


Vocal Acoustic Analysis: ANN Versos SVM in Classification of Dysphonic Voices and Vocal Cords Paralysis

João Paulo Teixeira, Research Centre in Digitalization and Intelligent Robotics (CEDRI) and Applied Management Research Unit (UNIAG), Instituto Politécnico de Bragança, Bragança, Portugal

Nuno Alves, Instituto Politécnico de Bragança, Bragança, Portugal

Paula Odete Fernandes, Applied Management Research Unit (UNIAG), Instituto Politécnico de Bragança, Bragança, Portugal

 <https://orcid.org/0000-0001-8714-4901>

ABSTRACT

Vocal acoustic analysis is becoming a useful tool for the classification and recognition of laryngological pathologies. This technique enables a non-invasive and low-cost assessment of voice disorders, allowing a more efficient, fast, and objective diagnosis. In this work, ANN and SVM were experimented on to classify between dysphonic/control and vocal cord paralysis/control. A vector was made up of 4 jitter parameters, 4 shimmer parameters, and a harmonic to noise ratio (HNR), determined from 3 different vowels at 3 different tones, with a total of 81 features. Variable selection and dimension reduction techniques such as hierarchical clustering, multilinear regression analysis and principal component analysis (PCA) was applied. The classification between dysphonic and control was made with an accuracy of 100% for female and male groups with ANN and SVM. For the classification between vocal cords paralysis and control an accuracy of 78,9% was achieved for female group with SVM, and 81,8% for the male group with ANN.

KEYWORDS

ANN, Classification, Feature Selection, Hierarchical Clustering, HNR, Jitter, Multilinear Regression Analysis, PCA, Shimmer, SVM, Vocal Acoustic Analysis, Voice Pathologies

1. INTRODUCTION

Vocal Acoustic Analysis is often used for voice disorders assessment and diagnose (Bielamowicz et al., 1996; Brockmann-Bauser, 2011; Pylypowich, & Duff, 2016; Salhi, Mourad, & Cherif, 2010; Teixeira & Fernandes, 2015). The advantage of such techniques relies on the non-invasive character of the exam when compared with current practice in medicine, for example, laryngoscopy or stroboscopic exams (Brockmann-Bauser, 2011).

Both laryngoscopy and stroboscopic exam consist in inserting a thin tube into the throat or into the nostrils. Stroboscopy is painless, an office-based procedure done with topical anaesthesia. It is a special method used to visualize vocal fold vibration (Hirano, 1974). It uses a synchronized, flashing

DOI: 10.4018/IJEHMC.2020010103

This article, originally published under IGI Global's copyright on January 1, 2020 will proceed with publication as an Open Access article starting on January 25, 2021 in the gold Open Access journal, International Journal of E-Health and Medical Communications (converted to gold Open Access January 1, 2021), and will be distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

light passed through a flexible or rigid telescope. The flashes of light from the stroboscope are synchronised to the vocal fold vibration at a slightly slower speed, allowing the examiner to observe vocal fold vibration during sound production in what appears to be slow motion. The resulting video depicts video-stroboscopic examination of the vocal folds.

This incision technique will always be necessary to confirm or even support chirurgical operations on the vocal folds or in the larynx/pharynx.

Although voice disorders may be diagnosed by an auditory perceptual analysis made by the otolaryngologist, this may lead to different results depending on the practitioner experience (Teixeira & Fernandes, 2014).

It is common in daily life of primary care facilities the people complain about hoarseness in their voices. The dysphonia affects 30% of adults and 50% of older adults. This disease modifies voice quality and has a significant impact on life quality. This also represents a significant economic burden. In patients with a progressive pathology, it is important to do a diagnosis as fast as possible for the sake of having access to better treatment and prognosis (Pylypowich & Duff, 2016).

There are several acoustic parameters extracted from speech signal processing useful to identify the vocal pathology, yet no parameter alone is able to classify between healthy or pathologic voice.

Teixeira and Fernandes (2015) analysed the statistical significance of Jitter, Shimmer and HNR parameters for dysphonia detection. A statistical analysis was performed over the three parameters for the vowels /a/, /i/ and /u/ at three different tones, high, low and normal. In this work, Jitter and Shimmer are suggested as good parameters to be used in an intelligent diagnosis system of dysphonia pathologies.

To test this analysis, it is necessary to apply an intelligent tool and some reduction dimension and feature selection techniques. Feature selection is intended to select the best subset of predictors. The feature selection problem arises from large datasets who may contain redundant information and variables that have little or no predictive power (May, Dandy, & Maier, 2011). The correct choice of input features leads to a small subset that may boost/improve the performance when intelligent tools are used.

Henríquez et al. (2009) studied the usefulness of six nonlinear chaotic measures based on nonlinear dynamics theory in the discrimination between two levels of voice quality: healthy and pathological. The studied measures are first and second order Rényi entropies, the correlation entropy and the correlation dimension. The values of the first minimum of mutual information function and Shannon entropy were also studied. Two databases were used to assess the usefulness of the measures: a multi-quality and a commercial database (MEEI Voice Disorders). A classifier based on standard neural networks was implemented in order to evaluate the measures proposed. Global success rates of 82.5% (multi-quality database) and 99.7% (commercial database) were obtained. This difference in performance highlights the importance of having a controlled speech acquisition process.

In Forero et al. (2015), several parameters of glottal signal were used to identify nodule, unilateral paralysis or healthy voices. The database, obtained from a speech therapist, was composed by records of voices from 12 speakers with nodule, 8 speakers with vocal fold paralysis and 11 speakers with normal voices. Three different classifiers were used, an Artificial Neural Network, a Support Vector Machine (SVM) and Hidden Markov Model. The best accuracy, 97.2%, was reached using glottal signal parameters and MFCC's with SVM classifier.

