

# A Bayesian Neural Network Approach to Compare the Spectral Information from Nasal Pressure and Thermistor Airflow in the Automatic Sleep Apnea Severity Estimation

Gonzalo C. Gutiérrez-Tobal, *Member, IEEE*, Julio de Frutos, Daniel Álvarez, Fernando Vaquerizo-Villar, *Student Member, IEEE*, Verónica Barroso-García, Andrea Crespo, Félix del Campo, and Roberto Hornero, *Senior Member, IEEE*

**Abstract**— In the sleep apnea-hypopnea syndrome (SAHS) context, airflow signal plays a key role for the simplification of the diagnostic process. It is measured during the standard diagnostic test by the acquisition of two simultaneous sensors: a nasal prong pressure (NPP) and a thermistor (TH). The current study focuses on the comparison of their spectral content to help in the automatic SAHS-severity estimation. The spectral analysis of 315 NPP and corresponding TH recordings is firstly proposed to characterize the conventional band of interest for SAHS (0.025-0.050 Hz.). A magnitude squared coherence analysis is also conducted to quantify possible differences in the frequency components of airflow from both sensors. Then, a feature selection stage is implemented to assess the relevance and redundancy of the information extracted from the spectrum of NPP and TH airflow. Finally, a multiclass Bayesian multi-layer perceptron (BY-MLP) was used to perform an automatic estimation of SAHS severity (no-SAHS, mild, moderate, and severe), by the use of the selected spectral features from: airflow NPP alone, airflow TH alone, and both sensors jointly. The highest diagnostic performance was reached by BY-MLP only trained with NPP spectral features, reaching Cohen's  $\kappa = 0.498$  in the overall four-class classification task. It also achieved 91.3%, 84.9%, and 83.3% of accuracy in the binary evaluation of the 3 apnea-hypopnea index cut-offs (5, 15, and 30 events/hour) that define the four SAHS degrees. Our results suggest that TH sensor might be not necessary for SAHS severity estimation if an automatic comprehensive characterization approach is adopted to simplify the diagnostic process.

## I. INTRODUCTION

The sleep apnea-hypopnea syndrome (SAHS) is a prevalent disease characterized by recurrent events of complete absence (apneas) and/or significant reduction (hypopneas) of airflow during sleep [1]. Apneas and hypopneas cause overnight inadequate gas exchange that derives into restless sleep and triggers inflammatory processes [1, 2]. These imply both life's quality worsening and health deterioration of affected people, leading them to cognitive impairment and depression, as well as to a higher

chance for cardiovascular disease suffering [2, 3]. Furthermore, SAHS has been recently associated with a higher cancer incidence [4].

Nocturnal polysomnography (PSG) is the overnight standard test to diagnose SAHS and establish its severity [1]. During PSG, multiple physiological signals, such as electrocardiogram, electroencephalogram, electrooculogram, blood oxygen saturation, and airflow are monitored and recorded [1]. This technical complexity, which requires a hospital night stay for patients, derives into high costs and consumed time [5]. Therefore, simplified alternatives have been previously evaluated to avoid PSG drawbacks.

Airflow signal has been commonly studied for this purpose since it is involved in both apnea and hypopnea definitions [6]. The American Academy of Sleep Medicine (AASM) recommends the use of two sensors to record airflow during PSG: nasal prong pressure (NPP) and thermistor (TH) [6]. The former has a quadratic response to the measured signal that may derive into overestimation of apnea events when airflow decreases [7]. Conversely, TH signal is an indirect measure of airflow based on the different temperature of inhaled and exhaled air, which is not sensible enough to small airflow changes, as in hypopneas [7]. These limitations cause that physicians use TH to score apneas and NPP to score hypopneas [6]. In spite of these limitations, multiple studies have shown the usefulness of single-channel airflow to detect SAHS; regardless it has been acquired from TH or NPP [8-11]. Many of these studies focus on the automatic detection of SAHS, and its severity, based on a comprehensive characterization of airflow [8, 9]. To the best of our knowledge, however, none of these has directly compared the performance of airflow from each sensor by the use of NPP and TH recordings acquired at the same time.

