M. Sandra Paniagua, Carlos J. Pérez, Fernando Calle-Alonso, Carmen Salazar (2020). An Acoustic-Signal-Based Preventive Program for University Lecturers' Vocal Health, Journal of Voice, Vol 34 (1), pages 88-99, 10.1016/j.jvoice.2018.05.011

An acoustic-signal based preventive program for university lecturers' vocal health

M. Sandra Paniagua[a], Carlos J. Pérez[b], Fernando Calle-Alonso[b], and Carmen Salazar[c]

[a] Departamento de Enfermería, Universidad de Extremadura, Mérida, Spain.

[b] Departamento de Matemáticas, Universidad de Extremadura, Cáceres, Spain.

[c] Servicio de Otorrinolaringología, Hospital San Pedro de Alcántara, Cáceres, Spain.

Summary: Introduction. Professional activities of university lecturers involve continued and sustained use of the voice, leading in many cases to increased risk of developing voice disorders. Risk identification followed by the fast application of preventive or corrective measures is a key issue in this context.
Objective. Define and implement a preventive program for the vocal health of university lecturers by using acoustic features automatically extracted from voice recordings to identify risk groups and manage preventive or corrective actions.
Material and Methods. A total of 170 subjects, aged between 18 and 65, were recruited at the San Pedro de Alcántara Hospital and at the University of Extremadura in Cáceres (Spain). They formed three groups: one of 25 people suffering from vocal fold nodules, another of 25 healthy people, and the third of 120 university lecturers. Medical history and voice status assessment was performed, and voice recordings were made following a research protocol. A feature extraction, selection, and classification procedure was applied to the voice recordings to provide the best predictors for discriminating between pathological and healthy voices. The model parameters were then used to determine the lecturers' probability of suffering vocal fold nodules or other pathologies with similar dysphonic speech. These probabilities were used to classify the lecturers into three risk groups: low, medium, and high. These groups were taken as the basis to assign the lecturers to a primary, secondary, or tertiary prevention level. Different preventive or corrective actions were applied for each prevention level.
Results. The best set of predictors comprised sample entropy, correlation dimension, pitch period entropy, glottal noise excitation, and sex, achieving an overall accuracy of 92% with a random forest classifier. They all showed statistically significant differences between vocal fold nodules and healthy groups ($p<0.05$). Three out of the four best acoustic features were nonlinear, showing the importance of nonlinear dynamics for clinical practice. The model parameters were applied to the predictors of the lecturers so as to assign them to the different risk groups, leading to 60.8% (73 out of 120) of the lecturers in the low-risk group, 29.2% (35 out of 120) in the medium-risk group, and 10% (12 out of 120) in the high-risk group. The prevention levels were assigned on the basis of this classification and the medical history and laryngological evaluation of some specific subjects. A statistically significant association was found between the voice status and the assigned prevention level ($p<0.001$), with there being a clear dependence relationship (Cramér's V = 0.630).
Conclusion. It is feasible to develop and apply a preventive voice program for university lecturers that is aided by features automatically extracted from voice recordings. As the program progresses, it is expected that the information automatically provided for the assignment to prevention levels will become ever more precise. The method proposed can be extended to other voice professionals and other voice disorders.
Keywords: Acoustic feature; Preventive program; Teacher; University teaching staff; Voice care; Vocal fold nodule.

1. INTRODUCTION

Misuse or overuse of the vocal folds can damage vocal function. Some larynx problems may be caused by straining the vocal cords or injuring them by excessive speaking, clearing, coughing, screaming, singing, or talking too loudly or too softly. Dysphonia occurs after excessive and poorly executed speech through which the voice's normal tone is lost. This is especially relevant for voice professionals such as teachers, telemarketers, singers, actors, presenters, coaches…, whose work requires continued and sustained use of the voice [1, 2].