Markaki and Stylianou (2011) explored the information provided by a joint acoustic and modulation frequency representation, referred to as modulation spectrum, for detection and discrimination of voice disorders. The initial representation is first transformed to a lower dimensional domain using higher-order singular value decomposition (HOSVD). For voice pathology detection an accuracy of 94.1% was achieved using one SVM as classifier.

In Panek et al. (2015) a vector made up of 28 acoustic parameters was evaluated using Principal Component Analysis (PCA), kernel principal component analysis (kPCA) and an auto-associative neural network (NLPCA) in four kinds of pathology detection (hyperfunctional dysphonia, functional

dysphonia, laryngitis and vocal cord paralysis) using the /a/, /i/ and /u/ vowels, spoken at a high, low and normal tones. The results show best efficiency levels of around 100%.

Al-Nasheri et al. (2016) investigated different frequency bands using correlation functions. The authors extracted maximum peak values and their corresponding lag values from each frame of a voiced signal by using correlation functions as features to detect and classify pathologic samples. Three different databases were used, Arabic Voice Pathology Database (AVPD), Saarbrücken Voice Database (SVD) and Massachusetts Eye and Ear Infirmary (MEEI). A Support Vector Machine was used as classifier. For the detection of pathology, an accuracy of 99.8%, 90.9% and 91.1% was achieved for the three databases respectively. In the classification of the pathology task an accuracy of 99.2%, 98.9% and 95.1%, respectively, was achieved for the three databases.

In Sellam and Jagadeesan (2014), an attempt was made to analyse and to discriminate pathologic voice from normal voice in children using different classification methods. The classification of pathologic voice from normal voice was implemented using Support Vector Machine (SVM) and Radial Basis Functional Neural Network (RBFNN). Several acoustic parameters were extracted such as the signal energy, pitch, formant frequencies, mean square residual signal, reflection coefficients, Jitter and Shimmer. The best accuracy results were obtained by RBFNN with, 91%, and for the SVM 83%.

The artificial neural networks are among the most used classifiers for this kind of task although SVM is also used often (Al-Nasheri et al., 2016; Forero et al., 2015; Henríquez, et al., 2009; Markaki & Stylianou, 2011; Panek et al., 2015; Sellam & Jagadeesan, 2014).

Cordeiro (2017) presented a set of experiments to identify the best set of features from the vocal tract (MFCC, Line Spectral Frequencies (LSF), Mel-Line Spectral Frequencies (MLSF) and first peak of the spectral envelop) and the best classifiers amongst SVM and Gaussian Mixture Models (GMM) for the identification of pathologic voices. He achieved an accuracy of 84.4% for the identification between 3 groups (healthy subjects, subjects with physiological larynx pathologies - vocal fold nodules and edemas, and subjects with neurological larynx pathologies - unilateral vocal fold paralysis). He also used Regression Trees to the pathological voice recognition based on formant analysis and harmonic-to-noise ratio with 95% recognition rate.

In Teixeira and Gonçalves (2014) an algorithm was presented to automatically extract the jitter, shimmer and HNR features. The accuracy of measurements was compared with the ones extracted with Praat software (Boersma & Weenink, 2009) and showed better accuracy for synthesized speech signals and similar values as the Praat software for real signals.

Teixeira, Fernandes and Alves (2017) published the classification of dysphonic voices. A set of Jitter, Shimmer, and HNR parameters extracted from 3 sustained vowels at different tone levels was analysed. Three methods were used to reduce the features dimension set to be used in an ANN to classify between control and dysphonic voices. In this present research, the same dataset and methodology were used, but an additional classification tool was experimented with. The SVMs were compared with the ANNs. In this work, besides the dysphonia pathology, also the vocal cords paralysis pathology is used.

Next section describes the methodology. In this section the database is presented, as well as the set of features, a brief description of the used pathologies and the ANN and SVM architectures, the methods used to reduce the features dimension and the description of the used measures to evaluate the performance of the model. Section 3 presents the results obtained for the classification of each pathology by gender using ANN and SVM with selected features by the feature selection methods. Finally, section 4 presents the conclusions.

2. METHODS AND METHODOLOGY

The Saarbrücken Voice Database (SVD) (Barry & Pützer, 2007) was used in this study. For each subject, one segment of speech record was used for sustained vowels /a/, /i/ and /u/ for High, Low and Mid/Neutral tones in a total of 9 speech segments. Each segment of speech consists of a steady state

sustainable pronunciation of the respective vowel. For each speech segment a set of jitter, shimmer and HNR parameters, detailed below, was determined using the algorithm developed by Teixeira and Gonçalves (2016). This algorithm extracts a set of Jitter parameters (jitta, jitter, rap and ppq5), Shimmer (ShdB, Shim, apq3 and apq5) and HNR (Harmonic to Noise Ratio). A subset of the control subjects was selected in order to have similar distribution between gender and age of each pathologic group.

Two voice pathologies were used separately in this study, dysphonia and vocal cords paralysis (VCP). Each pathology was compared to control subjects. The control subjects consist of voice segments of healthy persons.

The classification in healthy/pathological voice was carried out for women and men separately. The number of samples taken from control group was the same of the pathologic group under examination (Panek et al., 2015). More details of the number of samples and distribution of ages can be seen in Table 1. A similar set of databases for dysphonia pathologic group was used in (Teixeira et al., 2017).

