two distinct aspects: (i) expert-system agreement in regard to the location of apneic events, and (ii) classification of apneic events as apneas or hypopneas.

Specifically, since every apneic pattern has associated three different degrees of membership with respect to categories apnea, hypopnea and false positive, assignment of an apneic pattern (AP) to a concrete category uses a maximum criterion as described in Chapter 5. That is, for (i) an apneic event is confirmed if:

$$Apneic\ event \Leftrightarrow \max\left\{\mu_{\text{apnea}}(\text{AP}), \mu_{\text{hypopnea}}(\text{AP})\right\} \geq \mu_{\text{falsepositive}}(\text{AP})$$

and for (ii):

$$Apnea \Leftrightarrow \max\left\{\mu_{\text{hypopnea}}(\text{AP}), \mu_{\text{false positive}}(\text{AP})\right\} \leq \mu_{\text{apnea}}(\text{AP}),$$

$$Hypopnea \Leftrightarrow \max\left\{\mu_{\text{apnea}}(\text{AP}), \mu_{\text{false positive}}(\text{AP})\right\} \leq \mu_{\text{hypopnea}}(\text{AP})$$

For the first goal temporal positioning of the detected events is crucial. However temporal location of events is difficult from the point of view of implementing validation. Any given event can be represented by a segment located in time and demarcated by a start point and an end point. In this regard, in order to determine an exact matching, it is very difficult to compare two segments in terms of an exact coincidence according to their starting and ending points. Let us assume, for example, that precision of the temporal scale is that of real numbers and that the starting point for two segments is the same. Given the respective ending points, end_point_A and end_point_B, let us assume that these differ by 0.3 seconds. A human scorer might then consider that these two segments represent a temporal match –i.e. they refer to the same event, yet from the computational perspective, and taking into account precision of real numbers, they cannot be considered to be the same as they do not coincide exactly in time. On the other hand, it is reasonable to admit a certain difference Δ in the location of the starting and ending points when comparing two segments to see if they belong to the same event. However assigning a value to Δ is in itself a problem as one has to consider the different combinations between the starting and ending points which also depend on the concrete situation.

The proposed solution is similar to that applied for the validation of EEG arousals, and it consists in the segmentation of the temporal axis into classifiable—and therefore comparable—units. Again, the concept of epoch can be taken as an arbitrary unit of time in the context of sleep studies. Thus, by locating the mid-point of the segment that represents the event, this can be unequivocally assigned to an epoch, and the validation process can be implemented by comparing epochs annotated, respectively, by either the system or the standard reference (see Figure 6.7).

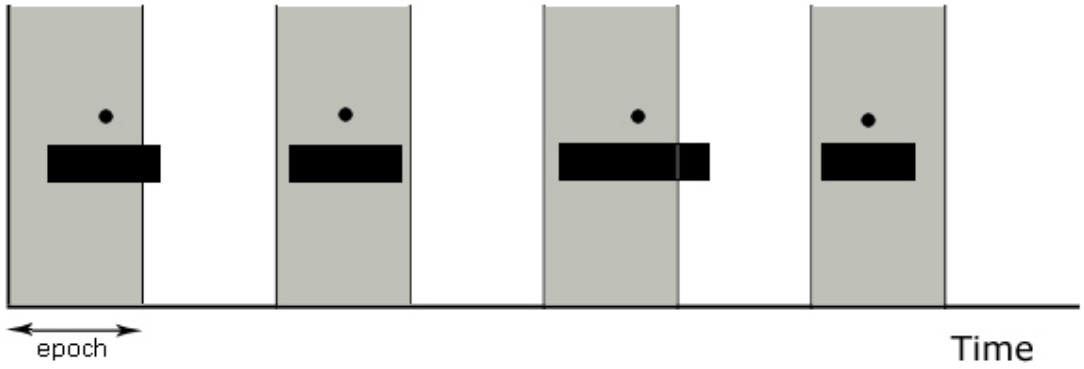


Figure 6.7. Each black rectangle represents an apneic event over a time scale (the x axis). The circles represent the mid-points that locate apneic events unequivocally in an epoch (in grey)

Therefore, similarly to the case of the assignment of arousal patterns to a specific epoch, to implement the validation process the concept of epoch is used in order to decide on the existence of an apneic event within an epoch. The former process enables us to construct the contingency table as in section "*Agreement ratios*" of the current chapter and to calculate the number of TP, TN, FP and FN accordingly. For example in the cases where both system and standard reference detect an apneic event in the same epoch then TP is scored; if an apneic event is detected by the system but it is not marked by the standard reference then a FP is computed, and so on.

Once an apneic event is localized in a concrete epoch through its mid-point, then validation regarding its concrete type can also be assessed. For this purpose, epochs in which there is positive agreement on the location of apneic events are used in order to validate the system with respect to classification of the events as apneas or hypopneas. On the other hand, since validation is performed for epochs, in order to correctly validate apneic event classification, only epochs with just one apneic event are considered. For validation purposes, for example, it would not be clear whether an epoch annotated by experts as featured by an apnea and by a hypopnea should be actually considered as apnea or as hypopnea.

Validation results and analysis of data are shown on the corresponding section of Chapter 7.

6.3.4. Apneic events classification

As it has been described throughout Chapter 5, three different classes of apneic events are considered from our system: obstructive, central and mixed. However, as it has also been introduced in this chapter, standard reference used for validation only considers classification as obstructive or central over apnea events. Specifically, no distinction is made between mixed apneas and obstructive apneas. Therefore, for validation purposes events classified as mixed for our system will be considered within the obstructive category. Hypopnea events are not either classified by the standard reference (they are considered as obstructive by default). Therefore system validation with respect to classification of origin of the apneic event is limited to the classification of events of type apnea. For these events, again, the maximum criterion is applied for classification of the confirmed apneic event into a concrete category, and therefore:

$$Obstructive \Leftrightarrow \max\left\{\mu_{\text{obstructive}}(\text{AP}), \mu_{\text{mixed}}(\text{AP})\right\} \geq \mu_{\text{central}}(\text{AP})$$

$$Central \Leftrightarrow \max\left\{\mu_{\text{obstructive}}(\text{AP}), \mu_{\text{mixed}}(\text{AP})\right\} \leq \mu_{\text{central}}(\text{AP})$$

Similarly to the validation tests scheduled for evaluation of the detection phase, in this case, the type of validation measures to be used comprises calculation of agreement ratios for each of the previous apnea categories: obstructive and central. It has to be taken into account that obviously, only the set of true positives for the class apnea can be evaluated for this classification. Thus, the number of cases to be evaluated depends on the number of true positives for the apnea category marked in the validation of the preceding detection phase (see previous section). Results of the validation process and the corresponding data analysis are shown in the corresponding section of the subsequent Chapter 7.

6.3.5. Final patient diagnosis

The objective of this validation test is to assess reliability of the system with regard to final diagnosis of the patient. In this respect a first aspect of interest is to calculate final AHI values for each of the 26 recordings in the IVS set. Comparison on the obtained values is then performed against the respective AHI values according to the

standard reference. Since comparison involves numerical data (not categorical), validation in this respect is carried out using the methods described in previous subsection "*Measures involving numerical data*".

In addition to the analysis of the AHI, consistency of the TST as well as validation considering ApI and HI indices separately is also assessed. Validation from a categorical perspective is performed by taking into consideration the respective classifications of syndrome severity associated to the AHI. For that purpose AHI is segmented according to the categories established in Table 5.4 (see Chapter 5, "*Diagnosis generation*"). The comparative is performed through the calculation of the corresponding kappa index and the analysis of the respective severity distributions.

Similar analysis is performed for the respective syndrome classifications according to the prevalent type of apneic event in the patients, i.e. obstructive, central or mixed. Respective AHI_{obs} and AHI_{cen} between the system and the standard reference are compared using numerical methods, and then from the categorical perspective according to the corresponding nominal classification. Criteria for syndrome classification according to the previous indices can be consulted in subsection "*Diagnosis generation*" of Chapter 5. Again, the analysis is repeated by taking into account final indices for positional SAHS (AHI_s and AHI_{ns}).

Final computed indices for ArI are also compared and analyzed from the numerical perspective. Finally, additional comparative validation is performed with regard to the relation between apneic events and arousals, the distribution of apneic events over the different sleep stages, and the correlation between different estimations of sleep fragmentation.

Results and data analysis concerning final validation can be found in subsection "*Final patient diagnosis*" of Chapter 7.

6.4 Summary of this chapter

This chapter introduces the set of validation measures and the design of the tests used for the validation of the system described through the previous chapters.

An introduction to the validation process is firstly given, specially focusing in the validation of intelligent systems. In this respect different types of procedures and measures are presented which are firstly structured into two great groups, depending on their adequateness to handle categorical or numerical data. Within each group further classification is performed, for example, according to pair-wise measures, group measures and agreement ratios. The different types of validation measures are then described in more detail. Measures that operate over categorical data include agreement index, kappa analysis, chi-squared test for homogeneity, sensitivity, specificity, false positive ratio, false negative ratio, specific agreements, and area under ROC curve. Among measures operating over numerical data, several statistical tests of interest are introduced for the analysis of two numerical distributions such as Student's and Wilcoxon tests, linear and Spearman's correlation, and group measures for the comparison of several numerical distributions such as ANOVA analysis, Kruskal-Wallis and intraclass correlation coefficient. Additionally, in order to perform model selection over methods based on machine learning approaches, a general methodological framework is introduced describing specific mechanisms and metrics.

The chapter continues detailing the design of the validation tests, beginning with the available data and explaining the followed procedures according to each case. Validation data is taken from real PSG recordings from the Sleep Heart Health Study. The recordings are annotated by expert scorers which are taken as the gold standard. Thus, in general, validation is carried out by confronting system's outputs with expert's annotations in the recordings. In order to perform a more structured validation, the process is performed by splitting the main task into several subtasks of interest: detection of EEG arousals and feature selection, sleep staging, apneic events detection (*apnea*/*hypopnea*/*false positive*), apneic events classification (*obstructive*/*mixed*/*central*) and final diagnostic. Specific validation procedure and measures used in each case are respectively described throughout the final part of the chapter.

6.5 References

[1] RM. O'Keefe, O. Balci, and EP. Smith, "Validating expert system performance," *IEEE Expert magazine*, vol. 2, no. 4, pp. 81-90, 1987.

[2] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Phychological Measurement*, vol. 20, pp. 37-46, 1960.

[3] JR. Landis and GG. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159-174, 1977.

[4] AP. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, 1997.

[5] JA. Hanley and BJ. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29-36, 1982.

[6] E. Mosqueira-Rey and V. Moret-Bonillo, *Validación de sistemas inteligentes*.: Tórculo Edicións, 2001.

[7] JH. McDonald, *Handbook of biological statistics*, 2nd ed. Baltimore, Maryland: Sparky House Publishing, 2009.

[8] MG. Kendall and A. Stuart, *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*.: Griffin, 1973.

[9] PE. Shrout and JL. Fleiss, "Intraclass correlations: uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, no. 2, pp. 420-428, 1979.

[10] Y. Hochberg and AC. Tamhane, *Multiple Comparison Procedures*.: John Wiley & Sons, 1987.

[11] SF. Quan et al., "The Sleep Heart Health Study: design, rationale and methods," *Sleep*, vol. 20, no. 12, pp. 1077-1085, 1997.

[12] B. Kemp, A. Värri, AC. Rosa, KD. Nielson, and J. Grade, "A simple format for exchange of digitalized polygraphic recordings," *Electroencephalography and Clinical Neurophysiology*, vol. 82, pp. 391-393, 1992.

[13] Sleep Heart Health Study SRC, "Sleep Heart Health Study. Reading center manual of operations," Case Western Reserve University, Tech Report VMLA-039-02, 2002.

[14] C. Iber, S. Ancoli-Israel, A. Chesson, and SF. Quan, "The AASM Manual for the scoring of sleep and associated events: rules, terminology and technical specifications," American Academy of Sleep Medicine, Westchester, IL, 2007.

[15] The Atlas Task Force, "EEG Arousals: scoring rules and examples," *Sleep*, vol. 15, pp. 174-184, 1992.

7. RESULTS

In this chapter presentation and analysis of the validation results is performed according to the validation tests designed in the preceding chapter. In this regard sections are organized in arrangement to the corresponding subsections of Chapter 6. Specifically, the chapter is organized according to the following validation subtasks: (1) detection of EEG arousals and feature selection, (2) sleep staging, (3) detection and differentiation of apnea and hypopnea events, (4) classification of apneic events as obstructive, central or mixed, and (5) final patient diagnosis. Further details on the results of the different validation tests are given throughout the subsequent sections.

7.1. Identification of EEG arousals

A) Detection of EEG arousals using machine learning models

In accordance with validation strategy described in subsection "*Detection of EEG arousals*" of Chapter 6, two first experiments are carried out to assess (i) classification capabilities of the classifiers over TS_{eq} set and (ii) within the method over the expanded set TS+D. Table 7.1 and Table 7.2 respectively show the results for the two experiments. In order to provide statistical significance to the results the procedure was repeated 30 times. Therefore the results showed in Table 7.1 and Table 7.2 represent average values on each measure. Note that the first experiment (see Table 7.1) pursues a preliminary comparison of the classifiers under ideal conditions where the number of classes is balanced. The best results on each measure are stressed in bold.

Table 7.1. Results using various classifiers trained with the balanced set TReq and tested in the balanced testing set TSeq. AUC = Area Under ROC Curve

	TReq vs. TSeq			
Model	Error	Sensitivity	Specificity	AUC
Linear discriminant	0.275	0.667	0.784	0.725
Quadratic discriminant	0.271	0.680	0.779	0.730
SVM	0.199	0.781	**0.820**	0.801
ANN	**0.188**	**0.812**	0.812	**0.812**

It can be seen from the results in Table 7.1 that both the SVM and ANN demonstrate better results than the linear and quadratic discriminants. Both offer similar results, thus pair-wise comparison between the measures of the two models is performed using the non-parametric Wilcoxon test in order to check for statistically significant differences between model's outputs. Based on this test it can be concluded that the SVM has better specificity than the ANN, whereas the ANN has better sensitivity (*p-value = 1.4x10*$^{-3}$). Normally in medicine one desires the most sensitive test. In addition, taking into account that both, classification error and AUC, provide a measure on the overall performance, the ANN can be considered the best model in this case (*p-value = 1.4x10*$^{-3}$).

Subsequently, Table 7.2 shows the results for the performance of the method on each of the classifiers tested over the whole classification stage. The best results for each measure are similarly marked in bold.

Table 7.2. Results from the method using various classifiers trained with the balanced set TReq. AUC = Area Under ROC Curve

	TReq vs. TS+D (method)			
Model	Error	Sensitivity	Specificity	AUC
Linear discriminant	0.229	0.739	0.777	0.758
Quadratic discriminant	0.238	0.751	0.764	0.757
SVM	**0.198**	0.840	**0.794**	**0.817**
ANN	0.218	**0.868**	0.765	**0.817**

The results show a similar trend to those in Table 7.1, with the SVM and the ANN achieving the best results. An increase in the sensitivity is seen while specificity decreases. There is also a slight increase on the classification error. This can be

expected due to the new unbalanced scenario in which the proportion of actual events decreases in comparison with the number of non-arousal epochs, therefore increasing the number of false positives. However, it is remarkable that an increase in the average performance (sensitivity and specificity) occurs reflected in a higher AUC index when compared to Table 7.1. Since both classifiers perform equal in terms of AUC (*p-value = 1*), again it is necessary to decide if it is preferable to choose the more sensitive or the more specific method. As previously stated, it is chosen the one with the highest sensitivity since in medicine one normally prefers to over detect the disease, instead of under detecting it. Thus, the ANN is the model selected to act as the classifier since it is more sensitive (*p-value* = $6.34x10^{-5}$) than the SVM. Nevertheless, the preference for ANN can only be tentatively conjectured given the small difference between the ANN and SVM, and the limited number of patients used in the test set. In fact, lower classification error is achieved by the SVM.

The choice of a balanced set in order to train the models could be a discussion point, taking into account that in a real full recording the classes are, in fact, unbalanced. However what happens if one trains the models with the TR set, i.e. with unbalanced classes, is that the classifiers tend to learn only the majority class overlooking the other. Table 7.3 and Table 7.4 are provided with the results of such training. Data have been obtained by exactly the same procedure followed in Table 7.2 and Table 7.3 but this time using the TR set to train the models.

It is clear that the results obtained are much worse, and although at first sight an improvement on the error is perceived in Table 7.4 with respect to Table 7.2, one is aware of the drastic drop in sensitivity because the classifier is learning only one class. Results in the balanced testing set of Table 7.3 confirm this hypothesis, and it is also confirmed that training the classifiers with TR_{eq} is the best option.

Table 7.3. Results using various classifiers trained with the unbalanced set TR and tested in the balanced testing set TSeq. AUC = Area Under ROC Curve

	TR vs. TSeq			
Model	Error	Sensitivity	Specificity	AUC
Linear discriminant	**0.290**	0.539	0.881	**0.710**
Quadratic discriminant	**0.290**	**0.553**	0.867	**0.710**
SVM	0.331	0.368	0.971	0.669
ANN	0.332	0.356	**0.977**	0.667

Table 7.4. Results from the method using various classifiers trained with the unbalanced set TR. AUC = Area Under ROC Curve

	TR vs. TS+D (method)			
Model	Error	Sensitivity	Specificity	AUC
Linear discriminant	0.181	0.619	0.858	0.739
Quadratic discriminant	0.186	**0.644**	0.847	0.746
SVM	0.107	0.491	**0.972**	0.732
ANN	**0.105**	0.509	0.971	**0.740**

B) Feature selection on the detection of EEG arousals

In accordance with the validation procedure depicted throughout the corresponding subsection of Chapter 6, first step to study the benefits of feature selection methods involved the calculation of a reference value for the used subset of 10 PSG recordings.

Thus, in order to obtain a reference value, the classifiers (ANN and SVM) are firstly trained using the full set of 42 features. Table 7.5 shows the results achieved using as measure the classification error, i.e. the proportion of misclassified patterns. It can be shown that resulting values are in accordance with those obtained in Table 7.1 for the proposed method and evaluated using 5 recordings acting as the testing set (see previous subsection "*Detection of EEG arousals using machine learning models*"). Lower values of error with respect to values of Table 7.5 are to be expected by the application of feature selection techniques.

Table 7.5. Proportion of misclassified patterns using the full set of 42 features over the TS set

SVM	ANN
0.196	0.194

Results for filters methods

The candidate subsets obtained using filter methods are shown in Table 7.6. Candidate subsets for each filter method are constructed taking into account the 20th first ranked features in each case. For an easier display, steps down through rankings are taken two-by-two on the features for the individual filters. Feature numbers (see Table 5.2 to identify the specific features) are provided in order to describe the subset. Each row includes the features of the previous row plus (+) the corresponding two new ones.

Table 7.6. Candidate subsets for the individual filters

	Candidate Subsets		
Number of features	InfoGain	OneR	Relief
2	38,39	39,37	41,40
4	+28,37	+38,21	+38,37
6	+3,21	+3,41	+39,28
8	+10,20	+28,23	+10,5
10	+23,41	+10,20	+23,18
12	+2,19	+40,5	+35,2
14	+1,5	+2,42	+20,8
16	+30,40	+12,1	+13,15
18	+12,33	+19,30	+33,14
20	+13,31	+13,31	+36,17

Using the previous candidate subsets, SVM and ANN are trained using the TR dataset and evaluated over the TS dataset. Figure 7.1 and Figure 7.2 show the error achieved by the filters on the different candidate subsets on the external test set (TS) using the SVM and ANN classifiers, respectively. The dotted line represents the error value achieved with the complete set of 42 features.

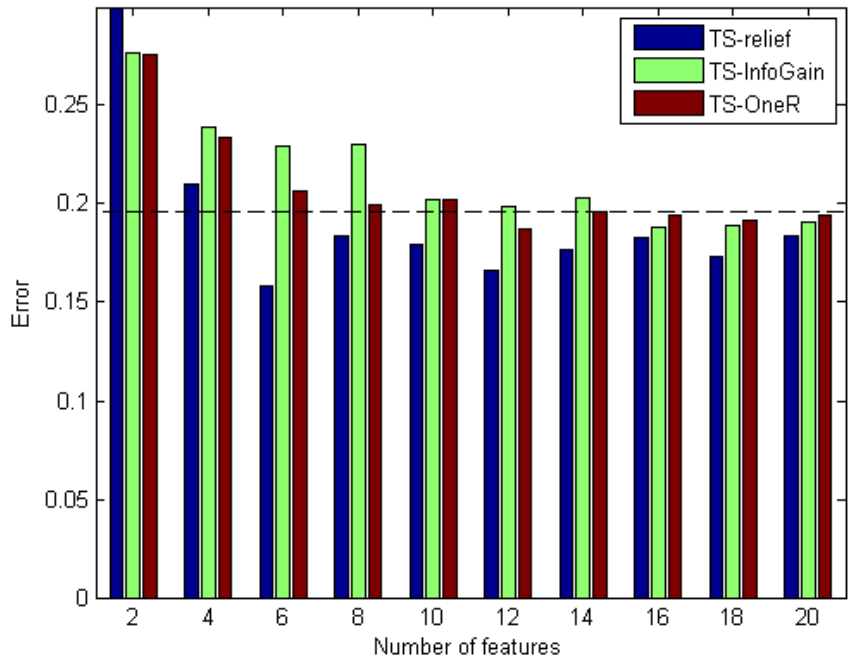


Figure 7.1. Error achieved by the three filter methods checked, using SVM as classifier. The dotted line represents the error achieved by the SVM using the full set of 42 features

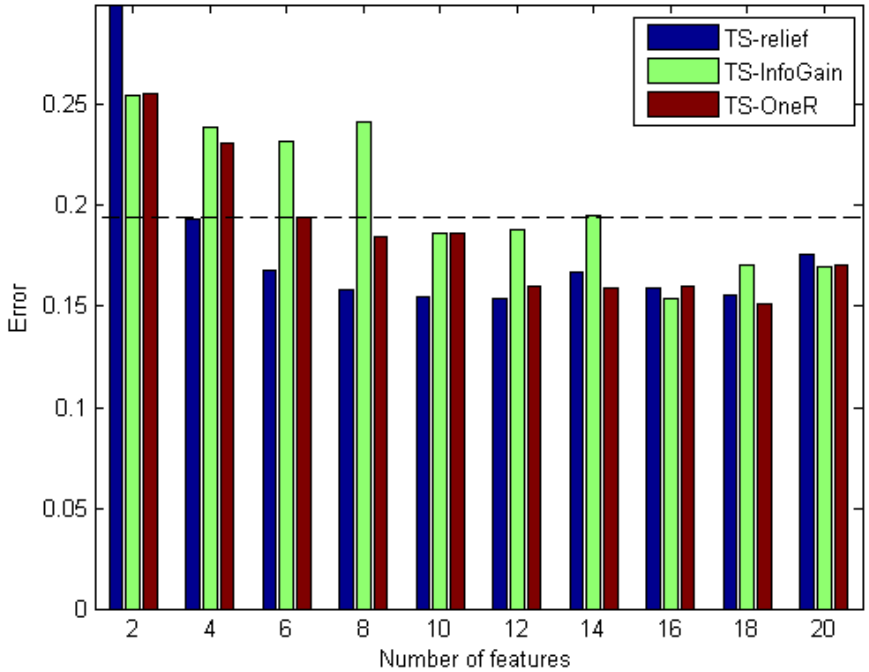


Figure 7.2. Error achieved by the three filter methods checked, using ANN as classifier. The dotted line represents the error achieved by the ANN using the full set of 42 features

In order to investigate filter combinations, the union and the intersection of the candidate subsets on Table 7.6 are performed. The resulting candidate subsets together with their corresponding predictive power using the ANN and the SVM are shown, respectively for the union and the intersection, in Table 7.7 and Table 7.8. In both tables, due to the union and intersection operations (see Chapter 6, eq. 6.1 and 6.2) with the i^{th} top-ranked features on the filters application, the resulting number of elements for each step is not constant, as in the individual filters. Therefore the results cannot be included in the previous figures but in form of tables, where in order to make a comparative study, the errors of the rows can be contrasted with those of the filter approaches using a similar number of features. In Table 7.7 and Table 7.8 the results with a lower error with respect to the reference subset (see Table 7.5) are shown in boldface. For the lowest error achieved, the results are highlighted with a grey background.

Table 7.7. Filters' candidate subsets evaluation for the union of the top ranked features on filters. Errors under the reference value are shown in boldface, and the best values of the table are displayed with grey background

Number of features	*Candidate subsets*	*SVM Error*	*ANN Error*
5	37,38,39,40,41	0.197	**0.185**
6	+28	**0.158**	**0.168**
7	+21	**0.183**	**0.158**
8	+3	**0.177**	**0.150**
9	+10	**0.178**	**0.166**
12	+5,20,23	**0.187**	**0.160**
13	+18	**0.180**	**0.156**
15	+2,35	**0.174**	**0.156**
16	+19	**0.179**	**0.171**
19	+1,8,42	**0.180**	**0.162**
22	+12,13,30	**0.188**	**0.180**
23	+15	**0.187**	**0.162**
24	+33	**0.191**	**0.178**
25	+14	**0.190**	**0.172**
27	+36,31	**0.192**	**0.158**
28	+17	**0.193**	**0.184**

Table 7.8. Filters' candidate subset evaluation for the intersection in terms of misclassified patterns. Errors under the reference value are shown in bold, while the best ones are displayed with a grey background

Number of features	*Candidate subsets*	*SVM Error*	*ANN Error*
1	38	0.279	0.279
2	+37	0.279	0.277
3	+39	0.279	0.254
4	+28	0.239	0.239
6	+10,23	0.230	0.226
7	+41	0.221	**0.190**
10	+2,5,20	0.210	**0.191**
11	+40	**0.191**	**0.170**
12	+13	0.203	**0.160**

Table 7.9 summarizes the best results achieved for each of the methods described above. Comparing the results obtained when using SVM, Relief is the best individual filter in both number of features and level of error. Related to the combination of methods, in this case, the union of filters achieves the same result as Relief, while the intersection achieves higher values for the error and even using more features. Using ANN, the union of filters gets the best values of all the trials: individual filters and their combinations. In this case, it is not clear the election of the best individual filter. On one hand, OneR obtains the lowest value for the error, but on the other hand, Relief uses the smallest set of features. In any case, Table 7.9 clearly denotes that the union of filters leads to the best performance results without increasing the number of features used.

Table 7.9. Best results achieved by each filter method and their combinations in the TS dataset. Best results are those showing the lowest percentage of misclassified events –error-. The number of features related with the specific error is showed as well

	SVM		*ANN*	
Filter method or combination	Error	Number of features	Error	Number of features
Relief	0.158	6	0.154	12
InfoGain	0.188	16	0.154	16
OneR	0.187	12	0.151	18
Union	0.158	6	0.150	8
Intersection	0.191	11	0.160	12

Results for the wrappers methods

In the case of wrappers the results for the individual methods are shown in Table 7.10, Table 7.11, Table 7.12 and Table 7.13. The procedure used to obtain these results has been described in section "*Feature selection on the detection of EEG arousals*" of Chapter 6. Table 7.10 and Table 7.11 show the results for the SVM-based wrapper, for the forward search and the backward search, respectively. Similarly, Table 7.12 and Table 7.13 show the results for the ANN-based wrapper. Notice that, independently of the classifier employed as part of the wrapper, SVM and ANN are trained and tested using the different subsets of obtained features. Error measures in the tables comprise the results achieved using the test set. It is also important to remember that a 10-fold cross-validation has been performed to select the features, and so different subsets of features may be selected at each fold. The first column in all the tables contains the number of times (NTS) a feature has been selected within the 10-fold, -being 10 the maximum number of times and 0 the minimum. Initially, only features that have been selected for all the folds are chosen, and subsequently, features that are selected in 9, 8, etc. folds are included. The second column in these tables shows the features involved in the resulting candidate subset. Analogously to the tables when using filter methods, each row includes the features contained in the previous row plus (+) the corresponding new ones. When the resulting feature subset remains unaltered from one row to the next one, the row is omitted. The number of features on the resulting candidate subset is indicated in the third column (NF). The next two columns show the error achieved in the TS test using, respectively, the SVM and the ANN classifiers. As in the case of filters, values of error below the reference are boldfaced, whereas the value with the lowest error is showed with a grey background.