Teaching staff constitute one of the most vulnerable groups. There are five factors directly affecting their vocal load: phonation time, intensity, tone, environmental noise level, and vocal rest time. Other indirect factors that can have an influence are the individual's characteristics (including sex because of the size of the larynx and the buccal cavity), habits (smoking, sedentary lifestyle, poor hydration, inadequate nutrition…), and stress. In the development of their activity, teachers often overuse their voices, leading to increased risk of developing voice disorders [3]. Roy et al. [4] reported that teachers had a greater prevalence of dysphonia than non-teachers, and that the former were also significantly more likely than the latter to have consulted a physician or speech therapist regarding a voice disorder. Women not only had a greater lifetime prevalence of voice disorders than men, but also a greater prevalence of chronic voice disorders. Martins et al. [5] present a review of such disorders in teachers, including their main risk factors and their repercussion on professional activities. In contrast to the numerous studies that have identified risk factors that may lead to voice disorders in teachers in general, there is a minimal literature examining such factors specifically for university teaching staff [6]. University lecturers encounter similar risks of voice disorders as do primary or secondary school teachers, despite there being some differences in the conditions of their teaching practice [7, 8]. Indeed, teaching in general can be considered to be a high-risk occupation with regard to voice disorders.

There are several diseases related to the vocal folds, including nodules, polyps, Reinke's edema, and paralysis. Some of them require only speech therapy, while others require surgery. Detection of the pathology is the first step in correctly managing voice disorders. These diseases are commonly diagnosed by using invasive methods like videolaryngostroboscopy or fibre-optic nasendoscopy. In particular, the diagnosis and monitoring of the different diseases related to the voice depend on the human factor and specialized medical equipment. However, objective assessment using acoustical analysis can also be useful to help clinicians in their decision-making [9]. These techniques are non-invasive and low-cost, and can be used as a method for pre-diagnosis in many situations, although they can not substitute clinicians.

In recent years, many techniques have been developed to aid the diagnosis and monitoring of voice-related diseases. Baghay-Ravary and Beet [10] provided an overview of the state of the art in the automatic classification of voice signals for clinical purposes. These techniques are included in the general tools of computer-aided diagnosis systems, a broad concept that integrates signal processing, artificial intelligence, and statistics, using computation techniques to help health professionals in their decision-making processes [11]. They seek to maximize the information that can be extracted automatically from speech recordings by using quantitative techniques that involve pattern recognition. The phonations considered in the protocols are transformed by means of a microphone into an electrical signal that is subsequently quantified by applying feature extraction algorithms [12]. The features extracted are then used in pattern recognition algorithms to classify subjects by similarity [13].

Many studies based on acoustic parameters have been performed to discriminate pathological from healthy voices (see for example [9, 14, 15] and references therein). Acoustic parameters have also been used to discriminate teachers from the general population [16] and to assess teachers' voice quality [17]. Most of these studies use acoustic features based on detection algorithms of pitch marks (such as jitter or shimmer) which assume a linear source-filter speech model. However, the existence of nonlinear phenomena in the speech production process has been established theoretically and experimentally. According to Titze [18], voice signals can be divided into three types. Type 1 signals are nearly periodic, Type 2 signals contain strong modulations or bifurcations, and Type 3 signals are irregular and aperiodic. Traditional linear perturbation methods of voice signal analysis are only appropriate for Type 1 signals. These linear speech processing methods do not account for the two main biophysical symptoms of voice disorders – complex nonlinear aperiodicity, and turbulent non-Gaussian randomness [19]. For this reason, features based on nonlinear dynamics are of particular relevance to clinical practice.