Pathology sets (dysphonia and VCP) include all subjects available in the SVD, meanwhile the subjects of the respective control groups were selected in order to have the most similar possible age. Anyhow, the available control subjects are mostly younger than pathologic subjects turning difficult to have very similar age in control and pathologic groups. The standard deviation between pathologic and control groups is mainly similar except in the female control dysphonia/dysphonia. The effect of ageing on voice quality is recognised, anyhow, the authors believe that the small difference in the average and standard deviation age groups does not affect the study results.

2.1. Dysphonia

Dysphonia is a medical term meaning disorder (dys-) of voice (-phonia) (Teixeira & Fernandes, 2015). The airflow moving through the vocal cords originates the human voice. Voice is different from speech, which is modulated by the pharynx, tongue and oral cavity (Pylypowich & Duff, 2016). Although there are many causes of dysphonia, it can be characterised by a disturbance in the phonation mechanism causing alteration in voice pitch. Voice dysfunction is not a disease by itself but can be a symptom of an underlying pathology.

2.2. Vocal Cords Paralysis

Vocal cord paralysis is a voice disorder that occurs when one (unilateral) or both (bilateral) vocal folds do not open or close properly. Unilateral paralysis is a common disorder, while bilateral paralysis is rarer and life-threatening.

The vocal cords are two elastic bands present in the larynx just above the trachea. When they are breathing, they remain distant and in swallowing they are closed. However, in the production of voice, the air coming from the lungs causes them to vibrate oscillating between the open and closed position.

Table 1. Gender and age distribution of the subjects in the chosen subset of SVD

	# Subjects		Margin of Years Old		Average (Standard Deviation) Years Old	
	Female	Male	Female	Male	Female	Male
Control Dysphonia	41	29	19-56	20-69	24.8 (7.32)	41.2 (18.7)
Dysphonia	41	29	18-73	11-77	45.6 (14.8)	48.7 (18.0)
Control VCP	126	69	18-84	18-69	31.0 (15.9)	34.8 (15.8)
VCP	126	69	21-79	23-81	55.8 (12.4)	59.1 (14.4)

In cases of paralysis, the vocal chords may remain open leaving the airways and lungs unprotected. This type of pathology can either occur after trauma to the head, neck or chest as well as in people with neurological problems such as multiple sclerosis, Parkinson's disease or who have suffered a stroke.

Symptoms may manifest as hoarseness, breathiness, trouble breathing, wheezing and swallowing problems. There may also be changes in voice quality such as loss of volume or fundamental frequency.

Bilateral vocal cord paralysis refers to the neurologic causes of bilateral vocal fold immobility and specifically refers to the reduced or absent function of the vagus nerve or its distal branch, the recurrent laryngeal nerve. Vocal fold immobility may also result from mechanical derangement of the laryngeal structures, such as the cricoarytenoid joint (Netter, 2014).

2.3. Parameters

Jitter, shimmer and HNR parameters were extracted with the algorithm developed by Teixeira and Gonçalves (2015). Jitter is defined as the periodic variation from cycle to cycle, and shimmer relates to the magnitude variation of the glottal period. A perspective of jitter and shimmer can be seen in Figure 1. Patients with lack of control of the vibration of vocal folds have tendency to have higher values of jitter. Reduction of glottal resistance and mass lesions causes a variation in the magnitude of the glottal period correlated with breathiness and noise emission, causing higher shimmer. The jitter and shimmer can be measured usually by four different forms. Jitter can be measured as absolute (jitta), relative (jitter), Relative Average Perturbation (rap) and the Period Perturbation Quotient (ppq5), according to Equations 1 to 4. Shimmer can be measured as absolute value in dB (ShdB), as relative value (Shim), as Amplitude Perturbation Quotient in 3 cycles (apq3) and as Amplitude Perturbation Quotient in 5 cycles (apq5), as Equations 5 to 8:

$$jitta = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i-1}| \quad (1)$$

$$jitter(relative) = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i-1}|}{\frac{1}{N} \sum_{i=1}^N T_i} \times 100 \quad (2)$$

$$rap = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} \left| T_i - \frac{1}{3} \sum_{n=i-1}^{i+1} T_n \right|}{\frac{1}{N} \sum_{i=1}^N T_i} \times 100 \quad (3)$$

$$ppq5 = \frac{\frac{1}{N-1} \sum_{i=2}^{N-2} \left| T_i - \frac{1}{5} \sum_{n=i-2}^{i+2} T_n \right|}{\frac{1}{N} \sum_{i=1}^N T_i} \times 100 \quad (4)$$

where T_i is the i glottal period lengths and N is the number of glottal periods.

$$ShdB = \frac{1}{N-1} \sum_{i=1}^{N-1} \left| 20 * \log \left(\frac{A_{i+1}}{A_i} \right) \right| \quad (5)$$

$$Shim = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |A_{i+1} - A_i|}{\frac{1}{N} \sum_{i=1}^N A_i} \times 100 \quad (6)$$

$$apq3 = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} \left| A_i - \left(\frac{1}{3} \sum_{n=i-1}^{i+1} A_n \right) \right|}{\frac{1}{N} \sum_{i=1}^N A_i} \times 100 \quad (7)$$

$$apq5 = \frac{\frac{1}{N-1} \sum_{i=2}^{N-2} \left| A_i - \left(\frac{1}{5} \sum_{n=i-2}^{i+2} A_n \right) \right|}{\frac{1}{N} \sum_{i=1}^N A_i} \times 100 \quad (8)$$

where A_i is the i peak-to-peak glottal magnitude and N the number of periods.