Table 7.10. Candidate subsets for the SVM wrapper with forward search and their performance on the TS set. NTS = Number of times selected on the 10-fold; NF = number of features in the candidate subset

NTS	Candidate subset	NF	SVM Error	ANN Error
10	3,39,40,41	4	**0.173**	**0.172**
9	+18	5	**0.173**	**0.168**
7	+5,28,34	8	**0.163**	**0.153**
6	+17,35	10	**0.175**	**0.151**
5	+16	11	**0.172**	**0.171**
4	+4,21,25,31,36	16	**0.179**	**0.165**
3	+8,9,23,27,29	21	**0.183**	**0.182**
2	+10,15,24	24	**0.176**	**0.176**
1	+1,2,6,12,14,20,22,26,30,37	34	**0.187**	**0.160**

Table 7.11. Candidate subsets for the SVM wrapper with backward search and their performance on the TS set. NTS = Number of times selected on the 10-fold; NF = number of features in the candidate subset

NTS	Candidate subset	NF	SVM Error	ANN Error
10	18,39,40,41,42	5	**0.194**	**0.186**
9	+12,14,15,33,34	10	**0.195**	0.195
8	+13,22,28,35	14	**0.179**	**0.171**
7	+1,5,9,16,17,36	20	**0.183**	**0.174**
6	+3,20,27,31	24	**0.184**	**0.170**
5	+4,6,8,29,30,32	30	**0.188**	**0.170**
4	+7,26	32	**0.186**	**0.179**
3	+10,21,23,24	36	**0.187**	**0.191**
2	+2,11,25	39	**0.194**	**0.177**

Table 7.12. Candidate subsets for the ANN wrapper with forward search and their performance on the TS set. NTS = Number of times selected on the 10-fold; NF = number of features in the candidate subset

NTS	Candidate subset	NF	SVM Error	ANN Error
10	28,39,40,41	4	**0.160**	**0.162**
9	+3	5	**0.165**	**0.165**
8	+8,13,21,31	9	**0.185**	**0.168**
7	+2,5,9,11,15,26,27,30,33,35,36,42	21	**0.179**	**0.172**
6	+10,12,17,18,19,24,34,37,38	30	**0.191**	**0.179**
5	+4,6,20,29	34	0.197	**0.176**
4	+1,7,14,16,22,23	40	**0.193**	0.205
3	+25,32	42	**0.196**	**0.194**

Table 7.13. Candidate subsets for the ANN wrapper with backward search and their performance on the TS set. NTS = Number of times selected on the 10-fold; NF = number of features in the candidate subset

NTS	Candidate subset	NF	SVM Error	ANN Error
10	3,6,10,11,17,19,20,21,23,26,32,35,36,38,39,40,41	17	**0.180**	**0.168**
9	+1,7,12,15,16,18,22,25,27,28,31,34,37,42	31	**0.193**	**0.188**
8	+2,4,13,14,29,30,33	38	**0.196**	**0.188**
7	+5,8,9	41	0.200	**0.179**
6	+24	42	**0.196**	**0.194**

Comparing forward and backward search strategies (i.e. Table 7.10 versus Table 7.11 and Table 7.12 versus Table 7.13), it is clear that the forward strategy obtains better performance results while using a reduced set of features. Besides, it can be checked that, in general, the best results are achieved using the ANN classifier, independently on the classifier used for construction of the wrapper (see Table 7.10).

Finally, as the best results are obtained using the forward search strategy, only those wrappers are used for the combination using the union and the intersection, i.e., wrappers which results are shown in Table 7.10 and Table 7.12. Table 7.14 and Table 7.15 show the results for the union and the intersection of wrappers, respectively. Note that, after intersecting or adding features, the number of subsets obtained is different and, therefore, the resulting number of rows in these tables can vary with regard to the number of rows in Table 7.10 and Table 7.12 (see Chapter 6, eq. 6.1 and 6.2).

Table 7.14. Candidate subsets for the intersection of the individual wrappers (ANN and SVM) in the forward search. NTS = Number of times selected on the 10-fold; NF = number of features in the resulting subset

NTS	Candidate subset	NF	SVM Error	ANN Error
10	39,40,41	3	**0.192**	**0.190**
9	+3	4	**0.173**	**0.172**
7	+5,28	6	**0.172**	**0.161**
6	+18,28,34,17,35	11	**0.175**	**0.171**
4	+16,4,21,31,36	16	**0.179**	**0.171**
3	+25,8,9,23,27,29	22	**0.183**	**0.183**
2	+10,15,24	25	**0.178**	**0.178**
1	+1,2,6,12,14,20,22,26,30,37	35	**0.185**	**0.185**

Table 7.15. Candidate subsets for the union of the individual wrappers (ANN and SVM) in the forward search. NTS = Number of times selected on the 10-fold; NF = number of features in the resulting subset

NTS	Candidate subset	NF	SVM Error	ANN Error
10	39,40,41,3,28	5	**0.164**	**0.156**
9	+18	6	**0.162**	**0.162**
8	+8,13,21,31	10	**0.189**	**0.163**
7	+5,34,2,9,11,15,26,27,30,33,35,36,42	23	**0.184**	**0.173**
6	+17,10,12,19,24,37,38	30	**0.191**	**0.179**
5	+16,4,6,20,29	35	0.200	**0.172**
4	+25,1,7,14,22,23	41	0.196	0.200
3	+32	42	0.196	**0.184**

Overall, the best performance results for wrappers are obtained using a ANN classifier after selecting features using a SVM-based wrapper (0.151 in Table 7.10). This error value is not outperformed by the union or the intersection combinations. However, it is important to mention that the union obtains a similar value (0.156 in Table 7.15), but using less number of features (5 instead of 10). So, similarly to filter results, the union of different wrappers achieves almost top performance results while using fewer features than the individual methods.

In the discussion about the determination of the best feature selection method, it is necessary to have two factors under consideration: the improvement of error and the number of features used to obtain that error. Normally, better accuracy is obtained by including more features in the selected subset. However, it may be preferable to achieve a reasonable accuracy while trying to reduce the number of features to the minimum. To be taken into mind is that, in the field of medicine, normally the major interest is in the prediction capabilities of the test, rather than in the amount of information needed. Supposing that no extra cost is associated with the amount of features –as it is in this case- it can be concluded that obtaining better accuracy is preferred.

After the previous considerations, in the case of filters, the method which achieves the lowest error is the union (0.150 error / 8 features). In the case of wrappers, the obtained results also corroborate the union as a better combination method than the intersection. For wrappers, however, the best absolute values are achieved using the ANN with forward search (0.151 error / 10 features). In any case the union is in fact fairly close with half of the features (0.156 error / 5 features).

In conclusion the improvement on the performance can be shown in both the results using the ANN and the SVM as classifiers, regardless of the feature selection method used. For filters, the best method is the union regardless of the used classifier. The results on wrappers show that the selection of features performs better, in general, following a forward search strategy and using the ANN as classifier (independently of the use model used as wrapper). The union of the features has also showed to perform well. Trying to make a decision on the use of filters or wrappers, the results are not conclusive in terms of error. However, considering the higher computational requirements of wrappers, filters seem to be a more adequate method. After the previous considerations, in the case of filters, the method which achieves the lowest error is the union (0.150 error / 8 features). It is important to note the considerable reduction achieved in the percentage of error compared with the results using the whole set of features (0.196 for SVM and 0.194 for ANN). In addition the great reduction in the number of features is achieved (for union of filters 8 versus 42, thus over 80% reduction with respect to the original set).

C) Final validation of EEG arousals detection using the IVS set

As it has been shown in the preceding subsections, referring to Table 7.2, a method for the identification of arousal events in PSG recordings has been reported achieving, respectively, a sensitivity and specificity of 0.868 and 0.765, or equivalently 86,8% and 76.5%. These results have been obtained using the full set of 42 features over a set of 20 patients, divided into 15 recordings acting as a training set, and 5 recordings as an independent testing set. The former data however was obtained by repeating training 30 times in order to provide statistical relevance and to avoid initialization effects for the ANN, thus allowing comparison to decide on the best classification model. In this respect, and according to the obtained data, the best model has proved to be the ANN classifier.

Feature selection methods were applied in order to get rid of possible data redundancy. Study for the reduction of the number of features was carried out using a subset of 10 patients, which was also organized in a training set containing 2/3 of the total patterns and another with the remaining 1/3 used as testing set. A subset of 10

patients from the original set of 20 patients was used due to the high computational cost requirements of the wrapper methods. Results have shown that, independently of the applied feature selection technique, lower classification error can be achieved while reducing the number of necessary features. Lower absolute error in this respect has been achieved using 8 features (see Table 7.9). These results have been obtained over the 1/3 testing set, however the actual performance of the resulting classifier (using the reduced number of features) has not been tested within the general method.

Thus, in order to assess final validation results of the method, final training process with the ANN model is scheduled using TR_{eq} as training set, while using TS_{eq} as the validation test for the stopping criterion. In this regard, the ANN classifier is trained using the reduced number of features, while error in TS_{eq} is below of training error in $TR_{eq,}$ so that preventing possible overfitting. Once trained, final validation results are obtained on an epoch-by-epoch basis using the independent validation set (IVS) of 26 recordings (see "*Design of the validation tests*" in Chapter 6).

Subsequently, Table 7.16 shows the validation results in the said IVS set. As it has been stated throughout the corresponding section of Chapter 6, none of these recordings was included in the training or testing sets used for parameter configuration of the methods. In Table 7.16 the first column references the recording number and the second column indicates the number of classifiable epochs. Total number of micro-arousal events localized by standard reference and system are shown respectively in columns three and four. Number of TPs, FNs, TNs and FPs are in the subsequent columns. Remaining four columns contain computed values for agreement index, sensitivity, specificity and AUC that correspond to each recording. Last row in Table 7.16 contains total values (summation over the 26 recordings) and the corresponding overall indexes shown in bold.

Table 7.16. Validation of the epoch-based validation for detection of EEG arousals. RN = Record Number; SR = Standard Reference; TP = True Positive; FN = False Negative; TN = True Negative; FP = False Positive; AgrI = Agreement Index; Sens = Sensitivity; Spec = Specificity; AUC = Area Under ROC Curve

EEG Arousal Event Location											
RN	**Epochs**	**Arousal events**		**TP**	**FN**	**TN**	**FP**	**AgrI**	**Sens**	**Spec**	**AUC**
		SR	*System*								
200088	1200	22	33	11	10	1157	22	0.973	0.524	0.981	0.753
200259	1233	262	258	208	48	927	50	0.921	0.813	0.949	0.881
200386	1312	198	179	118	76	1057	61	0.896	0.608	0.945	0.777
200532	1020	229	80	58	169	771	22	0.813	0.256	0.972	0.614
200568	1170	120	170	87	29	971	83	0.904	0.750	0.921	0.836
200929	1320	406	364	291	115	841	73	0.858	0.717	0.920	0.818
201249	1140	338	305	234	101	734	71	0.849	0.699	0.912	0.805
201294	1200	347	335	282	64	801	53	0.903	0.815	0.938	0.876
201394	1364	48	91	36	20	1253	55	0.945	0.643	0.958	0.800
201824	1168	231	217	153	76	875	64	0.880	0.668	0.932	0.800
202275	1140	124	119	88	36	986	31	0.941	0.710	0.970	0.840
202666	1120	50	66	39	11	1043	27	0.966	0.780	0.975	0.877
202733	1200	258	288	227	31	881	61	0.923	0.880	0.935	0.908
202956	1200	296	191	160	136	873	31	0.861	0.541	0.966	0.753
203249	1260	225	275	162	55	930	113	0.867	0.747	0.892	0.819
203294	1050	42	87	38	4	959	49	0.950	0.905	0.951	0.928
203494	1260	166	168	118	46	1046	50	0.924	0.720	0.954	0.837
203645	1364	235	97	75	159	1108	22	0.867	0.321	0.981	0.651
203798	1119	308	192	141	159	768	51	0.812	0.470	0.938	0.704
204135	1200	158	105	81	76	1019	24	0.917	0.516	0.977	0.746
204452	1110	33	68	31	2	1040	37	0.965	0.939	0.966	0.953
204480	1320	72	70	39	31	1219	31	0.953	0.557	0.975	0.766
205813	940	104	160	84	18	762	76	0.900	0.824	0.909	0.866
205948	1030	40	41	15	25	954	26	0.950	0.375	0.973	0.674
206040	1320	260	180	135	124	1016	45	0.872	0.521	0.958	0.739
206181	1320	278	296	213	64	960	83	0.889	0.769	0.920	0.845
Total	*31080*	*4850*	*4435*	*3124*	*1685*	*24951*	*1311*	***0.904***	***0.650***	***0.950***	***0.800***

Taking a look to the results of Table 7.16, it can be shown that in general the overall number of micro arousal events detected over the total number of epochs is quite similar (4850 for the standard reference and 4435 for the system). Wilcoxon paired sign rank test does not detect significant differences among the individual recordings (*p*-value 0.446). However data suggest a slight tendency toward underestimation of the number of arousal events, which is confirmed by taking into consideration the overall indices of sensitivity and specificity, respectively with values 0.650 and 0.950. Overall performance indexes measured through agreement index and AUC show respective values of 0.904 and 0.800. Individual differences can be shown in this respect among some of the recordings, however the general trend maintains. The lowest values of sensitivity are found for recordings 200532, 206645 and 205948, all below 0.4. These recordings also show the lowest values for AUC dragged by the values in sensitivity.

The uneven proportion of arousal events over these recordings (according to the standard reference 22%, 17% and 4% respectively) as well as the presence of high sensitivity in other recordings (for example, recording 204452 shows 0.939 sensitivity with 3% of arousals, recording 202733 shows 0.880 sensitivity with 22% of arousals), suggest that there is not a direct relation between the proportion of arousal events and sensitivity of the system. Next Figure 7.3 confirms this hypothesis. In the upper side of the figure both linear correlation –Person's *r*- and Spearman's *rho* coefficients are shown. Their respective *p*-values for testing the null hypothesis H_0: $r = 0$ and H_0: $rho = 0$, are displayed between brackets. Black straight line represents the linear regression line obtained using the *Least Mean Square* (LMS) method. According to the obtained *p*-values (0.7 for *r*, 0.71 for *rho*) it is clear that no significant correlation exists between the proportion of arousal epochs and the sensitivity of the detection method.

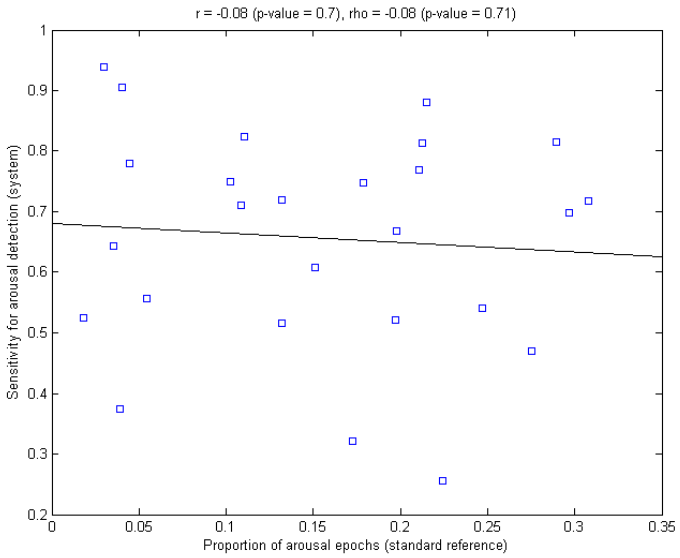


Figure 7.3. Sensitivity for arousal detection as a function of the proportion of arousal epochs in the recording. No significant correlation is perceived

It is noticeable that when comparing these results with those obtained in Table 7.2 (which measured the method using the full set of 42 features) the general trend seems to invert: a more sensitive method has been formerly obtained while now the method tends to be more specific; the conclusion is, therefore, that the higher initial dimensionality was effectively causing an increase in the number of false positives. Nonetheless, improvement of the classification error is demonstrated by comparing the two tables (0.218 for ANN in Table 7.2, 0.096 Table 7.16), pointing out to the fact that,

effectively, general performance of the method increases. Moreover, improvement of the classification error in the method outperforms the results achieved over the balanced dataset used to test the different feature selection methods (see Table 7.9 in previous subsection "*Feature selection on the detection of EEG arousals*"). This is explained by the joint action of the trained classifier once integrated in the method.

On the other hand, the general trend observed in Table 7.16 actually resembles more to that obtained in Table 7.4. In Table 7.4 model comparison was performed in the method using the unbalanced training set TR and the full set of 42 features. Since the discussion taken in the corresponding section ruled out the use of TR as training in favor of TR_{eq}, one may wonder if final results in Table 7.16 can be regarded as acceptable. However, in the case of the results in Table 7.4, the resulting classifier barely detected an event based on learned discriminative capabilities. It basically learned the majority class (non arousal event) and positive examples were detected almost by chance. Comparison of Table 7.1 and Table 7.3 also support this fact. On the contrary, in Table 7.16 improvement over classification error, sensitivity and AUC indexes is obtained with respect to Table 7.4. Moreover, the balanced number of false negatives and false positives in Table 7.16 (1685 and 1311 respectively) suggest that discrepancy in the detection can be explained due to time location of the arousal events, and not because lack of detection capabilities.

Subsequent Figure 7.4 represents the distributions of the validation indexes over the 26 recordings. It can be shown in the figure that maximum dispersion corresponds to sensitivity index (mean±std. = 0.657±0.178) while specificity remains as the most stable among the different recordings (mean±std. = 0.949±0.025). Distributions for the agreement index (mean±std. = 0.904±0.045) and AUC (mean±std. = 0.803±0.084) confirm the results of Table 7.16.

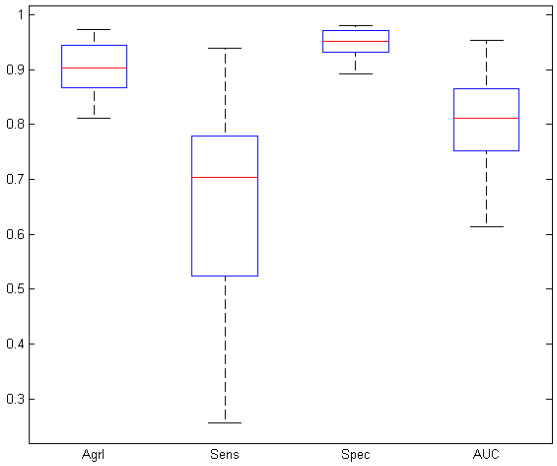


Figure 7.4. Box plot showing distribution of agreement ratios over the individual recordings

7.2. Sleep Staging

Results for the validation of the discrete hypnogram obtained using the method described in subsection "*Hypnogram generation*" of Chapter 5 are shown in this section. In this regard Figure 7.5 presents a box plot for the sensitivities, specificities and the values for AUC, obtained for the four considered sleep stages and for the 26 patients in the IVS set. Subsequently Table 7.17 shows the means and the standard deviations for each one of the corresponding sets. Results in Table 7.17 are represented in the form *mean±std.deviation*.

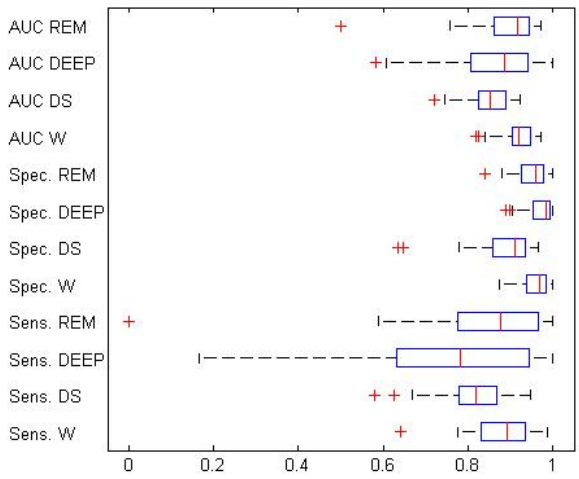


Figure 7.5. Box plot representing the distribution of Sensitivity (Sens), Specificity (Spec) and Area Under ROC Curve (AUC) for the different sleep states

Table 7.17. Validation results between expert and system (mean ± std. deviation)

	Awake	Drowsy Sleep	DEEP	REM
Sens.	0.88±0.07	0.81±0.08	0.75±0.23	0.84±0.19
Spec.	0.96±0.04	0.89±0.08	0.97±0.03	0.95±0.04
AUC	0.92±0.04	0.85±0.05	0.86±0.11	0.89±0.11

As it can be seen from data, the method works especially well in the discrimination of wakefulness and REM sleep (average sensitivity/specificity of 0.88/0.96 and 0.84/0.95 respectively). Sensitivity and specificity slightly decreases for drowsy sleep (0.81/0.89). However, the lowest value is achieved for the sensitivity with respect to stage DEEP (0.75) while on the other hand the highest specificity is achieved (0.97). Indeed, attending to AUC values, it can be said that best results are obtained regarding W and REM (AUCs of 0.92 and 0.89), followed by DEEP (AUC = 0.86) and DS (AUC = 0.85). Similar trends can be shown in the distributions of Figure 7.5 in which the thinnest dispersions are observed among the specificities, whereas higher variability spreads over the sensitivity of the system. Especially remarkable is the dispersion in the sensitivity of DEEP stage, according to standard deviation of Table 7.17. Presence of some outliers can also be shown among the different phases. These outliers are in general spread among the different subjects, i.e. they are not related to specific tough patients. An exception constitutes the outlier present in sensitivity of phase REM which corresponds to a patient with abnormal REM activity causing an additional outlier in the specificity of DS sleep.

Subsequent Table 7.18 presents the accumulated contingency table over the 26 patients. Summing values in the main diagonal and dividing by the total number of cases leads to an overall agreement of 0.838. Agreement indexes considering each sleep stage separately yield values of 0.93 for W, 0.86 for DS, 0.95 for DEEP, and 0.93 for REM. On the other hand, the percent positive agreement can be measured for a particular stage, by considering the number of epochs in which at least one in the scoring pair (system or standard reference) indicated that stage. In this case, obtained values are 0.82 for W, 0.72 for DS, 0.55 for DEEP and 0.63 for REM. These values showed that in relative terms, the highest discrepancies involve classification of DEEP and REM.

Table 7.18. Accumulated contingency table. In parenthesis associated frequency ratio is indicated

REFERENCE \ SYSTEM	W	DS	DEEP	REM
W	12208 (0.31)	1076 (0.03)	18 (< 0.01)	576 (0.01)
DS	983 (0.02)	14357 (0.36)	1062 (0.03)	1152 (0.03)
DEEP	16 (< 0.01)	666 (0.02)	2293 (0.06)	25 (< 0.01)
REM	87 (< 0.01)	749 (0.02)	36 (< 0.01)	4509 (0.11)

Calculation of Cohen's kappa index over Table 7.18 results in $\kappa = 0.76$ (observed agreement $p_0 = 0.84$, agreement due to chance $p_c = 0.33$). Figure 7.6 shows the distribution of the individual kappa indexes over the 26 recordings (mean±std = 0.75±0.07).

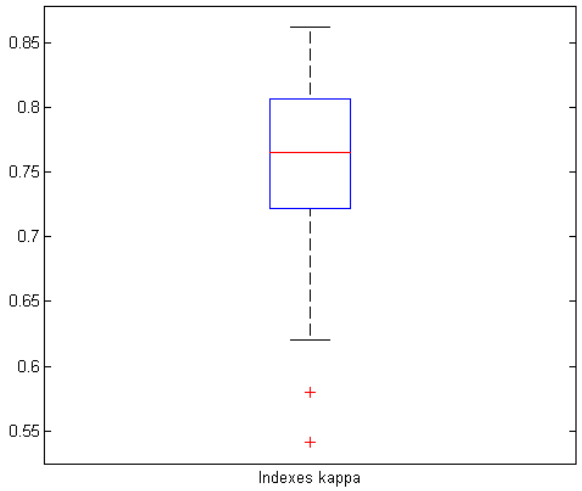


Figure 7.6. Box plot representing the distribution of value of kappa index for the different subjects

As it has been stated in subsection "*Pair-wise measures*" of Chapter 6, one has to be cautious at the time of interpreting the kappa index when the number of events classified in each category by the experts is not the same. As it is here the case, determination of the maximum possible value of kappa (κ_M) is performed in order to account for the actual marginal distributions. Thus, taking into account formula for maximum value possible of kappa resulting value of κ_M is 0.95, and therefore it can be concluded that current value of κ is $\frac{\kappa}{\kappa_M} = 0.80$ times larger than the maximum possible value given the circumstances.

From Table 7.18, the deviating scorings for each pair of sleep stages can be calculated by summing the upper and lower triangular matrices around the main diagonal. Figure 7.7 represents the resulting pair-wise percentages of overall mismatches.

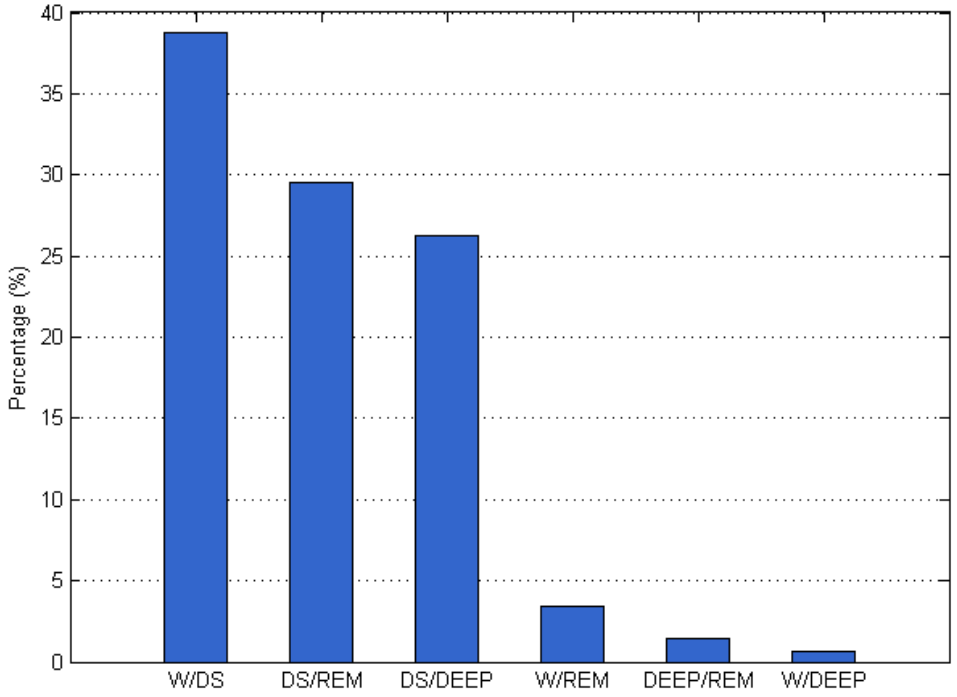


Figure 7.7. Pair-wise epoch deviating scorings expressed as percentage of overall mismatches

According to Figure 7.7 most of the discrepancies in the scoring are found between stages W and DS which account for 38.74% of total misclassifications. Almost as common of as the discrepancy between W and DS, it is that between DS and REM (29.52% of total misclassifications), and between DS and DEEP (26.25% of total misclassifications). It is noticeable that among these three groups, DS is always present. Therefore it might be concluded that the major difficulty resides in the correct identification of stage DS. Nevertheless, it should also be taken into account that DS is by far the most commonly scored sleep stage (by both the system and the standard reference). Therefore, it is normal that most of the discrepancies involve this stage. In fact, by taking into account the proportion of misclassified epochs, it has been revealed that DEEP is the stage with the highest discrepancy on the positive agreement (0.55), followed by REM (0.65). In other words, regarding DEEP, system and standard

reference agree 55% of the time, whereas in the 45% of the remaining cases, either the system or the standard reference, classify the epoch as a different stage. However, this is barely reflected by overall indexes since the number of epochs with deep sleep only represents around 7% of the total count.

7.3. Apneic events detection

Two tables are presented here which show the results for apneic events detection regarding capabilities of the system to localize the apneic event (Table 7.19) and, once localized, capabilities to classify the event either as apnea or as hypopnea (Table 7.20).

It has to be taken into account that since two complementary categories are considered for each table (event/no event in Table 7.19, apnea/hypopnea in Table 7.20) a TP in one category implies a TN for the complementary and vice versa. Analogously, given this duality, false positives (FPs) and false negatives (FNs) for one category result, respectively, in FNs and FPs for the other. Therefore, for reasons of brevity, in both tables only breakdown of one category is provided in the terms of TPs, FNs, TNs and FPs (apneic events and apnea, respectively).

Table 7.19 summarizes the validation results for the IVS set comprising 15540 minutes of polysomnographic recordings. Each recording represents a patient affected to some degree by SAHS. Organization of the table is similar to that for the epoch-based validation of the detected arousal events (see Table 7.16 arousal events). Accordingly, in Table 7.19 the first column references the recording number and the second column indicates the number of classifiable epochs. Total number of apneic event localized by the standard reference and the system are shown respectively in columns three and four. Number of TPs, FNs, TNs and FPs are in the subsequent columns. Remaining four columns contain computed values for agreement index, sensitivity, specificity and AUC that correspond to each recording. Last row in Table 7.19 contains total values (summation over the 26 recordings) and the corresponding overall indexes are shown in bold.

