

# Detection of Severe Obstructive Sleep Apnea through voice analysis

Jordi Solé-Casals<sup>1</sup>, Cristian Munteanu<sup>2,3</sup>, Oriol Capdevila Martín<sup>3</sup>, Ferrán Barbé<sup>4,5</sup>, Joaquín Durán-Cantolla<sup>5,6</sup>, Carlos Queipo<sup>7</sup>, José Amilibia<sup>8</sup>

<sup>1</sup>*Data and Signal Processing Group, University of Vic – Central University of Catalonia, Sagrada Família 7, 08500 Vic, Spain. email: jordi.sole@uvic.cat*

<sup>2</sup>*Universidad de Las Palmas de Gran Canaria, Campus de Tafira, 35017 Las Palmas, Spain*

<sup>3</sup>*Coocelea, S.L., Mestre Garriga 10, 08500 Vic, Spain*

<sup>4</sup>*Respiratory Department IRBLleida, Av. Alcalde Rovira Roure, 80, 25198 Lleida, Spain*

<sup>5</sup>*CIBERES, ISCIII, Sinesio Delgado 20, Madrid, Spain*

<sup>6</sup>*Sleep Unit, Service of Pneumology, Hospital Txagorritxu, Servicio Vasco de Salud-Osakidetza, José Achótegui s/n, Vitoria-Gasteiz, Spain*

<sup>7</sup>*Respiratory Department, Sleep Unit - Hospital Universitario Marqués de Valdecilla, Avda. Valdecilla nº 25, 39008 Santander, Spain*

<sup>8</sup>*Respiratory Department, Sleep Unit - Hospital Universitario de Cruces, Plaza de Cruces, s/n 48903 Barakaldo – Bizkaia, Spain*

**Abstract:** This paper deals with the potential and limitations of using voice and speech processing to detect Obstructive Sleep Apnea (OSA). An extensive body of voice features has been extracted from patients who present various degrees of OSA as well as healthy controls. We analyze the utility of a reduced set of features for detecting OSA. We apply various feature selection and reduction schemes (statistical ranking, Genetic Algorithms, PCA, LDA) and compare various classifiers (Bayesian Classifiers, kNN, Support Vector Machines, neural networks, Adaboost). S-fold crossvalidation performed on 248 subjects shows that in the extreme cases (that is, 127 controls and 121 patients with severe OSA) voice alone is able to discriminate quite well between the presence and absence of OSA. However, this is not the case with mild OSA and healthy snoring patients where voice seems to play a secondary role. We found that the best classification schemes are achieved using a Genetic Algorithm for feature selection/reduction.

**Keywords:** *Obstructive Sleep Apnea, voice processing, Genetic Algorithms, feature reduction.*

# 1 Introduction

Obstructive Sleep Apnea Hypoapnea Syndrome (OSA for short) is a common sleep disorder that manifests itself by daytime sleepiness caused by a cease in breathing occurring repeatedly during sleep, often for a minute or longer and as many as hundreds of times during a single night. OSA is associated with a reduced-caliber upper airway, and repetitive effects of apneas and hypopneas include oxygen desaturation, reductions in intrathoracic pressure, and central nervous system arousals [1]. Diagnosis of the sleep condition is based on the calculation of the apnea–hypopnea index (AHI) which measures the frequency of reductions in airflow associated with upper-airway collapse or narrowing that occurs with the state change from wakefulness to sleep [1]. The gold standard procedure to determine the AHI is polysomnography, however it is a quite costly methodology [2]. No other measure has proven to be superior to AHI in assessing the overall effect of obstructive sleep apnea. Nevertheless, there is no common consensus between laboratories regarding its definition. Other metrics such as the number or frequency of arousals during a night sleep might be considered an equally good indicator of OSA [1]. Thus, seeking alternative methods of diagnosis that are simpler and more cost effective is fully motivated, and in recent years it was advocated that voice may play a central role into detection of OSA syndrome. Preliminary findings on speech disorder in OSA have been reported firstly in [3] employing a rather small sample (39 subjects) and subjective results of acoustic evaluation of voice changes in OSA, followed by a study [4] on a bigger sample (252 patients) giving again only subjective judgement results. An attempt to a more objective evaluation study was given in [5]. To discriminate between OSA patients and controls, the authors apply spectral analysis to vowels, but again the sample taken into account is small (28 subjects). Recently, in **[Error! No s'ha trobat l'origen de la referència.]** and **[Error! No s'ha trobat l'origen de la referència.]** the authors show the importance of using voice as a discriminatory factor for detection of severe sleep apnea employing Gaussian Mixture Models on phrases (in **[Error! No s'ha trobat l'origen de la referència.]**) and on vowels (in **[Error! No s'ha trobat l'origen de la referència.]**). However, the authors recognize the need for a wider training and validation sets. So far, either due to small samples or subjective judgements, it is hard to quantify up to what extent or under what circumstances we might consider voice as a good discrimination measure between OSA and healthy subjects. Recent efforts such as [6] try to model the upper-airway in OSA subjects as compared to controls by employing computational fluid dynamics models, and they conclude that there is a clear tendency to closure of the upper-airway in OSA. As the

upper-way coincides in part with the vocal tract, the thinning of the lumen and tendency to closure experienced in OSA do suggest that there may be an identifiable dysfunction in voice also.

## 2 Method

### 2.1 Subjects

We have 376 subjects that undertook this study, both controls (proven healthy subjects) as well as snoring OSA suspects, mild OSA and severe OSA patients, 123 women and 253 men, with ages comprised between 18 and 82. This cross sectional data has been pooled from several state hospitals in Spain (namely from Vitoria, Lleida, Cruces and Valdecillas). The diagnosis for each patient was confirmed by specialized medical staff through polysomnography (PSG) or through respiratory polygraphy (RP) whenever PSG was not available. For the present study we consider  $AHI \leq 5$  as controls (healthy subjects) and  $AHI \geq 30$  as severe OSA patients, which is in agreement with the recommendations made by the American Academy of Sleep Medicine [9]. For the purpose of clarity, along the present study, we call these subjects *extreme cases*, while in-between we may have mild OSA, or snoring non-OSA patients. Thus, among the total of 376 available cases we extract a group of 127 controls and a group of 121 severe OSA with the following characteristics:

(Table 1)

### 2.2 Voice database

Speech was recorded using an AKG Perception 100 condenser microphone, a Digidesign M-box<sup>®</sup> sound card (Avid Audio), and a sound acquisition software by Pro Tools<sup>®</sup> (Avid Audio). The microphone was held 20 cm away from the subject's mouth, by a technician designated for this task. The audio signal was sampled at 44.1 kHz with 16 bits per sample, and recording was done for two distinct positions for each subject: upright or seated ('A' position) and supine or stretched ('E' position). Before each recording session, during 3 minutes the patient was kept as comfortable as possible in order to induce a relaxation feeling as stress is known to affect voice [10]. The room's ambient was kept quiet, in dim comfortable light and no external noise. Each subject was asked to emit the 5 vowels present in Spanish language that are: /a/, /e/, /i/, /o/, /u/ in a sustained fashion for at least 4 seconds each. Additionally, the patients were asked to utter the following sentence (in Spanish): \De golpe nos quedamos a oscuras\. Between each utterance a silence gap of 2-3s was enforced through the recording

protocol. The reason for using two distinct uttering positions ('A' and 'E') was that as gravity and head position affect differently the vocal tract when seated and when stretched, the sound properties also change [11, 12]. Therefore, we add a second source of information per patient besides the utterance in the more common position (seated). To the best of our knowledge, this is the first attempt to detect OSA through voice analysis that uses this idea. All recordings are done by technicians from the sleep units in the 4 hospitals participating in the study, all technicians being "blind" with respect to the outcome of the experiment.