Acoustic features could be used to define a preventive voice care program for university teaching staff. From the point of view of occupational health, the field of voice disorder prevention has received scant attention. Indeed, the public administrations and insurance companies in many countries have not even recognized many voice disorders as corresponding to the field of occupational medicine [20]. For example, in Spain only vocal fold nodules are recognized as an occupational disease for voice professionals in the Royal Decree 1299/2006. This disorder is one of the commonest vocal pathologies among teachers [21, 22]. Its recognition requires identifying risks and establishing preventive measures. A critical review of surveillance practices with respect to voice disorders in teachers was presented by Santana et al. [23]. Recently, De Oliveira et al. [24] analysed the effectiveness of a program for teachers' vocal health. Kyriakoy et al. [6] recommended the implementation of preventive vocal hygiene education programs for university teaching staff. Although several programs have been established and reported in the scientific literature, to the best of the authors' knowledge, no preventive voice care program for teaching staff based on acoustic features has yet been implemented.

The implementation of a preventive program would help to reduce the negative consequences of vocal load by helping diagnosis in a non-invasive way. In this context, early diagnosis is essential, since it facilitates the mitigation of symptoms and the optimization of treatment for timely recovery. Moreover, there could be continuous follow-up of the teachers. Also, the economic costs of sick leave could be substantially reduced, since voice problems are one of the main causes of work absenteeism among teachers. In this sense, Lages et al. [25] presented a review of absenteeism due to voice disorders in teachers from 2005 to 2015.

This paper describes an acoustic-signal based preventive program that was designed and implemented for the teaching staff of the University of Extremadura in its Program of Disease Prevention and Health Promotion. The first step was to provide the university's lecturers with an assessment of their risk of vocal fold nodules or other pathologies with similar dysphonic speech. To this end, several features were extracted from voice recordings of people suffering from nodules and of people with good vocal health to train a model with which to provide a probability for each lecturer. Then, in accordance with these probabilities, the lecturers were divided into three risk groups (low, medium, and high risk), with each group having different recommendations of actions aimed at learning patterns of vocal hygiene and vocal technique, re-educating the voice, or even referring the subject to a speech therapist or to an otorhinolaryngologist.

The rest of this paper is set out as follows. Section 2 presents the study design, the information on the participants, the protocol and equipment, the feature extraction and selection, and the statistical data analyses applied. The proposed acoustic-signal based preventive voice care program is described in Section 3. Section 4 presents the results of applying various classifiers, and their analysis. This is followed by a discussion in Section 5 which explores the significance of the results, and describes the limitations of the study. Finally, the conclusions are outlined in Section 6.

2. MATERIAL AND METHODS

2.1. Study design

This is a case-control study. Recruitment was performed at the San Pedro de Alcántara (SP) Hospital and at the University of Extremadura in Cáceres (Spain).

The research protocol was approved by the Bioethics Committees of the SP Hospital and of the University of Extremadura. All participants were properly informed and freely provided their consent.

2.2. Participants

A total of 170 persons participated in the study. Fifty of them constituted two groups: one of 25 people suffering from vocal fold nodules, and the other of 25 healthy people. In addition, 120 people were recruited from the teaching staff of the University of Extremadura. The general eligibility criteria for participation were to be volunteers, native Spanish speakers, aged from 18 to 65, and to properly perform the phonation task in the research protocol. The specific criteria for each group are described in the following paragraphs.

The vocal fold nodule (VFN) group comprised 25 people (21 women and 4 men), with mean (standard deviation) age of 42.2 (11.5) years. They were selected among the volunteers who attended the voice disorder program at the SP Hospital from January to June 2017. They underwent a medical examination consisting of a laryngological evaluation by an otorhinolaryngologist. The existence of vocal fold nodules was clearly confirmed as being the only voice pathology.

The healthy control (HC) group comprised 25 people (21 women and 4 men), with mean (standard deviation) age of 42.8 (9.9) years. They were selected among university administration staff volunteers with no voice pathology who attended the annual health check-up of the Service of Disease Prevention and Health Promotion of the University of Extremadura from January to November 2017. An otorhinolaryngologist performed a laryngological evaluation, finding that they had good vocal health status. Also, they had never suffered from any voice pathology or used the voice professionally (as teacher, singer, coach…). They were selected to match the sex and (approximately) age distributions of the diseased group.