The Harmonic to Noise Ratio (HNR), Equation 9, provides an indication of the overall periodicity of the voice signal by quantifying the ratio between the periodic (harmonic part) and aperiodic (noise) components. This parameter is usually measured as an overall characteristic of the signal. The overall value of the HNR of the signal varies because different vocal tract configurations involve different amplitudes for the harmonics:

$$HNR = 10 * \log_{10} \frac{AC_v(T)}{AC_v(0) - AC_v(T)} \quad (9)$$

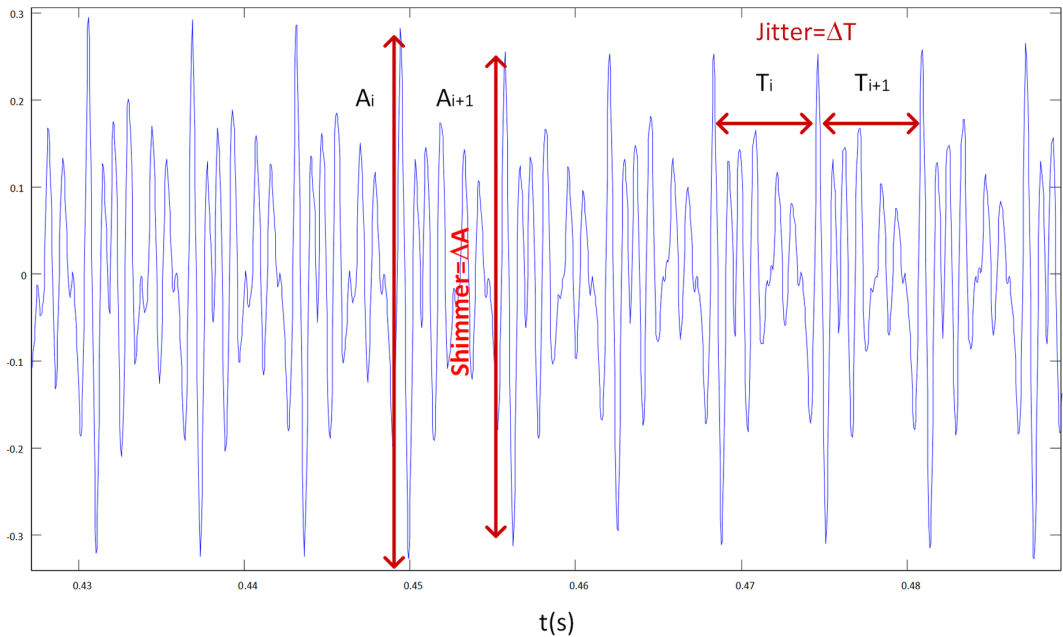
where $AC_v(0)$ is the total energy of the signal and $AC_v(T)$ is the energy of the first harmonic.

2.4. Artificial Neural Network (ANN)

For the ANN classifier, a Multilayer Perceptron (MLP) structure, trained with the back-propagation algorithm was used. Different typologies were examined with a different number of neurons in the hidden layer to seek the best generalization performance. The dataset was divided into three subsets, train, validation and test sets. The quantity of each dataset was 70%, 15%, 15%, respectively.

The ANN is composed of weights and bias trying to adapt to the desired output. The model has one neuron in the output layer. The output target was composed of zeros and ones. The output given by the ANN is not always exactly zero or one and so it had to be post-processed to be zero or one. For this process, a threshold of 0.5 was experimentally established for the output. The number of input nodes and hidden nodes is different for each model.

Figure 1. Jitter and Shimmer perturbation measures in speech signal of a sustained vowel /a/



2.5. Support Vector Machines

A support vector machine (SVM) is a type of intelligent tool based on minimizing structural risk. They can be used to solve classification and regression problems. The main idea of SVM is to construct hyperplanes as the optimal separation surface between positive and negative examples in a binary classification context (Almeida, 2010; Sellam & Jagadeesan, 2014).

Since problems are not always linear, it is necessary to transform the data so that it can be linearly separated. For this separation, the SVM use Kernel methods that make a non-linear transformation to the data for a multi-dimensional space where it will be an image of the data that allows a linear separation (Cruz, 2007).

Among the most used kernel methods are the linear, polynomial, radial basis function (RBF) and multi-layer perceptron (MLP). In the training of the SVM, the hyperplane parameters are adjusted so that the distance of the hyperplane to the data is maximum. The SVM also has another set of parameters called hyper-parameters of which the kernel function is dependent as the constant C of the boundary lines that border the hyperplane, the width of the Gaussian kernel and the degree of the polynomial kernel, among others (Ben-Hur & Weston, 2010). The choice of kernel can be important for the success of the SVM.

The implementation of the Support Vector Machine was done using two main functions to train the SVM and determine the predictive power of the classifier. The SVM requires the input matrix to be divided into two subsets, the training and the test. The percentage used for training was 85% and 15% for test.

All combinations of input parameters of the training function were experimented. The combinations refer to the different Kernel types, parameters associated with these Kernels, and different methods that allows to find the separation hyperplane. The precision, sensitivity and specificity under the test set were recorded for all experiments. The precision, sensitivity and specificity are in accordance with the next section.

Table 2. Confusion matrix used in the analysis.

		Results From Classification	
		Healthy	Pathology
Diagnosed	Healthy	True Positive (TP)	False positive (FP)
	Pathology	False Negative (FN)	True Negative (TN)
		Sensitivity = $TP/(TP+FN)$	Specificity = $TN/(TN+FP)$

2.6. Performance Evaluation

In order to evaluate the results, a confusion matrix was used and sensitivity, specificity and accuracy measures were calculated, as presented in Table 2 and Equation 10:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (10)$$

2.7. Feature Selection

The selection of input features is a fundamental consideration in identifying the optimal functional form of statistical models. The task of selecting input features is common to the development of all statistical models, and is largely dependent on the discovery of relationships within the available data to identify suitable predictors of the model output (May, Dandy & Maier, 2011). The process intends to explain the data in the simplest way eliminating redundant features. Applied to regression analysis, this implies that the smallest model that fits the data is the best. Unnecessary parameters will add noise to the estimation of other quantities. It is the attempt to avoid collinearity, caused by having too many variables giving the same information. Feature selection allows to save time and money reducing the problem dimension, turning the system more efficient from the computational point of view (Guyon & Elisseeff, 2003).