## **2.3 Voice features**

A total of 253 features per patient were extracted from the utterance of 5 vowels and a sentence in two distinct positions. The rationale behind choosing the following listed features is that most of these measures have been previously employed for detection or characterization of pathological voice. Our working hypothesis is that severe OSA may present abnormalities in the voice production, such as increased nasality, harshness or dullness, which is also in agreement with previous findings (see [3, 4, 5, **Error! No s'ha trobat l'origen de la referència.**]). The features may be grouped as follows.

### *2.3.1 Formant and pitch based*

For each vowel we compute the second formant using the classical algorithm of root finding for the Linear Predictive Coefficient polynomial [13], with a previous octave-jump filtering step. Next, we extract the Mean Frequency (MF), Coefficient of Variation in Frequency (CVF), Jitter Factor (JF), Relative Average Perturbation (RAP), Mean Bandwidth (MBW) and Coefficient of Variation of the Bandwidth (CVBW). Definitions of these measures are given for example in [14, 15]. Voice pitch is extracted for each vowel employing an improved autocorrelation method given in [16]. The postprocessing octave-jump filtering stage and the features extracted from pitch are exactly the same as in the case of the second formant.

### *2.3.2 Time domain analysis*

The time signal (one signal for each vowel and each subject position) yields a set of features that are pitch-synchronous in that we take as a reference signal the pitch extracted in section 2.3.1. The features (see [17] for detailed definitions) are the Mean Intensity/Amplitude (MIA), the Coefficient of Variation of the Intensity/Amplitude, the Shimmer of the signal Intensity (SIA) and a measure of the perturbation in the signal amplitude: Amplitude Perturbation Quotient (APQ).

### *2.3.3 Voice harshness and turbulence analysis*

The first measure employed is related to the content of harmonics present in voice (versus non-harmonics content, denoted as noise) and is commonly designated as Harmonics to Noise Ratio (HNR). To compute HNR we took a well-established frequency method described in [18] among other more basic variants such as [19, 20]. A particularly useful feature as turned-out to be from results obtained (see section 3) is the MHNR: the mean HNR computed at the beginning (approximately the first second) of vowel \a|. Other measures are the Soft Phonation Index (SPI) and the Voice Turbulence Index (VTI). VTI measures the turbulence components caused by incomplete or loose adduction of the vocal folds; SPI evaluates the poorness of high-frequency harmonic components that may be an indication of loosely adducted vocal folds during phonation. In our implementation we compute SPI and VTI according to definitions in [14] but employing the improved algorithm in [18] to calculate the intra-harmonic and inter-harmonic energies present in the voice signal.

### *2.3.4 Linear prediction analysis*

Based on a linear predictions analysis on the voice signal, we extracted the Pitch Amplitude (PA) and Spectral Flatness Ratio (SFR) with methods described in [21]. PA measures the dominant peak of the residual signal auto-correlation function, and SFR quantifies the flatness of the residue signal spectrum.

### *2.3.5 Dynamical systems analysis*

To account for significant nonlinear and non-Gaussian random phenomena present in disordered sustained vowels we employ two features inspired by dynamical system analysis performed on the voice signal. These features were introduced in [22]. The authors apply state-space recurrence analysis to produce an entropy measure  $H_{norm}$ , and Fractal scaling analysis that yields a measure called Detrended Fluctuation Analysis (DFA).

### *2.3.6 LTAS based*

So far, we introduced features computed on sustained vowels. Next, we present features extracted from phrase analysis. The core analysis method of the sentence was Long-Term Average Spectrum (LTAS). In [23, 24] the authors focus on the use of LTAS to quantify voice quality, and therefore we find LTAS as a suitable (and quite simple) method for detecting a decline in voice quality for severe OSA. Based on LTAS we extract the following

features: the Absolute Spectral Slope (SLOPE\_LTAS), statistical measures: spectral centroid (CENTRAL\_LTAS), spectral spread (SPREAD\_LTAS), spectral skewness (SKEWNESS\_LTAS), spectral kurtosis (KURTOSIS\_LTAS). Next, we have the spectral roll-off (ROLLOFF\_LTAS) which, as the SLOPE\_LTAS measure, quantifies the energy decay at higher frequencies. Finally, we have two measures computed on 5 frequency bands of the LTAS: the Spectral Flatness Ratio (SFR1÷5) and Spectral Crest (SC1÷5); the frequencies bands are: 175 – 500 Hz, 500 – 1000 Hz, 1000 – 2000 Hz, 2000 – 3000 Hz, and 3000 – 4000 Hz.

The nomenclature used for the features is as follows: for vowels we have  $[measure]_V[position]_{[vowel]}$  as in, for example, SFR\_VE\_O, while for the phrase we have  $[measure]_F[position]$ , as in, for example, SC2\_ltas\_FA.

## 2.4 Classification problem

In order to quantify the utility of voice in detecting OSA, we focus primarily on the binary classification problem of the extreme groups: the control group and the severe OSA group. If voice were to be considered an important factor in detecting OSA, then it should discriminate well at least the most extreme categories.

### 2.4.1 Classifiers

The discrimination power is measured through experiments we perform with several classifiers.

The first classifier employed was a classical Multi Layer Perceptron (MLP) Neural Network trained with the Back Propagation technique with an adaptive learning rate [26]. As discussed in section 2.5 we will perform a feature input-space reduction to 5 dimensions. We choose a two hidden layer MLP with  $n_i:n_{h1}:n_{h2}:n_o$ , where the number of inputs  $n_i = 5$ , the number of nodes on the first hidden layer  $n_{h1} = 10$ , the number of nodes on the second hidden layer  $n_{h2} = 5$ , and the number of output nodes  $n_o = 2$  (as we have two classes). The activation function (transfer function) is the hyperbolic tangent sigmoid function. The inputs suffered a pre-normalization step, that is: all features values were linearly mapped to  $[-1, +1]$ ; The MLP runs for  $T_{MLP} = 750$  iterations (epochs) found sufficient to achieve a low classification error margin and a good generalization for most of the runs.

Next, we apply a Support Vector Machine [27] classifier which is a powerful kernel-based classification paradigm. We used the simple linear kernel variant SVM (SVM<sub>lin</sub>) that

performs a linear discrimination, and the non-linear kernel variant ( $SVM_{\text{poly}}$ ) which employs a polynomial kernel of degree 3, capable of finding nonlinear decision boundaries between classes.

AdaBoost [28] is a classifier that combines several weak classifiers (in our implementation these weak classifiers are decision trees) to produce a powerful classification scheme with good generalization capabilities. AdaBoost is quite successful in modern face recognition applications [29].

We also employed a k-Nearest Neighbour (KNN) classification strategy [32] where the number of neighbours was taken to be 5.

Finally, we checked the performance of a classical Bayesian Classification (BC) scheme that uses a multivariate Gaussian model for the distribution of each class, assuming independence between features (a diagonal covariance matrix for the model, implying a linear decision boundary) [32, 28].