Finally, the 120 lecturers selected (45 women and 75 men) had a mean (standard deviation) age of 48.9 (8.5) years. They worked as full time lecturers at the University of Extremadura, with a mean (standard deviation) teaching experience of 20.0 (11.6) years, and a mean (standard deviation) annual teaching load of 6.5 (2.4) hours per week. They were selected among volunteers who attended the annual health check-up of the Service of Disease Prevention and Health Promotion of the University of Extremadura from January to November 2017. An occupational medicine doctor analysed their medical histories. None of the volunteers satisfying the general eligibility criteria were discarded, independently of their vocal health status.

2.3. Protocol and equipment

Each participant filled out a short questionnaire for assessment of part of the general and specific eligibility criteria, and provided such information as sex, age, vocal health status, previous surgical interventions… The laryngological evaluation of the VFN and HC groups was performed by an otorhinolaryngologist using videostroboscopy, and an occupational medicine doctor followed the entire process, referring lecturers to specialized care when needed.

The voice recordings were made in two quiet (but not acoustically prepared) rooms in the SP Hospital and in the University of Extremadura. The vocal task was the phonation of the /a/ vowel at a comfortable pitch and loudness, as constantly as possible. This phonation had to be kept up for at least 5 seconds and in one breath. This procedure was repeated thrice per individual for the purpose of subsequent averaging. A portable computer with an external sound card (TASCAM US322) and a headband microphone (AKG 520) featuring a cardioid pattern was used. The digital recording was performed at a sampling rate of 44.1 kHz and a resolution of 16 bits/sample using the Audacity software package (release 2.0.5).

2.4. Feature extraction and selection

Ten acoustic features were considered, including linear and nonlinear ones. The linear features were a local pitch perturbation measure (jitter) and an amplitude perturbation measure (shimmer), both traditional measures based on the fundamental frequency [12], and a feature related to glottal closure – the glottal-to-noise excitation ratio (GNE) that attempts to quantify the amount of voice excitation by vocal-fold oscillations versus excitation by turbulent noise [26]. Although this last feature is linear, its calculation is not based on any prior estimation of the fundamental frequency, which is always a critical step in the presence of pathology.

Nonlinear features are of particular relevance to clinical practice, because severely pathological dysphonic voices are precisely the ones that are most likely to present highly nonlinear and random phenomena, whereas healthy voices are closer to the linear source-filter model. Three entropy-based features were considered – permutation entropy (PermutEn), fuzzy entropy (FuzzyEn), and sample entropy (SampleEn) [27] – together with D2 (correlation dimension) [28], RPDE (recurrence period density entropy) [19], DFA (detrended fluctuation analysis) [19], and PPE (pitch period entropy) [29].

Sex is one major differentiating factor for voice disorders. Women are more prone to be affected by voice disorders than men [30]. The most likely reason is the anatomic and physiological difference between both sexes. The smaller cross-sectional mass and greater tension of the female vocal folds produce a higher fundamental frequency, which renders female voice users more prone to voice disorders [12]. Sex was therefore also included as a potential predictor of voice disorders.

The feature extraction process was implemented and run in Matlab. For each voice feature, the three replicates per individual were averaged, providing a matrix of 170 rows (one per individual) and 11 columns (one per acoustic feature plus sex).

After feature extraction, a variable selection procedure was implemented to select the optimal subset of predictors based on the features of the VFN and HC groups by applying Fisher's discriminant ratio [9]. The dimensionality of the input space was reduced in accordance with Occam's razor or the parsimony principle, according to which the model with fewest features with predictive power should be given preference [31]. This avoids unnecessary or redundant predictors that may add noise, at the same time as reducing the computational effort.