Table 7.19. Results for the validation regarding the location of apneic events. RN = Recording Number; SR = Standard Reference; TP = True Positive; FN = False Negative; TN = true negative; FP = False Positive; AgrI = Agreement index; Sens = Sensitivity; Spec = Specificity; AUC = Area Under ROC Curve

Apneic Event Location											
RN	**Epochs**	**Apneic events**		**TP**	**FN**	**TN**	**FP**	**AgrI**	**Sens**	**Spec**	**AUC**
		SR	*System*								
200088	1200	162	136	87	75	989	49	0.897	0.537	0.953	0.745
200259	1233	645	630	580	65	538	50	0.907	0.899	0.915	0.907
200386	1312	359	380	306	53	879	74	0.903	0.852	0.922	0.887
200532	1020	481	497	396	85	438	101	0.818	0.823	0.813	0.818
200568	1170	139	137	67	72	961	70	0.879	0.482	0.932	0.707
200929	1320	471	523	419	52	745	104	0.882	0.890	0.878	0.884
201249	1140	444	440	342	102	598	98	0.825	0.770	0.859	0.815
201294	1200	554	541	491	63	596	50	0.906	0.886	0.923	0.904
201394	1364	120	135	49	71	1158	86	0.885	0.408	0.931	0.670
201824	1168	394	441	325	69	658	116	0.842	0.825	0.850	0.838
202275	1140	346	332	251	95	713	81	0.846	0.725	0.898	0.812
202666	1120	77	79	34	43	998	45	0.921	0.442	0.957	0.699
202733	1200	275	337	223	52	811	114	0.862	0.811	0.877	0.844
202956	1200	406	420	342	64	716	78	0.882	0.842	0.902	0.872
203249	1260	413	390	318	95	775	72	0.867	0.770	0.915	0.842
203294	1050	58	65	41	17	968	24	0.961	0.707	0.976	0.841
203494	1260	486	534	438	48	678	96	0.886	0.901	0.876	0.889
203645	1364	585	548	463	122	694	85	0.848	0.791	0.891	0.841
203798	1119	543	499	463	80	540	36	0.896	0.853	0.938	0.895
204135	1200	281	377	249	32	791	128	0.867	0.886	0.861	0.873
204452	1110	46	59	26	20	1031	33	0.952	0.565	0.969	0.767
204480	1320	60	88	45	15	1217	43	0.956	0.750	0.966	0.858
205813	940	37	40	19	18	882	21	0.959	0.514	0.977	0.745
205948	1030	92	102	51	41	887	51	0.911	0.554	0.946	0.750
206040	1320	636	673	567	69	578	106	0.867	0.892	0.845	0.868
206181	1320	306	352	251	55	913	101	0.882	0.820	0.900	0.860
Total	*31080*	*8416*	*8755*	*6843*	*1573*	*20752*	*1912*	***0.888***	***0.813***	***0.916***	***0.864***

It can be shown taking a look to results in Table 7.19 that, in general, the overall number of apneic events detected throughout the total 15540 minutes of sleep is quite similar (8705 for the standard reference and 8454 for the system). Wilcoxon paired sign rank test does not detect significant differences among the individual recordings at the 0.05 significance level (*p*-value 0.067). Therefore, although individual differences can be shown in this respect among some of the recordings, the general trend maintains. Nonetheless, taking a look to the agreement ratios it can be shown a general tendency toward specificity of the system (0.916) whereas slightly lower sensitivity (0.813) is achieved. These differences are statistically significant (Wilcoxon rank sum *p*-value of 5.56×10^{-7}). When looking at the individual recordings, lower sensitivity is accompanied by higher variability (mean±std. = 0.738±0.159) when compared to specificity (mean±std. = 0.910±0.044).

General performance measured through agreement index and AUC also remain less sensitive to individual inter-variability (mean±std. of 0.889±0.040 and 0.824±0.068 respectively). Overall computed indexes show values of 0.888 for agreement and 0.864 for AUC, thus showing a more robust operation of the system from the general perspective. Similar trends can be observed from the box plot displayed in Figure 7.8 which shows the larger tails over sensitivity and AUC distributions. Recordings 201394, 202666 and 200568 show the lowest values for sensitivity and AUC over the 26 recordings (actually AUC values are dragged by the low sensitivities for these recordings).

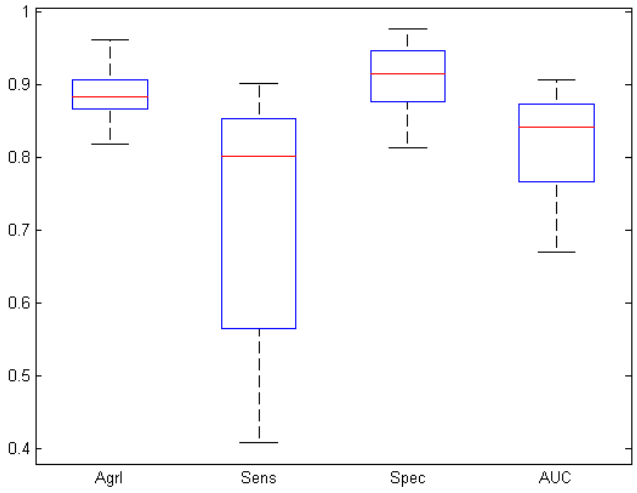


Figure 7.8. Box plot showing distribution of agreements ratios over individual recordings

It is noticeable that, for example, in the recording 201394 the number of epochs containing apneic event according to the standard reference is 135 while the number of non-apneic epochs is 1364, which represents a proportion of 0.099 of apneic epochs. Similar situation can be observed for recordings 200568 and 202666, both with sensitivity below 0.5 and in which proportion of apneic epochs is 0.117 and 0.068 respectively. In fact, a general tendency can be observed in which sensitivity of the system for the detection of apneic events decays as the proportion of apneic events in the recording does, as it can be seen in the upper left plot of Figure 7.9.

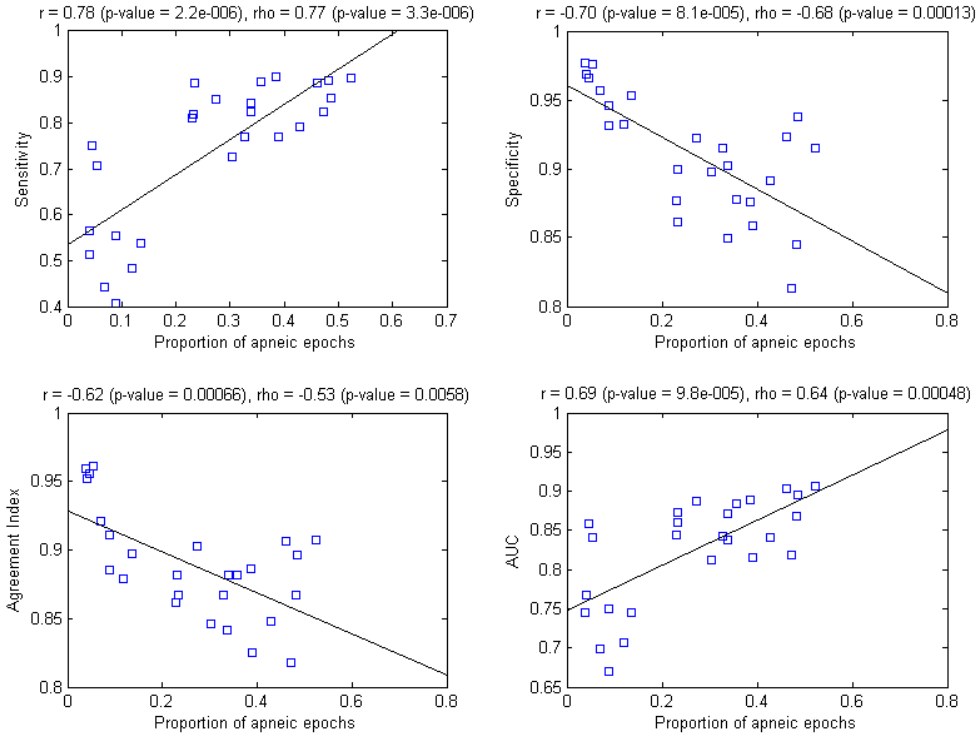


Figure 7.9. Evolution of the agreement ratios for the detection of apneic epochs as a function of the proportion of apneic epochs in the recording

In Figure 7.9 the correlation between the proportion of apneic events in the recording (according to standard reference) and the evolution of the validation indexes is investigated. Significant correlations (taking α of 0.05) according to all indexes are found that show opposite trends between sensitivity and AUC (positive correlation) and specificity and agreement index (negative correlation).

As it has been pointed out, according to the evolution of the sensitivity index in Figure 7.9, it can be concluded that the higher the proportion of apneic events in the recording the higher the sensitivity of the system for its detection. On the other hand it can be seen in Table 7.19, that for these recordings the number of FN is approximately the same as the number of FP. Therefore, while the net count of apneic events localized by both the system and the standard reference is similar, it seems that there is discrepancy about their concrete temporal instants of occurrence. A look to the respective number of apneic epochs supports this interpretation since no relevant differences have been found between the experts and the system (see Table 7.19, columns 3 and 4). A contrary effect can be seen on the specificity index as a function of

the proportion of apneic epochs in the recording: the more apneic events in the recording the lesser the specificity of the system (see Figure 7.9 upper right). In order to study the causes of these trends, evolution of the proportions of TPs, FNs, TNs and FPs can be studied separately (see Figure 7.10). In the figure it can be observed that the increase produced in sensitivity is due to the significant (r = 0.99, p-value = 8.1x10^{-24}, *rho* = 0.99, p-value = 1.7x10^{-7}) correlation between the proportion of TPs and the SAHS severity (the proportion of apneic epochs in the recording is an indirect measure of AHI). This increase in the sensitivity index even compensates the moderate increasing trend in the proportion of FNs (correlation is lesser but still significant here, r = 0.63, p-value = 6.4x10^{-4}, *rho* = 0.62, p-value = 7.1x10^{-4}).

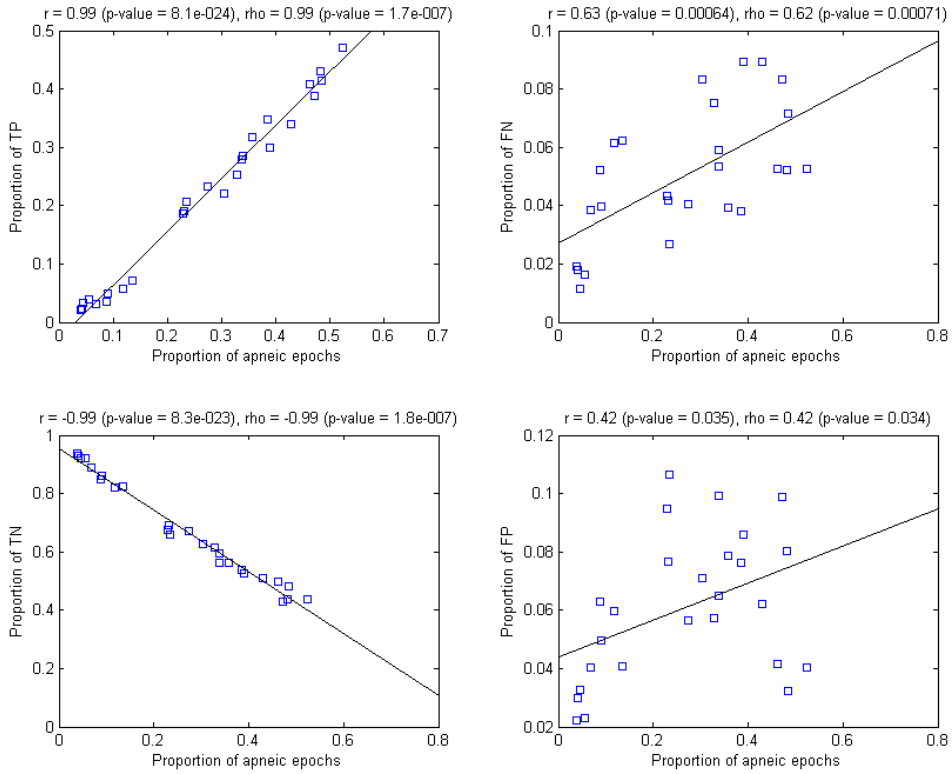


Figure 7.10. Individual evolution of the proportion of True Positives (TP), False Negatives (FN), True Negatives (TN) and False Positives (FP) as a function of the proportion of apneic epochs in the recordings

Having that in mind, a possible explanation is that when the number of apneic events in the recording is low –the subject has low SAHS severity or even does not evidence significant SAHS at all- then also the severity or the *clearness* of the associated events is more reduced: that is, predominant apneic events are of the very

mild hypopnea type. A plot of the proportion of apnea events (with respect to hypopnea) versus the AHI confirms the hypothesis that the proportion of apnea events increases with the severity of the syndrome (see Figure 7.11, significant correlation is found). The former implies that accurate detection of these events for the human scorer is more difficult due to limited visual precision of the human eye, which increases subjectivity in the marking of events, therefore reducing agreement in the detection between the human scorer and the system. As the severity of the syndrome increases, it also does the sharpness of the associated events, which increments the proportion of true positives. On the other hand, this has a counter effect since the sleep pattern and the sleep in general becomes more unstable, which slightly affects to the increasing proportion of FPs and FNs of the system (although in a much more lesser rate compared to the rise in the proportion of TPs).

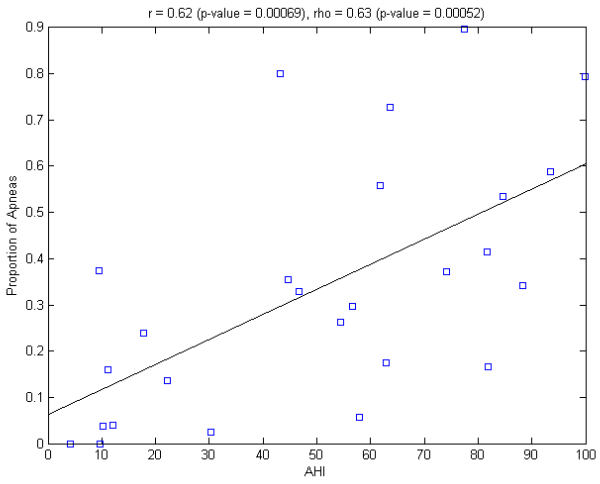


Figure 7.11. Proportion of apneas from the total of apneic events as a function of the Apnea-Hypopnea Index (AHI). Significant correlation is evidenced at the 0.05 significance level

In fact, the drop in the specificity as a function of the AHI is explained as a consequence of the previous. Indeed, because of the increasing number of epochs with presence of apneic event, a reduction in the number of possible TN occurs (there are less negative cases since the number of positive cases has increased), and this causes a FP to weight more in the computation of the specificity index. Hence, the reduction in the proportion of TNs, and the increase in the proportion of FPs, due to the rupture of the normal breathing, causes the observed pattern in the evolution of the specificity index.

Correlation patterns observed over the agreement index and AUC can be explained by the above mentioned trends on the sensitivity and the specificity. What it may look contradictory is that although both indexes can be regarded as general performance indexes, opposite trends are observed with the associated SAHS severity. Effectively the AUC index, as it is calculated as a one point operating curve, it directly depends on the corresponding associated values of sensitivity and specificity. It actually is an average of the two values. As it can be observed in Figure 7.9, the observed positive correlation for the evolution of the sensitivity index is higher -and it is so its associated significance- than the observed negative correlation for the specificity index in the same figure. Therefore average value compensates the fall in the specificity and it results in significant positive correlation for AUC (r = 0.69, p-value = $9.8\text{x}10^{-5}$, *rho* = 0.64, p-value = $4.8\text{x}10^{-4}$). Note that values for specificity are always above the 0.8 while dispersion associated to sensitivity ranges from 0.4 until 0.9 (see again Figure 7.9). The former explains that AUC follows a similar trend to sensitivity.

On the other hand agreement index shows negative correlation. This tendency although slightly less pronounced than the one observed for AUC is still significant (r = -0.62, p-value = $6.6\text{x}10^{-4}$, *rho* = -0.53, p-value = $5.8\text{x}10^{-3}$). Therefore it can be concluded that although sensitivity and AUC of the analysis increase among severe SAHS patients, without concerning about the type of the error, the possibility of committing mistakes (i.e. the total number of misclassified epochs in the detection) slightly increases. As it has been mentioned, the main cause for this effect is the instability of the sleep pattern associated to the severity patient, which increases the observed discrepancy between the system and the standard reference.

In subsequent Table 7.20 epochs in which there is positive agreement in the apneic events location (TP column in Table 7.19) are used in order to validate the system with respect to characterization of events either as apneas or hypopneas. Since validation is performed for epochs, in order to correctly validate apneic event characterization, only epochs with just one apneic event are considered. For validation purposes, for example, it would not be clear whether an epoch containing both apnea and hypopnea events should be considered as an apnea or as a hypopnea. This explains why the second column in Table 7.20 has fewer epochs for classification than the total number of true positives given in Table 7.19. Sensitivity and specificity are given in Table 7.20 for

apneas also with dual values for hypopneas. As expected, respective values are complementary.

Table 7.20. Validation results for characterization of apneic events as apneas or hypopneas. RN = Recording Number; SR = Standard Reference; TP = True Positives; FN = False Negatives; TN = True Negatives; FP = False Positives; AgrI = Agreement Index; Sens = Sensitivity; Spec = Specificity; AUC = Area Under ROC Curve

Apnea/Hypopnea Characterization											
RN	**Epochs**	**Apnea**						**Hypopnea**		**AgrI**	**AUC**
		TP	**FN**	**TN**	**FP**	**Sens**	**Spec**	**Sens**	**Spec**		
200088	83	2	0	77	4	1.000	0.951	0.951	1.000	0.952	0.975
200259	554	409	31	94	20	0.930	0.825	0.825	0.930	0.908	0.877
200386	259	51	17	175	16	0.750	0.916	0.916	0.750	0.873	0.833
200532	318	39	14	250	15	0.736	0.943	0.943	0.736	0.909	0.840
200568	63	11	4	48	0	0.733	1.000	1.000	0.733	0.937	0.867
200929	409	347	20	24	18	0.946	0.571	0.571	0.946	0.907	0.758
201249	308	53	1	225	29	0.981	0.886	0.886	0.981	0.903	0.934
201294	449	139	14	276	20	0.908	0.932	0.932	0.908	0.924	0.920
201394	44	6	0	32	6	1.000	0.842	0.842	1.000	0.864	0.921
201824	306	156	15	125	10	0.912	0.926	0.926	0.912	0.918	0.919
202275	206	49	12	122	23	0.803	0.841	0.841	0.803	0.830	0.822
202666	34	0	0	34	0	---	1.000	1.000	---	1.000	---
202733	180	140	4	26	10	0.972	0.722	0.722	0.972	0.922	0.847
202956	323	118	2	158	45	0.983	0.778	0.778	0.983	0.854	0.881
203249	300	10	7	268	15	0.588	0.947	0.947	0.588	0.927	0.768
203294	26	1	0	24	1	1.000	0.960	0.960	1.000	0.962	0.980
203494	427	307	4	80	36	0.987	0.690	0.690	0.987	0.906	0.838
203645	402	71	96	225	10	0.425	0.957	0.957	0.425	0.736	0.691
203798	421	201	24	164	32	0.893	0.837	0.837	0.893	0.867	0.865
204135	243	81	5	143	14	0.942	0.911	0.911	0.942	0.922	0.926
204452	25	2	2	20	1	0.500	0.952	0.952	0.500	0.880	0.726
204480	40	14	1	24	1	0.933	0.960	0.960	0.933	0.950	0.947
205813	19	0	0	19	0	---	1.000	1.000	---	1.000	---
205948	51	1	1	47	2	0.500	0.959	0.959	0.500	0.941	0.730
206040	507	267	31	185	24	0.896	0.885	0.885	0.896	0.892	0.891
206181	234	73	4	137	20	0.948	0.873	0.873	0.948	0.897	0.910
Total	*6231*	*2548*	*309*	*3002*	*372*	***0.892***	***0.890***	*0.890*	*0.892*	***0.891***	***0.891***

For the apnea category, Table 7.20 shows an overall sensitivity of 0.892 and a specificity of 0.890. As previously indicated, the complementary nature of the categories causes a duality in the interchange of the values associated with the sensitivity and specificity indices (0.890 and 0.892, respectively) when considering the hypopnea category. Global agreement index and AUC values are 0.891 and 0.891 respectively, so that it can be concluded that in general terms the system shows stability among the different agreement ratios in the apnea/hypopnea discrimination task. In fact Kruskal-Wallis test reported no statistical differences among the distributions of the different indexes (p-value = 0.433).

Dispersions associated to the indexes can be shown in Figure 7.12 when looking at the inter-individual differences for the apnea category. Cells containing missing values have been excluded for obtaining the resulting distributions. The graph shows the highest variability regarding sensitivity with the lowest values found for recording 203645 (0.425 sensitivity) and recordings 204452 and 205948 (both with 0.5 sensitivity values). However for these last two recordings only two and one positive cases are respectively available in the recording. On the other hand, specificity (hypopnea sensitivity) for the three previous recordings is always over 0.95 (0.957, 0.952 and 0.959 respectively). The outlier present in the specificity distribution corresponds to recording 200929 (0.571 specificity) achieving conversely good sensitivity (0.946). As in the case for apneic event localization, performance measured through agreement index and AUC show more stability over individual recordings (mean±std. of 0.899±0.048 and 0.861±0.080 respectively, while 0.844±0.177 is obtained for sensitivity and 0.878±0.102 for specificity for the apnea category). The outlier present among the agreement indexes corresponds to recording 203645 that also shows the highest number of false negatives for apnea category and the lowest value of sensitivity.

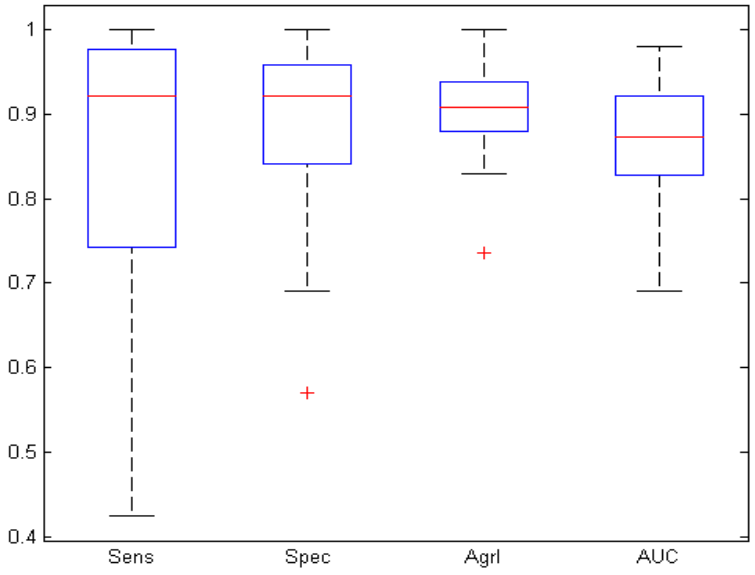


Figure 7.12. Box plot showing distribution of agreements ratios over individual recordings for characterization of apneic events as apnea

Exploring the possible existence of correlation between the proportion of apnea events and the different indexes, it has been found significant decrease in the specificity for the apnea class (see Figure 7.13, upper right) as the proportion of apnea events in the sample (with respect to hypopnea events) increases. However no relevant effect is interpreted on the operation of the classifier due to this fact. Indeed as the proportion of apnea events in the sample increases, a good classifier will tend to increase the number of true positives while keeping the number of false positives and false negatives low. On the other hand, because only two kinds of events are included in the sample (apneas or hypopneas) as the number of positive cases increases (in this case, apneas) the number of possible true negatives (hypopneas) consequently decreases. As it has been commented for the case of apneic event localization, this increases the effect of any false positive in the specificity index. Figure 7.14 confirms this trend showing strong and inverse correlation among the proportion of true positives and of true negatives. However at the same time no significant increase in the proportion of false positives and false negatives is perceived. This confirms the classifier is behaving robustly, being the decrease in the specificity of Figure 7.13 just a collateral effect of the non-significant increase of the proportion of false positives, that when the proportion of true negatives (hypopnea events) decreases –as explained, because of the increasing proportion of apnea events in the sample- augment their relative weight in the computation of the specificity index.

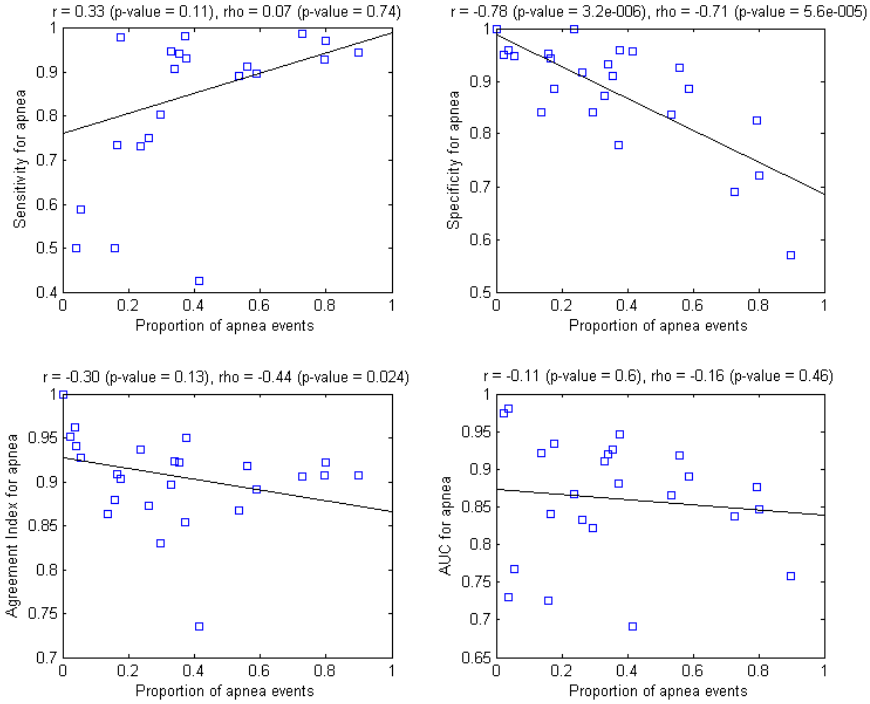


Figure 7.13. Evolution of the agreement ratios as a function of the proportion of apnea events in the sample (true positively detected apneic events) versus proportion of hypopnea events

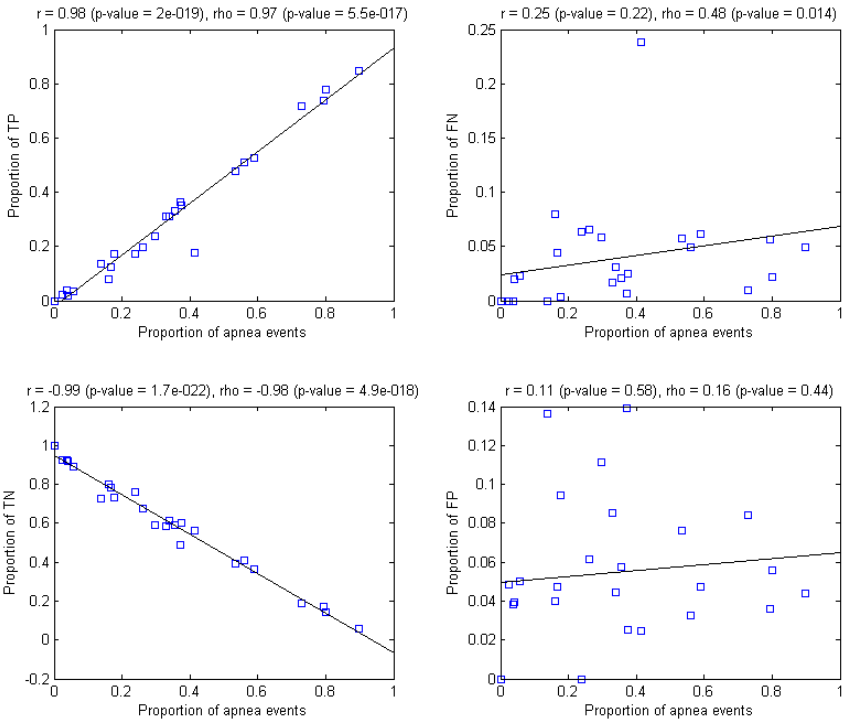


Figure 7.14. Individual evolution of the proportion of True Positives (TP), False Negatives (FN), True Negatives (TN) and False Positives (FP) as a function of the proportion of apnea events in the sample

When studying the effects of SAHS severity (proportion of apneic epochs in the recording) in the characterization of the detected events as apneas or hypopneas, slight significant negative correlation is found affecting specificity and agreement indexes (see Figure 7.15, upper right and lower left). Attending to complementary study of individual proportions of TP, FN, TN and FP, the interpretation here is related to that given when studying the relationship between SAHS severity and localization of apneic events. In this respect, as stated before, an increase in the severity of the SAHS is related with higher occurrence of apnea events with respect to hypopneas (see Figure 7.11). This causes the increase in the proportion of TP and the reduction in the proportion of TN, as observed in Figure 7.16, in accordance with the trends in Figure 7.10. However, differently from Figure 7.10, in this case the correlation found in the proportion of TPs and TNs is less pronounced, which is expected according to the trend shown in Figure 7.15. This causes that the slight positive trend (barely significant, almost no significant correlation is found) present over the proportions of FNs and FPs, affects more to the indexes of sensitivity (that does not show significant positive correlation) and specificity (that shows significant negative correlation at the 0.05 significance level) in Figure 7.15. In addition, the slight concurrent increase in the proportion of FNs and FPs, also affects to the global proportion of misclassified events. This is reflected in the slight negative correlation found for the agreement index (see Figure 7.15, lower left). On the other hand, the reasons for this concurrent increase in the proportion of FNs and FPs are more difficult to interpret: the concurrent increase shows no clear bias in the misclassification. In this regard, on possible interpretation may be that, when severity of the SAHS increases, subjectivity of the clinician in the classification of the apneic events also increases. The reason, perhaps, is that with higher number of apneic events in the recording, the clinician spends less time in the classification of every single event.