#### 2.4.2 Crossvalidation

In order to obtain a good estimate of the classifier's performance on a relatively reduced set of patterns, as the one employed in our study, we may first perform a crossvalidation process and then draw suitable conclusions on the mean classification errors obtained. We employ an  $S$ -fold crossvalidation method [28] that consists of dividing the ordered set of patterns into  $S$  contiguous chunks containing approximately the same number of patterns each, and then performing  $S$  training-testing experiments as follows: for each chunk  $i \in \{1, 2, \dots, S\}$  we hold the current chunk for testing the classifier and we perform training on the remaining  $S-1$  chunks, recording the results. We repeat the  $S$  training-testing experiments for a number of  $\tau$  trials, each trial starting with a random permutation of the whole set of patterns. The main result of each training-test experiment is the Correct Classifications Rate ( $CCR$ ) expressed as a percentage. The  $S$ -fold crossvalidation yields a matrix of  $\tau \times S$  of results from each training-testing experiment. We denote the matrix as **CCR**. The process is identical for all classifier but the MLP. It is well-known that neural networks are prone to get stuck in local minima of the error surface as basically they perform a gradient-descent or other similar local optimization with respect to the free parameters (weights, biases) [26]. Therefore, for a given set of training and test patterns it is important to perform several trials with different (usually random) starting points (values for weights and biases) and take into account the best run. In our case, the  $S$ -fold crossvalidation for the MLP performs  $\xi$  runs of the neural network with

randomly taken starting points (random initialization), for each of the  $\tau \times S$  training-testing experiment. After one such experiment we record only the best run in  $\xi$  runs. Thus, the matrix of recorded results **CCR** will still be  $\tau \times S$  dimensional. For all experiments we took  $S = 5$ ,  $\tau = 50$ , and  $\xi = 20$ .

## 2.5 Feature reduction

Due to the high number of features employed in our study, which is 253, and the relatively low number of available subjects (248), in order to avoid the curse of dimensionality [30] (i.e. a uniform and sufficiently dense sampling in such high dimensional spaces, requires a huge number of data/patients), we must reduce the dimensionality of the feature space. We do so using two strategies: feature selection (find a small number of representative features) and feature combination (apply a transformation to the input feature space to produce a reduced output feature space). In all cases we perform a strong reduction from 253 to 5 variables (i.e. a 5-dimensional feature vector).

### 2.5.1 Feature Ranking

The first method used to reduce the dimensionality of the feature space is a selection scheme that first ranks all features according to a statistical test of the discrimination power of each feature. Discrimination refers to the values each feature may take for the two classes involved in the comparison: control group and severe OSA. We observed that most of the features for both classes have a distribution that deviates significantly from the normal distribution and moreover they present outliers (Fig. 1a). Therefore, the test employed should not rely on normality assumptions, and we choose for a nonparametric test that is the two-sample unpaired Wilcoxon test (also known as the Mann-Whitney u-test) [31]. The method ranks the features in the entire set  $\Phi$  of 253 features using the independent evaluation criterion for binary classification. This yields a number  $Z$  for each feature which is the absolute value of the u-statistic. Moreover, we outweigh the  $Z$  values using the following equation:

$$Z_{\text{final}} = Z \cdot (1 - \alpha \rho) \quad (1)$$

where  $\alpha \in [0, 1]$  a parameter of the method and  $\rho$  is the Pearson cross-correlation coefficient between the candidate feature and all previously selected features. We took  $\alpha = 0.9$ , that is we outweigh the significance statistic, meaning that features that are highly correlated with the features already picked are less likely to be included in the output list. Finally, we sort in

decreasing order all features upon  $Z_{\text{final}}$ , taking the 5 features which correspond to the top 5  $Z_{\text{final}}$  values.

### 2.5.2 Genetic Algorithms-based feature selection

Genetic Algorithms (GAs) as part of the wider field of Evolutionary Algorithms (EAs) are population-based, stochastic search and optimization methods inspired by the natural evolution process [33]. The populations consist of a fixed number of potential solutions to the optimization problem, called “chromosomes”. That is:

$$P = \{x_i | i = \overline{1 \dots N}\} \text{ and } x_i = (x_{i1}, \dots, x_{il}), x_{ij} \in [v_{lb_j}, v_{ub_j}] \subset \mathbf{R}, \forall i = \overline{1 \dots N}, \forall j = \overline{1 \dots l} \quad (2)$$

with  $N$  the size of the population  $P$  and  $x_i$  the chromosomes in  $P$  defined (for the present application) as vectors of integer genes  $x_{ij}$ ;  $v_{lb_j}$  and  $v_{ub_j}$  represent the lower and upper bound respectively of the genes’ values. Each chromosome  $x_i$  bears a utility score  $F(x_i)$  called *fitness* in direct relationship with the optimization criterion. It is expected that by repeated application of *selection* of the best chromosomes and variation operators called *crossover* and *mutation* to the whole population, the algorithm *evolves* such as the average fitness of the chromosomes increases/decreases (maximization/minimization). The final populations contain the optimal or near optimal solutions.

For feature selection purpose, each gene corresponds to the index of a feature in  $\Phi$ , thus it is an integer between 1 and 253 (i.e.  $v_{lb_j} = 1, v_{ub_j} = 253, \forall j$ ), and  $l = 5$  as we want to reduce the dimension of the feature set to 5. That is, the GA seeks the best combination of 5 unique features from the entire set of available features  $\Phi$ , according to an optimization criterion (fitness function). The termination criterion of the algorithm is the expiration of the maximum number of generations the GA is let to run ( $T_{\text{max}}$ ).

Selection is a probabilistic mechanism which chooses the best individuals (i.e. minimum fitness) with some probability from the current generation and passes them to the next generation. We have adopted a binary tournament selection scheme [33] due to its constant selection pressure over time [34]. We prevent losing the best individuals from the population [35] by an elitist replacement of the 5 worst individuals in each generation with the 5 best individuals in the previous generation. We used a rather high number of elites (i.e. 5) as we adopt a relative high mutation rate as well (see end of this subsection).

The *fitness function* we propose is in direct relationship with the classification performance. We choose the fitness function for a given chromosome  $x$  (i.e. a given

combination of  $l$  features) to be proportional to the Error Rate (ER%) obtained after measuring how well a classifier discriminates the two classes using the features  $x$ . As we perform minimization we seek the best combination of  $l$  features that minimizes ER or equivalently minimizes the quantity  $100 - CCR\%$ . We evaluate the performance of the classifier by performing an  $S$ -fold crossvalidation as described in section 2.4.2. The fitness function should penalize the repetition of features in a chromosome  $x$  – we seek a vector of  $l$  *unique* features. It should also penalize a high variation of the  $CCR$  values in the  $S$ -fold crossvalidation for chromosome  $x$ , as we seek, besides high  $CCR$  values, a *reduced variation* of  $CCR$  between training/testing experiments in the crossvalidation. That is,  $f(x)$  should increase substantially if we encounter repetitions of the features and should increase mildly with the variance of the  $CCR$  results after  $S$ -fold crossvalidation. The fitness (minimization) is taken as:

$$f(x) = 100 - \overline{\text{vec}(\mathbf{CCR}(x))} + \underbrace{\omega \cdot \text{std}(\text{vec}(\mathbf{CCR}(x)))}_{\text{penalty term1}} + \underbrace{\text{rep}(x) \cdot e^{\text{rep}(x)}}_{\text{penalty term2}}, \quad x \in P \quad (3)$$

where  $f(x)$  is the fitness of feature vector  $x$ , “vec” represents the operator that stacks the matrix columns into a vector, the upper horizontal bar is the average operator,  $\omega$  is the weight of the first penalty term which is the standard deviation of  $CCR$ , “rep” is the repetition operators that counts how many repeated features occur in the feature vector  $x$ , and is used for the second penalty term. By increasing  $f(x)$  through the penalty terms, due to the selection effect in the population, such “bad” chromosomes tend to disappear after several generations of the GA.