2.5. Statistical data analysis

The data obtained from the acoustic feature extraction and the questionnaire were analysed statistically using R software [32]. Means, standard deviations, box plots, and contingency tables were used for the descriptive comparison of variables. When applicability conditions (normality, or normality and heteroskedasticity) were assumable, independent t-tests or an ANOVA were used. Otherwise, Mann-Whitney or Kruskal-Wallis tests were applied. In the case of applying an ANOVA or Kruskal-Wallis tests, in cases of significance, Bonferroni tests or Mann-Whitney tests with Bonferroni corrections were applied to make pairwise comparisons. For qualitative variables, when the applicability conditions were satisfied, chi-squared tests were used to analyse statistically significant differences between proportions or statistical associations. Otherwise, exact Fisher tests were used. Cramér's V was used to measure the strength of the statistical associations when they were significant. Two-sided p-values less than 0.05 were considered statistically significant.

Several of the usually best classifiers in machine learning were applied to the selected variables to estimate the probability of having vocal fold nodules or other pathologies with similar dysphonic speech. Individual and ensemble models were considered [33]. Specifically, these classifiers were K-Nearest Neighbour (KNN), Random Forest (RF), Vector Generalized Linear Model Adjoint Categories (VGLAC), Elastic-Net Regularized Generalized Linear Model (ERGLM), Averaged Neural Network (ANN), Least Squares Support Vector Machine with Polynomial Kernel (SVMP), Bagged Classification and Regression Trees (BCRT), Boosted Logistic Regression (BLR), Bayesian Additive Regression Trees (BART), Conditional Inference Random Forest (CRF), and Random Generalized Linear Model Ensemble (GLME).

The model's generalization performance was validated by stratified k-fold cross-validation [34]. This technique involves randomly splitting the dataset into k equally sized subsets preserving the same percentage for each health status. A single subset is retained as the validation data to test the model, while the remaining k-1 subsets are used as training data. This process is repeated k times, with each one of the k subsets used exactly once as the validation data. Finally, the k results are averaged to produce overall statistical measures of the classifier's performance such as accuracy, area under the ROC curve (AUC), sensitivity, and specificity. Since the study has 50 subjects, taking k=5 led to subsets with 10 subjects (5 from VFN, and 5 from HC).

A model comparison was performed, and the classifier with best accuracy under the 5-fold cross-validation framework was selected. The model parameters obtained with information from the VFN and HC groups were applied to the predictors of the lecturers, providing, for each of them, a probability of having vocal fold nodules or other pathologies with similar dysphonic speech. These probabilities were used to classify the lecturers into three risk groups in the acoustic-signal based preventive voice care program to be described in the next section.

3. ACOUSTIC-SIGNAL BASED PREVENTIVE VOICE CARE PROGRAM

Those responsible for educational centres should establish a strategy for identifying risks produced by the use of the voice, and implement the appropriate preventive measures, including training of the teaching staff. Each country has its own regulatory framework for the prevention of occupational risks, and there is no common consensus. However, there are certain international guidelines on occupational health surveillance, such as Directive 89/391/EEC in Europe and the ILO Convention 155.

The evaluation of vocal risks is usually performed empirically based on factors such as seniority in the position, classroom reverberation, class time, and vocal rest time. The lack of formal preventive programs negatively influences the vocal health of teaching staff. Motivated by this, an acoustic-signal based preventive voice care program was proposed and implemented in the Service of Disease Prevention and Health Promotion of the University of Extremadura. The program had two parts. In the first, the lecturers were classified into three risk groups, and, in the second, prevention levels were assigned and the corresponding actions were applied.

The risk assignment was based on information obtained from the acoustic features of subjects from the VFN and HC groups. A classification model was trained with the acoustic characteristics of these two groups, and then the parameters obtained were applied to the predictors of each lecturer, providing their probability of suffering from vocal fold nodules or other pathologies with similar dysphonic speech. Then, according to these probabilities, the lecturers were divided into three risk groups: low, medium, and high risk. Lecturers with a probability below 0.5 were assigned to the low-risk group, between 0.5 and 0.9 to the medium-risk group, and above 0.9 to the high-risk group. These cut-off points were chosen based on previous experiments carried out in the Program of Disease Prevention and Health Promotion of the University of Extremadura. This acoustic-feature based procedure provides a triage method that is non-invasive and low-cost. Of course, as with any other computer-aided diagnosis system, it is not error-free, but it does provide a useful tool for medical staff in their decision-making.