Three methods were used, the first two of feature selection and the last one of dimension reduction.

2.7.1. First Method - Hierarchical Clustering (HC)

Hierarchical clustering was used in the first method (Rokach & Maimon, 2005). The basic idea of this method is to create groups with variables most correlated. Then, only one parameter is selected from each group by its Euclidian distance. The parameter with the greater Euclidian distance is selected. The Euclidian distance is calculated for all parameters and between the classes being compared, healthy or pathologic.

2.7.2. Second Method - Multilinear Regression Analysis (MRA)

In the second method, a multilinear regression analysis was applied (Rokach & Maimon, 2005). It is a systematic method for adding and removing terms from a multilinear model based on their statistical significance in a regression. An initial model is created and the static significance is evaluated. At each step, one variable is added or removed based on this analysis. Resulting in an output model that will feed the neural network.

2.7.3. Third Method - Principal Components Analysis (PCA)

In the third method was applied a technique called Principal Components Analysis (PCA) (Smith, 2002). It is a statistical technique that uses the mathematical concepts like standard deviation, covariance, eigenvalues and eigenvectors. First, it is necessary to subtract the mean from each data dimensions. This produces a data set whose mean is zero, called data adjusted. Next, the eigenvectors and eigenvalues are calculated from covariance matrix. It is necessary to decide how many principal components to pick. The principal components are determined and in the output the eigenvalues are already ordered, it is just necessary to calculate the cumulative percentage of these values. Therefore, the first few eigenvectors corresponding to 90% or 95% of the cumulative percentage are selected. This means that the first few eigenvectors explain 90 or 95% of the data. Finally, the adjusted data is multiplied by the inverse of the eigenvectors matrix selected. To get results closer to the real the mean value is calculated only for the train set and subtracted to validation and test set. For the two pathologies, seven principal components were selected.

3. ANALYSIS OF RESULTS

The classification in healthy or pathological was carried out separately for each pathology and for women and men also separately. The number of samples taken from the control group is the same as the pathological group under examination. A vector of 9 parameters (4 of Jitter, 4 of Shimmer and HNR) times three tones (High, Low and Normal), times three vowels (/a/, /i/ and /u/) was created using the algorithm developed by Teixeira and Gonçalves (2016). Once the features vector was made up 81 variables, some dimension reduction and feature selection techniques were applied. This reduces the processing time and shows up the variables capable to distinguish between healthy and pathologic subject. An ANN and the SVM were used for the classification task. For the ANN, different typologies were experimented with a different number of neurons in the hidden layer to find the best generalization model. For the SVM different kernel and values of the hyper-parameters were tested. These alternatives of the ANN and the SVM were experimented with the selected features made with hierarchical clustering method, multilinear regression analyses methods and the components of the PCA method for each pathological group.

3.1. Feature Selection Analysis

Each pathologic group and its counterpart of controls was analysed by the feature selection methods HC and MRA to select the features along the initial set of 81 features. Table 3 presents the selected features resultant from the methods HC and MRA applied for the dysphonia and vocal cords paralysis pathologies, by gender. The top line presents the pathologic group, feature selection method and gender. The left column presents the features extracted from each speech file, the number of features selected for each method and the best accuracy in the test set given by the ANN and SVM classification models. Inside the table, the selected speech file for each feature is presented in the format VT (V- vowel, T- tone). For instance, the method MRA for dysphonia male group selected the jitter feature extracted from vowel /a/ low tone and Rap from vowel /a/ high tone and achieved an accuracy of 100% in the test set.

The feature selection method with the best accuracy achieved was clearly the MRA. This method, generally, selected a small number of features. The features selected more often are: jitta for the three vowels and three tones, shim also for the three vowels and tones, and HNR for the three vowels and tones except for /u/ at high tone. This can be justified because /u/ vowel has high frequency components that are more mixed with noise components, and at high tones, the harmonic components can be more difficult to separate from noisy components. The other features were used more rarely but with improved results by the MRA selection method.

Table 3. Set of selected features for HC and MRA feature selection methods by pathology and gender

Features	Pathology/Method/Gender							
	Dysph HC Female	Dysph HC Male	Dysph MRA Female	Dysph MRA Male	VCP HC Female	VCP HC Male	VCP MRA Female	VCP MRA Male
Jitta	ul, ih, an, un, in, ah, al	in, an, il, ih, al, uh			al, un, ah, an, ul, il	un, il, ul, ih, an, ah	an	ah
Jitter				al				
Rap				ah				in
Ppq5							un	
ShdB								
Shim	il, ul, in, ih, an	ah, al, an, in	in			un, ah, al	uh	
Apq3			an, ah, uh				ah	
Apq5			in					
HNR	ih, il, ul, al	al, ul, ah, un	al		al		al, in	an
# features	16	14	6	2	7	9	6	3
Best accuracy	83.3	70.0	100.0	100.0	73.7	72.7	78.9	81.8

3.2. Dysphonia/Control Classification

Although, the accuracy has been calculated for validation and training set, only the test set accuracy is analysed here. This means that the accuracy was measured in a set that was never seen during training stage. Anyhow, this test set complies only 15% of the subject in each case.

The dysphonia/control classification was made separately for female and male subjects with the ANN and SVM.