The variation operators are: Uniform Crossover (AX) defined in [36] and applied to pairs of chromosomes with some probability  $P_c$ . Mutation (flip mutation) [33] replaces, with some probability  $P_m$ , the gene's value at a given locus  $j$  with a random value in  $[v_{lb_j}, v_{ub_j}]$ .

The GA population is initialized as follows: half of the chromosomes in the population, chosen at random, get their initial gene's values by picking randomly for each gene the index of a feature in the top 62 (approximately a quarter of the features in  $\Phi$ ) of previously ranked features. Features are ranked according to the decreasing ordered set of  $Z_{\text{final}}$  values as described in section 2.5.1. Thus, we assure that at least half of the population has been initialized with good features, the rest of the population being initialized with random values in  $[v_{lb_j}, v_{ub_j}]$ . Random initialization is a standard procedure for GA that allows for a wide exploration of the search space from the first generation.

The parameterization of the GA in all our experiments is the following:  $N = 50$  individuals,  $T_{\max} = 100$  generations,  $P_c = 0.9$ ,  $P_m = 0.2$ ,  $\omega = 0.5$ ; the  $S$ -fold crossvalidations when calculating the fitness have  $S = 5$ ,  $\tau = 10$ , and in the case of the neural network (MLP)  $\xi = 5$ .

### 2.5.3 Feature combination

We may reduce the dimensionality of the feature space by performing linear transformation and taking the most important components. We adopt Principal Component Analysis (PCA) which is a well-known statistical technique that has been widely used in data analysis and compression (for example, articles such as [37] and textbooks such as [28] present reviews of the method). The goal of the method is the compression of a high-dimensional input data into a lower dimensional space, without loss of relevant information. To capture the main features of the data set, PCA is looking for directions along which the dispersion or variance of the point cloud is maximal. These "principal" directions form a subspace of lower dimension than the original input space. The projection of the data onto the respective subspace will yield a transformation similar to compression, which minimises the loss of information according to the Minimum Mean Square Error criterion. In our case, we perform the transformation over the 253-dimensional feature space and take only the first 5 principal components.

Fisher’s Linear Discriminant Analysis (LDA) is a well known dimensionality reduction scheme [32] that projects the patterns onto a lower dimensional subspace such that the classes become “more separable” according to a criterion (maximization) called the Fisher Linear Discriminant.

## 3 Results

### 3.1 Discriminating potential of voice features

We may analyze the discriminating potential of voice features (section 2.3) by looking at the top 5 features as yielded by the ranking method described in 2.5.1. These features (in decreasing order of the  $Z_{\text{final}}$  values) are the following: 'MEAN\_HNR\_VA\_A', 'VTI\_VE\_A', 'MBW\_formant2\_VA\_E', 'MBW\_formant2\_VE\_I', 'MF\_formant2\_VE\_U'. The best feature is therefore 'MEAN\_HNR\_VA\_A' that passes (favours the alternative hypothesis) the Wilcoxon two-sampled test of difference in medians with a good  $p$ -value,  $p = 2.09 \cdot 10^{-10}$  (the null hypothesis states that medians are equal for the two groups – control and severe OSA –

and we reject the null hypothesis at a 1% significance level with a quite small  $p$ -value,  $p < 0.01$ ). From the boxplot in Figure 1, the difference between the distribution of 'MEAN\_HNR\_VA\_A' is apparent. Moreover, from histograms in Figure 1a and Figure 1b it is apparent that distributions for the two groups seem to depart from normality and outliers occur. This justifies the use of a non-parametric robust statistical test, in the first place (i.e. Wilcoxon test). Furthermore, besides the mean and standard deviation values we gave the median values as well, less affected by outliers and heavy tail skewed distributions. It is relevant to note that the use of 'MEAN\_HNR\_VA\_A' was inspired by studies that try to discriminate between normal voice and sleepy voice [38], as we consider that severe OSA patients may exhibit certain fatigue in voice. Looking at the next 4 features and applying the same statistical test, it follows that for all features, the medians between groups cannot be considered equal, and this is a strong assertion judging by the very small  $p$ -values obtained ( $p < 0.00001$ ):  $2.057 \cdot 10^{-12}$  ('VTI\_VE\_A'),  $1.4843 \cdot 10^{-8}$  ('MBW\_formant2\_VA\_E'),  $2.77 \cdot 10^{-8}$  ('MBW\_formant2\_VE\_I'),  $1.9 \cdot 10^{-6}$  ('MF\_formant2\_VE\_U'). Even though 'VTI\_VE\_A' has a smaller  $p$ -value than 'MEAN\_HNR\_VA\_A', the ranking method outweighed this feature as it was found to be correlated to 'MEAN\_HNR\_VA\_A'. Statistics performed indicate that, at least taking into account the top 5 ranked features, voice may be considered *distinct* between the extreme groups: control and severe OSA.

### 3.2 Classifier comparison

The results of the  $S$ -fold crossvalidation for all classifiers and feature reduction schemes are given in Table 2. It follows that the best strategy in terms of  $CCR$  (average 82.04%), Sensitivity (average 81.74%) and Specificity (average 82.40%) is the Bayesian Classifier (BC) with featured selected by the GA (denoted as BC-GA for short). BC-GA also achieves the smallest standard deviations of the  $CCR$ /Sensitivity/Specificity triplet among all classifiers and all feature selection methods. The second best strategy is the SVM with linear kernel ( $SVM_{lin}$ ) and features selected by the GA, and the third best is the MLP with a dimensionality reduction through LDA. GA achieves the best or close to best 5 features for each classifier, and therefore is the best feature selection scheme, while LDA is the best feature combination method.

A closer look at the results of the BC – GA (Fig. 2) indicates that Sensitivity (Fig. 2a) and Specificity (Fig. 2b) present a skewed distribution with longer tails for smaller than 70% values (the skewness is -0.47 for Sensitivity and -0.21 for Specificity), therefore we might consider that the most representative (probable) values for the Sensitivity and Specificity are

closer to the mode than to the mean of the distribution (Sensitivity  $\approx$  83%, and Specificity  $\approx$  88%).

(Figure 1)

It is instructive to see what are the 5 features selected by the GA in the case of BC-GA: 'CVF\_pitch\_VE\_A', 'VTI\_VE\_A', 'MBW\_formant2\_VE\_E', 'APQ\_VE\_I', 'SC2\_ltas\_FA'. One of the features is related to the phrase analysis and is the Spectral Crest of the LTAS for the second frequency band 500 – 1000 Hz ('SC2\_ltas\_FA', see section 2.3.6). Thus, both vowel processing and phrase processing are important in decision making.

The best triplet (*CCR*, Sensitivity and Specificity) in a single training – test run of the BC-GA during the *S*-fold crossvalidation is *CCR* = 96%, Sensitivity = 96% and Specificity = 96%.

Next, we consider an ensemble of the three best classifiers: BC-GA, SVM<sub>lin</sub>-GA and MLP-LDA. We apply majority vote for the outputs of the classifiers to decide the class to which each pattern pertains. For the whole ensemble we apply the *S*-crossvalidation process as before and get a slight improvement of the results as given in Table 3. The best triplet (*CCR*, Sensitivity and Specificity) in a single training – test run of the ensemble classifier during the *S*-fold crossvalidation is *CCR* = 95.91%, Sensitivity = 92.85% and Specificity = 100%.