The actions for lecturers were based on three types of prevention: primary, secondary, and tertiary. Subjects assigned to the low-risk group are the potential users of primary prevention. It includes measures related to personal habits and the use of the voice. The recommendations regarding personal habits are aimed at taking care of vocal hygiene (not smoking, drinking water, avoiding drinks that are too cold, healthy food, regular exercise, maintaining a sufficient sleep rhythm…) [35]. The recommendations regarding the use of the voice consist of counteracting vocal overexertion, as well as avoiding the factors triggering and favouring problems. It is necessary to be aware of warning signs such as fatigue at the end of the week, changes in the tone of voice, and burning sensation and clearing of the throat. The aim is to avoid the first episode by learning patterns of vocal hygiene and vocal technique. This task is carried out through a specific course within the Faculty Training Program of the University of Extremadura.

Secondary prevention is aimed at those lecturers who present a preclinical stage, and therefore involves acting once the symptoms are appearing. Subjects assigned to the medium-risk group are the potential users of secondary prevention, which is centred on the teaching of vocal technique by a speech therapist in group sessions. The objectives of this prevention level are the control of tone and volume, as well as those of the primary prevention. On the one hand, the program offers learning and training activities to achieve the optimal tone and internalize it as the habitual tone. On the other hand, the subjects learn to control the volume without making variations in the tone (tone-volume independence), avoiding overload of the vocal folds and vocal overexertion. In this way, standardized speech patterns are built up.

Tertiary prevention is applied to lecturers diagnosed with vocal fold nodules or other pathologies with similar dysphonic speech, including those who have undergone surgery and need rehabilitation. Subjects assigned to the high-risk group are the potential users of tertiary prevention. The diagnosis must be finally confirmed by an otorhinolaryngologist. The aim is to avoid greater loss of vocal function, establishing vocal re-education as the main measure to achieve the release of laryngeal tension. This level of prevention includes some of the actions of previous levels plus personalized monitoring by a speech therapist. The main objective is the correction of altered physiology through relaxation, breathing, and articulation techniques. Relaxation in vocal therapy is aimed at achieving muscle control by performing exercises to detect tension and stiffness located in the muscles (mainly in the face, neck, and shoulder girdle). The breathing techniques are focused on the differentiation between normal breathing (with active expiration) and phonic breathing (with passive expiration) and training. With re-education techniques of the articulation organs, accurate articulation of vowels and consonants should be acquired so as to help the voice be clear, while maintaining flexibility in the mobility of the jaw and the tongue.

The procedure of assignation to risk groups helps both to prevent the first episode of vocal problems and to uncover new pathological cases. However, these measures must be complemented with others that need to be implemented by those who are responsible for such matters in the University. The main measures might be:

- Preventive measures of a technical nature. Measures that take advantage of technical advances for the design of more efficient constructions and tools. It is necessary to design classrooms with optimal acoustic levels, without excessive reverberation, isolated from outside noise, with good air quality, and with adequate temperature and humidity conditions. The installation of microphones, which help transmit sound waves, and the use of electronic boards, which avoid the use of chalk, are two important measures.
- Organizational preventive measures. Measures designed as part of an appropriate rational organization of work that allows teaching hours to be distributed favouring pauses and vocal breaks, and includes strategies to avoid stress. Training by means of specific courses

and campaigns of awareness of the importance of voice care are two important organizational aspects. Lastly, it is essential to ensure continued provision of economic and human resources for the periodic checks of general and vocal health within the preventive programs.

The next section presents the results of the proposed preventive voice care program applied in the Service of Disease Prevention and Health Promotion of the University of Extremadura.