In this preliminary study, we hypothesize that two airflow channels are not necessary to help in SAHS severity estimation when a comprehensive signal characterization approach is conducted. Hence, we firstly propose a spectral analysis of NPP and TH recordings simultaneously acquired from 315 subjects. This analysis includes the computation of the power spectral density (PSD) from NPP and TH [12], the magnitude squared coherence (MSC) estimation between each pair of airflow recordings from the same subject [13], as well as the characterization of the SAHS conventional spectral band of interest by means of a feature extraction stage. Then, a feature selection phase is implemented in order to evaluate the relevance and redundancy of the features extracted from the PSDs of each airflow sensor. The fast correlation-based filter (FCBF) selection method is used for this purpose [14]. Finally, we use a Bayesian multi-layer

This research was supported by the projects 158/2015 of "Sociedad Española de Neumología y Cirugía Torácica", TEC2014-53196-R of "Ministerio de Economía y Competitividad (MINECO)" and FEDER, and VA037U16 of "Consejería de Educación de la Junta de Castilla y León". F. Vaquerizo-Villar is granted with the project PEJ-2014-P-00349 from MINECO and the University of Valladolid.

G. C. Gutiérrez-Tobal, V. Barroso-García, F. Vaquerizo-Villar, and R. Hornero, are with the Biomedical Engineering Group, Universidad de Valladolid, Spain (e-mail: gonzalo.gutierrez@gib.tel.uva.es).

J. de Frutos, D. Álvarez, Andrea Crespo, and F. del Campo are with the Hospital Universitario Río Hortega of Valladolid, Spain (e-mail: fsas@telefonica.net).

perceptron artificial neural network (BY-MLP) approach to evaluate the spectral information from NPP, TH, and both sensors simultaneously when classifying each subject under study into one out of the four SAHS severity degrees: healthy (no-SAHS), mild, moderate, and severe [3].

## II. SUBJECTS AND SIGNALS

In this study, 315 subjects were involved. All of them were derived to the sleep unit of the Hospital Universitario Rio Hortega of Valladolid (Spain) due to suspicion of SAHS. They underwent an overnight conventional PSG test (E-series, Compumedics) to reach a diagnosis. According to the AASM recommendations, the apnea-hypopnea index (apneas and hypopneas per hour of sleep, AHI) was used to establish the presence and severity of SAHS [6]. Thus, an apnea was scored when a 10 seconds or more decrease of at least 90% was detected in TH airflow amplitude [6]. Similarly, a hypopnea was scored when a 10 seconds or more decrease of at least 30% was detected in NPP amplitude, and it was accompanied of a 3% or more decrease in blood oxygen saturation or an arousal [6]. According to these rules, the physicians diagnosed 39 no-SAHS subjects ( $AHI < 5$  events/hour), 91 mild-SAHS subjects ( $5 \leq AHI < 15$  e/h), 69 moderate-SAHS subjects ( $15 \leq AHI < 30$  e/h), and 116 severe-SAHS subjects ( $AHI \geq 30$  e/h). Table I shows the demographic and clinical data of the population under study. The whole dataset was randomly divided into a training (60%) and a test set (40%). The former was used to conduct exploratory analyses as well as to train the BY-MLP models, whereas the latter was only used for validation purposes. The Ethics Committee of the Hospital Universitario Rio Hortega approved the protocol and all the participants gave their informed consent. NPP and TH airflow recordings were simultaneously obtained during the PSG of each subject. Both signals were acquired at a sample rate of 128 Hz.

## III. METHODS

### A. Spectral Analysis

1) *Power Spectral Density and Feature Extraction.* The analysis in the frequency domain is justified by the recurrent nature of apneic events. In this regard, PSD has already shown its usefulness to obtain SAHS-related information from airflow signal [8-10]. Hence, PSDs were obtained from all NPP and TH recordings. Welch's periodogram (Hamming window of  $2^{15}$  points, 50% overlap, and discrete Fourier transform of  $2^{16}$  points) was used to estimate PSD since it is suitable to deal with non-stationary signals [12].

Six spectral features were extracted from each NPP and TH recording, i.e., 12 features were finally obtained from each patient. All of these features were acquired from the conventional SAHS spectral band of interest in airflow signals, 0.025-0.050 Hz [8, 9], which corresponds to the most common range of apneic event duration (20 to 40 seconds) [1]. Thus, maximum amplitude ( $MA$ ), minimum amplitude ( $mA$ ), mean ( $Mf1$ ), standard deviation ( $Mf2$ ), skewness ( $Mf3$ ), and kurtosis ( $Mf4$ ) were computed from this band in each NPP and TH PSDs.