Table 4 presents the best results between several experiments with ANN. This table shows the number of input nodes, hidden and output layers [I,H,O], the transfer function in hidden and output layers (TF-H, TF-O), the training function (TR func), the correlation coefficient (R) and Accuracy determined for the test set. The columns of the table display for each case of the selected parameters method and for female and male groups the best architecture of the ANN concerning the number of nodes, transfer function and training function. The best architecture was selected using the performance along the training and validations sets together. The final performance is presented in the table for the test set. Concerning the number of features, 4 situations were considered: using the all set of features, features selected by HC and MRA, and a vector created by the PCA method.

The transfer function used in the ANN are the tangent sigmoidal (tansig), logarithmic sigmoidal (logsig) and linear (purelin). The experimented training functions were the Levenberg-Marquardt (trainlm) based on Marquardt (1963), Resilient back-propagation (trainrp) based on Reidmiller and Braun (1993), and scale conjugate gradient (trainscg) (Moller, 1993).

Concerning the female group, the best accuracy was 100% using the MRA feature selection method. This method reduced from 81 to 6 features and yet achieved an excellent accuracy. The other situations of input features used higher number of features but accuracy remained at 83%. For the male group, the best accuracy was 90%, achieved again by the MRA and PCA methods. Anyhow, the MRA used only 2 features against the 7 used by PCA.

Table 4. Dysphonia, ANN

Input	All features		HC		MRA		PCA	
	Female	Male	Female	Male	Female	Male	Female	Male
Arch. [I,H,O]	[81,20,1]	[81,10,1]	[16,10,1]	[14,20,1]	[6,15,1]	[2,15,1]	[7,15,1]	[7,10,1]
TF-H	tansig	logsig	tansig	logsig	tansig	logsig	tansig	tansig
TF-O	purelin	purelin	purelin	purelin	purelin	purelin	purelin	purelin
TR func.	trainlm	trainlm	trainlm	trainlm	trainscg	trainlm	trainlm	trainscg
R	0.67	0.50	0.71	0.41	1.00	0.82	0.67	0.82
Accuracy	83.3	70.0	83.3	70.0	100.0	90.0	83.3	90.0

Table 5 displays the performance of the SVM for female and male dysphonia groups. The table presents the sensitivity, specificity and accuracy determined under the test set for best kernel and parameters of SVM using all features, features selected by MRA and parameters determined by PCA. The performance achieved by HC features selection method was very poor compared with the other method and therefore it was excluded from the table.

For female and male groups, the best performance was achieved again by the MRA feature selection method and with different kernel and parameters of the SVM model for female and male groups. The accuracy was of 100% using only 6 and 2 features for female and male groups, respectively. For this model, the sensibility and specificity were also 100% for both genders.

As a conclusion, for dysphonia/control classification the set of features selected for female and male groups achieve 100% accuracy using the best ANN and SVM, and for male group achieved 90% accuracy with the best ANN and 100% with the best SVM.

According to Table 3, the best set of features for the female group is:

- Shim: /i/ vowel at normal tone
- Apq3: /a/ vowel at normal and high tones; /u/ vowel at high tone
- Apq5: /i/ vowel at normal tone
- HNR: /a/ vowel at low tone

For the male group:

- Jitter: /a/ vowel at low tone
- Rap: /a/ vowel at high tone

Table 5. Dysphonia, SVM

Input	All Features		MRA		PCA	
	Female	Male	Female	Male	Female	Male
Kernel	linear	linear	linear	Gaussian	Gaussian	linear
Parameters	C = 0.1	C = 0.1	C = 0.1	S = 0.1, C = 0.2	S = 2, C = 10	C = 1
Sensitivity	100.0	100.0	100.0	100.0	83.3	100.0
Specificity	83.3	75.0	100.0	100.0	83.3	75.0
Accuracy	91.7	87.5	100.0	100.0	83.3	87.5

3.3. Vocal Cords Paralysis/Control Classification

The same type of analysis is presented for vocal cords paralysis using ANN in Table 6 and using SVM in Table 7.

Considering the results presented in Table 6 with ANN for the female group the best accuracy was 76.3% using the all set of features. The feature selection method did not get any improvements in the performance of the ANN.

For the male group, the best accuracy was achieved using the MRA feature selection method and PCA. In both cases, the best accuracy was 81.8% for test set. The MRA selected only 3 features to use in the input layer of the ANN. The PCA changes the features dimension of 81 features to 7 new features.

Table 7 presents the best results using the SVM for VCP/control classification. The data follows the same structure described for Table 5. Once more, the HC feature selection method resulted in poor results not presented here.

Concerning the results for the female group, the best accuracy was 78.9% achieved when it was used the all set of features and repeated with the feature selected by MRA method.

For the male group, the best accuracy was 80.0% achieved with the 7 parameters of PCA.

Considering both ANN and SVM models the best results for the female group was 78.9% accuracy achieved using SVM and 76.3% with the ANN. For the male group, the best accuracy was 81.8% attained by the ANN, and 80.0% achieved by SVM.