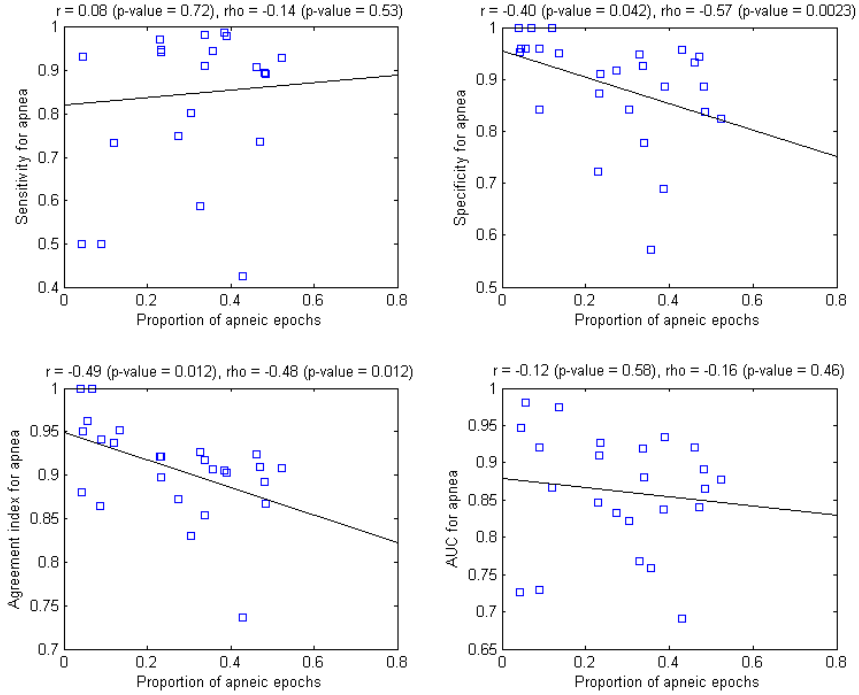


Figure 7.15. Evolution of the agreement ratios for apnea category as a function of the proportion of apneic epochs in the recording

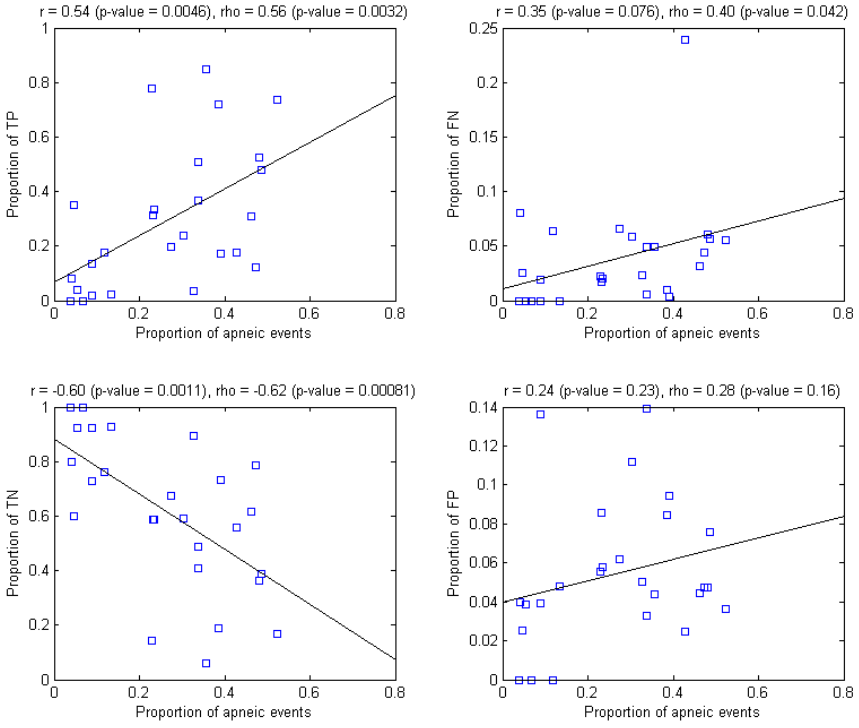


Figure 7.16. Individual evolution of the proportion of True Positives (TP), False Negatives (FN), True Negatives (TN) and False Positives (FP) for apnea category as a function of the proportion of apneic epochs in the recording

7.4. Apneic events classification

In accordance with the considerations made in the corresponding section of Chapter 6, following Table 7.21 shows validation results achieved using the set of 26 patients with regard to apneic event classification as obstructive or central. In this respect it has to be remarked that only the set of true positives for the class apnea (see Table 7.20 of the preceding section) can be evaluated for classification. This justifies the number of epoch in the second column of Table 7.21. In addition, hypopnea events by standard reference are systematically classified as obstructive, and therefore the analysis of system's classification output for these events has no interest for validation. On the other hand it has also been mentioned that since standard reference only makes distinction between obstructive and central events, for validation purposes, mixed events by the system are included within the obstructive category. Computed indexes as well as table organization are identical to that of Table 7.20.

Table 7.21. Results of the validation for apneic event classification. RN = Recording Number; TP = True Positives; FN = False Negatives; TN = True Negatives; FP = False Positives; Sens = Sensitivity; Spec = Specificity; AgrI = Agreement Index; AUC = Area Under ROC Curve

Obstructive/Central Classification											
RN	**Epochs**	**Obstructive**						**Central**		**AgrI**	**AUC**
		TP	**FN**	**TN**	**FP**	**Sens.**	**Spec.**	**Sens.**	**Spec.**		
200088	2	2	0	0	0	1,000	---	---	1,000	1,000	---
200259	409	136	91	157	25	0,599	0,863	0,863	0,599	0,716	0.731
200386	51	31	5	9	6	0,861	0,600	0,600	0,861	0,784	0.731
200532	39	22	5	11	1	0,815	0,917	0,917	0,815	0,846	0.866
200568	11	11	0	0	0	1,000	---	---	1,000	1,000	---
200929	347	325	11	10	1	0,967	0,909	0,909	0,967	0,965	0.938
201249	53	53	0	0	0	1,000	---	---	1,000	1,000	---
201294	139	136	3	0	0	0,978	---	---	0,978	0,978	---
201394	6	6	0	0	0	1,000	---	---	1,000	1,000	---
201824	156	30	11	85	30	0,732	0,739	0,739	0,732	0,737	0.735
202275	49	27	3	13	6	0,900	0,684	0,684	0,900	0,816	0.792
202666	0	0	0	0	0	---	---	---	---	---	---
202733	140	8	13	113	6	0,381	0,950	0,950	0,381	0,864	0.665
202956	118	104	0	0	14	1,000	0,000	0,000	1,000	0,881	0.500
203249	10	10	0	0	0	1,000	---	---	1,000	1,000	---
203294	1	0	0	1	0	---	1,000	1,000	---	1,000	---
203494	307	97	67	109	34	0,591	0,762	0,762	0,591	0,671	0.677
203645	71	57	4	4	6	0,934	0,400	0,400	0,934	0,859	0.667
203798	201	63	29	80	29	0,685	0,734	0,734	0,685	0,711	0.709
204135	81	7	2	51	21	0,778	0,708	0,708	0,778	0,716	0.743
204452	2	2	0	0	0	1,000	---	---	1,000	1,000	---
204480	14	14	0	0	0	1,000	---	---	1,000	1,000	---
205813	0	0	0	0	0	---	---	---	---	---	---
205948	1	1	0	0	0	1,000	---	---	1,000	1,000	---
206040	267	219	16	26	6	0,932	0.813	0.813	0,932	0,918	0.872
206181	74	59	15	0	0	0,797	---	---	0,797	0,797	---
Total	*2549*	*1420*	*275*	*669*	*185*	*0,838*	***0.783***	***0.783***	*0,838*	***0,820***	***0.811***

Analysis of the results reveals a system more sensitive to the classification of obstructive events (0.838) than with respect to central events (0.783). Once again since class categories are complementary, the trend swaps when looking at overall indexes for specificities. General AUC value obtained for classification is 0.811 while agreement index is 0.820.

Distributions of the agreement ratios over the individual recordings are shown in Figure 7.17, taking as reference values for the obstructive category. Note that cells with missing values in Table 7.21 have been omitted. Means and standard deviations associated to each index are 0.867±0.171 for sensitivity, 0.720±0.259 for specificity, 0.886±0.117 for agreement index and 0.741±0.111 for AUC. Wilcoxon test applied over sensitivity and specificity distributions confirmed relevant differences between the two

distributions at the 0.95 confidence level (p-value = 0.031). Therefore, it can be said that there is significant tendency to the detection of obstructive events. With regard to the presence outliers, it is noticeable the one in the specificity distribution that corresponds to recording 202956 in which 0 specificity is obtained. According to the standard reference, over a total of 118 apnea events in this recording sample, 104 are obstructive and 14 are central, however the system classifies all the events as obstructive (perfect sensitivity). Another extreme case for specificity is the recording 203645, in which 6 of a total of 10 central events where misclassified (specificity of 0.4). In this recording the number of obstructive events is 61 of which 57 where correctly classified as obstructive (0.934 sensitivity). The outlier for sensitivity distribution corresponds to recording 202733 in which the number of obstructive events is 21 while the number of central events is 119. In this recording respective sensitivity and specificity are of 0.381 and 0.950. In general when the number of central events in the recording is higher, specificity index seems to improve, suggesting that sensitivity for central events increases as the proportion of central events does.

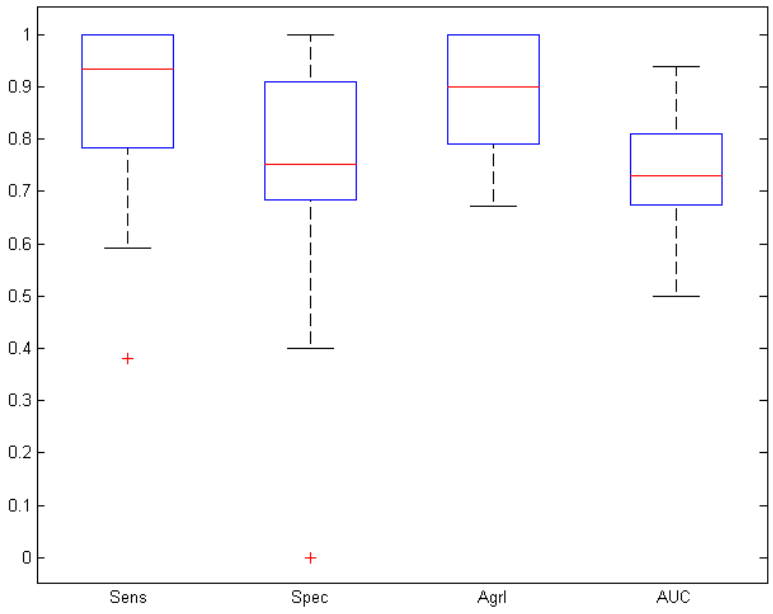


Figure 7.17. Box plot showing distribution of agreement ratios over the individual recordings regarding classification of apneic events as obstructive

Next Figure 7.18 represents the evolution of the different agreement ratios with respect to the proportion of obstructive events in the individual recordings. The trends in the figure confirm the previous hypothesis, that when the number of central events in the recording is high, specificity index (sensitivity for central events) is the highest. However, the found correlation does not show statistically significant values for specificity at the 0.05 significance level (see Figure 7.18, upper right). On the other hand, significant positive correlation is found for the sensitivity index and the agreement index as the proportion of obstructive events approximates to one. This confirms that, in general, the classifier tends to detect obstructive events, while there is slight underestimation of the actual number of central events. As previously commented, detection of central events improves when the proportion of obstructive events decreases; however the increase in the agreement index confirms that major classification performance (less number of misclassified events) is achieved when the prevalent event in the recording is of obstructive type.

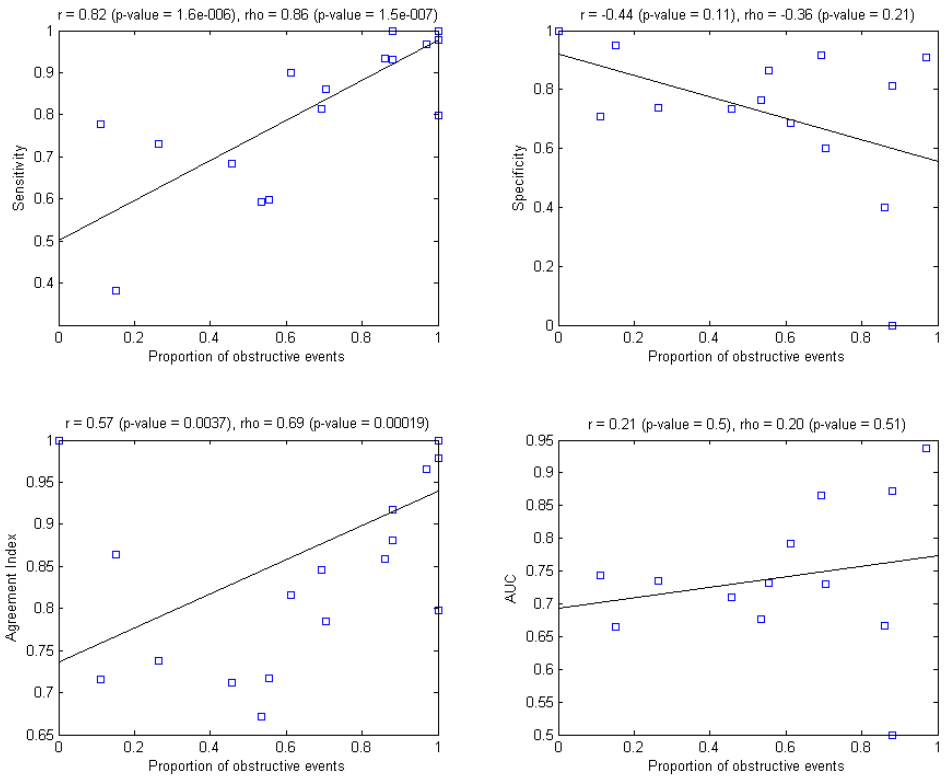


Figure 7.18. Evolution of the agreement ratios for obstructive category as a function of the proportion of obstructive events in the sample (positively detected apnea events)

7.5. Final patient diagnosis

In the following Table 7.22 final computed AHI indexes of both system and standard reference are shown for each of the 26 patients. The corresponding linguistic label for the severity classification is assigned according to linguistic scales expressed in Table 5.4 (see Chapter 5, "*Diagnostic generation*"). In the table the respective associated sleep times (TST) are also indicated for the evaluation of the corresponding AHI.

Table 7.22. Final computed AHIs and corresponding severity classification. RN = Recording Number, TST = Total Sleep Time, AHI = Apnea-Hypopnea Index

RN	System			Standard reference		
	TST	AHI	Severity	TST	AHI	Severity
200088	4:13:30	32.9	SEVERE	5:27:30	30.23	SEVERE
200259	6:44:30	97.16	SEVERE	6:36:30	99.87	SEVERE
200386	6:57:00	70.36	SEVERE	6:42:00	54.48	SEVERE
200532	7:18:00	84.52	SEVERE	6:01:30	81.83	SEVERE
200568	7:36:30	18.8	MILD	7:44:00	17.72	MILD
200929	6:45:30	79.01	SEVERE	6:04:30	77.53	SEVERE
201249	6:58:30	72.54	SEVERE	7:06:00	62.96	SEVERE
201294	6:52:00	84.61	SEVERE	6:19:30	88.22	SEVERE
201394	6:38:00	21.41	MODERATE	5:26:00	22.27	MODERATE
201824	7:05:00	67.48	SEVERE	6:25:00	61.71	SEVERE
202275	6:37:00	62.12	SEVERE	6:14:00	56.63	SEVERE
202666	7:31:00	10.78	MILD	7:15:30	9.64	NOT SIGNIFICANT
202733	7:19:30	59.39	SEVERE	6:23:30	43.18	SEVERE
202956	5:56:00	77.87	SEVERE	5:28:30	74.16	SEVERE
203249	7:08:30	57.13	SEVERE	7:14:00	57.93	SEVERE
203294	5:38:00	11.01	MILD	5:41:00	10.21	MILD
203494	8:34:30	63.91	SEVERE	7:40:30	63.58	SEVERE
203645	6:38:00	90.6	SEVERE	7:36:00	81.71	SEVERE
203798	6:33:30	82.34	SEVERE	6:29:30	84.57	SEVERE
204135	6:55:00	55.37	SEVERE	6:19:30	44.74	SEVERE
204452	5:12:00	10.96	MILD	5:03:00	11.09	MILD
204480	6:25:30	12.63	MILD	6:24:00	9.53	NOT SIGNIFICANT
205813	6:12:29	5.96	NOT SIGNIFICANT	6:37:00	4.08	NOT SIGNIFICANT
205948	5:28:00	16.12	MILD	6:17:30	12.08	MILD
206040	7:59:30	91.6	SEVERE	7:06:30	93.41	SEVERE
206181	7:00:00	53.29	SEVERE	6:37:00	46.7	SEVERE
Total	174:16:59			168:19:30		

As it can be seen from Table 7.22, according to system diagnosis, from the total set of 26 patients, 18 patients were diagnosed with severe SAHS, 1 was classified as moderate, 6 as mild, and 1 patient was found as not having relevant SAHS evidence. According to standard reference 18 were considered to be severe patients, 1 as

moderate, 4 as with mild SAHS, and 3 were considered as normals. Figure 7.19 shows the comparative in form of bar diagram. Overall, both distributions are not statistically different ($\chi^2_{(3)}$ = 1.4, *p*-value 0.705). According to the classification results of Table 7.22 calculation of the kappa index leads to a κ of 0.839 with κ_M of 0.839.

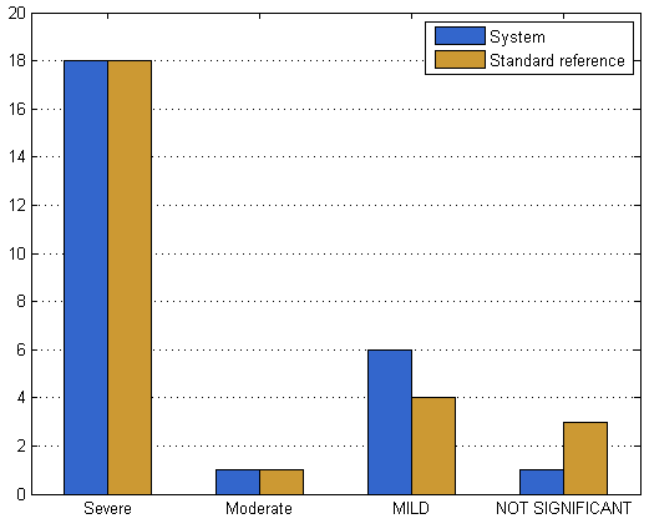


Figure 7.19. Diagnosis severity distribution

Only 2 patients (7%) received different severity considerations (recordings 20266 and 204480). In these cases, however, minimum differences were obtained on the final computed AHI indexes: 10.78 (system) vs. 9.64 (standard reference) for one patient (recording number 202666), and 12.63 (system) vs. 9.53 (standard reference) for the other (recording number 204480). In this respect, while there is discussion on the appropriateness of using AHI as the mere criteria to diagnose a patient with SAHS (see discussion in the next Chapter), it is clear that the use of arbitrary thresholds (in this case AHI > 10) to differentiate patients from normal subjects can affect the final classification. Next Figure 7.20 shows the result of applying different cut-offs to the respective AHI values of both the system and the standard reference. The resulting number of patients is indicated. Regardless on the discussion of the appropriateness of setting a fix threshold for the classification, the figure shows that system's estimation of the AHI is consistent with the output provided by the standard reference ($\chi^2_{(11)}$ = 0.622, *p*-value = 1).

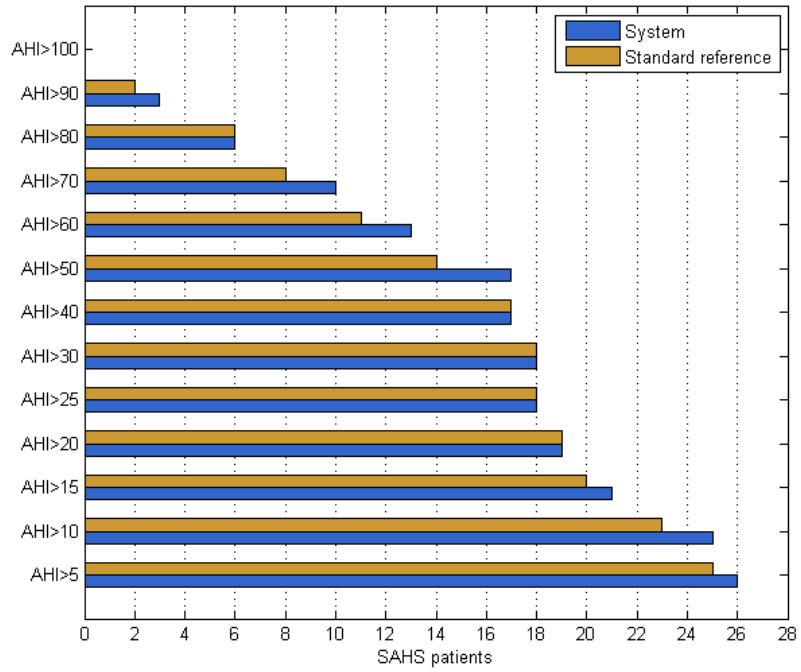


Figure 7.20. Number of patients using different AHI cut-offs for the system (blue) and for the standard reference (orange)

Taking into account AHI values for the individual recordings, Figure 7.21 shows the box plot of the respective dispersions. Respective associated mean and standard deviation are 53.46±30.49 for the system and 50.00±30.46 for the standard reference. Wilcoxon rank sum test shows no significant differences among the medians of the distributions (*p*-value = 0.564), however when paired comparison is performed, significant differences are found at the 0.05 significance level (*p*-value = 0.003). Since no significant differences were found on the number of apneic events detected for each recording (see Table 7.19 in subsection "*Apneic events detection*"), these differences are attributed to the effect of discrepancies in TST, which indeed are found be statistically relevant (Wilcoxon paired sign test *p*-value of 0.026). Agreement estimated using the Intraclass Correlation Coefficient (ICC) shows a value of 0.979 (*p*-value = 0). Inspecting individual differences, maximum absolute difference is found for recordings 202733 (16.21 difference, with 59.39 AHI for system and 43.18 for the standard reference) and 200386 (15.88 difference, with 70.36 AHI for the system and 54.48 for the standard reference). Nevertheless given the associated severities the differences do not have consequences on the diagnosis.

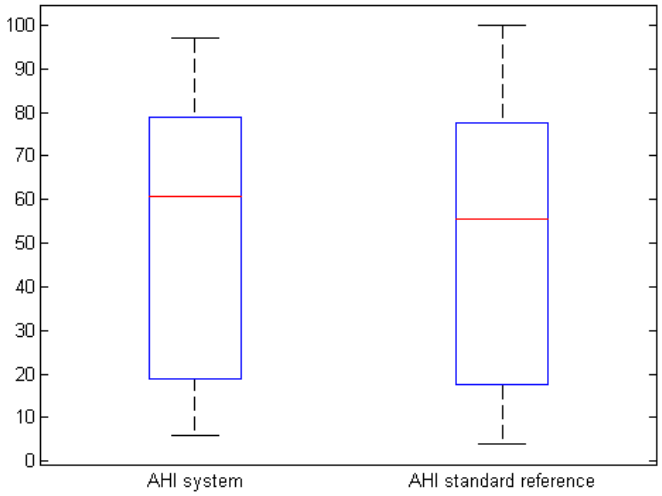


Figure 7.21. Box plot showing AHÍ distributions of the 26 patients for both the system (left) and standard reference (right)

Subsequent Table 7.23 breaks down AHI indexes and shows respective computed ApI and HI indexes for the 26 recordings. Distribution of the respective indexes can be shown in the box plots of the Figure 7.22. Again no significant differences have been found among the respective medians of the distributions of ApI and HI between the system and the standard reference (*p*-values of 0.564 and 0.721, respectively). Significant differences are found when paired comparison is performed at 0.95 confidence level between ApI distributions (*p*-value of 0.022). As stated before, differences are attributed to discrepancies on estimation of TST. No significant differences are found, on the other hand, in the respective HI distribution (*p*-value of 0.199). Computation of ICC leads to agreement of 0.965 for ApI ratings (*p*-value of 1.11×10^{-16}), and 0.951 for HI (*p*-value of 1.01×10^{-14}). Maximum absolute differences regarding ApI have been found in the recording 203645 (15.74 difference, with ApI 17.94 for system and 33.68 for the standard reference). For the HI maximum discrepancy is also found for the same recording (24.63 difference with 72.66 HI for system and 48.03 for the standard reference). These results confirms those observed in Table 7.21 of subsection "*Apneic events detection*" in which recording 203645 was considered an outlier in terms of characterization of the apneic events as apneas and hypopneas. Overall respective AHI indexes for this recording are 90.6 for the system and 81.71 for the standard reference, therefore differences between respective ApI and HI indexes compensate for the AHI.

Table 7.23. Distribution of ApI and HI indexes among the recordings for the system and the standard reference

RN	System		Standard Reference	
	ApI	**HI**	**ApI**	**HI**
200088	2.13	30.77	0.37	29.86
200259	70.9	26.25	75.66	24.21
200386	17.27	53.09	12.54	41.94
200532	13.15	71.37	13.61	68.22
200568	1.56	17.24	1.29	16.42
200929	68.51	10.51	69.47	8.07
201249	17.78	54.77	10.28	52.68
201294	27.52	57.09	26.72	61.5
201394	2.86	18.54	2.21	20.06
201824	32.33	35.15	29.61	32.1
202275	23.27	38.84	13.96	42.67
202666	0	10.78	0	9.64
202733	48.19	11.19	35.05	8.14
202956	37.08	40.79	25.39	48.77
203249	5.18	51.95	2.9	55.02
203294	1.07	9.94	0.18	10.03
203494	46.88	17.03	42.08	21.5
203645	17.94	72.66	33.68	48.03
203798	42.24	40.1	41.59	42.98
204135	19.08	36.29	14.07	30.67
204452	0.77	10.19	0.79	10.3
204480	3.42	9.21	3.28	6.25
205813	0	5.96	0	4.08
205948	0.78	15.34	0.16	11.92
206040	46.17	45.42	50.79	42.63
206181	15.57	37.71	13.45	33.25

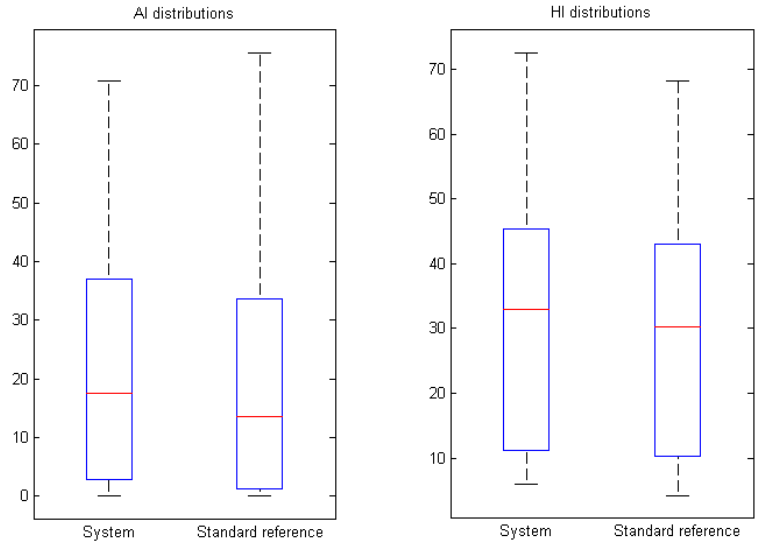


Figure 7.22. Box plots showing distributions of ApI and HI indexes over the individual for both the system and the standard reference

Next, Table 7.24 shows diagnosis output regarding syndrome characterization according to the predominant event type in the patient. In the table the respective AHI indexes for the two categories of apnea events considered (obstructive and central) are shown for the system and for the standard reference. Again, it has to be remarked that mixed events by the system are considered within the obstructive category. On the other hand note that in this case hypopneas classification is included for the computation of the corresponding indexes. In this respect it has to be taken into account that hypopnea events are systematically classified by the standard reference as obstructive, while classification by the system follows the same classification procedure as for every apneic event, regardless of its type (see subsection "*Classification of apneic events*" of Chapter 5). Considerations made with respect to assignment of the concrete syndrome class can be consulted in subsection "*Apneic events classification*" of Chapter 6.