Next, we check the performance on “in-between cases”, that are snoring patients and mild OSA. These cases have an AHI between 5 and 30. We check the performance of the BC-GA on a group of 128 patients, where we consider AHI = 15 as the border between non-OSA (below this threshold) and OSA (above this threshold). We have 65 non-OSA and 63 OSA patients, 24 women and 104 men, with ages comprised between 21 and 76, mean 48, median 48 and mode 42. BMI ranges from 18 to 47, mean 29, median 28 and mode 26. We train the BC-GA on *all* 248 extreme cases (patterns considered in previous sections) and test the classifier on the 128 *intermediate* cases only.

(Table 2)

(Figure 2)

(Table 3)

We get a  $CCR = 70.31\%$ , with a Sensitivity = 73.01% and a Specificity = 67.69%. This is a clear drop in performance with respect to validation on extreme cases, meaning that it is difficult to discriminate between milder OSA and non-OSA/snoring patients based solely on knowledge acquired from the voice in the extreme cases group. Moreover, we may perform an  $S$ -fold crossvalidation ( $S = 5$ ,  $\tau = 50$ ) on the intermediate patients alone. In this case, we get the results presented in Table 4, which again show a dramatic drop in performance. We may conclude at this point that it is hard to build an efficient classifier using the intermediate cases alone, and is preferable to build the classifier on the extreme cases (this assures, at least, a good recognition rate for the extreme cases:  $CCR$  above 80%) and a recognition rate ( $CCR$ ) for the intermediate validation cases around 70%. For completion, we also perform a  $S$ -fold crossvalidation on *all* 376 patterns (union between extreme and intermediate cases). Results are given in Table 5.

The classifier trained on extreme cases achieves the best results when validated on extreme cases, and significantly worse results when validated on intermediate cases. A potential means of sieving-out the intermediate cases prior to application of the classifier (in a screening scenario, for example) would be the use of simple parameters from the patient's medical or clinical record, readily available standard measures such as age, BMI, neck circumference, Epworth sleepiness scale (EPW), blood pressure, etc. Numerous studies such as [39, 40, 41, 43] are investigating the relationships and correlations between such standard measures and the OSA. We may perform a quick check of these relationships by computing the Pearson linear correlation coefficient ( $\rho$ ) between basic measures and AHI on our body of 248 extreme cases. We obtain  $\rho = 0.60$  ( $p$ -value  $< 10^{-25}$ ) for the correlation between AHI and age,  $\rho = 0.49$  ( $p$ -value  $< 10^{-20}$ ), for the correlation between AHI and BMI, and  $\rho = 0.29$  ( $p$ -value  $< 10^{-20}$ ), for the correlation between AHI and EPW. These values indicate that we can rely on the assumptions that these measures (especially age) are correlated to AHI, therefore they may be employed to cull-out potential intermediate cases. For example, cases with ages between 30-45, relatively low or medium BMI, medium EPW values, evidence of snoring during nightsleep, may be discarded as intermediate cases prior to the application of the classifier. Such intermediate cases need a deeper analysis, and fusion with other more traditional sources of information (such as the established means of diagnosis: RP, or PSG).

Finally, it is worth mentioning that more complex classifiers such as  $SVM_{poly}$  or KNN that attain nonlinear decision boundaries have less generalization capabilities than simpler classifiers such as BC or  $SVM_{lin}$  that use linear decision boundaries. The explanation is that such complex classifiers seem to present *overfitting*, in that they are capable of learning very well the training patterns with all incorporated noise and spurious information, but the complex decision border is not able to classify well the new test patterns. Simpler (linear) decision borders seem better for the current distribution of patterns. Actually, by looking at several runs in the  $S$ -fold crossvalidation of the MLP, we found that many times overfitting of the neural networks [44] occurs: when monitoring the learning curve, at some point in time as the learning error keeps decreasing the classification error of the test patterns starts to increase. For such runs we performed early-stopping, that is we stopped learning for an epoch less than  $T_{MLP}$  when the error on the test set began increasing.

## 4 Conclusions

The present study focuses on voice alone as a primary discriminating source of information between healthy subjects and severe OSA. Both statistical analysis on several voice extracted features, as well as performance of several classifiers indicate that voice has a clear potential to detect severe OSA among healthy subjects. The performance of the classifiers has been estimated using robust statistical techniques ( $S$  – fold crossvalidation) while counting with a relatively large body of subjects (i.e. 248), larger than most of the present studies analyzing the relationship between voice and OSA. The group of subjects involved in our experimental design increases to 376, when including the intermediate cases as well. We may get a better grasp on the relationship between OSA and voice by looking at the extreme cases that also have a clear-cut diagnosis. The results in terms of  $CCR$ , Sensitivity and Specificity, all above 80% for several classifiers point out the good potential of voice as a discriminating factor between healthy subjects and severe OSA.

Careful analysis on subjects with different degrees of OSA reinforced our prior belief that voice may act as a good discriminating factor for most of the severe cases. However, for intermediate cases where upper-airway closure may not be so pronounced (thus voice not much affected), we cannot rely on voice alone for making a good discrimination between OSA and non-OSA.

Analysing the features discovered by the feature reduction methods, we conclude that both vowel and phrase features are useful (more vowel features are selected, however) and both uttering positions as well, with more features selected from the stretched ('E') uttering.

The GA feature selection method proved to be the best reduction scheme that is well adapted to the classifier, and that achieves the best *CCR*, Sensitivity and Specificity with a small variance of these results due to the specifically designed fitness function (see eq. 3), for almost all cases involved in comparison. The GA is capable of discovering useful associations between voice features, and that are not apparent beforehand, the degree of utility being in direct relationship to the classifier performance.

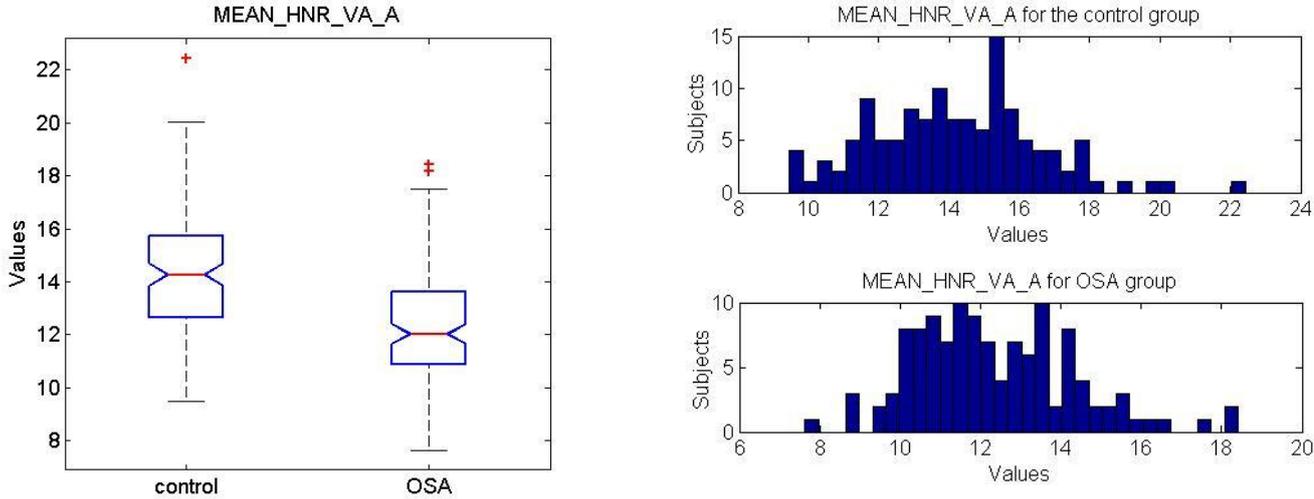
Feature selection is a crucial stage in our design as there are many features that can be extracted from voice and speech but there is no apriori knowledge regarding the most discriminant to be employed in the detection of the OSA cases. Therefore, we would like to highlight the use of GAs as one of the most innovative aspects in the present study. GAs have turned out to be the perfect choice when it comes to salient feature discovery, achieving good adaptation with the classification tools employed.