4. RESULTS

First, the results of a comparative statistical study of the VFN and HC groups will be presented. Then, the learning process is performed with information extracted from these pathological and healthy subjects to provide an assignment of the teaching staff as to their risk of having vocal fold nodules or other pathologies with similar dysphonic speech.

4.1. Vocal fold nodules vs controls

Once the VFN group had been established, the HC group was selected to exactly match by sex and approximately match by age. There were no statistically significant differences between the mean ages of the two groups ($p=0.834$). Neither were there statistically significant differences between the proportions of smokers and non-smokers ($p=1.000$), with both groups having 16% of smokers.

The self-assessment of voice health was done using the reduced version of the Voice Handicap Index (VHI-10) [36]. This index ranges from 0 to 40, with 0 being the best voice perception and 40 the worst. There were statistically significant differences between the VHI-10 scores of the VFN and HC groups ($p<0.001$), with the members of the VFN group having a worse perception of their voices. The mean (standard deviation) of VHI-10 for the VFN group was 16.4 (10.8), and for the HC group was 1.7 (2.7).

In accordance with the feature selection procedure and classification methods described previously, the best subset of variables consisted of SampleEn, D2, PPE, GNE, and sex, which achieved an accuracy of 0.92 (95% confidence interval, 0.86-0.98) with the RF classifier. Both the sensitivity and the specificity were also 0.92. Increasing the number of acoustic features did not increase the overall accuracy. Table 1 shows the statistical measures of the classifier performance.

Classifiers	**Accuracy**	**AUC**	**Sensitivity**	**Specificity**
KNN	0.76	0.80	0.56	0.96
RF	0.92	0.94	0.92	0.92
VGLAC	0.76	0.87	0.72	0.80
ERGLM	0.82	0.92	0.76	0.88
ANN	0.74	0.86	0.72	0.76
SVMP	0.84	0.90	0.80	0.88
BCRT	0.90	0.92	0.84	0.96
BLR	0.86	0.90	0.90	0.92
BART	0.84	0.86	0.76	0.92
CRF	0.88	0.94	0.80	0.96
GLME	0.89	0.95	0.89	0.88

Table 1. Statistical measures of the classifier performance with the selected predictors.

Statistically significant differences were found between the acoustic features of the VFN and HC groups. Table 2 presents a descriptive summary and the p-values for the four acoustic features, and Figure 1 shows the corresponding box plots.

		Mean	Median	Std. Dev.	Min	Max	p
SampleEn	VFN	0.31	0.30	0.13	0.15	0.71	0.018
	HC	0.24	0.23	0.05	0.17	0.33	
D2	VFN	2.51	2.55	0.32	1.81	2.90	<0.001
	HC	2.18	2.25	0.29	1.60	2.55	
PPE	VFN	0.31	0.78	0.24	0.01	0.82	0.002
	HC	0.11	0.09	0.12	0.01	0.53	
GNE	VFN	0.91	0.91	0.03	0.86	0.98	<0.001
	HC	0.96	0.97	0.17	0.92	0.99	

Table 2. Descriptive statistics and p-values for the VFN and HC comparison.