Table 7.24. Results for syndrome characterization according predominant nature of the apneic event. RN = Recording Number; AHI_{obst} = Obstructive Apnea-Hypopnea Index; AHI_{cen} = Central Apnea-Hypopnea Index

RN	System			Standard reference		
	AHI_{obs}	**AHI_{cen}**	**Syndrome type**	**AHI_{obs}**	**AHI_{cen}**	**Syndrome type**
200088	32.66	0.24	OSAHS	30.23	0	OSAHS
200259	54.14	43.02	MIXED OSAHS	67.49	32.28	MIXED OSAHS
200386	65.18	5.18	OSAHS	50	4.48	OSAHS
200532	80.96	3.56	OSAHS	81.16	0.66	OSAHS
200568	18.8	0	OSAHS	17.72	0	OSAHS
200929	75.17	3.85	OSAHS	77.2	0.33	OSAHS
201249	71.83	0.72	OSAHS	62.96	0	OSAHS
201294	83.74	0.87	OSAHS	88.06	0.16	OSAHS
201394	20.65	0.75	OSAHS	22.27	0	OSAHS
201824	50.4	17.08	MIXED OSAHS	40.52	21.19	MIXED OSAHS
202275	50.33	11.79	MIXED OSAHS	49.89	6.74	OSAHS
202666	10.51	0.27	OSAHS	9.64	0	OSAHS
202733	15.29	44.1	CSAHS	13.14	30.04	CSAHS
202956	77.87	0	OSAHS	71.23	2.92	OSAHS
203249	56.85	0.28	OSAHS	57.79	0.14	OSAHS
203294	9.94	1.07	OSAHS	10.03	0.18	OSAHS
203494	38.48	25.42	MIXED OSAHS	44.56	19.02	MIXED OSAHS
203645	86.98	3.62	OSAHS	75.92	5.79	OSAHS
203798	58.86	23.48	MIXED OSAHS	63	21.57	MIXED OSAHS
204135	42.94	12.43	MIXED OSAHS	32.25	12.49	MIXED OSAHS
204452	10.77	0.19	OSAHS	10.89	0.2	OSAHS
204480	12.63	0	OSAHS	9.53	0	OSAHS
205813	5.96	0	OSAHS	4.08	0	OSAHS
205948	16.12	0	OSAHS	12.08	0	OSAHS
206040	84.21	7.38	OSAHS	89.33	4.08	OSAHS
206181	50.43	2.86	OSAHS	46.7	0	OSAHS

As it can be seen from Table 7.24, according to the classification made for the system, from the total set of 26 patients, 19 patients were diagnosed with OSAHS, 6 were classified as with mixed OSAHS (relevant presence of central events) and one patient was diagnosed with pure CSAHS. From the part of the standard reference 20 patients were diagnosed with OSAHS, 5 with mixed OSAHS and 1 with CSAHS. Figure 7.23 shows the comparative in form of bar diagram. Calculation of the associated kappa index for the classification leads to a $\kappa = 0.839$ with $\kappa_M = 0.839$. Test for distribution homogeneity shows no significant differences ($\chi^2_{(2)} = 0.117$, p-value 0.943).

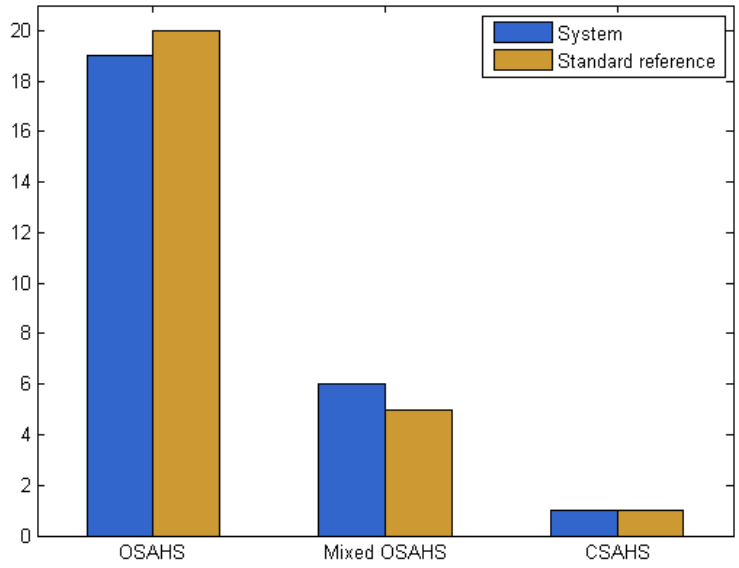


Figure 7.23. Distribution of recordings according to syndrome classification

Only 1 patient (recording 202275) was considered under different syndrome type categories. Taking a look to the respective computed indexes (AHI_{obs} of 50.33 and AHI_{cen} of 11.79 for the system, AHI_{obs} of 49.89 and AHI_{cen} of 6.74 for the standard reference) no great differences seem to be present. Specifically, for this recording the respective proportions of central events with respect to the number of obstructive events are of 0.23 for the system and 0.16 for the standard reference. Recall from section "*Diagnostic generation*" of Chapter 5 that the limit for class transition was arbitrarily set to 0.2, and therefore the resulting discrepancy in the classification. As previously stated for the case of severity classification, probably a more smooth transition between class boundaries might contribute to relax this kind of abrupt transitions.

Subsequent Figure 7.24 shows the distribution of the respective indexes. No significant differences have been found between the medians of the respective distributions, neither for AHI_{obs} nor for AHI_{cen} populations (*p*-values of 0.721 and 0.292, respectively). When paired test is considered, the null hypothesis of zero difference between paired recordings is accepted with regard to AHI_{obs} distributions (*p*-value 0.1996) and rejected for AHI_{cen} (*p*-value 0.0108). Noticeable is the high number of outliers in the upper tails of AHI_{cen} distributions. Basically, what this is telling us is that, as expected, patients with high prevalence of central events are less frequent among the population. Respective ICC indexes for the distributions are 0.974 (*p*-value 0) for the ratings for AHI_{obs}, and 0.934 (*p*-value 1.03×10^{-13}) for the ratings with respect to AHI_{cen}.

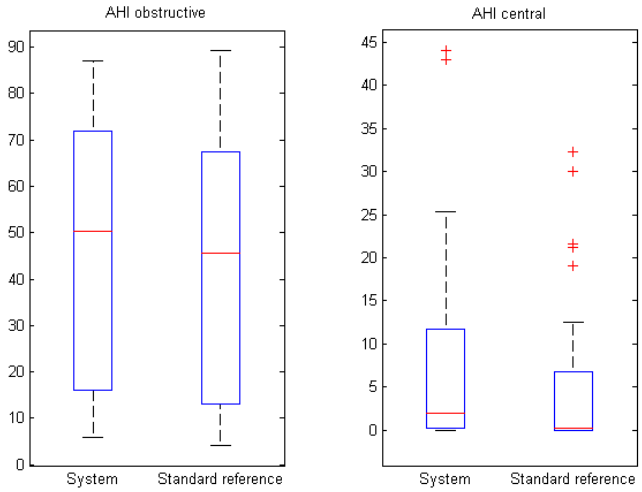


Figure 7.24. Box plots showing distributions of AHI_{obs} and AHI_{cen} indexes over the recordings for both the system and the standard reference

Maximum absolute differences regarding AHI_{obs} have been found in the recording 200386 (15.18 difference, with AHI_{obs} 65.18 for system and 50.00 for the standard reference). For this recording the associated proportion of central events with respect to the number of obstructive events is 0.08 and 0.09 (respectively for system and standard reference). OSAHS classification by the system is therefore, in spite of the difference, clear and consistent with that of the standard reference. On the other hand, for the AHI_{cen}, maximum discrepancy is found for recording 202733 (14.06 difference, with 44.10 AHI_{cen} for system and 30.04 for the standard reference). In this case associated proportions of central events with respect to the number of obstructive events are 2.88

and 2.29 (respectively for system and standard reference). Final classification of this recording has been of CSAHS type both by the system and the standard reference. In fact, this is the only pure CSAHS patient in the population.

Classification according to sleeping position of the patient is shown in subsequent Table 7.25. Criterion described in subsection "*Diagnostic generation*" of Chapter 5 has been applied in order to obtain the corresponding classification labels.

Table 7.25. Results for syndrome characterization according predominant sleep position during occurrence of the apneic event. RN = Recording Number; AHI_{ns} = Non-Supine Apnea-Hypopnea Index; AHI_s = Supine Apnea-Hypopnea Index

RN	System			Standard reference		
	AHI_{ns}	AHI_s	Classification	AHI_{ns}	AHI_s	Classification
200088	0	32.9	SUPINE POSITIONAL	0	30.23	SUPINE POSITIONAL
200259	15.13	82.03	SUPINE POSITIONAL	15.44	88.44	SUPINE POSITIONAL
200386	70.07	0.29	NON-SUPINE POSITIONAL	54.48	0	NON-SUPINE POSITIONAL
200532	56.03	28.49	NON POSITIONAL	60.08	21.74	NON-SUPINE POSITIONAL
200568	18.14	0.66	NON-SUPINE POSITIONAL	17.07	0.65	NON-SUPINE POSITIONAL
200929	38.77	40.25	NON POSITIONAL	39.18	38.35	NON POSITIONAL
201249	45.16	27.38	NON POSITIONAL	40.42	22.54	NON POSITIONAL
201294	79.66	4.95	NON-SUPINE POSITIONAL	83	5.22	NON-SUPINE POSITIONAL
201394	0	21.41	SUPINE POSITIONAL	0	22.27	SUPINE POSITIONAL
201824	55.2	12.28	NON-SUPINE POSITIONAL	50.81	10.91	NON-SUPINE POSITIONAL
202275	57.43	4.69	NON-SUPINE POSITIONAL	53.1	3.53	NON-SUPINE POSITIONAL
202666	8.25	2.53	NON-SUPINE POSITIONAL	5.92	3.72	NON POSITIONAL
202733	22.66	36.72	NON POSITIONAL	18.93	24.25	NON POSITIONAL
202956	0.34	77.53	SUPINE POSITIONAL	0	74.16	SUPINE POSITIONAL
203249	11.06	46.07	SUPINE POSITIONAL	11.2	46.73	SUPINE POSITIONAL
203294	4.79	6.21	NON POSITIONAL	2.46	7.74	SUPINE POSITIONAL
203494	10.5	53.41	SUPINE POSITIONAL	7.43	56.16	SUPINE POSITIONAL
203645	64.37	26.23	NON-SUPINE POSITIONAL	58.95	22.76	NON-SUPINE POSITIONAL
203798	66.02	16.32	NON-SUPINE POSITIONAL	70.55	14.02	NON-SUPINE POSITIONAL
204135	30.8	24.58	NON POSITIONAL	24.66	20.08	NON POSITIONAL
204452	2.12	8.85	SUPINE POSITIONAL	0.2	10.89	SUPINE POSITIONAL
204480	1.09	11.54	SUPINE POSITIONAL	0.16	9.38	SUPINE POSITIONAL
205813	3.87	2.09	NON POSITIONAL	2.72	1.36	NON POSITIONAL
205948	16.12	0	NON-SUPINE POSITIONAL	12.08	0	NON-SUPINE POSITIONAL
206040	12.01	79.58	SUPINE POSITIONAL	11.68	81.74	SUPINE POSITIONAL
206181	37.14	16.14	NON-SUPINE POSITIONAL	32.04	14.66	NON-SUPINE POSITIONAL
Total	27.95	25.51		25.87	24.29	

According to data in Table 7.25, system's output comprises 9 subjects classified as supine positional, 10 patients as non-supine positional while the remaining 7 were classified as non positional. Standard reference's classification is composed of 10 supine positional patients, 10 non-supine positional and 6 non positional. Figure 7.25 shows the comparative in form of bar diagram.

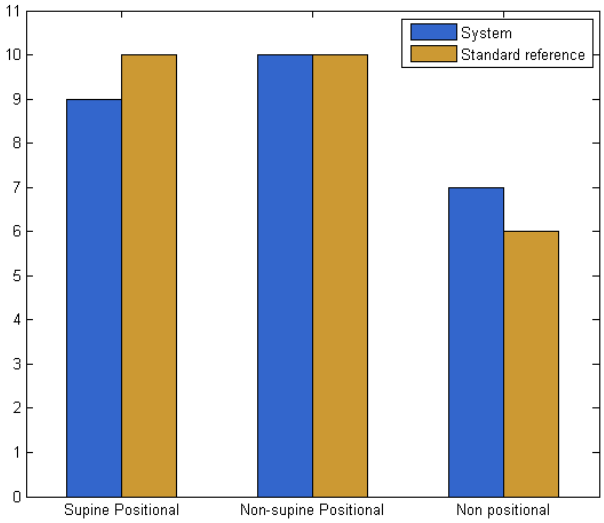


Figure 7.25. Distribution of recordings according to positional syndrome classification

From the total set of 26 patients, 3 of them show discrepancy in terms of the sleeping position associated to SAHS. In recording 200532 (AHI_{ns} of 56.03 and AHI_s of 28.49 for the system, AHI_{ns} of 60.08 and AHI_s of 21.74 for standard reference), the ratio $\frac{AHI^{NS}}{AHI^{S}}$ is 1.97 (*non-positional*) for the system and 2.76 (*non-supine positional*) for the standard reference. The limit that defines class boundaries has been set to 2 (see Chapter 5, "*Diagnostic generation*"). For the recording 202666 (AHI_{ns} of 8.25 and AHI_s of 2.53 for the system, AHI_{ns} of 5.92 and AHI_s of 3.72 for standard reference) the respective $\frac{AHI^{NS}}{AHI^{S}}$ ratios are 3.26 (*non-supine positional*) for the system and 1.59 (*non-positional*) for the standard reference. Finally for recording 203294 (AHI_{ns} of 4.79 and AHI_s of 6.21 for the system, AHI_{ns} of 2.46 and AHI_s of 7.74 for standard reference) the respective ratios are 0.77 (*non-positional*) for the system and 0.32 (*positional*) for the standard reference. Recall that the limit between supine positional and non positional SAHS for this ratio has been set to 0.5. Calculation of the associated kappa index for the positional

SAHS classification leads to a κ = 0.824 with κ_M = 0.941. Test for distribution homogeneity shows no significant differences ($\chi^2_{(2)}$ = 0.1296, *p*-value 0.9373).

Distribution of the AHI_{ns} and AHI_s indexes over the 26 recordings is shown in Figure 7.26. No significant differences have been found between the respective distributions neither for AHI_{ns} (*p*-value 0.749) nor for AHI_s populations (*p*-value 0.749). Difference between paired observations show statistical significant differences for AHI_{ns} distribution (*p*-value 0.007), while the null hypothesis is accepted for the ratings of AHI_s at the 0.05 significance level (*p*-value 0.087). With regard to this last, some statistical differences can also be perceived from the point of view of the outliers. Specifically within AHIs distributions, recordings 200259 (AHI_s of 88.44), 202956 (AHI_s of 74.16) and 206040 (AHI_s of 81.74) are considered outliers at the 0.05 significance level over the scorings of the standard reference. The corresponding AHI_s values for these recordings according to system's ratings are very similar (respectively, 82.03, 77.53 and 79.58). No remarkable consequences on the diagnosis are perceived beyond this statistical singularity. ICC values in this case report 0.985 agreement for AHI_{ns} scorings and 0.989 for scorings of AHI_s (both with *p*-value 0). The indexes corroborate the high agreement in the rating between system and standard reference.

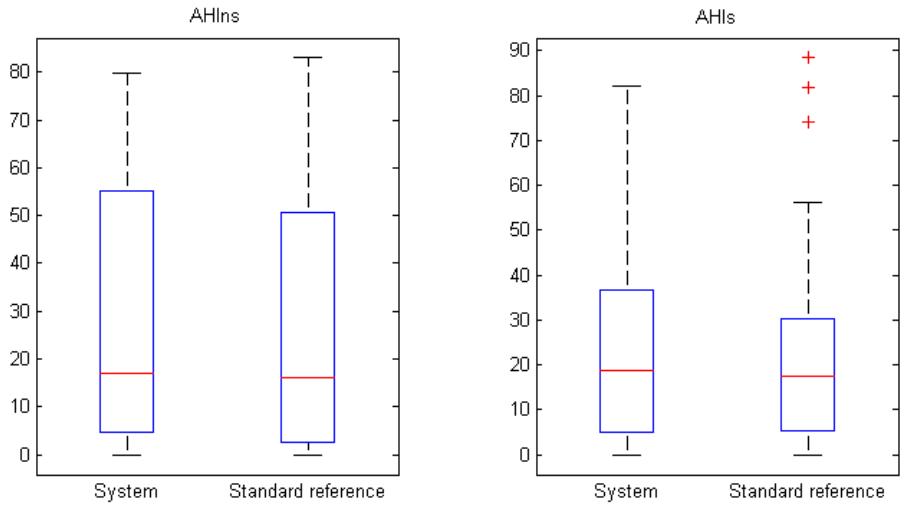


Figure 7.26. Box plots showing distributions of AHI_{ns} and AHI_s indexes over the recordings for both the system and the standard reference

Maximum absolute differences regarding AHI_{ns} have been found for recording 200386 (15.59 difference, with AHI_{ns} 70.07 for system and 54.48 for the standard reference). No supine events are scored for the standard reference and only 0.29 of AHI_s is obtained for the system. Therefore despite the difference in the AHI_{ns} indexes the lack of supine effect in SAHS for this recording is clear. On the other hand, with regard to AHI_s, maximum difference is found for recording 202733 (12.47 difference, with 36.72 AHI_s for system and 24.25 for the standard reference). In this case associated $\frac{AHI^{NS}}{AHI^{S}}$ ratios are of 0.61 for the system and 0.78 for the standard reference, both above the 0.5 limit criterion established to discard positional effects in the patient.

Final arousal indexes obtained by the system and the standard reference are represented in subsequent Table 7.26. Resulting TST is also indicated in order to interpret the corresponding ArI values.

Table 7.26. Final computed arousal indexes; RN = Recording Number; TST = Total Sleep Time; ArI = Arousal Index

	System		Reference	
RN	**TST**	**ArI**	**TST**	**ArI**
200088	4:13:30	7.83	5:27:30	4.03
200259	6:44:30	38.32	6:36:30	39.65
200386	6:57:00	25.76	6:42:00	29.55
200532	7:18:00	10.96	6:01:30	38.06
200568	7:36:30	22.32	7:44:00	15.52
200929	6:45:30	53.93	6:04:30	66.92
201249	6:58:30	43.78	7:06:00	47.61
201294	6:52:00	48.79	6:19:30	54.93
201394	6:38:00	13.72	5:26:00	8.83
201824	7:05:00	30.64	6:25:00	42.54
202275	6:37:00	17.98	6:14:00	19.32
202666	7:31:00	8.78	7:15:30	6.90
202733	7:19:30	39.32	6:23:30	40.42
202956	5:56:00	32.19	5:28:30	54.15
203249	7:08:30	38.51	7:14:00	31.11
203294	5:38:00	13.10	5:41:00	7.39
203494	8:34:30	19.61	7:40:30	21.65
203645	6:38:00	14.62	7:36:00	30.92
203798	6:33:30	29.31	6:29:30	47.51
204135	6:55:00	15.18	6:19:30	25.01
204452	5:12:00	13.08	5:03:00	6.53
204480	6:25:30	10.91	6:24:00	11.25
205813	6:12:29	25.81	6:37:00	15.72
205948	5:28:00	7.50	6:17:30	6.37
206040	7:59:30	22.52	7:06:30	36.53
206181	7:00:00	42.29	6:37:00	42.01

Distribution of the corresponding ArI values over the individual recordings is displayed in Figure 7.27. The respective mean and standard deviations are 24.87±13.69 for the system, and 28.86±17.91 for the standard reference. Wilcoxon test accepts the null hypothesis of distributions having equal medians with *p*-value of 0.552. When paired design is checked the null hypothesis is also accepted with *p*-value of 0.118 (significance level at 0.05). In difference with AHI, in this case discrepancy on TST does not affect significance of the paired test due to the increase in the standard deviation of the individual differences for each recording (mean±std of the differences was 3.45±5.24 for AHI, and it is -3.98±9.79 for ArI). On the other hand, the negative sign in the average points out toward slight subestimation (although not significant) of the actual ArI value. With regard to the absolute discrepancies over the individual recordings, maximum difference has been found for recordings 200532 (27.1 difference, with ArI 10.96 for system and 38.06 for the standard reference) and 202956 (21.96 difference, with ArI 32.19 for the system and 54.15 for the standard reference). Calculation of ICC index shows absolute agreement of 0.792 (*p*-value 1.44×10^{-7}).

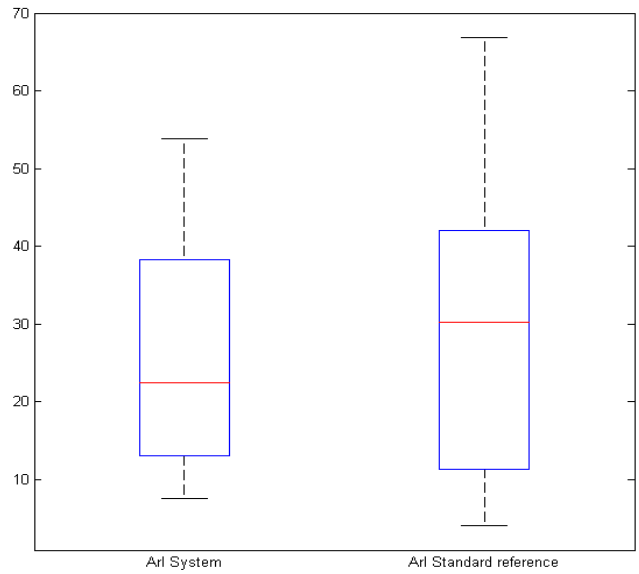


Figure 7.27. Distribution of the values of ArI among the individual recordings for system (left) and standard reference (right)

Finally, some additional results are presented with regard to the relation between apneic events and arousals, the distribution of apneic events over the different sleep stages and the correlation between different estimations of sleep fragmentation.

In the following Figure 7.28, a diagram is presented showing the percentage of apneic events associated with the occurrence of an arousal event. Apneic events have been broken down in obstructive apneas, central apneas and hypopneas, according to classification by standard reference. Again, recall that mixed events are included within the obstructive category. Criteria for association of arousals with respiratory events have been described in Chapter 5 and they establish that the distance between the apneic event and the subsequent arousal should be less than 5 seconds. In Figure 7.28 "Total Events" refers to the total number of apneic events regardless of its concrete type, and the last bars ("Total Arousals") refer to the corresponding percentage of arousals associated with apneic event. Exact percentage values have been superimposed over the corresponding bars.

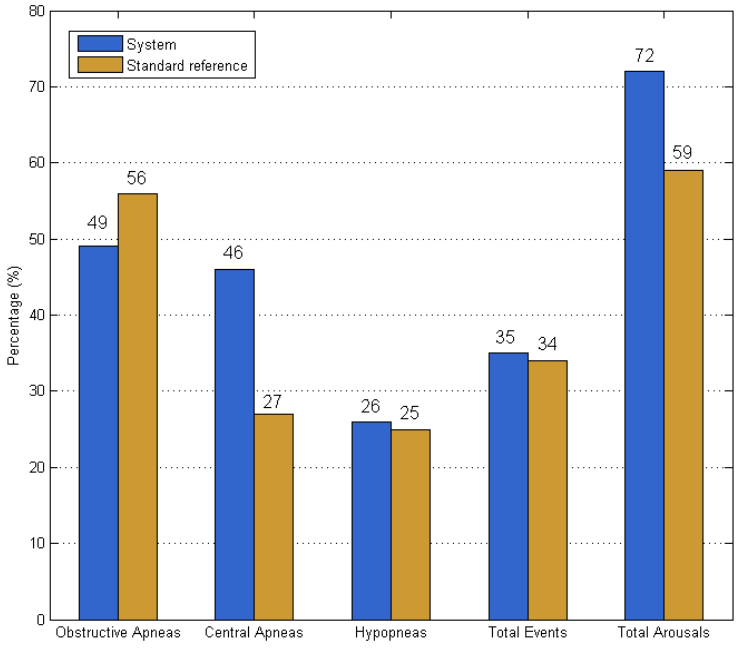


Figure 7.28. Percentages of apneic events ending with EEG arousal events

Taking a look to the graph, it can be seen that the proportion of total apneic events with terminating arousal is very similar (35%, 34%, Wilcoxon sign rank test *p*-value 0.253). When exploring individual differences regarding the different types of apneic events, similar proportion maintains for hypopnea events (26%, 25%, *p*-value 0.057). On the other hand higher diversity is found regarding apneas in general and particularly in the case of central apneas. No statistical significance has been found, for the case of

obstructive apneas (*p*-value 0.433). However, in the case of central apneas the median of the paired differences is found to be statistically different from zero at the 0.05 significance level (p-value 0.021). It is also noticeable that both, for the system and the standard reference, more than the half of the arousal events have an apneic origin ("Total arousals" in Figure 7.28). The proportion in this case is higher according to the system (72% vs. 59%), and is found to be statistically significant (*p*-value 7.69×10^{-4}).

Distribution of the apneic events over the different sleep stages is shown in Figure 7.29. Results are shown taking into account obstructive apneas (upper left), central apneas (upper right), hypopneas (lower left) and the full set of apneic events (lower right).

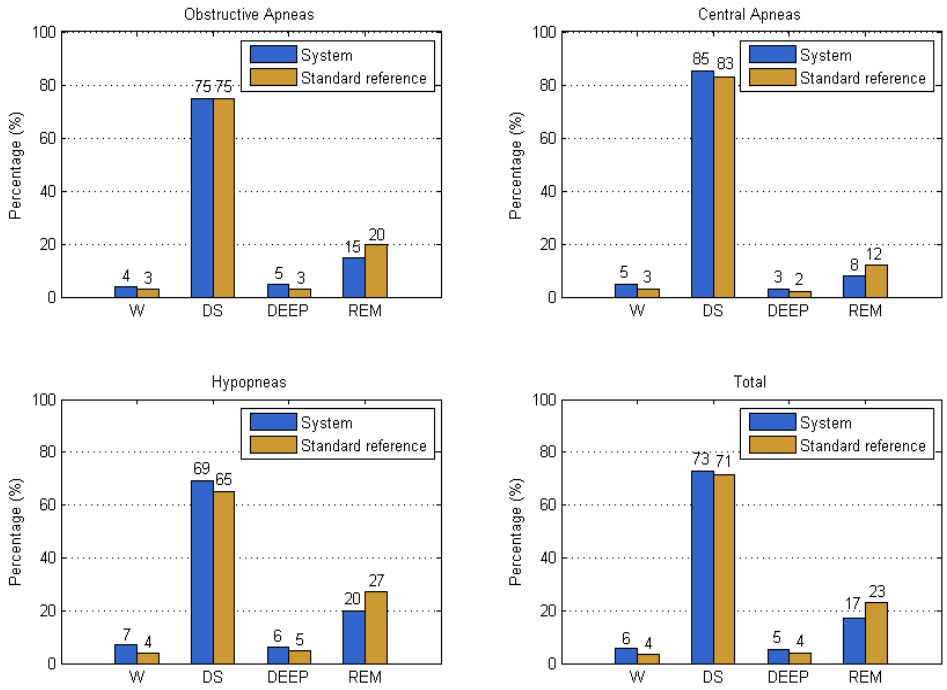


Figure 7.29. Distribution of apneic events over the sleep stages for the system and the standard reference

The general trend is repeated with independence of the concrete type of apneic event. In general, visual inspection shows some underestimation on the number of apneic events in REM intervals, while this difference is distributed among the other stages. Differences are also significant at the statistical level ($\chi^2_{(3)}$= 33.10, *p*-value 3.07×10^{-7} for obstructive apneas; $\chi^2_{(3)}$ = 15.13, *p*-value 0.002 for central apneas; $\chi^2_{(3)}$ =

95.98, *p*-value 0 for hypopneas; $\chi^2_{(3)}$ 201.35, *p*-value 0 considering apneic events in general). Nevertheless, the general picture shows accordance with standard reference, standing out that most of the apneic events take place within light sleep periods (DS) or during REM. An observation has to be made regarding apneic events in the W (wakefulness) category: it is not about apneic events occurring during stable periods of wakefulness, but apneic events taking place within isolated W stages caused by sleep disruption, probably as a consequence of an associated arousal. No statistical differences have been found, in general, among the medians of different distributions at the 0.05 significance level. The exception is over the respective hypopnea distributions regarding REM (*p*-value 0.035). On the other hand, when paired comparison is performed on the individual recordings, the median of the differences is in general statistically different from zero, with the exception of the proportion distributions of central events. For reasons of brevity the concrete *p*-values are omitted in this case.

Finally, existence of possible correlation has been investigated among different markers for sleep breath disturbance in the system. According to design of the validation tests (see Chapter 6), the interest in this respect is the confrontation of the AHI with the rate of desaturations using different cut-offs and the ArI. The resulting plots can be seen in Figure 7.30. Correlation coefficients (Person's *r* and Spearman's *rho*) as well as their respective *p*-values for significance test are shown on the top of each graph.