For a screening application that detects severe OSA cases among healthy people we may employ an ensemble classifier that combines the output of various classifiers to yield a more robust decision. As seen from section 3.2 such an ensemble classifier achieves slightly better results than the best classifier (BC-GA). Moreover, fusion with other measures from the subject's medical record (i.e. sex, age, BMI, EPW, blood pressure) is expected to increase the overall performance. Such parameters are correlated with the AHI index and thus with the presence or absence of OSA, and may shed light into the suitable discrimination of the intermediate subjects as well (mild OSA, snoring subjects), subjects that are difficult to classify by voice analysis only. A multiclass approach, instead of a binary classification, is also expected to increase the classification performance. We might consider more than 2 classes, such as, for example: controls, healthy snoring subjects, mild-OSA, and severe-OSA, and we may make a differentiation between sexes, as well.

So far, results presented as an *S*-fold crossvalidation for several classifiers are by no means a substitute for a clinical validation study. Crossvalidation served us to better estimate the discriminating potential of voice, and the *expected* correct classification rate, sensitivity and specificity. Actually, during each training-testing experiment involved in the *S*-fold crossvalidation only a fifth of the total number of subjects (about 50, for the extreme cases problem) was employed for validation purposes, the rest being used to train the classifier. For

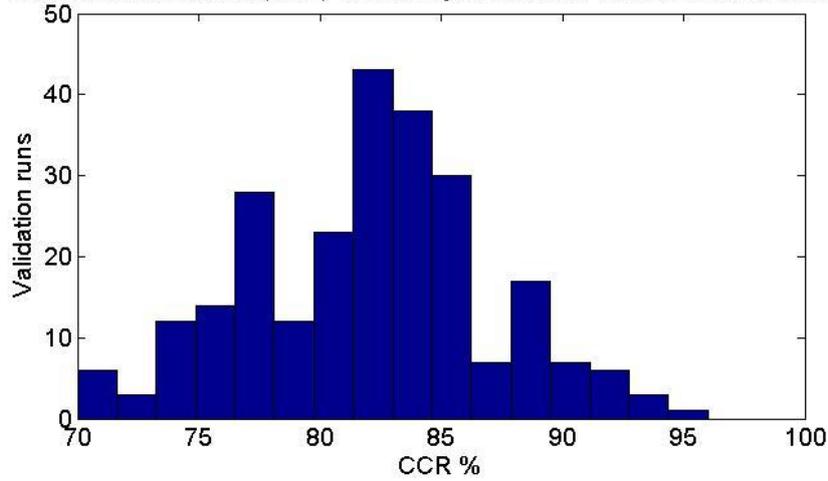
future work, we will seek to produce clinical validation results for a comprehensive body of new subjects, with an already trained classifier using the model developed in this paper.

### Figures



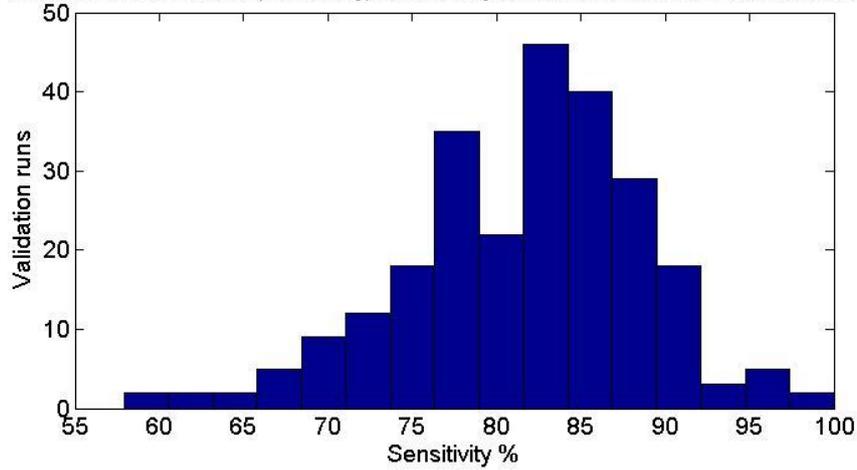
**Fig. 1** Boxplot a) and histograms b) of the MEAN\_HNR\_VA\_A features for the control and severe OSA group

Crossvalidation results (CCR) for the Bayes Classifier with GA feature reduction



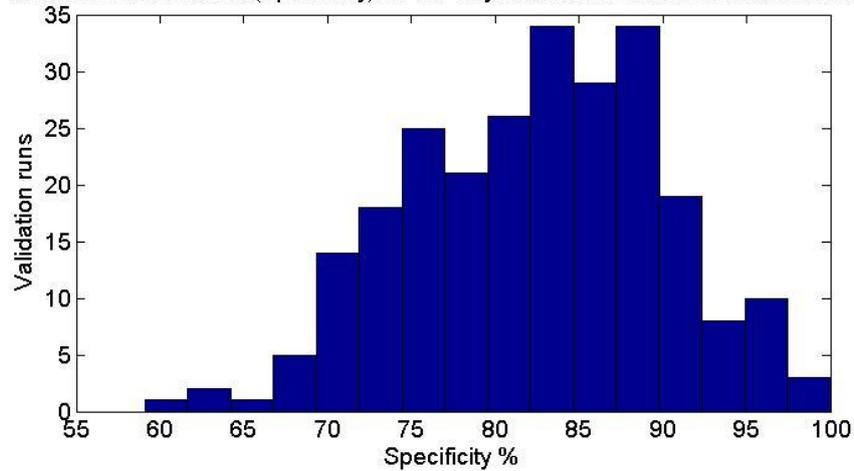
a)

Crossvalidation results (Sensitivity) for the Bayes Classifier with GA feature reduction



b)

Crossvalidation results (Specificity) for the Bayes Classifier with GA feature reduction



c)

**Fig. 2** Histograms of the a) CCR, b) Sensitivity and c) Specificity for the S-fold crossvalidation of the Bayesian Classifier with features selected by the GA.

**Table 1** Considered database for this study. Gender, Age (range, mean, median and mode) and Body Mass Index (range, mean, median and mode) of both groups are provided.

<b>Control group</b>	
Gender	48 men, 79 women
Ages	18÷64, mean 29.68, median 24, mode 21
Body Mass Index (BMI)	18÷64, mean 29.68, median 24, mode 21
<b>Severe OSA group</b>	
Gender	101 men, 20 women
Ages	28÷82, mean 54.04, median 55, mode 62
Body Mass Index (BMI)	23÷53, mean 32.56, median 31.2, mode 34.6

**Table 2** Results in terms of Average (AVG), Median (MED), Mode (MOD), Standard Deviation (STD) for the Correct Classification Rate (CCR), Sensitivity and Specificity for all classifiers and feature reduction methods