The best set of features for the female group is the all dataset or only the following features:

- Jitta: /a/ vowel at normal tone
- Ppq5: /u/ vowel at normal tone

Table 6. Vocal cords paralysis, ANN

Input	All Features		HC		MRA		PCA	
	Female	Male	Female	Male	Female	Male	Female	Male
Arch. [E,CE,S]	[81,10,1]	[81,15,1]	[7,15,1]	[10,15,1]	[6,20,1]	[3,25,1]	[7,10,1]	[7,15,1]
FTCE	tansig	logsig	tansig	logsig	tansig	tansig	tansig	logsig
FTS	purelin	purelin	purelin	purelin	purelin	purelin	purelin	purelin
FT	trainlm	trainlm	trainlm	trainlm	trainlm	trainscg	trainlm	trainlm
R	0.527	0.567	0.484	0.462	0.436	0.647	0.476	0.636
Accuracy	76.3	77.3	73.7	72.7	71.1	81.8	73.7	81.8

Table 7. Vocal cords paralysis, SVM

Input	All Features		MRA		PCA	
	Female	Male	Female	Male	Female	Male
Kernel	Polynomial	Gaussian	Polynomial	linear	Gaussian	Polynomial
Parameters	O=2, C=0.04	S=4, C=0.2	O=2, C=10	C=0.1	S=1, C=0.1	O=4, C=1
Sensitivity	84.2	80.0	89.5	90.0	84.2	80.0
Specificity	73.7	70.0	68.4	60.0	68.4	80.0
Accuracy	78.9	75.0	78.9	75.0	76.3	80.0

- Shim: /u/ vowel at high tone
- Apq3: /a/ vowel at high tone
- HNR: /a/ vowel at low tone and /i/ vowel at normal tone

For the male group:

- Jitta: /a/ vowel at high tone
- Rap: /i/ vowel at normal tone
- HNR: /a/ vowel at normal tone

A final consideration about the accuracy should be made. The accuracy was measured in the test set that consists of 15% of all dataset. Therefore, the test set consists of 12 subjects for the Dysphonia/female, 9 subjects for Dysphonia/male, 38 subjects for VCP/female and 21 subjects for VCP/male. An accuracy of 100% means that all the subjects of the test set of the pathology/gender get a correct diagnosis. However, since the number of subjects is not too high this result must be seen carefully.

4. CONCLUSION

Vocal acoustic analysis technique was applied to classify between dysphonic and healthy voices and between vocal cords paralysis and healthy voices. The classification was made for female and male with different models and architectures. The ANN and SVM were used as classifier tools. Different architectures of the ANN and SVM were experimented. The ANN and the SVM made the classification using a subset of features selected by hierarchical cluster method, multilinear regression analysis, or a set of new features obtained with the PCA method. These methods select a subset of features from the 4 measures of jitter, 4 measures of shimmer and HNR extracted from speech sound files with 3 vowels pronounced at three tones, in a total of 81 features.

The accuracy of the classification between pathologic/control measured in the test set was used to compare the models.

The best results for the classification between dysphonic/control female subjects was 100% of accuracy, achieved using ANN and SVM with the set of features selected by RMA method. For the male group, the best result was also 100% of accuracy, obtained by the SVM with the MRA features.

The MRA method reduced from 81 to 6 features for female and 2 features for male groups. The SVM showed to be very powerful to classify dysphonic voices of both genders.

The vocal cords paralysis showed to be more difficult to classify. The same methodology achieved an accuracy of 78.9% for the female group with SVM, and 81.8% for the male group with ANN.

The MRA method reduced the 81 features to 6 features for the female group with the same accuracy, and to only 3 features for the male group.

Comparing the ANN and SVM models both proved to be adequate tools to classify between control and pathologic voices with high accuracy.

Some future challenges in this work consists in extend the development of the diagnosis system to including other pathologies and then classify the pathology. The major limitation concerns with the short length of the speech database with clinical labelled pathologies.

ACKNOWLEDGMENT

The authors thank the HCIST 2017 conference for the invitation to publish this article.

REFERENCES

- Al-nasheri, A., Muhammad, G., Alsulaiman, M., & Ali, Z. (2016). Investigation of Voice Pathology Detection and Classification on Different Frequency Regions Using Correlation Functions. *Journal of Voice*, 31(1), 3–15. doi:10.1016/j.jvoice.2016.01.014 PMID:26992554
- Almeida, N. C. (2010). Sistema inteligente para diagnóstico de patologias na laringe utilizando máquinas de vetor de suporte. Universidade Federal do Rio Grande do Norte. Retrieved from <https://repositorio.ufrn.br/jspui/handle/123456789/15149>
- Barry, W. J., & Pützer, M. (2007). Saarbruecken Voice Database. *Stimmdaten Bank*. Retrieved from http://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4
- Ben-Hur, A., & Weston, J. (2010). A User's Guide to Support Vector Machines. *SourceForge*. Retrieved from <http://pymml.sourceforge.net/doc/howto.pdf>
- Bielamowicz, S., Kreiman, J., Gerratt, B., Dauer, M., & Berke, G. (1996). Comparison of Voice Analysis Systems for Perturbation Measurement. *Journal of Speech and Hearing Research*, 39(1), 126–134. doi:10.1044/jshr.3901.126 PMID:8820704
- Boersma, P., & Weenink, D. (2009). Praat Manual: doing phonetics by computer. 5.1.18. [Computer program]. Retrieved from http://www.fon.hum.uva.nl/praat/download_win.html
- Brockmann-Bauser, M. (2011). Improving jitter and shimmer measurements in normal voices [PhD Thesis]. Newcastle University.
- Cordeiro, H. T. (2017). Reconhecimento de patologias da voz usando técnicas de processamento da fala. Retrieved from <https://run.unl.pt/handle/10362/19915>
- Cruz, A. (2007). Data Mining via Redes Neurais Artificiais e Máquinas de Vetores de Suporte.
- Forero, L. A., Kohler, M., Vellasco, M., & Cataldo, E. (2015). Analysis and Classification of Voice Pathologies Using Glottal Signal Parameters. *Journal of Voice*, 30(5), 549–556. doi:10.1016/j.jvoice.2015.06.010 PMID:26474715
- Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3, 1157–1182. Retrieved from <http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf>
- Henríquez, P., Alonso, J. B., Ferrer, M. A., Travieso, C. M., Godino-Llorente, J. I., & Díaz-di-María, F. (2009). Characterization of Healthy and Pathological Voice Through Measures Based on Nonlinear Dynamics. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6), 1186–1195. doi:10.1109/TASL.2009.2016734
- Hirano, M. (1974). Morphological structure of the vocal cord as a vibrator and its variations. *Folia Phoniatrica*, 26(2), 89–94. doi:10.1159/000263771 PMID:4845615
- Markaki, M., & Stylianou, Y. (2011). Voice Pathology Detection and Discrimination Based on Modulation Spectral Features. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), 1938–1948. doi:10.1109/TASL.2010.2104141
- Marquardt, D. W. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2), 431–441. doi:10.1137/0111030
- May, R., Dandy, G., & Maier, H. (2011). Review of Input Variable Selection Methods for Artificial Neural Networks. In *Methodological Advances and Biomedical Applications*. doi:10.5772/16004
- Moller, M. F. (1993). A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. *Neural Networks*, 6(4), 525–533. doi:10.1016/S0893-6080(05)80056-5
- Netter, F. (2014). *Atlas of Human Anatomy* (6th ed.). Philadelphia, PA: Saunders Elsevier.
- Panek, D., Skalski, A., Gajda, J., & Tadeusiewicz, R. (2015). Acoustic Analysis Assessment in Speech Pathology Detection. *International Journal of Applied Mathematics and Computer Science*, 25(3), 631–643. doi:10.1515/amcs-2015-0046