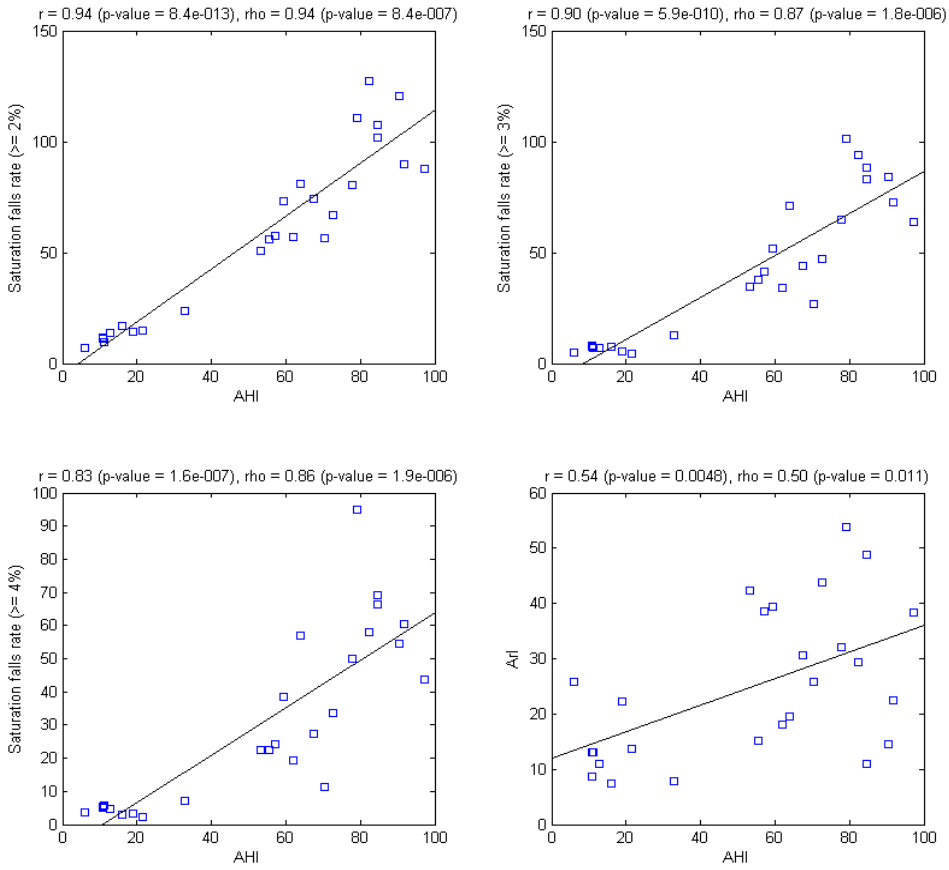


Figure 7.30. Scatter plots showing correlation between AHI and different markers for sleep breath disruption

According to Figure 7.30, significant correlation has been found in all the cases, being the maximum correlation found between AHI and the rate of desaturations higher than 2% (Figure 7.30, upper left). This correlation slightly decreases as the cut-off for relevant desaturations increases. The lowest correlation has been found between AHI and ArI (Figure 7.30, lower right), thus meaning that the number of desaturations per hour of sleep is a better estimator of the AHI than the ArI. It has to be said that all detected falls in the saturation signal are used for the calculation, regardless of their association, or not, with an apneic event. Correlation between AHI and ArI for the standard reference is also calculated and the resulting graph is displayed in subsequent Figure 7.31.

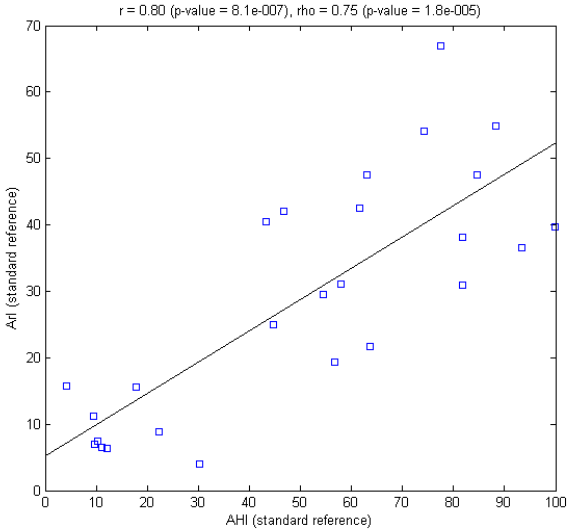


Figure 7.31. Scatter plot of AHI values and ArI values according to standard reference

The found correlation in this case is higher than the one obtained for the system (compare Figure 7.31 with lower right subplot of Figure 7.30). The reason is probably because of the underestimation by the system of the total number of arousal events as the severity of SAHS increases. Effectively, the study of the sensitivity index for arousal detection as a function of the AHI of the standard reference shows a decreasing factor (see Figure 7.32). However the negative correlation is not significant at the 0.05 significance level. On the other hand, one has to realize that different estimation of the TST between system and standard reference might also be affecting the resulting ArI indexes.

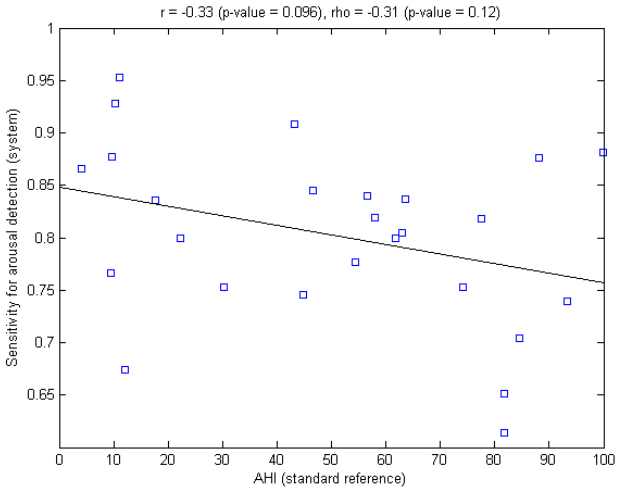


Figure 7.32. Scatter plot between AHI values according to standard reference and sensitivity index for detection of arousal epochs by the system

No study for the comparison of desaturation rates for the standard reference has been made due to unreliable scoring of desaturations (see subsection "*Design of the validation test*" in Chapter 6).

7.6. Summary of this chapter

In this Chapter presentation and analysis of the results for validation of the constructed system to aid the physician in the diagnosis of SAHS have been carried out. Design of the validation strategy has been developed throughout Chapter 6 which has been scheduled into several validation subtasks focusing in five key aspects according to available annotations from the standard reference: (1) detection of EEG arousals, (2) characterization of the sleep macrostructure, (3) detection and differentiation of apnea and hypopnea events, (4) classification of apneic events as obstructive, central or mixed, and (5) final diagnosis.

A set of 26 patients has been drawn from the SHHS database and used as independent validation set among the different validation subtasks in order assess real performance of the system.

8. DISCUSSION, CONCLUSIONS AND FUTURE WORK

8.1. Discussion

From decade of the 1970s in which Guilleminault formally defined the sleep apnea-hypopnea syndrome, many approximations have been developed with the aim of aiding the physician in the diagnostic task. Such an interest initially emerged from the difficulty associated to the manual analysis of the voluminous data recorded from patient's biosignals during sleep. The long duration of the PSG recording, and the number and the complexity of the involved signals, makes of the analysis of the PSG to be a very tedious and time consuming task for the clinician.

Incorporation of the health management information systems into the medical centers represented an important first step in order to solve many of the initial problems. The simple fact of being able to carry out an offline analysis over the digitalized signal yet represents an important evolution. Advantages are many, and include the amount of saved paper, reduction of space requirements for archiving, exchangeability of data, easiness in the annotation of the detected events, support of visual revision of the PSG (for example allowing visualization of the signals over different time and amplitude scales), and in general, favoring of the documentation process and reporting. However, even with the support of computers on the previous tasks, automation of the manual PSG analysis continued to be the key milestone in the development of computer programs to support the physician in the diagnosis of SAHS.

As with the evolution in computer's computational power and the development of new techniques for signal analysis, supporting tools started to appear, that automated, at least in part, the analysis task. In fact, once the technology gap for the requirements of a

full PSG analysis has been overcome, there also started to appear approximations aimed at constituting comprehensive solutions for a full PSG analysis. In fact, nowadays there are already commercial products in this line (see Chapter 3). The reality is, however, that the use in practice of these systems from the medical personnel is, in general, very scant.

In the recent years the field of sleep studies has experienced an increasing demand because of the growing recognition of sleep as one of the fundamental pillars for a good quality of life. Today, the increasing interest in sleep medicine demands for an important technology support, which due to the complexity of the PSG analysis, necessarily involves the definitive introduction of automatic analysis systems to aid the physician in the diagnosis. However, as stated before, the developed systems up to this time still present certain drawbacks which have caused their use in practice, yet to be very low and restricted (see Chapter 1, "*Background*"). These two factors (increasing demands in the analysis and shortages of the current developments) favor that research lines in the development of these systems continue today to be an open area of interest, and motivate the development of this doctoral thesis.

Throughout Chapter 3 an analysis of the state-of-art regarding existent approaches in the field of automatic SAHS diagnosis has been performed. In this chapter drawbacks of preceding and current systems have been addressed (see Chapter 3, "*Critical Analysis*"). The conclusion is that there is still a lack of real comprehensive approaches that adequately handle the analysis from both the respiratory and the neurophysiological perspective. In addition, existent approaches rely excessively in the use of fixed protocols and categorical classifications. As a consequence, they are unable to deal with imprecision in data and to account for variability due to subjectivity of the human decisions.

On the basis of these shortages this doctoral thesis addresses the development of an intelligent system to aid the clinician in the diagnosis of SAHS. The system follows a comprehensive approach: it does not limit to the analysis of the respiratory signals, but neurophysiological activity is used as a contextual framework for the interpreting the respiratory events. Several signal processing and artificial intelligence techniques are integrated for this purpose, with prominent use of fuzzy logic as supporting technology

in the reasoning processes, allowing adequate handling of data imprecision, avoiding categorical judgments and enabling explanatory mechanisms. The main objective is a global characterization of the respiratory disease in the sleep context by mimicking physician's diagnostic procedures.

The system assumes data to be taken from an EDF file according to a certain montage that includes the most common set of PSG signals. The use of an open and standard format such as EDF is aimed at minimizing the problem of format diversification. Input acquisition module can deal with variations in the amplitude – dynamical range- or in the sampling rate of the signals to adapt to different recording configurations. Scaling and resampling operations can also be applied in the case of being necessary. In the our case, the concrete set of signals as well as their respective configuration lay in the montage of the standard reference that has been used for the validation of the system (see Chapter 5 "*Description of the inputs*"). On the other hand, if the available set of signals varies, then partial readaptation of the analysis algorithms might be necessary to adapt them for the new input montage.

From a functional perspective and according to its comprehensive philosophy, the system is structured into two great functional groups that interact with each other through common interfaces (see Chapter 5, "*System's architecture*"). Each of the two groups is in charge of carrying out the respective analysis of the neurophysiological and the respiratory activity.

Neurophysiological analysis has as its main objective the construction of the hypnogram and the detection of transient EEG micro-arousal related to the occurrence of apneic events.

The method developed for the evaluation of the macrostructure of sleep comprises mainly three processing tasks. The first one acts over the raw signals in order to extract relevant features, that include computation of the main characteristic EEG frequencies, EOG movements, characterization of the EMG, and the detection of important sleep transient events such as sleep spindles and K-complexes. The set of extracted features are subsequently fed into a reasoning stage organized as four fuzzy submodules, each one involving a different state of sleep (W, DS, DEEP, REM). A degree of membership

representing the continuous evolution of the corresponding sleep state is obtained as the output for each submodule. In the method, continuous means second-by-second. Of course, in the context of the processing of digital signals it would never be possible to achieve a real continuous output, but on the other hand, it can be asymptotically approximated by increasing the rate of the analysis (only limited by the sampling frequency of the digitalized signal).

The main objective of the obtained continuous representation is to overcome limitations of epoch-based methodologies such as AASM and R&K regarding their low temporal resolution and the unnatural classification of sleep based on the labeling of discrete epochs [**1**] [**2**] [**3**]. By using fuzzy logic, categorical classifications can be avoided and soft transitions can be exploited. Such properties allow us approximating the continuous evolution of the sleep though its different states. On the other hand it is also an objective to keep as much as possible interpretability of the system. In this respect, the fuzzy approach permits us to implement medical knowledge in form of fuzzy rules, close to human language, which facilitates understandability and allows explanatory capabilities. These rules can be consulted, interpreted and eventually edited by the final user. Moreover, follow up of the inferential process can be performed by tracking the reasoning process through the set of activated rules according to the given input. Explanation of the outputs is therefore enabled by listing the set of activated rules which, as previously mentioned, are implemented using linguistic labels in natural language.

Finally, a third processing step is applied, consisting of a series of postprocessings, in order to go back over the classical epoch-based hypnogram. Actually, in the developed system, this dual representation of the sleep macrostructure (continuous or epoch-based) is available. An important reason to keep the epoch-based representation - beyond allowing validation to be performed- is that, according to current standards, physicians are still use to *watch* at the sleep in an epoch scale. Thus to keep the system usable for physicians, it has been decided to allow both representations, continuous and epoch-based, to coexist in the application. In this manner, the clinician can keep at using an epoch-based scale, however having available information on the intra-epoch evolution.

There are a number of approaches throughout the literature dealing with the issue of achieving reliable automatic sleep scoring (see Chapter 5, "*Hypnogram generation*"). Reported results vary considerably depending on the concrete technique, the validation methodology or the condition of the individuals. In the approximations analyzed in the review by Penzel and Conradt, for example, when involving normal subjects [**4**] [**5**] [**6**] [**7**] reliability varies between 70% and 90%. On the other hand when agreement is measured in disturbed sleep there is a general drop ranging from 65% to 87% [**8**]. Based on just one EEG channel the work of Flexer et al. developed a probability-based sleep stager accounting for three different stages: W, DEEP and REM. Validation results regarding recordings from two different laboratories show accuracy of 79% for W, 82% for deep and 68% for REM (group A, 20 training recordings, 20 testing recordings) and 25% for W, 87% for deep and 61% for REM (group B, 14 training recordings, 14 testing recordings). Both groups included only healthy subjects [**9**]. In Schaltenbrand et al. an automatic sleep scoring system based on ANNs is trained over a set of 12 recordings and tested on a set of 11 recordings [**10**]. Agreement between visual consensus and the system was 80.6% with the lowest agreement involving classification of stage 1 and 3. The authors further improved their system and presented a bigger study including different disorders [**11**] and obtained 82.3% agreement on average, with higher accuracy (84.5%) for the healthy group. Roberts and Tarassenko developed a method based on the analysis of nine healthy subjects that derived automatic hypnogram generation. They did not provide quantifiable validation results but they were concerned about the adequateness of using an epoch-based hypnogram because of the poor temporal resolution and the limitation imposed by having discrete stages [**12**]. Later they presented a system based on autoregressive modeling for feature extraction and an artificial neural network, giving continuous measures in parallel with automatically calculated sleep stages [**4**]. This system was then used for comparison with manual sleep staging in clinical practice by Caffarel et al. [**13**] over a set of 114 patients. In addition 28 of the 114 studies were also examined by an additional human scorer in order to assess human interrater reliability. This study is of special interest for comparative results since validation is carried out using the four categories considered by the system object of this doctoral thesis: wake, light-sleep (here referred as DS), deep sleep and REM. Results of the validation show overall agreement of automatic and manual scoring for the 114 studies of poor $\kappa = 0.305$. In the subgroup of 28 randomly selected studies, the overall agreement of automatic and manual scoring was again

relatively low (κ = 0.331). In this work human inter-scorer reliability was also assessed and found to be higher than that achieved by the system (κ = 0.641).

Investigation on the magnitude of agreement between human scorers is very important in order to identify a target value for accuracy of automated scoring systems. In fact, in the pathway toward the development of an automatic scoring system with reliable generalization capabilities, agreement between the system and one particular human should not overpass this target level. Otherwise the system would lose generalization by resembling too much to a particular scorer, increasing disagreement with another one.

In the work of Anderer et al. [**14**] validation of an e-health solution for automatic classification according to R&K was performed on a large database of PSG recordings. Validation on a set of 286 PSGs involving both healthy subjects (n = 94) and controls (n = 49) reported 78.3% of overall agreement (κ = 0.71) between the system and expert human consensus scoring. When quality control was added (affecting less than 1.1% of total epochs) overall agreement improved to 79.6% (κ = 0.72). In this study interrater human variability was found to be 76.9% (κ = 0.68). In a later publication by the authors, the system was modified to comply with the new AASM rules and validation was presented over 72 recordings (56 healthy subjects, 16 patients). Overall epoch-by-epoch agreement between computer-assisted and the human expert scoring slightly increased to 82% (κ = 0.76) for the AASM version [**15**]. Agreement between human scorers regarding the AASM standard in this study was 82.1% (κ = 0.76). Another interesting work is the one by Pittman et al. [**16**] which investigated the assessment between automatic analysis of PSGs and two human scorers in a population of 31 subjects with suspected sleep-disorder breathing. With respect to sleep stage scoring, agreement between the system and the human scorers was 77.7% (κ = 0.67) for scorer A and 73.3% (κ = 0.61) for scorer B. Interscorer reliability yield to 82.1% agreement (κ = 0.73). The sleep scoring scale comprised five different stages: wake, 1, 2, delta and REM. When stages 1 and 2 were merged into a unique 1+2 (DS) stage, agreement between human scorers increased to κ = 0.80, whereas agreement between system and human scorers improved to κ = 0.71 when comparing scorer A, and κ = 0.65 for comparison with scorer B.

Among additional studies that assessed human interscorer variability, the following ones can be highlighted [**17**] [**18**] [**19**] [**20**]. The work of Danker-Hopfe et al. [**17**] is based on the assessment of interrater variability according to the R&K and AASM standards. Evaluation was performed using the SIESTA database, which was also used for validation of the system presented in [**15**]. Similarly, kappa values yield to $\kappa = 0.76$ for AASM criteria and $\kappa = 0.68$ for R&K. In Norman et al. [**19**] evaluation of interobserver agreement was performed on a set of 62 recordings (52 patients) involving 5 experienced scorers from 5 different clinical centers. Results according to the R&K standard showed an overall epoch-by-epoch agreement between scorers of 73% (range 67-82%). Agreements were higher in the normal subset (mean 76%, range 65-85%) than in a subset of 38 patients with SDB (mean 71%, range 65-78%). The conclusion is that the level of agreement in sleep stage assignments varies between scorers, by diagnosis and by recording. In Basner et al. [**18**] interrater agreement according to R&K method was assessed on a database of 60 PSG recordings (20 OSASs patients) involving three clinical centers. Interscorer variability in this case resulted in $\kappa = 0.72$. Finally, the study of Whitney et al. [**20**] is of special interest since it assessed the reliability for the SHHS database. Procedure involved the R&K method and interscorer comparisons on epoch-by-epoch sleep staging showed values of kappa in the range 0.81-0.83. Intrascorer reliability yielded to kappa values in the range 0.79-0.87. In comparison with other studies, agreement values in this case are relatively higher than the previous reported values for the R&K standard. Nevertheless, procedures to maximize reliability have been implemented by rigorous training of scorers, systematic reporting of signal quality, and explicit formulation of scoring rules [**21**]. In fact, according to Norman et al. this agreement likely represents a maximal value for interobserver agreement which may also be influenced by the smaller proportion of subjects with sleep disordered breathing in the population [**19**].

With regard to the proposed approach within the developed system, validation has been performed using an independent set of 26 recordings. Comparison has been carried out using the discrete output from the system to expert's manually staging of sleep. Attending AUC indexes best results were achieved regarding discriminating capacities on wakefulness (AUC of 0.92) and REM sleep (AUC of 0.89). AUC value for deep sleep was 0.86 and for this stage, the highest specificity was obtained (0.97), however at the cost of lower sensitivity (0.75). On the other hand drowsy sleep characterization

showed slightly lower values of agreement with expert's scorings (AUC of 0.85) while maintaining similar values of sensitivity and specificity. According to overall general agreement, in 84% of the epochs, human and system agreed in the classification. Analysis by adjusting possible agreement occurring by chance on the 26 patients, established an overall inter-rater agreement for the method of $\kappa = 0.76$. This value can be classified as *substantial* ($0.61 \leq \kappa \leq 0.80$) according to the linguistic scale provided by Landis and Koch [**22**], which is in accordance with the general reported values on human inter-rater reliability (see above).

Attending to individual epoch misclassifications, pair-wise comparative has reported the highest discrepancy to be found involving stages W/DS, DS/REM and DS/DEEP. These three groups account for 94% of total discrepancy, respectively 38.74%, 29.52% and 26.25%. The common factor that stage DS is present among all groups, showed that with regard to epoch-based sleep staging, most of the discrepancies between system and standard reference involve DS characterization. On the other hand one has to take into account that DS is by far the sleep stage in which patients spent most of their time (44.09% of the time according to standard reference, 42.32% of the time according to system's epoch-based hypnogram). Therefore it is normal that in absolute terms, most of the errors are localized regarding this period. The former agrees with Basner et al. [**18**] who reported that kappa values correlate positively with the amount of time spent in the respective sleep stage. In fact, when looking at the relative error for each sleep stage, it has been revealed that DEEP was the stage with the highest discrepancy on the positive agreement (0.55).

In the work of Danker-Hopfe et al. [**17**] deviating scorings between human scorers were analyzed with regard to both AASM and R&K procedures. In order to allow comparison of the results, for the R&K method stages S3 and S4 were grouped over a common SWS stage, equivalent to N3. In this study the combination of deviating scorings between SWS/N3 and S2/N2, and between S1/N1 and S2/N2, account for more than 60% of all discrepancies for both standards (AASM and R&K). Subsequent relevant misclassifications involved W/S1 and W/N1, S1/N1 and REM, and S2/N2 and REM. Having in mind that in the developed system, DS involves S1/N1 and S2/N2, a similar trend is suggested with regard to validation results obtained for the system. Moreover, if the combination of N1/N2 and S1/S2 into synthetic DS (DS_{syn}) is

performed, then for AASM method 50.48% of total misclassifications are obtained for DS_{syn}/N3, 28.53% for W/DS_{syn} and 18.77% for DS_{syn}/REM, which account for 97.78% of overall discrepancies. Analogously for R&K, 43.94% of total misclassifications are obtained for DS_{syn}/SWS, 30.73% for W/DS_{syn} and 23.28% for DS_{syn}/REM, representing 97.95% of total discrepancy. This trend resembles much more to the one obtained for the proposed system, regardless of the relative order. However one should consider the proportion of the total time spent in each stage to evaluate this order (something not reported in the work of Danker-Hopfe). With regard to individual sleep stages, analysis of the agreement between two experts revealed the highest agreement for stage REM, followed by W, N3/SWS, N2/S2 and S1/N1.

A similar analysis can be performed over the study of Whitney et al. [**20**] which assessed the reliability of sleep staging for the SHHS database. In this study S1 has the greatest discrepancies between the scorers (average 22.91% positive agreement), followed by delta (deep, with average 69.14% positive agreement), REM (average 77.52% positive agreement), and N2 (78.75% positive agreement), being stage W the most reliably discriminated (average 88.79% positive agreement). This resembles much more the trend by the implemented system. In fact, by taking combination of S1/S2 into synthetic DS (DS_{syn}), then the respective trend becomes DEEP (the highest discrepancy, 69.14%), followed by REM (77.52%), DS_{syn} (79.64%) and W (88.79%), which is the same pattern obtained in for the system (see "*Sleep Staging*" in Chapter 7). Taking the combination of deviating scorings between the different experts, discrepancies between SWS/S2 and S1/S2, account for almost 50% of all discrepancies. Subsequent relevant misclassifications involved S2/REM (17.58%), W/S2 (12.62%) and W/S1 (12.09%). Again, if the combination of S1/S2 is taken into synthetic DS (DS_{syn}), then for 36.63% of total misclassifications are obtained for DS_{syn}/SWS, 30.56% for W/DS_{syn} and 28.75% for DS_{syn}/REM, accounting in total for 95.94% of overall discrepancies. Recall for the system these three discrepancies also accounted for 94% of total discrepancies (see above).

Taking into account the results on the SHHS database, it can be said that overall the system behaves as one human expert more. In other words, system's output regarding discrete epoch classification is comparable to that made by a human expert. This support the idea that a reliable epoch-based hypnogram can be obtained from the

continuous fuzzy output of the system, and it suggests that the continuous representation preserves the information contained in the discrete hypnogram (in fact constituting a superset).

The general approach for the detection of EEG arousals consists of a method comprising three different stages: (i) firstly, a signal processing stage for extraction of features along the different channels occurs. Two central derivations of EEG and submental EMG are used. In this respect a frequency-based analysis is performed throughout the EEG decomposing the signal into its main frequency bands. A marker based on changes into alpha-beta range is used for localization of possible arousal events and the EMG signal is analyzed searching for amplitude changes; (ii) a set of relevant features is extracted and a grouping process of the individual detected events is performed. Correlation of the individual events in time is solved by using some temporal aggregation rules. Grouped events form arousal patterns characterizing time intervals in the PSG, where the occurrence of an arousal event is possible; (iii) finally, a classification task is performed in order to classify those epochs associated with the occurrence of characteristic patterns. Several models were compared to be integrated into this classification stage using a first set of 42 features, and the ANN model proved to be the classifier with the best discriminative power.

Due to the high number of inputs features (42) existence of possible redundancy is then explored and feature selection methods are applied at this respect. The study pursued a dual objective: (i) on one hand, to check whether feature selection could improve the results obtained using the original set of 42 features. Several feature selection methods including various filters and wrappers approaches were tested. The aim is to check if either it is possible to reduce the number of needed features, or even to improve classification results by reducing the effects of possible data redundancy. On the other hand (ii) it was also of interest to investigate the best approach for feature selection. In this respect, besides the comparison of the individual filter and wrapper methods, a combinatory approach was explored through the union and the intersection of the features selected by the individual methods.

For this study two different machine learning methods (an ANN and a SVM) were used to check if improvement can be achieved independently of the used classifier.

Taking a look to the results (see subsection B in "*Identification of EEG arousals*", Chapter 7) feature selection was proved to be adequate, because the number of features was drastically reduced while also reducing the classification error. Moreover, it can be checked that, independently of the classifier and the feature selection method used, lower errors are obtained than employing the whole set of features. In this respect, while obtaining the same performance with fewer features is an indicator for the presence of irrelevant information in the initial input set, the concurrent improvement in the performance implies that this irrelevancy was actually preventing the machine learning algorithms for achieving better prediction, and therefore blocking the learning process. Cancellation of this noise therefore allowed for a better performance of the classifier.

To be mentioned is that, apart from classical methods –based on filters and wrappers- the combination of individual methods through the union and the intersection of different candidate subsets was investigated. In this respect, it was found that the union of the different rankings of relevant features provided, in general, the best results. This is a surprising result as one can expect that if a feature is selected by all filters, then it should have associated a high predictive power. This *a priori* reasoning effectively should support the intersection as a better method than the union. In fact, this is the assumption of Aguilar-Ruiz et al. in [**23**], nevertheless the results show that their method is not able to obtain an adequate prediction.

Actually, after analysis of the results, the method that was proved to perform the best to reduce the number of features was the union of individual filters. Final results showed considerable reduction in the percentage of error compared with the results using the whole set of features (from 0.196 for SVM and 0.194 for ANN, to 0.150 for the ANN). In addition the great reduction in the number of features was achieved, 8 versus 42, thus over 80% reduction with respect to the original set.

After application of feature selection, the final resulting method was validated using an independent validation set composed of 26 recordings. Resulting agreement ratios show a sensitivity and specificity of 0.650 and 0.950 respectively, with overall agreement index of 0.904 and AUC of 0.800.

Detecting events in a multi-channel environment is a complicated task as one has to deal with time correlation of the individual events over the different channels. Developed method for detection of EEG-related arousals carries out the analysis in PSGs using the information available using three different channels of the PSG: EEG derivations (C3/A1 and C4/A2) and submental EMG. These three channels are those directly involved in the detection of EEG arousals according to the standard medical procedure [**24**]. Methods that work with single channel have been proposed [**25**] [**26**], however, they are unable to perform an accurate detection owing to the fact that they are discarding relevant information. For example detection of the arousal events in REM needs the detection of amplitude changes over the EMG signal to be done. Therefore, although working with two EEG derivations, the method presented in [**27**] should not be also reliable for the detection of EEG arousals in REM since it does not account for an EMG derivation.

In section "*Identification of EEG Arousals*" of Chapter 5 some other methods [**28**] [**29**] [**30**] have been reported being able to work in several channels. In this category, the first method mentioned was the one proposed by De Carli et al. [**30**], in which out of a total of 11 recordings, 8 were used as a test. After two experts' consensus for the definition of the validation criteria was obtained, two different data sets were considered, the first including only "definite arousals" whereas the second one included also a set of "possible arousals". Average sensitivity and specificity were reported respectively to be 89.60% and 74.46% for the first case, and 86.41% and 88.35% for the second one.

In the work of Sugi et al. [**28**] results are reported on 8 male patients in which parameter determination was done in a quarter of the whole available data, therefore, training set and testing sets were not independent. Validation was performed individually over each recording showing an average sensitivity of 86% with an average rate of false positives of 3.8%. Nevertheless the fact that the unit of measure for TN (30 seconds) was different from the unit for the TP, FN and FP (detection time of 1.28 sec), makes it difficult to directly compare these results with the method proposed here.

Finally in Shmiel et al. [**29**] the number of patients involved in the study is 26, of which 20 are independently used as the testing set. Only sensitivity and positive prediction value measures are reported, (respectively 75.2% and 76.5%), on the set of 20 patients used as a test. Again, these numbers are calculated as the average of the validation results over the individual recordings. Besides the EEG and the EMG, for these results, pulse and oxygen saturation signals of the patient are used. When only using information concerning the EEG and the EMG, the results in this work achieve an average sensitivity of 63% and mean positive predictive value of 82%. The approach makes also use of data mining techniques, however the philosophy is conceptually different to the one proposed here: the use is aimed at discovering hidden relationships among the different channels. In the case of the method within the proposed system, instead, a more AASM-likely procedure is followed and machine learning approaches are used to carry out feature selection of the events detected over the different channels.