			<b>Ranked</b>	<b>PCA</b>	<b>LDA</b>	<b>GA</b>
<b>MLP</b>	CCR [%]	AVG	78.98	77.37	79.98	79.76
		MED	79.59	77.77	80	80
		MOD	76	78	82	80
		STD	4.64	4.95	5.03	4.80
	Sensitivity [%]	AVG	75.01	73.77	83.75	77.13
		MED	77.27	76.92	86.36	80.76
		MOD	83.33	80	87.5	88
		STD	13.17	14.56	10.4	15.04
	Specificity [%]	AVG	75.58	76.64	78.3	77.07
		MED	77.27	78.26	89.32	78.26
		MOD	81	76.92	81	88
		STD	11.46	13.82	10.59	11.18
<b>SVM<sub>lin</sub></b>	CCR [%]	AVG	74.22	72.66	77.16	81.10
		MED	74	73.46	77.55	81.63
		MOD	76	74	78	80
		STD	6.12	5.02	5.38	5.41
	Sensitivity [%]	AVG	72.29	71.97	74.9	77.87
		MED	72.72	72	75	77.77
		MOD	75	66.66	75	75
		STD	9.6	8.56	8.73	8.52
	Specificity [%]	AVG	76.34	73.55	79.14	84.74
		MED	76.92	74	79.31	85.71
		MOD	78.26	74	81	88.88
		STD	8.87	9.45	6.91	7.14
<b>SVM<sub>poly</sub></b>	CCR [%]	AVG	65.35	68.74	74.05	72.87
		MED	66	69.38	74	73.46
		MOD	66	70	74	70
		STD	6.67	6.12	5.54	6.13
	Sensitivity [%]	AVG	63.76	70.14	71.82	72.75
		MED	65	70.83	71.19	73.79
		MOD	66.66	70	71	75
		STD	10.51	9.18	8.92	9.39
	Specificity [%]	AVG	67.06	67.65	76.44	73.11
		MED	66.66	68.18	77.77	73.07
		MOD	66.66	66.66	78.57	73
		STD	10.46	9.07	8.82	9.6

**Table 3** Results in terms of Average (AVG), Median (MED), Mode (MOD), Standard Deviation (STD) for the Correct Classification Rate (CCR), Sensitivity and Specificity for the ensemble classifier (BC-GA + SVM<sub>lin</sub>-GA + MLP-LDA)

		<b>Ensemble</b>
CCR [%]	AVG	82.85
	MED	82
	MOD	83
	STD	4.83
Sensitivity [%]	AVG	81.49
	MED	81.48
	MOD	84
	STD	7.57
Specificity [%]	AVG	84.69
	MED	85.71
	MOD	87
	STD	6.52

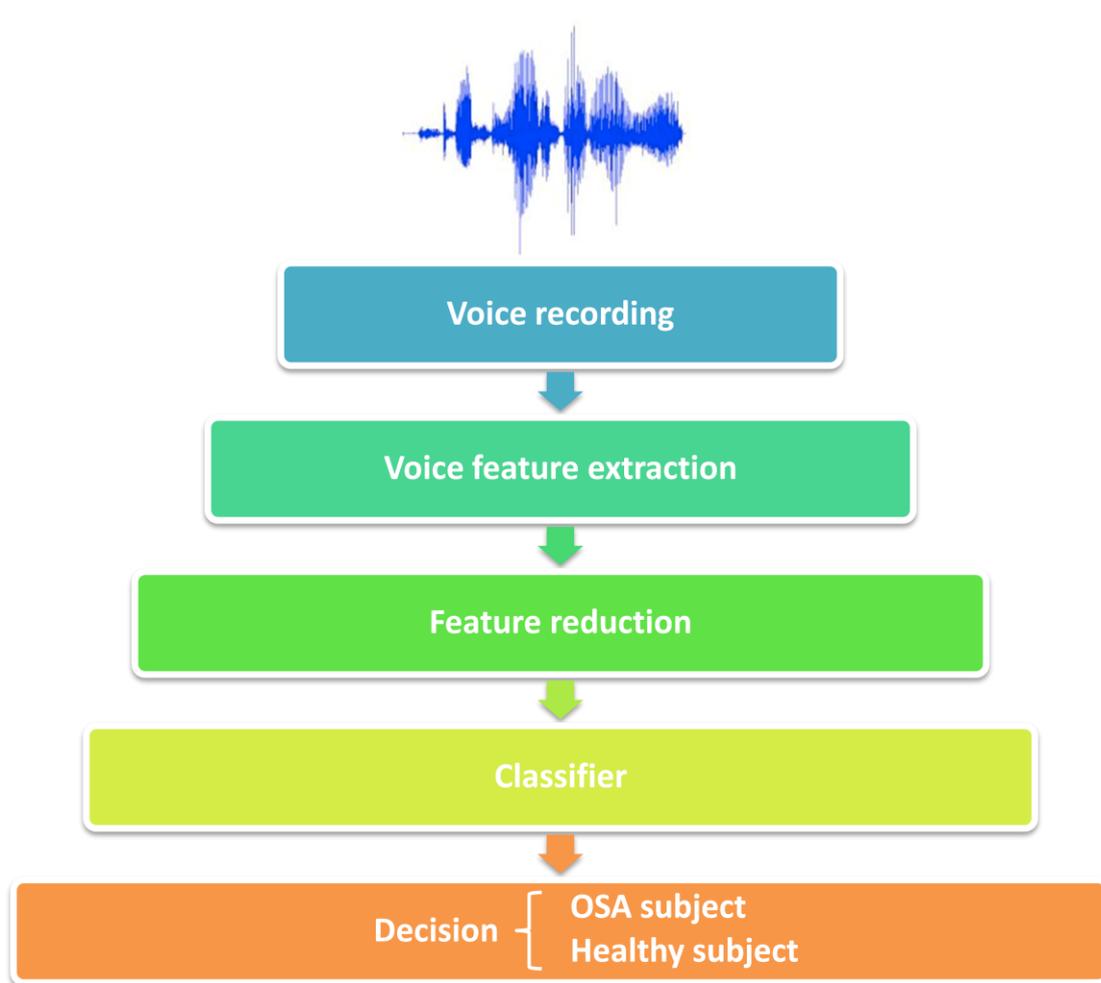
**Table 4** Results in terms of Average (AVG), Median (MED), Mode (MOD), Standard Deviation (STD) for the Correct Classification Rate (CCR), Sensitivity and Specificity for the BC-GA classifier on the 128 intermediate cases.

		<b>BC-GA</b>
CCR [%]	AVG	64.23
	MED	64.69
	MOD	69.23
	STD	8.26
Sensitivity [%]	AVG	55.36
	MED	55.55
	MOD	60
	STD	14.34
Specificity [%]	AVG	72.94
	MED	73.33
	MOD	69.23
	STD	11.68

**Table 5** Results in terms of Average (AVG), Median (MED), Mode (MOD), Standard Deviation (STD) for the Correct Classification Rate (CCR), Sensitivity and Specificity for the BC-GA classifier on the 128 intermediate cases + 248 extreme cases.