- Pylypowich, A., & Duff, E. (2016). Differentiating the Symptom of Dysphonia. *The Journal for Nurse Practitioners*, 12(7), 459–466. doi:10.1016/j.nurpra.2016.04.025
- Reidmiller, M., & Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPRO algorithm. In *Proceedings of the IEEE International Conference on Neural Networks*. doi:10.1109/ICNN.1993.298623
- Rokach, L., & Maimon, O. (2005). Clustering Methods. In *Data Mining and Knowledge Discovery Handbook* (pp. 321–352). New York: Springer-Verlag; doi:10.1007/0-387-25465-X_15
- Salhi, L., Mourad, T., & Cherif, A. (2010). Voice Disorders Identification Using Multilayer Neural Network. *The International Arab Journal of Information Technology*, 177–185.
- Sellam, V., & Jagadeesan, J. (2014). Classification of Normal and Pathological Voice Using SVM and RBFNN. *Journal of Signal and Information Processing*, 5(1), 1–7. doi:10.4236/jsip.2014.51001 PMID:24513981
- Smith, L. I. (2002). A tutorial on Principal Components Analysis. Retrieved from http://www.iro.umontreal.ca/~pift6080/H09/documents/papers/pca_tutorial.pdf
- Teixeira, J. P., & Fernandes, P. O. (2014). Jitter, Shimmer and HNR classification within gender, tones and vowels in healthy voices. *Procedia Technology*, 16, 1228–1237. doi:10.1016/j.protcy.2014.10.138
- Teixeira, J. P., & Fernandes, P. O. (2015). Acoustic Analysis of Vocal Dysphonia. *Procedia Computer Science*, 64, 466–473. doi:10.1016/j.procs.2015.08.544
- Teixeira, J. P., Fernandes, P. O., & Alves, N. (2017). Vocal Acoustic Analysis - Classification of Dysphonic Voices with Artificial Neural Networks. *Procedia Computer Science*, 121, 19–26. doi:10.1016/j.procs.2017.11.004
- Teixeira, J. P., & Gonçalves, A. (2014). Accuracy of Jitter and Shimmer Measurements. *Procedia Technology*, 16, 1190–1199. doi:10.1016/j.protcy.2014.10.134
- Teixeira, J. P., & Gonçalves, A. (2016). Algorithm for jitter and shimmer measurement in pathologic voices. *Procedia Computer Science*, 100, 271–279. doi:10.1016/j.procs.2016.09.155

João Paulo Teixeira is graduated, Master and PhD in Electrical and Computers Engineering from the Faculty of Engineering of the University of Porto (FEUP). He is Professor since 1995 in the Polytechnic Institute of Bragança teaching in the areas of digital electronics, signal processing and rehabilitation engineering. He is author of 2 books, twelve chapter books, 30 articles and more than 50 conference papers, all with revision by his peers. His research interest areas are related with signal processing and signal modulation and forecasting. He has been working with artificial neural networks, speech signal processing, EEG signals for discrimination of Alzheimer's disease, ECG signal events detection and prediction of time series.

Nuno Alves completed his graduation in Biomedical Engineering and his Master Graduation in Biomedical Technology at Polytechnic Institute of Bragança.

Paula Odete Fernandes, PhD in Economics and Management, is a Professor of Management at the Polytechnic Institute of Bragança (IPB) - Portugal. She is a researcher and Scientific Coordinator of UNIAG (Applied Management Research Unit, since 2013) and researcher of Research Unit in Business Sciences (NECE-UBI, since 2010). She participated in 13 international and national projects I&D (5 as Responsible Researcher), supervised one PhD thesis and several master students (over 75 master theses). She published more than 190 publications in peer-reviewed journals, book chapters and proceedings, and has several communications at international scientific conferences and won 6 scientific awards. She serves as a member of Program Board and Organizing Committees for some Scientific Committees of International Conferences and has collaborated as a reviewer with several journals. Her current research interest is in tourism research, importance-performance analysis, management, artificial neural network, entrepreneurship, econometric modelling, marketing and strategic management, market research, corporate social responsibility and sustainable, higher education quality and applied research methods.