2) *Magnitude Squared Coherence.* MSC is a measure of correlation between each of the individual frequency components,  $f$ , of two signals [13]. In this work, we use it to study differences between the spectrums of each pair of NPP

TABLE I. DEMOGRAPHIC AND CLINICAL DATA OF THE SUBJECTS UNDER STUDY.

Data	All	Training	Test
Subjects (n)	315	189	126
Males (%)	71.4	73.0	69.0
Age (years)	$49.9 \pm 12.0$	$49.2 \pm 12.2$	$50.9 \pm 11.8$
BMI ( $\text{kg}/\text{m}^2$ )	$25.5 \pm 9.5$	$25.3 \pm 10.1$	$25.9 \pm 8.6$

BMI: Body Mass Index

and TH recordings acquired from the same patients. MSC is computed as [13]:

$$MSC(f) = \frac{|P_{x,y}(f)|^2}{P_{x,x}(f) \cdot P_{y,y}(f)}, \quad (1)$$

where  $P_{x,x}$  and  $P_{y,y}$  are the corresponding PSDs of each signal (NPP and TH) and  $P_{x,y}$  is the cross PSD between them.

### B. Feature selection

The FCBF automatic selection algorithm was implemented to evaluate relevance and redundancy of the features obtained from NPP and TH [14]. This method computes symmetrical uncertainty (a normalization of mutual information) between each feature and the AHI taken as a reference variable ( $SU_{F,AHI}$ ).  $SU$  is used to perform a relevance ranking by classifying higher those features with values closer to 1. Redundancy is established by computing  $SU$  between each pair of features ( $SU_{i,j}$ ), where feature  $i$  is ranked higher than feature  $j$ . Then, if  $SU_{i,j} \geq SU_{j,AHI}$ , feature  $j$  is considered redundant and discarded for subsequent analysis [14].

### C. Bayesian multi-layer perceptron

Multi-layer perceptron (MLP) artificial neural networks were used to automatically classify each subject into one out of the four SAHS-severity degrees. Vectors formed by the previously selected spectral features were used as patterns to characterize each subject and to train the networks. MLP architecture was arranged in three layers of neurons (input layer, hidden layer, and output layer), since this configuration is known to be able to provide universal approximations [15]. Logistic activation functions were used for the neurons as a common choice for classification tasks [15]. The Bayesian approach was used to estimate the weights connecting neurons each other. This method (BY-MLP) has been reported to provide higher performance in SAHS context than the maximum likelihood optimization approach [16].

### D. Statistical analysis

The extracted features did not pass the Lilliefors normality test. Hence, the non-parametric Spearman's  $\rho$  coefficient was used to evaluate the correlation between the features from NPP and TH with the AHI, the apnea index (AI), and the hypopnea index (HI). Cohen's  $\kappa$  was used to assess the diagnostic performance of BY-MLP in the four-class classification task. Sensibility (Se), specificity (Sp), and accuracy (Acc) were used to perform a binary evaluation of BY-MLP in each of the AHI cut-offs that limit the SAHS-severity degrees (5, 15, and 30 events/hour).

## IV. RESULTS

### A. Power Spectral Density and Magnitude Squared Coherence plots

Figures 1 a) and b) show the normalized PSDs (PSDn) of NPP and TH airflow averaged by SAHS severity group (training set). As expected, spectrum of airflow from both sensors is concentrated around the normal respiratory frequency at rest (15 breaths per minute, i.e., 0.250 Hz.) [17]. Consistent with previous studies [10], the PSD amplitudes of the band of interest increase with SAHS severity for both NPP and TH. In Figure 1 c), MSC quantifies these qualitative findings. High correlation is found ( $>0.8$ ) around the normal respiratory rate. However, low correlation values are reached ( $<0.3$ ) when comparing the spectral content of the frequencies from the two sensors in the band of interest.

### B. Spectral features relationship with apneic events

Table II shows Spearman's  $\rho$  coefficient of the spectral features from NPP and TH airflow with AHI, AI, and HI (training set). As can be observed, all the spectral features from NPP airflow reached higher  $\rho$  with AHI than the corresponding features from TH. Despite the AASM recommendation of using NPP to score hypopneas and TH to score apneas, all the spectral features showed higher correlation with AI than with HI, regardless the sensor they were acquired from. Furthermore, NPP spectral content showed higher correlations with apneas than TH. Nonetheless, features from both sensors reached significant correlations with apneas and hypopneas.