		<b>BC-GA</b>
CCR [%]	AVG	74.9
	MED	74.66
	MOD	73.33
	STD	4.92
Sensitivity [%]	AVG	71.39
	MED	71.42
	MOD	70
	STD	8.34
Specificity [%]	AVG	78.23
	MED	78.57
	MOD	80
	STD	5.99

## Graphical abstract



## References

1. Caples SM, Gami AS, Somers VK (2005) Obstructive Sleep Apnea. *Ann Intern Med* 142(3):187-197
2. Kushida CA, Littner MR, Morgenthaler T, Alessi CA, Bailey D, Coleman J, Friedman L, Hirshkowitz M, Kapen S, Kramer M, Lee-Chiong T, Loubé DL, Owens J, Pancer JP, Wise M (2005) Practice Parameters for the Indications for Polysomnography and Related Procedures: An Update for 2005. *Sleep* 28(4):499-521
3. Monoson PK, Fox AW (1987) Preliminary observation of speech disorder in obstructive and mixed sleep apnea. *Chest* 92:670-675
4. Fox AW, Monoson PK, Morgan CD (1989) Speech dysfunction of Obstructive Sleep Apnea. *Chest* 96:589 -595
5. Fiz JA, Morera J, Ahad J, Belsunces A, Ham M, Fiz JI, Jane R, Caminal PM, Rodenstein D (1993) Acoustic Analysis of Vowel Emission in Obstructive Sleep Apnea. *Chest* 104:1093-1096
6. José Luis Blanco, Luis A. Hernández, Rubén Fernández, Daniel Ramos (2013) Improving Automatic Detection of Obstructive Sleep Apnea Through Nonlinear Analysis of Sustained Speech. *Cognitive Computation*, December 2013, Volume 5, Issue 4, pp 458-472
7. Goldshtein E1, Tarasiuk A, Zigel Y. (2011) Automatic detection of obstructive sleep apnea using speech signals. *IEEE Trans Biomed Eng.* 2011 May;58(5):1373-82.
8. Lucey AD, King AJC, Tetlow GA, Wang J, Armstrong JJ, Leigh MS, Paduch A, Walsh JH, Sampson DD, Eastwood PR, Hillman DR (2010) Measurement, Reconstruction, and Flow-Field Computation of the Human Pharynx With Application to Sleep Apnea. *IEEE Trans Biomed Eng* 57(10):2535 – 2548
9. Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. The Report of an American Academy of Sleep Medicine Task Force (1999) *Sleep* 22 (5): 667-589
10. Tolkmitt FJ, Scherer KR (1986) Effect on Experimentally Induced Stress on Vocal Parameters. *J Exp Psychol Hum Percept Perform* 12(3):302-313
11. Kitamura T, Takemoto H, Honda K, Shimada Y, Fujimoto I, Syakudo Y, Masaki S, Kuroda K, Okuchi N, Senda M (2005) Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner. *Acoust Sci & Tech* 26(5):465-468
12. Buchaillard SI, Perrier P, Payan Y (2009) A biomechanical model of cardinal vowel production: muscle activations and the impact of gravity on tongue positioning. *J Acoust Soc Am* 126(4): 2033–2051
13. Rabiner LR, Schafer RW (1978) *Digital Processing of Speech Signals*, Prentice Hall
14. Deliyiski DD (1993) Acoustic Model And Evaluation of Pathological Voice Production. In: *Proceedings of EUROSPEECH'93*, Berlin, Germany, pp.1969-1972
15. Koreman J, Pützer M (1997) Finding Correlates of Vocal Fold Adduction Deficiencies. In: *PHONUS 3*, Saarbrücken, Institute of Phonetics, University of the Saarland, pp. 155-178
16. Talkin D (1995) A Robust Algorithm for Pitch Tracking (RAPT). In: Kleijn WB, Paliwal KK (eds) *Speech Coding & Synthesis*, Elsevier, Amsterdam, pp. 495-518
17. Baken RJ, Orlikoff RF (1999) *Clinical measurement of speech and voice*. 2nd ed. Singular, Thomson Learning

18. Qi Y, Hillman RE (1997) Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. *J Acoust Soc Am* 102(1):537–543
19. Muta H, Baer T, Wagatsuma K, Muraoka T, Fukuda H (1988) A pitch-synchronous analysis of hoarseness in running speech. *J Acoust Soc Am* 84(4):1292–1301
20. de Krom, G (1993) A cepstrum-based techniques for determining a harmonics-to-noise ratio in speech signals. *J. Speech Hear Res* 36:254–266
21. Parsa V, Jamieson DJ (2001) Acoustic Discrimination of Pathological Voice: Sustained Vowels Versus Continuous Speech. *J Speech Lang Hear Res* 44:327–339
22. Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM (2007) Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. *BioMed Eng OnLine* 6(23), available from: <http://www.biomedical-engineering-online.com/content/6/1/23>
23. Leino T (2009) Long-Term Average Spectrum in Screening of Voice Quality in Speech: Untrained Male University Students. *J Voice* 23(6):671-676
24. Boersma P, Kovacic G (2005) Spectral characteristics of three styles of Croatian folk singing. *J Acoust Soc Am* 119(3):1805–1816
25. Peeters G (2004) A large set of audio features for sound description (similarity and classification) in the CUIDADO project. IRCAM, Tech Rep
26. Haykin S (1999) *Neural Networks: A Comprehensive Foundation*. 2nd ed. Prentice Hall, pp. 178-277
27. Cristianini N, Shawe-Taylor J (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Trumpington Street, Cambridge, UK
28. Bishop, CM (2006) *Pattern Recognition and Machine Learning*, Springer-Science + Business Media, LLC
29. Viola P, Jones MJ (2004) Robust Real-Time Face Detection. *Int J Comput Vision* 57(2):137-154
30. Theodoridis S, Koutroumbas K (2003) *Pattern Recognition*. 2nd ed. Elsevier-Academic Press, Amsterdam, pp. 43
31. Kvam PH, Vidakovic B (2007) *Nonparametric Statistics with Applications to Science and Engineering*. Wiley-Interscience, A John Wiley & Sons, Inc., Publ., Hoboken, New Jersey, pp. 129-133
32. Duda RO, Hart PE, Stork DG (2001) *Pattern Classification*. 2nd ed. Wiley-Interscience, New York, N.Y.
33. Bäck T (1996) *Evolutionary Algorithms in Theory and Practice*, Oxford University Press.
34. Blicke T, Thiele L (1996) A comparison of selection schemes used in Evolutionary Algorithms. *Evolutionary Computation* 4(4):361-394
35. Bäck T, Hoffmeister F (1991) Extended selection mechanisms in genetic algorithms. In: Schaffer JD (ed) *Proceedings of the 4th International Conference on Genetic Algorithms*, Morgan Kaufmann, San Diego, CA, pp. 92-99
36. Syswerda G (1989) Uniform crossover in genetic algorithms. In: Belew RK, Booker LB (eds) *Proceedings of the Third International Conference on Genetic Algorithms*, Morgan Kaufmann, Fairfax, VA, pp. 2–9
37. Bezdek JC, Pal NR (1995) A note on self-organizing semantic maps. *IEEE Transactions on Neural Networks* 6(5):1029-1036

38. Krajewski J, Wieland R, Batliner A (2008) An acoustic framework for detecting fatigue in speech based Human-Computer-Interaction. In: Miesenberger K, Klaus J, Zagler W, Karshmer A (eds) *Computers Helping People with Special Needs*, Springer, Heidelberg, pp. 54-61
39. Bixler EO, Vgontzas AN, Ten Have T, Tyson K, Kales A (1998) Effects of age on sleep apnea in men: I. Prevalence and severity. *Am J Respir Crit Care Med* 157(1):144-148
40. Ware JC, McBryer RH, Scott JA (2000) Influence of sex and age on duration and frequency of sleep apnea events. *Sleep* 23(2):165-170
41. Mortimore IL, Marshall I, Wraith PK, Sellar RJ, NJ Douglas (1998) Neck and total body fat deposition in nonobese and obese patients with sleep apnea compared with that in control subjects. *Am J Respir Crit Care Med* 157(1):280-283
42. Young T, Peppard PE, Gottlieb DJ (2002) Epidemiology of Obstructive Sleep Apnea: A Population Health Perspective. *Am J Respir Crit Care Med* 165(9):1217-1239
43. Pang KP, Terris DJ (2006) Screening for obstructive sleep apnea: an evidence-based analysis. *Am J Otolaryngol* 27(2):112-118
44. Bishop CM (1995) *Neural Networks for Pattern Recognition*. Oxford University Press, pp. 11-14