Results of the implemented approach in the developed system are in the range of the results reported over literature. However, direct comparison with the preceding approaches is difficult due to different subject conditions and data processing methods. In this respect, it is known that characterizing the EEG and other physiological data differ among subjects. Moreover, considerable variability among different experts in the scoring of arousals has been reported [**31**] (see also discussion regarding final ArI indices below).

Analysis of the respiratory activity comprises processing of the signals related with the breathing function, which in accordance with input montage includes airflow measured with thermistor, plethysmography bands for the recording of thoraco-abdominal movements and oxygen saturation. Signal preprocessing includes the detection and characterization of overflow and loss of focus intervals, applying signal reconstruction where possible. Subsequent analysis of the signal comprises the identification of apneic intervals and the localization of descending (*possible desaturations*) and ascending (*possible resaturations*) intervals in the SaO_2 signal. For each individual event detected, features are then extracted which characterize the event with quantifiable data (see Chapter 5, "*Analysis of respiratory signals*").

Once all the respiratory signals have been analyzed, temporal correlation of the individual events is performed. The result of the temporal correlation is what it has been called *apneic pattern*. An *apneic pattern* defines a group of inter-related events as with potential diagnostic relevance. In this respect a first level of temporal constraints involves the events located in the respiratory signals. Temporal correlation is guided by the principle of physiological significance: saturation falling intervals detected over the SaO_2 signal are used as triggers, defining searching intervals to establish the corresponding cause-effect relationships with the corresponding apneic intervals.

The use of saturation events as the main triggering alert for the detection of apneic events relies mainly on its relative simplicity. The feature of interest in the saturation signal is relatively easy to extract because of its minimal features. It simply consists of a reduction and subsequent rise in level. This ease of recognition, together with its slow evolution in time (as compared with the remaining PSG signals) has obvious implications for its predictability and allows relatively simple artifact rejection mechanisms to be implemented (see Chapter 5). This simplicity, its high tolerance to noise, and the easiness of its interpretation, makes of the oxygen saturation signal, a reliable detector for the localization of apneic events. In fact, many works can be found over on the literature based on the use of the oxygen saturation signal as a simple apneic event detection method [**32**] [**33**]. The work of Taha et al. also started with the detection of desaturation and then analyzed the sum of RIP [**34**]. On the other hand, the correlation method also allows an apneic pattern to be formed without presence of significant fall in the saturation. This situation can be caused by presence of undetected artifacts in the saturation channel, or due to the chaining of several apneic events that, as a consequence of their proximity, do not reach to trigger two differenced desaturations.

Once temporal correlation of the respiratory events has been performed, evidence of the apneic pattern is then completed with information coming from sleep stages, detected artifacts, sleeping position, transient electroencephalographic events and information from light's state. Actually, the apneic pattern, as defined, constitutes the meeting ground for both the neurophysiological and the respiratory data.

Evaluation of the quantitative and qualitative information from the different events that integrate the apneic pattern is then evaluated using fuzzy inference. This approach contrasts with conventional approximations, which would use fixed protocols for evaluation of the individual features. These approximations excessively depend on the concrete setting of numerical limits. A typical example is the setting of a predefined threshold for the associated desaturation to score a hypopnea. For example if a 3% value is set, then an hypopnea will not be scored under the 3% desaturation even though the airflow reduction reflected over the respiratory channels may be more significant than other event with higher desaturation. In fact, higher variability is found between scorers and automatic scoring methods depending on the concrete setting [**20**] [**35**] [**36**]. In contrast, in the developed system, fuzzy reasoning allows implementation of reasoning processes based on approximation and similarity criteria. That means the apneic pattern is globally evaluated, not longer dependent on concrete numerical value of one feature, but based on its overall shape. Knowledge is implemented through IF-THEN fuzzy rules that use linguistic terms in natural language, and a degree of membership is obtained regarding each possible classification, i.e. considering the possibility of the pattern as being considered an *apnea*, a *hypopnea*, or to be discarded as a *false positive*. A similar process is scheduled for the subsequent classification of the apneic events as *obstructive*, *central* or *mixed*. Therefore categorical judgments are avoided, and besides, the use of linguistic labels enables follow up of the inferential process and explanation capabilities by tracking the set of activated rules according to the given input (see Chapter 5, "*Diagnostic generation*").

Final validation results of the system have been carried out considering different analysis levels. An epoch-by-epoch validation was firstly scheduled regarding capabilities for the location of the apneic events showing overall agreement of 0.89 over the 26 tested recordings, with sensitivity of 0.81 and specificity of 0.92. Not significant individual differences were detected over the recordings, however positive correlation was found regarding sensitivity of the system and severity of the syndrome, while negative correlation was found for detection accuracy (see Chapter 7 "*Apneic events detection*").

Further analysis suggested that when the proportion of apneic events in the recording is low, then the severity of the associated events is more reduced: that is,

predominant apneic events are of the very mild hypopneas. Therefore, accurate detection of these events for the human eye is more difficult, which increases subjectivity in the detection, thus agreement in the temporal detection is more reduced. At the same time, the proportion of negative cases is so high that total agreement tends to remain high. On the other hand, as the severity of the syndrome increases, it also does the sharpness of the associated events, which increments the proportion of true positives, and the sensitivity of the detection. In contrast, this has a counter effect since the sleep pattern and the sleep in general becomes more unstable, which slightly affects to the increasing proportion of FPs and FNs of the system. A general reduction in the agreement between human scorers has been reported in the literature as a function of the severity of the syndrome [**19**] [**17**] [**14**]. Therefore the system behaves as expected.

Not many studies can be found in the literature that show the results of an epoch-by-epoch validation regarding temporal localization of apneic events. In contrast, usually studies assessing reliability of computer approximations for SAHS diagnosis tend to present their results regarding final obtained AHI values [**37**] [**38**] [**39**], or they do not provide measurable units for scoring negative agreement [**34**] [**40**]. One of the few studies found that attempted such a task is the study of Pittman et al. [**16**], in which a 3x3 confusion matrix was created to identify epochs with 0, 1 or 2 events in a set of 31 PSG recordings from 3 different scorers (two human scorers M_1 and M_2 and one automatic algorithm A). Results showed 89.7% agreement between M_1 and A, 89.7% agreement between M_2 and A, and 94.9% agreement between M_1 and M_2. Only epochs previously scored as sleep by all 3 scorers were included in the analysis. In this study apnea was scored if airflow was absent for 10 seconds, and a hypopnea was scored if thoracoabdominal movement or airflow was reduced by 30% compared to baseline for at least 10 seconds with at least a 4% oxygen desaturation. With this criteria, mean RDI events in this population was 20.6 and 22.5 respectively for human scorers M_1 and M_2. As mentioned before, our system reported a mean agreement of 89%, however mean AHI in the population is 50 according to standard reference, which as stated before, is expected to reduce the possible maximum agreement. On the other hand, while a 4% criteria for scoring a hypopnea event in applied in the work of Pittman et al, in the SHHS database hypopnea events are scored on the basis of amplitude reduction criteria, regardless of the associated desaturation [**21**]. In this respect is well-known that interscorer agreement strongly depends on the criteria established for desaturation

threshold. The study of Whitney et al. for the SHHS database, for example, showed an increase of ICC between human scorers from 0.74 -without any required threshold for associated desaturations- up to 0.99 when a 4% threshold for desaturation was required.

Evaluation of discriminative capabilities of the developed system was also assessed on the characterization of apnea/hypopnea events, and their classification as obstructive/mixed/central. Results show, in the first case, overall sensitivity and specificity of 0.89, with AUC and agreement ratios also over 0.89. For the classification of apnea events, resulting sensitivity and specificity was 0.84 and 0.78 respectively for the obstructive category. AUC value for this case was 0.81 whereas agreement index was 0.82. Validation of apneic events classification required of mixed events from the system to be considered within the obstructive category. No attempts were performed by SHHS scorers neither to distinguish between obstructive and mixed, nor to classify hypopnea events [**21**].

According to the obtained results, robust characterization of the apneic event as apnea or hypopnea is achieved by the system (no significant bias has been detected favoring any particular class). On the other hand, in the case of apnea classification statistical significance was found between classification of obstructive and central events, with higher sensitivity for the detection of obstructive events (see Chapter 7, "*Apneic events classification*").

Results on the classification of apnea events as obstructive or central have to be carefully interpreted. Firstly it has been mentioned already that validation considering as obstructive, both actual obstructive but also mixed events of the system. Second, it has to be observed that in the population of 26 recordings used for validation, the proportion of obstructive events (1695 in total) with respect to that of central events (854) was 1.98. In this respect, and although the precise prevalence of central events in SAHS patients has not been described, studies have reported that incidence of obstructive events is much higher (around 10 times higher) when compared to the number of central events [**41**]. This has to be taking into account since a population with higher prevalence of obstructive events may lead to a significant increase in the general agreement index. In fact significant correlation was found between the proportion of obstructive events in the recording and the agreement index for classification.

In general, interrater reliability is higher on the scoring of apneas than hypopneas [**42**]. Clinical literature on the discrimination between apnea/hypopnea classes over individual events is scant, however there are several studies regarding variability of the associated AHI indexes by using different definitions of hypopnea. Actually regarding detection of hypopneas, considerable variability has been reported [**35**] [**20**]. There are also very few studies on the reliability of event subtypes (central, obstructive, mixed), although at least one study has demonstrated poor reliability for mixed apneas [**43**]. For a comparative analysis of the system's results with other computer-based approaches in the detection and classification of apneic events, the reader is referred to the corresponding discussion in Chapter 3.

In any case, besides the concrete validation results it is interesting to recall that fuzzy classification of events carried out by the system, provides of an intrinsic mechanism to deal with disagreement and interrater variability. Indeed, it has been stated that for validation purposes the maximum degree of membership associated with each detected apneic pattern is used as the final classification output (see Chapter 6, "*Design of the validation tests*"). However in practice for each formed apneic pattern the system does not hide the remaining classification possibilities. Moreover each apneic pattern is always associated with a certain degree of membership and its corresponding linguistic label for each one of the possible outputs; in the case of apneic event detection this characterization comprises categories of *apnea*, *hypopnea* and *false positive*. Then, each apneic event detected (independently if it has higher degree of membership for apnea or hypopnea) is assigned with a degree of membership regarding categories of *obstructive*, *mixed* and *central*. The former implies that in every moment the physician has weighted evidence pointing out at the possible characterizations of the apneic pattern under consideration. Therefore for doubtful apneic patterns –as for example in the case of mild hypopneas or mild apneas- the clinician has available individual evidence for evaluation. Based on this evidence the clinician can then decide on the final classification. Even more, besides the associated degree of membership, explanation of the outputs using natural language and complete quantifiable information can be accessed through the application interface (see Chapter 5, "*Main user interfaces*").

Final diagnosis evaluation of the system included analysis of different quantifiable parameters for sleep severity and syndrome characterization, as well as interrelation between neurophysiological and respiratory activity.

Results on syndrome severity estimation reported consistent results with those of the standard reference leading to κ of 0.839. Four different categories were considered at this respect by using widely extended segmentation of the corresponding AHI (see Chapter 5, "*Diagnosis generation*"). Based on this segmentation, only 2 out of 26 patients received different syndrome severity classifications, leading to 0.92 agreement similar to that obtained by Pittman et al. between human scorers and outperforming that obtained for man-machine agreement in the same study [**16**]. Moreover, minimal differences were found in the final AHI of the two recordings: 10.78 (system) vs. 9.64 (standard reference) for one patient, and 12.63 (system) vs. 9.53 (standard reference) for the other (recording number 204480). Actually, this is a good example in order to highlight the inconvenient of using discrete ranges to delimit the diagnostic categories. Indeed even though from the perspective of the associated severity, it is obvious that no significant differences exist between, for example, a patient with AHI of 25 and other with AHI of 26, it also does not between a patient with AHI of 10.78 and other with AHI 9.64. However the use of a fixed limit (in this case AHI > 10) in order to differentiate patients from normal subjects, causes in this case the first one to be classified as patient –with mild SAHS- and not the other. In fact, there is some discussion about the correct setting of the threshold limiting patients and normals (it usually ranges from 5 to 15), or even on the adequateness of using AHI as the sole parameter to diagnose a patient with SAHS [**44**] [**45**].

With respect to numerical comparison of the final AHI, ICC coefficient between the system and the standard reference was 0.98. In the work of Whitney this index was calculated over three human scorers for the SHHS database and obtained values ranged between 0.74 up to 0.99, depending on the concrete requirement for an associated desaturation or arousal for the scoring of an apneic event [**20**]. Other studies such as the one by Pittman et al. [**16**] reported ICC values in the range 0.95 and 0.99 for a 4% desaturation criterion. Results of the system are therefore, within the common ranges of human interrater agreement. Statistical analysis by means of paired test, however, showed some differences over individual recordings regarding AHI. On the other hand,

when ApI and HI were analyzed separately, only ApI reported significant paired differences. These differences were associated with discrepancies on TST between the system and the standard reference (see Chapter 7, "*Final patient diagnosis*"). As with regard to syndrome classification (OSAHS, mixed OSAHS or CSAHS) system's agreement with standard reference lead to $\kappa = 0.84$ and ICC of 0.97, with only 1 out of 26 patients under different syndrome categories. No similar studies have been found on the literature assessing reliability of such classifications. Similar agreement indices ($\kappa = 0.82$ and ICC of 0.98) were obtained for SAHS characterization according to positional effect. Such study revealed that in our validation population there was similar proportion among supine positional patients (9 out of 26) and non-supine positional (10 out of 26), whereas the was fewer number of patients absent of any positional effect (6 out of 26). Other studies have reported that the proportion of supine positional patients in OSAS is more than 50%. However, differences in severity and sample stratification may affect the final proportion [**46**].

Final computed arousal index (ArI) show a mean value of 24.87 within the population of 26 recordings. In this case no significant differences where found with respect to standard reference due to increase in standard deviation of the individual differences. Intraclass correlation coefficient leaded to agreement of 0.79 which can be regarded as high in comparison with interscorer ICC for the SHHS database that was 0.54 [**20**]. Other studies on interscorer reliability of ArI showed different values depending on the study. For example different definitions of arousals were assessed by Loredo et al. in 20 subjects with and without obstructive sleep apnea. Arousal scoring that used the AASM definition had ICC of 0.84, but the value dropped to 0.19 to 0.37 when shorter arousals were used [**47**]. Smurra et al. assessed the comparison of two different definitions of arousals in 20 patients with obstructive sleep apnea of varying severity. They reported ICC of 0.96 following AASM criteria [**48**]. Poor-to-moderate agreement in arousal identification has been reported in an interscorer reliability study of 15 accredited European laboratories. Using AASM definitions, they reported an overall kappa of 0.47. Agreement was best for arousals scored during deep sleep when the background EEG most contrasted with the faster frequencies or arousals, and for studies noted a priori to be easily classified [**49**]. An important factor to be taken into account over these studies is the presence of additional cues on the PSG being scored. Many of the previously cited studies included respiratory tracings and, while not

explicitly mentioned, probably also included ECGs. Thomas [**50**] assessed arousal scoring reliability using AASM criteria in 17 patients with obstructive sleep apnea syndrome. The event-by-event scoring agreement between scorers was 91%. However, when the respiratory tracings were removed from the recording, the agreement dropped to 59%.

Further interrelation between neurophysiological and respiratory activity was investigated with regard to the relation between apneic events and arousals, the distribution of apneic events over the sleep stages, and the correlation between different estimators of sleep fragmentation.

According to the obtained results 35% of total detected apneic events were associated with terminating arousal. The same results were obtained according to standard reference (34%). When looking at the different events separately, apneas were significantly more likely to be associated with EEG arousals than hypopneas. The former is in accordance with the results of Thomas for a study carried out in 17 patients with obstructive sleep disordered breathing [**50**]. Only for the case of central apneas, significant differences were found between the proportion of events associated with arousal between the system and the standard reference. Studying the total number of arousal events identified by the system, 72% had an apneic origin while this proportion decreased to 59% according to standard reference. Besides, significant correlation was found between ArI and AHI indexes, both for the system and the standard reference. These results support the use of EEG arousals as evidence criterion for the occurrence of apneic events. However, as expected, stronger correlation was obtained when using the rate of desaturations per hour of sleep. Similar results have been obtained in the study of Pitson and Stradling [**45**].

Finally, distribution of the apneic events over the different sleep stages showed that most of the events concentrate over DS and REM sleep. This distribution was independent of the concrete type of apneic events. Taking into account that DS accounts for more than 40% of sleep time and REM for about 14%, the former confirms REM as a stage of special prevalence of apneic events. Statistical differences were found regarding the exact distribution within each category between the system and the

standard reference; however the general picture reflects the common distribution reported over the clinical literature [**51**].

8.2. Future Work

Several future research lines and extensions can be proposed for the described system object of this doctoral thesis. Among these, the following ones can be highlighted:

Improvements to the sleep staging algorithm

Improvements regarding the sleep staging method are one of the lines of future work regarding the described system. In this respect main immediate efforts will focus on the accommodation of the epoch-based –discrete- hypnogram output to fulfill exact categories of the AASM standard. That is, decomposition of the current DS stage into N1 and N2. The former firstly implies decomposition of the current DS submodule into two new submodules and provide separate degrees of membership μ_{N1} and μ_{N2} allowing continuous characterization of the respective N1 and N2 states. Subsequently, post-processings have to be modified in order to incorporate the corresponding N1 and N2 epochs to the discrete hypnogram.

In addition, revision of the set of extracted features should be accomplished with the aim to detect additional transient events such as vertex sharp waves or slow eye movements (SEM). The motivation is to achieve a better characterization of the sleep process and the improvement of the inter-state discrimination capabilities.

Preliminary results concerning separation of DS into N1 and N2 have been evaluated using the same set of 26 patients used for system validation. No additional features were added by the time of submitting this document, i.e. the same set of extracted features and transient events are used as described throughout Chapter 5. Resulting accumulated contingency table, as shown in the corresponding section of Chapter 7, is shown below in Table 8.1.

Table 8.1. Accumulated contingency table showing preliminary results for sleep staging using the AASM's epoch-based classification

		SYSTEM				
REFERENCE		W	N1	N2	N3/DEEP	REM
	W	12331	405	640	41	461
	N1	387	564	573	5	325
	N2	350	1367	12221	927	835
	N3/DEEP	4	0	498	2475	23
	REM	157	52	665	22	4485

Calculation of Cohen's kappa over the previous table yields to $\kappa = 0.73$ ($\kappa_M = 0.94$) and total agreement index of 0.81. Thus, a slight decrease in the overall agreement is obtained, as expected, due to separation of DS into N1 and N2 (see Chapter 7, "*Sleep Staging*"). Percent of positive agreement leads to 83% for W, 16% for N1, 67% for N2, 62% for N3/DEEP and 64% for REM, that show the same trend as in the work of Whitney et al. for the SHHS database, although absolute agreements are slightly lower [**20**]. However it has to be taken into account that these are just preliminary results. The resulting sensitivities, specificities and AUC values are shown in the following Table 8.2.

Table 8.2. Agreement indexes for preliminary validation between expert and system using AASM classification (mean ± std. deviation). Sens = sensitivity; Spec = specificity; AUC = area under ROC curve

	Awake	N1	N2	N3/DEEP	REM
Sens.	0.89±0.08	0.34±0.19	0.77±0.12	0.77±0.21	0.83±0.12
Spec.	0.96±0.03	0.95±0.05	0.90±0.06	0.97±0.02	0.95±0.03
AUC	0.93±0.04	0.64±0.09	0.84±0.05	0.87±0.10	0.89±0.06

Again, previous results are still preliminary and are given only for guidance. Optimization of the corresponding post-processings for generation of the epoch-based hypnogram has not performed yet. More research is still needed prior to incorporation of these preliminary developments into the system, and future research will focus on this regard.

Additional research lines regarding the evaluation of patient's sleep in continuum may explore different approximations to characterize the underlying sleep processes

which can then be used for the inference of the output sleep stages. One interesting approach in this regard is the proposal of Kemp, which allows analysis of the sleep process as a consequence of neuronal feedback loops occurring between the brain cells [**52**]. According to this approach continuous sleep depth scale ranging from 0-100% can be obtained and reflecting the NREM sleep similar to delta power and an on/off switch for REM sleep on the basis of 1 sec intervals [**53**]. This proposal assumes that both NREM and a REM sleep processes can be simultaneously active, and it also allows characterization of transient EEG events as a consequence of the activity of these neuro-feedback loops. Besides, this model has shown nice properties such as being less sensitive to non-sleep related inter-subject variability effects, which suggest it as an interesting framework for the characterization the sleep micro-continuity.

Temporal correlation of events

Dealing with the time factor is of vital importance for the design of a comprehensive methodology for the analysis of the PSG. In accordance, throughout the development of the constructed system, several temporal correlation processes were scheduled for the detection of physiological patterns of clinical relevance over the different channels. For example, detection of EEG arousals involved the correlation of neurophysiological events over the EEG and EMG derivations (see Chapter 5, "*Identification of EEG arousals*"). Another example was the construction of apneic patterns in order to relate the different events of the respiratory channels, and subsequent integration of the neurophysiological information (see Chapter 5, "*Building apneic patterns: temporal event correlation*"). In this respect, even though the proposed mechanisms have proved to be effective, its implementation within the system can be considered of *ad-hoc* nature.

In this context, future work will address the extension of system's capabilities in the treatment of temporal information. The objective is the integration of a specific temporal reasoning module, in order to provide of systematic and powerful mechanisms to handle temporal relations between the different events in the diagnosis of SAHS. Specifically, research work is ongoing on the extension of the CTCN model [**54**] for the handling of imprecise information. Such extension represents a first step for its subsequent integration within the system. CTCN model, which states for *Causal*

Temporal Constraint Networks, has been developed as the result of a previous doctoral thesis of one of the members of the LIDIA research group [**55**]. Motivation emerged from the study of temporal aspects of SAHS after development of the SAMOA system (see Chapter 4) that revealed the importance of effective handling of the temporal information.

The CTCN model, which takes as a reference Meiri's general temporal constraint networks [**56**], is a general purpose temporal reasoning model. Main characteristics include: a) capabilities for processing quantitative, qualitative and causal constraints between temporal objects (points or intervals); b) the use of constraint satisfaction techniques to resolve reasoning tasks; c) enabling of objective causality (commonly accepted as public or semi-public knowledge) to be formalized; d) implementation of a single representation schema to represent temporal relationships that might arise between two events in a specific domain; or f) specification of temporal patterns for the inference of new knowledge.

The temporal handling of information is achieved by structuring of the information in different interpretation contexts, which are linked to each other through an inference mechanism in which the consistency of the temporal information is checked. This mechanism abstracts the information, ultimately producing further information with a high level of abstraction [**57**]. The inference mechanism is based on identifying temporal patterns in a context at two levels. The first level consists of temporal intervals –namely reference intervals- that establish the minimum temporal conditions to be satisfied by a set of events. The second level is composed of event subsets -referred to as inference intervals- within the reference intervals, from which the occurrence of new events of relevance to the information analysis can be inferred. In the inferential process, the temporal relations between the events, their degree of imprecision, and the timeless or static information, are all taken into account. The new events generated are fed back into the temporal information handling process, and any new patterns that may be identified in the different contexts are analyzed. The process terminates when no new patterns are encountered.

Based on the CTCN model, a working framework called TASAS has been developed and proved to be suitable for the diagnosis of SAHS [**55**]. In any case, and

abstracting an event from its temporal instant of occurrence, existentially, a temporal object in TASAS takes place in the domain in a binary form, that is, it exists (1 or true) or it does not (0, false). Therefore any interpretation or inference from itself should be carried out based on categorical decisions. Even though the representational CTCN model yet considered the possibility to deal with imprecise objects, the concrete mechanisms for the handling of this imprecision had not yet been defined. Therefore extension of CTCN to incorporate support for imprecise objects has been proposed as a first step for the integration of a CTCN-based temporal reasoning model into the system.

Research agenda in this respect includes: (i) the extension of CTCN model incorporating support for imprecision among the temporal objects, and (ii) in the definition of the temporal relationships (ii). Finally, (iii) the extended model has to be integrated within the system to act as temporal reasoning module in charge of the inferential processes for the detection of relevant events.

At the time of writing these lines, the first task on the research agenda has already been developed, but the description has been omitted from the doctoral manuscript since more work is still needed to achieve full integration of the CTCN model into the system. The interested reader is referred at consulting the corresponding paper in Appendix B "*A framework for handling fuzzy temporal events*" already submitted for publication.

Handling of variability and imprecision

Dealing with imprecision and variability in the medical diagnostics, but in artificial intelligence in general, is an interesting field for future research. In the developed system the fuzzy logic paradigm has been used as supporting framework to represent medical knowledge, and to implement the reasoning processes due to its nice properties for the concrete application domain, and enunciated throughout the text (see for example Chapter 3 "*Handling of imprecise information*" and "*Critical analysis*", and in general, Chapter 4).

In this line further investigation is aimed at continuing exploration of new methods for knowledge acquisition and model parameterization, in the line of the neuro-fuzzy modeling techniques discussed in the last parts of Chapter 4. On the other hand, it is the intention to extend application of fuzzy inference to additional parts of the system, for example, for the detection of transient events in the EEG, which at this time are still not characterized in terms of fuzzy degrees of membership.

In addition, future work will also assess the exploration of new representational schemas that handle data imprecision and that enable management of subjectivity and uncertainty in the reasoning processes. In this respect the framework that represents the emergent theory of *type-2* fuzzy sets offers an interesting paradigm that extends the idea of fuzzy sets to account for additional levels of uncertainty [**58**]. The basic idea is that of adding an extra –third- dimension to the representational form of the traditional fuzzy set (known under this theory as *type-1* fuzzy set) so that, instead of crisp grades of membership, the type-2 fuzzy set has grades of membership that are, at the same time, fuzzy. This allows stepping forward from the *modeling of words* to the *modeling of perceptions* [**59**]. An example for the case of SAHS may be the modeling of the concept of *severe SAHS*. Under the paradigm of type-1 fuzzy sets, a fuzzy set can be modeled representing the category *severe* so that given certain value of AHI *x*, then a certain degree of membership $\mu_{severe}(x)$ is obtained according to the given value of *x*. Let us say the value of *x* is 10, and that the resulting $\mu_{severe}(x)$ is 0.8 (thus a crisp value). Now, by using type-2 fuzzy logic, one might model a type-2 fuzzy set that given the value of *x*, instead of 10, returns a second-level type-1 fuzzy set $\mu_{severe,10}(y)$ that interprets the AHI value according to the age *y* of the patient. For example, the final membership value might be 0.8 if the age of the patient is 40, but 0.7 if the age of the patient is 60, or 0.6 if the age of the patient is 70. Beyond the concrete utility of the example, such a modeling will enable us to represent, under the same rule, different *perceptions* of the same concept, for example, depending if the subject is male or female, is an adult or a children, or in a general perspective, if the knowledge is acquired from one expert or another. It looks tempting, therefore, to explore the possibilities that this new representational framework may bring to handle variability and data imprecision in the field of SAHS.

Extension of the validation, use in clinical practice and technological transfer

Validation of the developed system has been carried out using an independent random sample of 26 recordings from the SHHS database. However, to move on from research scene to the clinical practice, more extensive validation of the system should be accomplished. In this respect, within the future work it is the intention to carry out further validation incrementing the sample of involved PSGs. One possible limitation of the current validation in this regard is that the set of PSG recordings in the SHHS database mostly comprise adult subjects. In fact average age range in the set of 26 patients is 68.5±7.7 (mean±std), which may bias the results of the validation for other population ranges.

Evaluation of the system on a different database will also contribute to assess generalization capabilities of the system. In fact, although SHHS database results from a multi-center cohort study, interscorer variability is expected to be higher when using a completely different dataset. The former may represent an additional challenge when using a database that involves a different set of signals. Indeed as stated in subsection "*Description of the inputs*" of Chapter 5, usually a problem with the design of automatic systems for PSG analysis is that they are highly dependent on the concrete signal specification and montage. Readapting the system to work on a different signal montage is also an interesting future line of development.

In this line, and although convergence to the AASM guidelines in clinical practice will still take some time -among other circumstances, because it implies substitution of previous existing hardware monitoring devices and training of current personnel in the new protocols- it is clear that adaptation of our system to fulfill AASM specifications should be addressed in the future. In this respect, it is important to comment that there are some differences between SHHS's recording montage (see Chapter 5 "*Description of the inputs*") and current AASM's recording specifications (see [**24**]). Particularly the absence of the nasal pressure sensor may influence the detection of hypopnea events. Indeed, characterization of the airflow in the SHHS study is done with the only use of a thermistor. However, as it has been commented in Chapter 2, pressure transducers produce an airflow estimation more sensitive to little airflow changes. Accordingly, besides the thermistor, AASM's guidelines recommend the adding of a nasal pressure

sensor for a better detection of the mild hypopneas. Adaptation of the software and test it over a database including recording of the nasal pressure would be of special interest. Other possible readaptations in this regard may include extension of the EEG channels beyond the use of central derivations, to include frontal and occipital excursions as recommended by the AASM. That would improve the characterization of the EEG activity and the detection of transient events. It is known, for example, that KCs tend to be maximal using frontal derivations, and that EEG-arousals can be better recorded over the occipital region [**24**]. Another difference is that the sampling rate specifications of the inputs do not always match AASM's recommended values. The former, however might not be critical since the input acquisition module of the developed system can already deal with variations in the amplitude –dynamical range- or in the sampling rate of the signals. In addition scaling and resampling operations can be applied in the case of being necessary. Therefore, it is assumed that the system will be able to analyze EDF recordings following different sampling rates –even those proposed by the AASM- without any problem.