### C. Feature selection: relevance and redundancy

The FCBF automatic selection method was applied to (training set): *i*) the 6 NPP spectral features; *ii*) the 6 TH spectral features; and *iii*) the 12 NPP and TH features. Two features were considered relevant and non-redundant in the case of NPP ( $Mf1$  and  $Mf4$ ) and TH ( $Mf2$  and  $Mf4$ ). By contrast, 3 features were selected when applying FCBF to the 12 spectral features ( $Mf1_{NPP}$ ,  $Mf2_{TH}$ , and  $Mf4_{NPP}$ ).

### D. Sleep Apnea Severity Classification

The three subsets of features selected in the previous stage (NPP, TH, and features from both sensors) were used to train BY-MLP networks with ability to assign the subjects into the SAHS severity degrees. Table III summarizes the diagnostic performance of BY-MLP (test set) when trained with these selected features from NPP (BY-MLP<sub>NPP</sub>), TH (BY-MLP<sub>TH</sub>), and both sensors (BY-MLP<sub>TH</sub>). According to the number of features and instances used for training the models, 10 neurons in the hidden layer were fixed in the three cases as reasonable choices for the sake of the bias-variance trade-off [15]. Since BY-MLP training depends on random initializations of the weights connecting the neurons, the results are reported as the median values of the performance in the test set after 100 training processes (with 95% confident interval). Thus, it can be observed that BY-MLP<sub>NPP</sub> substantially outperformed BY-MLP<sub>NPP-TH</sub> and BY-MLP<sub>TH</sub> in the overall 4-class classification task ( $\kappa = 0.498$ ), as well as in the binary assessment for the AHI cut-offs = 15 e/h (84.9 % Acc) and 30 e/h (83.3 % Acc). Additionally, only BY-MLP<sub>NPP-TH</sub> slightly outperformed BY-MLP<sub>NPP</sub> in the case of the AHI cut-off = 5 e/h (92.9% vs. 91.3%).

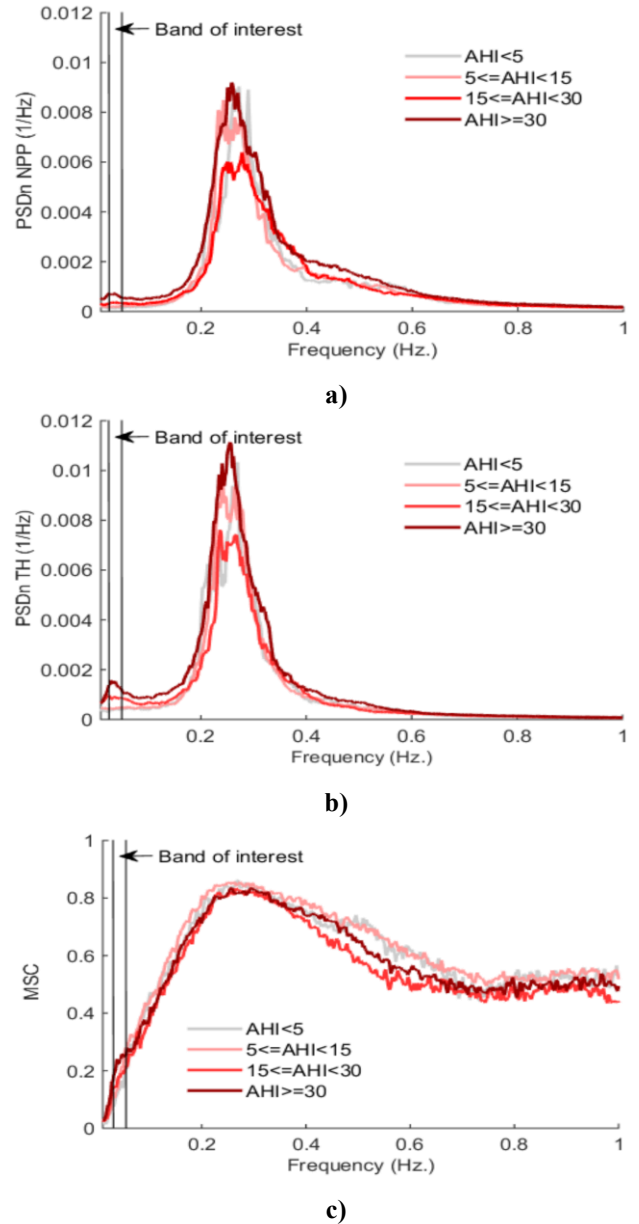


Figure 1. Averaged plots per SAHS-severity degree for a) PSDn of NPP airflow; b) PSDn of TH airflow and c) MSC comparing frequencial components of NPP and TH.