In an endeavor to fulfill these requirements, clinical evaluation of the system has been initiated already through ongoing collaboration with the local hospital *Complejo Hospitalario Universitario de A Coruña (CHUAC).* Main objective is to assess system's performance in real clinical practice and using a different set of patients from those of the SHHS database. Besides, with the aim to reach future technological transfer, the developed system is being currently integrated into a commercial solution under the name of MIASOFT (*Intelligent Software for the Monitoring of the Sleep Apnea/Hypopnea Syndrome*). More information on the MIASOFT project can be accessed through its website [**60**].

8.3. Conclusions

This doctoral thesis has addressed the development of a system to aid the clinician in the diagnosis of the Sleep Apnea-Hypopnea Syndrome (SAHS). The main objective has been to obtain a system modeling intelligent behavior of the human scorers, and to be able to reduce both, time and effort, required from the medical personnel in the visual inspection and the scoring of the PSG.

Main limitations of the current computer based approximations consist in the scarcity of comprehensive approaches to the diagnosis and the excessive use of fixed protocols and categorical classifications. Therefore these systems usually limit to offer partial solutions to the problem and they are unable to deal with data variability and human subjectivity. The developed system contributes in this regard, because of (i) its comprehensive philosophy, in which neurophysiological activity is used as a context for the interpretation of the respiratory events, and (ii) the implementation of mechanisms to handle data imprecision, which mimic human's diagnostic procedures under the principles of generalization and approximation. On the other hand, while man-machine discrepancy due to subjectivity and data imprecision is a problem for the final acceptation of the automatic scoring systems in real practice, it is clear that automatic systems which try to imitate visual scoring of the PSG cannot be improved very much beyond the agreement achieved between human scorers. Given the unavoidable subjectivity associated to the diagnostic analysis, a possible way of improvement should rely on the development of aiding tools that avoid categorical classifications and that produce judgments based on similarity criteria. The developed system serves of the fuzzy logic paradigm in order to give response to the previous issues, however without renouncing to the advantages of the automatic analysis in terms of savings in time and effort for the revision of the PSG.

Despite more research has to be done, the obtained results are in general accordance with those reported for the agreement between human scorers. In this respect the system can be considered to behave as one expert more in the diagnostic task. It can be concluded, therefore, that the main objectives of this doctoral thesis have been accomplished, and that the system can be effectively used as supporting tool to aid the clinician in the diagnosis of SAHS.

Schematically, fundamental conclusions of this research work are the following:

1. A system modeling intelligent behavior of human scorers has been developed to aid the clinician in the diagnosis of the Sleep Apnea-Hypopnea Syndrome

2. The system simplifies the analysis task of the PSG, reducing both time and effort needed from the medical personnel

3. The validation results have shown that the system behaves as one expert more with regard to the diagnostic results tested over real PSG recordings

4. Limitations of the current approaches for automatic diagnosis of SAHS have been addressed, concretely:

    a. The analysis procedure is scheduled integrating both neurophysiological and respiratory information, leading to a comprehensive diagnostic approach in which respiratory events are interpreted in the context of the sleep macro- and microstructure, as well as the additional signals from the PSG

    b. Handling of data variability and human subjectivity has been performed through the implementation of fuzzy analysis techniques, avoiding categorical judgments, developing reasoning mechanisms based on similarity and approximation, and providing of explanative capabilities of its results close to natural language

8.4. References

[1] SL. Himanen and J. Hasan, "Limitations of Rechtschaffen and Kales," *Sleep Medicine Reviews*, vol. 4, no. 2, pp. 149-167, 2000.

[2] S. Kubicki, WM. Herrmann, and L. Höller, "Critical comments on the rules by Rechtschaffen and Kales concerning the visual evaluation of EEG records," in *Methods of sleep research*, S. Kubicki and WM Herrmann, Eds. Stuttgart New York: Fischer, 1985, pp. 19-35.

[3] J. Hasan, "Past and future of computer-assisted sleep analysis and drowsiness assessment," *Journal of Clinical Neurophysiology*, vol. 13, pp. 295-313, 1996.

[4] J. Pardey, S. Roberts, L. Tarassenko, and J. Stradling, "A new approach to the analysis of the human sleep/wakefulness continuum," *Journal of Sleep Research*, vol. 5, pp. 201-210, 1996.

[5] BA. Geering, P. Achermann, F. Eggimann, and AA. Borbely, "Period-amplitude analysis and power spectral analysis: a comparison based on all-night sleep EEG recordings," *Journal of Sleep Research*, vol. 2, pp. 121-129, 1993.

[6] S. Kubicki, L. Höller, I. Berg, C. Pastelak-Price, and R. Dorow, "Sleep EEG evaluation: a comparison of results obtained by visual scoring and automatic analysis with the Oxford Sleep Stager," *Sleep*, vol. 12, pp. 140-149, 1989.

[7] A. Torodova, HC. Hofmann, and W. Dimpfel, "A new frequency based automatic sleep analysis: description of the healthy sleep," *European Journal of Medical Research*, vol. 2, pp. 185-197, 1997.

[8] T. Penzel and R. Conradt, "Computer based sleep recording and analysis," *Sleep Medicine Reviews*, vol. 4, no. 2, pp. 131-148, 2000.

[9] A. Flexer, G. Gruber, and G. Dorffner, "A reliable probabilistic sleep stager based on a single EEG signal," *Artificial Intelligence in Medicine*, vol. 33, pp. 199-207, 2005.

[10] N. Schaltenbrand, R. Lengelle, and JP. Macher, "Neural network model: application to automatic analysis of human sleep," *Computers and Biomedical Research*, vol. 26, pp. 157-171, 1993.

[11] N. Schaltenbrand et al., "Sleep stage scoring using the neural network model: comparison between visual automatic analysis in normal subjects and patients," *Sleep* , vol. 19, pp. 26-35, 1996.

[12] S. Roberts and L. Tarassenko, "New method of automated sleep quantification," *Medical and Biological Engineering and Computing*, vol. 30, pp. 509-517, 1992.

[13] J. Caffarel, GJ. Gibson, JP. Harrison, CJ. Griffiths, and MJ. Drinnan, "Comparison of manual sleep staging with automated neural network-based analysis in clinical practice," *Medical and Biological Engineering and Computing*, vol. 44, pp. 105-110, 2006.

[14] P. Anderer et al., "An E-Health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24x7 utilizing the Siesta database," *Neuropsychobiology*, vol. 51, pp. 115-133, 2005.

[15] P. Anderer et al., "Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: validation study of the AASM version of the Somnolyzer 24x7," *Neuropsychobiology*, vol. 62, pp. 250-264, 2010.

[16] SD. Pittman et al., "Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing," *Sleep*, vol. 27, no. 7, pp. 1394-1403, 2004.

[17] H. Danker-Hopfe et al., "Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard," *Journal of Sleep Research*, vol. 18, pp. 74-84, 2009.

[18] M. Basner, B. Griefahn, and T. Penzel, "Inter-rater agreement in sleep stage classification between centers with different backgrounds," *Somnologie*, vol. 12, pp. 75-84, 2008.

[19] RG. Norman, I. Pal, C. Stewart, JA. Walsleben, and DM. Rapoport, "Interobserver agreement among sleep scorers from different centers in a large dataset," *Sleep*, vol. 23, no. 7, pp. 901-908, 2000.

[20] CW. Whitney et al., "Reliability of scoring respiratory disturbance indices and sleep staging," *Sleep*, vol. 21, no. 7, pp. 749-757, 1998.

[21] Sleep Heart Health Study SRC, "Sleep Heart Health Study. Reading center manual of operations," Case Western Reserve University, Tech Report VMLA-039-02, 2002.

[22] JR. Landis and GG. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159-174, 1977.

[23] JS. Aguilar-Ruiz, F. Azuaje, and JC. Riquelme, "Data mining approaches to diffuse large B-Cell lymphoma gene expression data interpretation," in *Lecture Notes in Computer Science*. Zaragoza: Springer-Verlang, 2004, pp. 279-288.

[24] C. Iber, S. Ancoli-Israel, A. Chesson, and SF. Quan, "The AASM Manual for the scoring of sleep and associated events: rules, terminology and technical specifications," American Academy of Sleep Medicine, Westchester, IL, 2007.

[25] P. Gouveia, R. Oliveira, and R. Rosa, "Sleep Apnea related micro arousal detection with EEG analysis," in *7th Portuguese Conference on Biomedical Engineering*, 2003.

[26] SP. Cho, J. Lee, HD. Park, and KJ. Lee, "Detection of arousal in patients with respiratory sleep disorders using a single channel EEG," in *27th IEEE Engineering in Medicine and Biology Annual Conference*, 2005, pp. 2733-2735.

[27] A. Agarwal, "Automatic detection of micro-arousals," in *27th IEEE Engineering in Medicine and Biology Annual Conference* , China, 2005, pp. 1158-1161.

[28] T. Sugi, F. Kawana, and M. Nakamura, "Automatic EEG arousal detection for sleep apnea syndrome," *Biomedical Signal Processing and Control*, vol. 4, no. 4, pp. 329-337, 2009.

[29] O. Shmiel, T. Shmiel, Y. Dagan, and M. Teicher, "Data mining techniques for detection of sleep arousals," *Journal of Neuroscience Methods*, vol. 179, no. 2, pp. 331-337, 2009.

[30] F. De Carli, L. Nobili, P. Gelcich, and F. Ferrillo, "A method for the automatic detection of arousals during sleep," *Sleep*, vol. 22, no. 5, pp. 561-572, 1999.

[31] MH. Bonnet et al., "The scoring of arousal in sleep: reliability, validity, and alternatives," *Journal of Clinical Sleep Medicine*, vol. 3, no. 2, pp. 133-145, 2007.

[32] H. Rauscher, W. Popp, and H. Zwick, "Computerized detection of respiratory events during sleep from rapid increases in oxyhemoglobin saturation," *Lung*, vol. 169, no. 1, pp. 335-342, 1991.

[33] C. George, T. Millar, and M. Kryger, "Identification and quantification of apneas by computer-based analysis of oxygen saturation," *The American Review of Respiratory Diseases*, vol. 137, pp. 1238-1240, 1988.

[34] BH. Taha et al., "Automated detection and classification of sleep-disordered breathing from conventional polysomnography data," *Sleep*, vol. 20, no. 11, pp. 991-1001, 1997.

[35] M. Moser, B. Phillips, D. Berry, and L. Harbison, "What is hypopnea, anyway?," *Chest*, vol. 105, pp. 426-28, 1994.

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

[37] K. Polat, S. Yosunkaya, and S. Günes, "Comparison of different classifier algorithms on the automated detection of obstructive sleep apnea syndrome," *Journal of Medical Systems*, vol. 32, pp. 243-250, 2008.

[38] PR. Westbrook et al., "Description and validation of the Apnea Risk Evaluation System," *Chest*, vol. 128, no. 4, pp. 2166-2175, 2005.

[39] O. Amir et al., "An automated sleep-analysis system operated through a standard hospital monitor," *Journal of Clinical Sleep Medicine*, vol. 6, no. 1, pp. 59-63, 2010.

[40] A. Otero, P. Félix, and MR. Álvarez, "Algorithms for the analysis of polysomnographic recordings with customizable criteria," *Expert Systems with Applications*, vol. 38, pp. 10133-10146, 2011.

[41] TI. Morgenthaler, V. Kagramanov, V. Hanak, and PA. Decker, "Complex Sleep Apnea Syndrome: Is it a unique clinical syndrome?," *Sleep*, vol. 29, no. 9, pp. 1203-1209, 2006.

[42] S. Redline et al., "The scoring of respiratory events in sleep: reliability and validity," *Journal of Clinical Sleep Medicine*, vol. 3, no. 2, pp. 169-200, 2007.

[43] D. Bliwise, NG. Bliwise, HC. Kraemer, and W. Dement, "Measurement error in visually scored electrophysiological data: respiration during sleep," *Journal of Neuroscience Methods*, vol. 12, pp. 49-56, 1984.

[44] DJ. Levendowski et al., "Assessment of the test-retest reliability of laboratoy polysomnography," *Sleep Breath*, vol. 13, pp. 163-167, 2009.

[45] DJ. Pitson and JR. Stradling, "Autonomic markers of arousal during sleep in patients undergoing investigation for obstructive sleep apnoea, their relationship to EEG arousals, respiratory events and subjective sleepiness," *Journal of Sleep Research*, vol. 7, pp. 53-59, 1998.

[46] R. Wietske et al., "The role of sleep position in obstructive sleep apnea syndrome," *Eur Arch Otorhinolaryngol*, vol. 263, pp. 946-950, 2006.

[47] JS. Loredo, JL. Clausen, S. Ancoli-Israel, and JE. Dimsdale, "Night-to-night arousal variability and interscorer reliability of arousal measurements," *Sleep*, vol. 22, no. 7, pp. 916-920, 1999.

[48] MV. Smurra, M. Dury, G. Aubert, DO. Rodenstein, and G. Liistro, "Sleep fragmentation: comparison of two different definitions of short arousals during sleep in OSAS patients," *European Respiratory Journal*, vol. 17, pp. 723-727, 2001.

[49] MJ. Drinnan, A. Murray, CJ. Griffiths, and GJ. Gibson, "Interobserver variability in recognizing arousal in respiratory sleep disorders," *American Journal of Respiratory and Clinical Care Medicine* , vol. 158, no. 2, pp. 358-362, 1998.

[50] RJ. Thomas, "Arousals in sleep-disordered breathing: patterns and implications," *Sleep*, vol. 26, no. 8, pp. 1042-1047, 2003.

[51] R.D. Cartwright, F. Diaz, and S. Lloyd, "The effects of sleep posture and sleep stage on apnea frequency," *Sleep*, vol. 14, no. 4, pp. 351-353, 1991

.

[52] B. Kemp, AH. Zwinderman, B. Tuk, HAC. Kamphuisen, and JJL. Oberyè, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185-1194, 2000.

[53] B. Kemp, "A proposal for computer-based sleep/wake analysis," *Journal of Sleep Research*, vol. 2, pp. 179-185, 1993.

[54] Á. Fernández-Leal, V. Moret-Bonillo, and E. Mosqueira-Rey, "Causal temporal constraint networks for representing temporal knowledge," *Expert Systems with Applications*, vol. 36, no. 1, pp. 27-42, 2009.

[55] A. Fernández-Leal, Aspectos Temporales de la representación de conocimiento en el síndrome de apneas del sueño, 2006, Dissertation for the degree of Doctor of Philosophy in Information and Computer Science, University of A Coruña.

[56] I. Meiri, "Combining qualitative and quantitative constraints in temporal reasoning," *Artificial Intelligence*, vol. 87, no. 1, pp. 295-342, 1996.

[57] A. Fernández-Leal and V. Moret-Bonillo, "A framework for handling temporal knowledge in clinical diagnosis problems," in *ESBME*, 2006.

[58] JM. Mendel, "Type-2 fuzzy sets and systems: an overview," *IEEE Computational Intelligence Magazine*, pp. 20-29, 2007.

[59] JM. Mendel, RI. John, and F. Liu, "Interval type-2 fuzzy logic systems made simple," *IEEE Transactions on Fuzzy Systems*, vol. 14, no. 6, pp. 808-821, 2006.

[60] Laboratory for Research and Development of Artificial Intelligence. (2012) MIASOFT: Intelligent Monitoring of Sleep Apnea/Hypopnea Syndrome. [Online]. http://www.lidiagroup.org/miasoft

A. COST FUNCTIONS FOR THE MODELING OF NEURO-FUZZY SYSTEMS

A.1. Mean Squared Error (MSE)

MSE is a risk function corresponding to the expected value of the squared error loss or quadratic loss. MSE measures the average of the squares of the errors. It is the second moment (about the origin) of the error, and thus incorporates both the variance of the estimator and its bias.

Formally the MSE of an estimator $\hat{\theta}$ with respect to the estimated parameter θ is defined as:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$
$$= Var(\hat{\theta}) + (Bias(\hat{\theta}, \theta))^2$$

Like variance, mean squared error has the disadvantage of heavily weighting outliers. This is a result of the squaring of each term, which effectively weights large errors more heavily than small ones. This property, together with the assumption of gaussianity, undesirable in many applications, has led researchers to investigate the use alternative cost functions.

A.2. Minimum Error Entropy (MEE)

Entropy, which was introduced by Shannon [1], is a scalar quantity that provides a measure for the average information contained in a given probability function. As by definition, information is a function of the Probability Density Function (PDF), entropy as an optimally criterion, extends MSE since when entropy is minimized, all the moments of the error PDF (not only the second moments) are constrained.

Let X to be a random variable with probably distribution function f_x, then the Renyi's quadratic entropy [2] is defined as

$$H_2(X) = -\log \int f_x^2(x)dx\,. \tag{A.1}$$

Given a set of data points $\{x_i\}_{i=1}^N$ drawn from X, the Parzen window estimate of the PDF [3] is

$$\hat{f}_{x;\sigma}(x) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x - x_i) \tag{A.2}$$

where $\kappa_\sigma(x - x_i)$ is the Gaussian kernel

$$\kappa_\sigma(x - x_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x - x_i)^2}{2\sigma^2}) \tag{A.3}$$

and N is the number of data points and σ the kernel size.

The kernel size or bandwidth (σ) is a free parameter that must be chosen by the user using concepts of density estimation, such as Silverman's rule [4] or maximum likelihood. It has been experimentally verified that the kernel size affects much less the performance of ITL algorithms than density estimation [5], but a thorough treatment of this issue is beyond the scope of this section.

Substituting (A.3) into (A.2) and after some mathematical manipulations [6], an estimator for (A.1) is obtained:

$$\hat{H}_2(X) = -\log IP(X) \tag{A.5}$$

$$IP(X) = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\kappa_{\sqrt{2}\sigma}(x_j - x_i)\,. \tag{A.6}$$

IP(X) stands for Information Potential (IP), and represents average interactions among the data samples. Now, consider the error between the desired and the actual

outputs of the system $e = d - y$. It can be demonstrated that $H_2(e=0)$ is a global minimum and that the nonparametric estimator $\hat{H}_2$ preserves this property [6]. Thus minimization of Renyi's quadratic entropy of the error as a criterion for parameter optimization is feasible, and it was demonstrated to be effective on the training of several types of systems, including traditional adaptive filters, neural networks, and various algorithms in machine learning [7].

Using (A.1), (A.5) and (A.6), minimizing Renyi's quadratic entropy of the error becomes finding ξ such that

$$\begin{aligned} \xi &= \min_w \hat{H}_2(E) \\ &= \min_w(-\log \int \hat{f}^2(e)de) \\ &= \max_w \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\kappa_{\sqrt{2}\sigma}(e_j - e_i) \end{aligned} \tag{A.7}$$

since Renyi's quadratic entropy is a monotonic negative function of the information potential. This results in the so-called Minimum Error Entropy (MEE) criterion.

A.3. Maximum Correntropy Criterion (MCC)

Correntropy has been proposed from the ITL paradigm which incorporates both information of the distribution and time structure of the data [8] [9]. It has been shown that Correntropy involves not only the second-order moments about the data, but all the even-order moments [8]. These nice properties suggest the use of this measure as a new interesting cost function in the field of machine learning. It has been successfully applied to problems like robust regression [9], adaptive filtering [10], pitch detection in speech [11], non-linearity tests [12] or the detection of non-linear similarity in the EEG [13].

Let $\{\mathbf{x}_t, t \in \mathrm{T}\}$ to be a random process with T being an index set and $\{\mathbf{x}_t \in \Re^d\}$, then the generalized correlation function –correntropy- $V(t_1,t_2)$ is defined as a function from T x T into $\Re^+$ as

$$V(t_1,t_2) = \mathrm{E}[\kappa_\sigma(\mathbf{x}_{t_1} - \mathbf{x}_{t_2})], \tag{A.8}$$

where E[.] denotes mathematical expectation over the random process $\mathbf{x}_t$ and $\kappa_\sigma(.)$ is a positive definite kernel of bandwidth σ. In this paper we will assume $\kappa_\sigma(.)$ to be the Gaussian kernel

$$\kappa_\sigma(x-x_i)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-x_i)^2}{2\sigma^2}). \quad \text{(A.9 2)}$$

Using a series expansion for the Gaussian kernel, (1) can be expressed as

$$V(t_1,t_2)=\frac{1}{\sqrt{2\pi}\sigma}\sum_{n=0}^{\infty}\frac{(-1)^n}{2^n\sigma^{2n}n!}\mathrm{E}[\left\|\mathbf{x}_{t_1}-\mathbf{x}_{t_2}\right\|^{2n}] \quad \text{(A.10)}$$

and therefore one can see that it involves all the even-order moments of the difference $\mathbf{x}_{t_1}-\mathbf{x}_{t_2}$. If we take the term corresponding to *n*= 1 in (A.10), expanding the square in the expectation operator we get

$$\mathrm{E}[\left\|\mathbf{x}_{t_1}\right\|^2]+\mathrm{E}[\left\|\mathbf{x}_{t_1}\right\|^2]-2\mathrm{E}[\left\langle\mathbf{x}_{t_1},\mathbf{x}_{t_2}\right\rangle]=\sigma_{x_{t_1}}+\sigma_{x_{t_2}}-2R_x(t_1,t_2) \quad \text{(A.11)}$$

where $R_x(t_1,t_2)$ is the covariance function of the random process; thus it can be show that (A.8) includes the information provided by the conventional covariance function. It can also be demonstrated that (A.8) carries information about the quadratic Renyi's entropy of the input data [8]. This dual relation with the correlation measure and entropy brings to this measure the so-called name of Correntropy.

Now let's take a set of *N* learning patterns $\{(x_i,d_i)\}_{i=1}^N$ to be respectively the input and the desired output for a concrete problem, and let $\{y_i\}_{i=1}^N$ to be the corresponding system's classification output, where $x_i\in X$, $y_i\in Y$ and $d_i\in D$. We are thus interested in minimizing the error $e_i = y_i - d_i$ on the classification for the pattern x_i, $\forall x_i\in X$. Consider then

$$V(Y,D)=\mathrm{E}[\kappa_\sigma(Y-D)]. \quad \text{(A.12)}$$

Effectively maximizing (A.11) implies minimizing the difference between the random variables *Y* and *D* in terms of all their even-moments, thus reducing the error. Also in terms of information criteria, maximizing (A.11) turns out in reducing the entropy on the probability density function of the error. When applied to optimization problems –as in our case-, this leads to the Maximum Correntropy Criterion (MCC). In practice we work with a finite number of realizations of $\{(y_i, d_i)\}_{i=1}^{N}$ and therefore we use an estimator of correntropy

$$\hat{V}_{N,\sigma}(Y,D) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(y_i - d_i) \tag{A. 13}$$

yielding to be the resulting cost function:

$$J(E = Y - D) = \frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(e_i)\,. \tag{A. 14}$$

As for the case of IP, the kernel size σ is a free parameter that must be chosen by the user using concepts of density estimation, such as Silverman's rule [4]. In terms of the contribution to the cost function, when comparing the MSE in the space of errors, it has been shown [9] that Correntropy induces a new metric which is equivalent to the 2-norm distance if the data points are close to *e=0*, behaves similarly to the 1-norm as points get further and eventually approaches the zero-norm as they are far apart. On the other hand, in the case of MSE, all the samples in the error space contribute appreciably to the value of the cost function. This locality allows MCC to be less sensitive to outliers and more effective in cases where the distribution of the errors is far from gaussianity. In this context, the selection of σ acts as a compromise between estimation efficiency and sensitivity to outlier rejection.

A.4. References

[1] CE. Shannon, "A mathematical theory of communications," *Bell System Technical Journal*, vol. 27, pp. 379-423, 1948.

[2] A. Renyi, "Some fundamental questions of information theory," in *Selected papers of Alfred Renyi*. Budapest: Akademic Kiado, 1976.

[3] E. Parzen, "On the estimation of a probability density function and mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065-1076, 1962.

[4] BW. Silverman, *Density estimation for statistics and data analysis*. London, UK: Chapman and Hall, 1986.

[5] D. Erdogmus and JC. Principe, "From linear adaptive filtering to nonlinear information processing," *IEEE Signal Processing Magazine*, 2006.

[6] D. Erdogmus and JC. Principe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Transactions of Signal Processing*, vol. 50, no. 7, pp. 1780-1786, 2002.

[7] JC. Príncipe, *Information Theoretic Learning*.: Springer, 2010.

[8] I. Santamaría, PP. Pokharel, and JC. Príncipe, "Generalized Correlation Function: definition, properties and applications to blind equalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2187-2197, 2006.

[9] W. Liu, PP. Puskal, and JC. Príncipe, "Correntropy: properties and applications in non-gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286-5298, 2007.

[10] A. Singh and JC. Principe, "Using correntropy as a cost function in linear adaptive filters," in *International Joint Conference on Neural Networks*, Atlanta, USA, 2009.

[11] J. Xu and JC. Principe, "A pitch detector based on a generalized correlation function," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1420-1432, 2008.

[12] A. Gunduz, A. Hegde, and JC. Principe, "Correntropy as a novel measure for nonlinearity test," in *International Joint Conference on Neural Networks*, Vancouver, BC, Canada, 2006, pp. 1856-1862.

[13] J. Xu, H. Bakardjina, A. Cichocki, and JC. Principe, "A new nonlinear similarity measure for multichannel signals," *Neural Networks*, vol. 21, pp. 222-231, 2008.

B. RELEVANT PUBLICATIONS OF THE AUTHOR RELATED TO THE DOCTORAL THESIS

(Chronological order)

JCR Journals

[1] D. Álvarez-Estévez and V. Moret-Bonillo. Fuzzy Reasoning Used to Detect Apneic Events in the Sleep Apnea-Hypopnea Syndrome. Expert Systems with Applications, 36(4), pp. 7778-7785, 2009

[2] D. Álvarez-Estévez and V. Moret-Bonillo. Identification of Electroencephalographic Arousals in Multi-channel Sleep Recordings. IEEE Transactions on Biomedical Engineering, 58(1), pp. 54-63, 2011

[3] D. Álvarez-Estévez, N. Sánchez-Maroño, A. Alonso-Betanzos and V. Moret-Bonillo. Reducing Dimensionality in a Database of Sleep EEG Arousals. Expert Systems with Applications, 38(6), pp. 7746-7754, 2011

[4] D. Álvarez-Estévez, José M. Fernández-Pastoriza, E. Hernández-Pereira and V. Moret-Bonillo. A Method for the Automatic Analysis of the Sleep Macrostructure in Continuum. Expert Systems with Applications (IN PRESS), 2012

International Conferences

[1] D. Álvarez-Estévez and V. Moret-Bonillo. Dealing with Imprecision in the Sleep Apnea-Hypopnea Syndrome. Proceedings in the 20th IASTED International Conference on Artificial Intelligence and Soft Computing, ASC 2008, pp. 61-66, Palma de Mallorca, Spain, September 2008

[2] D. Álvarez-Estévez and V. Moret-Bonillo. Computer-Assisted Decision Support in the Sleep Apnea-Hypopnea Syndrome. Proceedings of the 4th European Conference of the International Federation for Medical and Biological Engineering (IFBME), 22, pp. 1121-1124, Antwerp, Belgium, November 2008

[3] D. Álvarez-Estévez and V. Moret-Bonillo. Model Comparison for the Detection of EEG Arousals in Sleep Apnea Patients. Lecture Notes in Computer Science 5517, pp. 997-1004, 2009. Proceedings of the 10th International Work-Conference on Artificial Neural Networks (IWANN 2009), Salamanca, Spain, June 2009

[4] D. Álvarez-Estévez, J.M. Fernández-Pastoriza and V. Moret-Bonillo. A Continuous Evaluation of the Awake Sleep State using Fuzzy Reasoning. Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5539-5542, Minneapolis, USA, September 2009

[5] D. Álvarez-Estévez, N. Sánchez-Maroño, A. Alonso-Betanzos and V. Moret-Bonillo. Filter-based Feature Selection for the Detection of Arousals in Sleep Studies. Proceedings of the XIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA), pp. 41-50, Sevilla, Spain, November 2009

[6] D. Álvarez-Estévez, J. C. Príncipe and V. Moret-Bonillo. Neuro-Fuzzy Classification using the Correntropy Criterion: application to Sleep Depth Estimation. Proceedings of the 2010 International Conference on Artificial Intelligence, ICAI'10, pp. 9-15, Las Vegas, NV, USA, July 2010

[7] D. Álvarez-Estévez, J. C. Príncipe and V. Moret-Bonillo. Information Theoretic Fuzzy Modeling for Regression. Proceedings of the 2010 IEEE World Congress on Computational Intelligence, FUZZ-IEEE, pp. 1979-1983, Barcelona, Spain, July 2010

[8] D. Álvarez-Estévez, José M. Fernández-Pastoriza, E. Hernández-Pereira and V. Moret-Bonillo. On the Continuous Evaluation of the Macrostructure of Sleep. Frontiers in Artificial Intelligence and Applications, 243, pp. 189-198. Proceedings of the 16th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2012), 2012

[9] A. Fernández-Leal, D. Álvarez-Estévez, José M. Fernández-Pastoriza and V. Moret-Bonillo. A Framework for Handling Fuzzy Temporal Events. Frontiers in Artificial Intelligence and Applications, 243, pp. 179-188. Proceedings of the 16th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2012), 2012