TABLE II. SPEARMAN'S COEFFICIENT BETWEEN THE SPECTRAL FEATURES AND EACH OF AHI, AI, AND HI.

Features	AHI	AI	HI
$MA_{NPP}$	<b>0.721</b>	<b>0.742</b>	<b>0.522</b>
$MA_{TH}$	0.588	0.580	0.429
$mA_{NPP}$	<b>0.687</b>	<b>0.713</b>	<b>0.486</b>
$mA_{TH}$	0.489	0.510	0.336
$Mf1_{NPP}$	<b>0.711</b>	<b>0.736</b>	<b>0.513</b>
$Mf1_{TH}$	0.556	0.560	0.397
$Mf2_{NPP}$	<b>0.651</b>	<b>0.674</b>	<b>0.464</b>
$Mf2_{TH}$	0.603	0.567	0.448
$Mf3_{NPP}$	<b>0.100</b>	0.081	-0.002
$Mf3_{TH}$	0.082	<b>0.122</b>	<b>0.029</b>
$Mf4_{NPP}$	<b>0.073</b>	<b>0.104</b>	0.017
$Mf4_{TH}$	0.043	0.101	<b>0.053</b>

In bold are those values of Spearman's coefficient higher than the corresponding from the other sensor.

TABLE III. DIAGNOSTIC PERFORMANCE OF BY-MLP TRAINED WITH THE SELECTED FEATURES FROM NPP, TH, AND BOTH SENSORS (NPP AND TH)

Models	Cut-off	Se (%)	Sp (%)	Acc (%)	$\kappa$
BY-MLP <sub>NPP</sub>	5 e/h	96.5 (96.5, 100.0)	41.7 (16.7,58.3)	91.3 (91.3, 92.9)	<b>0.498</b> (0.456,0.513)
	15 e/h	90.9 (89.6, 90.9)	75.5 (67.4, 75.5)	<b>84.9</b> (81.8, 84.9)	
	30 e/h	71.0 (67.7, 71.0)	95.3 (95.3, 95.3)	<b>83.3</b> (81.8, 83.3)	
BY-MLP <sub>TH</sub>	5 e/h	100.0 (100.0, 100.0)	0.0 (0.0, 0.0)	90.5 (90.5, 90.5)	0.325 (0.257,0.352)
	15 e/h	74.0 (67.5, 74.0)	73.5 (71.4, 79.6)	73.8 (71.4, 76.2)	
	30 e/h	69.2 (48.4, 66.1)	90.6 (89.1, 93.8)	77.0 (71.4, 78.6)	
BY-MLP <sub>NPP-TH</sub>	5 e/h	98.3 (96.5, 100.0)	41.7 (25.0,58.3)	<b>92.9</b> (92.1, 94.4)	0.449 (0.405,0.494)
	15 e/h	88.3 (87.0, 90.9)	71.4 (67.4, 75.5)	81.8 (80.2, 84.1)	
	30 e/h	66.1 (58.1, 71.0)	95.3 (93.8, 96.9)	81.0 (77.0, 82.5)	

## V. DISCUSSION AND CONCLUSIONS

In this study, a comparison of the spectral information of airflow obtained from NPP and TH sensors was conducted in the context of SAHS severity estimation. Although the averaged PSDs from both sensors showed a similar behavior, a MSC analysis revealed that the spectral content of NPP and TH is low correlated in the SAHS band of interest. This result pointed out to possible differences in the relationship with apneic events of the spectral information obtained from each sensor, which was subsequently supported by a correlation analysis conducted over the features extracted from the two of them. Thus, although spectral features from both sensors were significantly correlated with apneas and hypopneas, those derived from NPP showed higher correlation values with AHI, AI, and HI. In contrast to AASM recommendation, this suggests that NPP airflow might be able to reflect more accurately not only hypopneas-related information but apneas too.

The FCBF selection algorithm also revealed differences in the relevant and non-redundant features from each sensors, as well as supported some degree of complementarity in the information obtained from them (two features selected from NPP and one from TH). However, the BY-MLP models trained with the selected features from NPP, TH, and both sensors showed that the model only built with features from NPP reached the highest performance in SAHS severity estimation. Actually, the BY-MLP<sub>NPP</sub> reached high overall diagnostic ability ( $\kappa = 0.498$ ), as well as for the three AHI cut-offs binary evaluated, (91.3%, 84.9%, and 83.3% Acc for 5, 15, and 30 e/h, respectively). These results outperform those from a recent study focused on automatic SAHS severity estimation by the use of information from NPP single-channel airflow and AdaBoost as multiclass classification approach ( $\kappa = 0.432$ , 86.5%, 81.0%, and 82.5% Acc, respectively) [9]. However, further optimization of our BY-MLP approach is required (a search for the optimum number of hidden neurons) to be able to conduct a more reliable comparison.

Summarizing, we found differences in the SAHS-related spectral information provided by NPP and TH airflow. In contrast to scored-event recommended protocols, our analysis showed that the spectral information from NPP correlates more than TH's with both apneas and hypopneas. A BY-MLP model only trained with selected features from

NPP reached high diagnostic ability and widely outperformed similar ones trained with TH features, and with NPP and TH features jointly. These results suggest that TH airflow might be not necessary to simplify SAHS-severity estimation when an automatic comprehensive signal characterization approach is conducted.

## REFERENCES

- [1] D. J. Eckert and A. Malhotra, "Pathophysiology of adult obstructive sleep apnea," *Proc. Am. Thorac. Soc.*, vol. 5, pp. 144-153, 2008.
- [2] F. Lopez-Jiménez et al, "Obstructive Sleep Apnea," *Chest*, vol. 133, pp 793-804, 2008.
- [3] T. Young et al, "Epidemiology of Obstructive Sleep Apnea: A Population Health Perspective," *Am. J. Respir. Crit. Care Med.*, vol. 165, pp. 1217-39, 2002.
- [4] F. Campos-Rodriguez et al, "Association between obstructive sleep apnea and cancer incidence in a large multicenter spanish cohort," *Am. J. Respir. Crit. Care Med.*, vol. 187, pp. 99-105, 2013.
- [5] J. A. Bennet and W. J. M. Kinnear, "Sleep on the cheap: the role of overnight oximetry in the diagnosis of sleep apnoea hypopnoea syndrome," *Thorax* vol. 54, pp. 958-959, 1999.
- [6] R. B. Berry et al, "Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events," *J. Clin. Sleep Med.*, vol. 8(5), pp. 597-619, 2012.
- [7] A. BaHammam, "Comparison of nasal prong pressure and thermistor measurements for detecting respiratory events during sleep," *Respiration*, vol. 71(4), pp. 385-90, 2004.
- [8] G. C. Gutiérrez-Tobal et al, "Pattern recognition in airflow recordings to assist in the sleep apnoea-hypopnoea syndrome diagnosis," *Med. Biol. Eng. Comput.*, vol. 51, pp. 1367-80, 2013.
- [9] G. C. Gutiérrez-Tobal et al, "Utility of AdaBoost to detect sleep apnea-hypopnea syndrome from single-channel airflow," *IEEE Trans. Biomed. Eng.*, vol. 63, pp. 636-46, 2016.
- [10] H. Nakano et al, "Automatic Detection of Sleep-disordered breathing from a single-channel airflow record," *Eur. Respir. J.*, vol. 29, pp. 728-36, 2007.
- [11] A. BaHammam et al, "Evaluation of the accuracy of manual and automatic scoring of a single airflow channel in patients with a high probability of obstructive sleep apnea," *Med. Sci. Mon.*, vol. 17, pp. MT13-MT19, 2011.
- [12] P. D. Welch, "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on time Averaging Over Short, Modified Periodograms," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-15, pp. 70-73, 1967.
- [13] B. J. Roach and D. H. Mathalon, "Event-related EEG time-frequency analysis: an overview of measures and an analysis of early gamma band phase locking in schizophrenia," *Schizophrenia bulletin*, vol. 34(5), pp. 907-26, 2008.
- [14] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205-1224, 2004.
- [15] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.
- [16] J. V. Marcos, et al, "The classification of oximetry signals using Bayesian neural networks to assist in the detection of obstructive sleep apnoea syndrome," *Phys. Meas.*, vol. 31(3), pp. 375-94, 2010.
- [17] R. Farré, et al, "Accuracy of thermistors and thermocouples as flow-measuring devices for detecting hypopnoeas," *Eur. Respir. J.*, vol. 11(1), pp. 179-182, 1998.