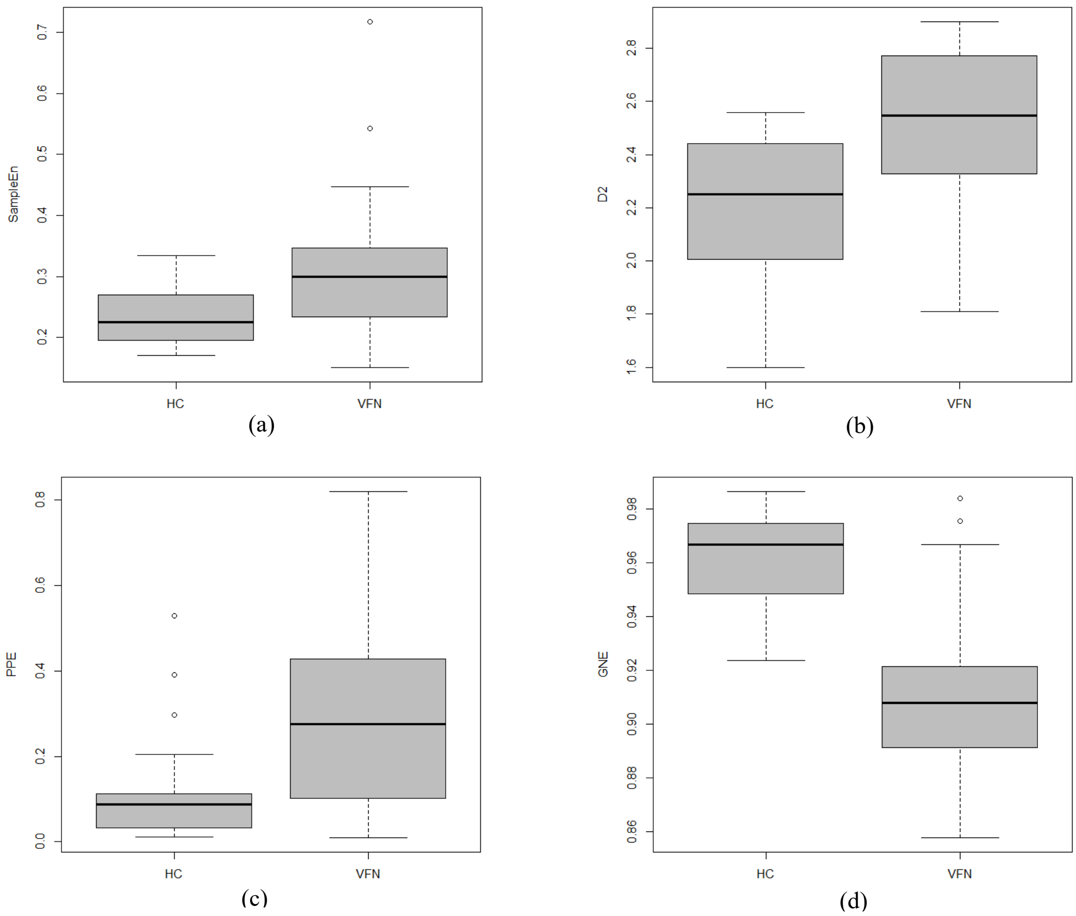

Figure 1. Box plots for the acoustic features: (a) SampleEn, (b) D2, (c) PPE, (d) GNE.

4.2. Risk assignment for teaching staff

Firstly, we shall present a brief description of some comparative results, and then follow this with the results of the learning process designed to assign the teaching staff to the different risk groups.

Statistically significant differences in the mean ages were found between the teaching staff and the VFN group ($p=0.004$) and between the teaching staff and the HC group ($p=0.008$), but not between the VFN and HC groups ($p=1.000$). The mean age (standard deviation) of the teaching staff was 48.9 (8.5) years, approximately 6 years older than the other two groups.

With regard to the smoking status, only 10% (12 out of 120) of the lecturers involved in the study were smokers. This percentage is slightly less than the 16% of the VFN and HC groups. However, the differences between the proportions of smokers were not statistically significant ($p=0.480$). Therefore, one can consider the smoking habits to be similar in the teaching staff, VFN, and HC groups.

The self-assessment of the teaching staff's voice health with VHI-10 gave a mean (standard deviation) of 3.5 (5.0) points. This mean (standard deviation) is between the 1.7 (2.7) of HC and 16.5 (10.8) of VFN. There were statistically significant differences in VHI-10 scores between the teaching staff and VFN ($p<0.001$) and between VFN and HC ($p<0.001$), but not between the teaching staff and HC ($p=0.172$). So, in general terms, the VHI-10 scores for lecturers are closer to those of the healthy subjects than to those of the subjects suffering from vocal fold nodules. Figure 2 displays the corresponding box plots for VFN, HC, and teaching staff. The outliers in the teaching staff case call for especial attention.

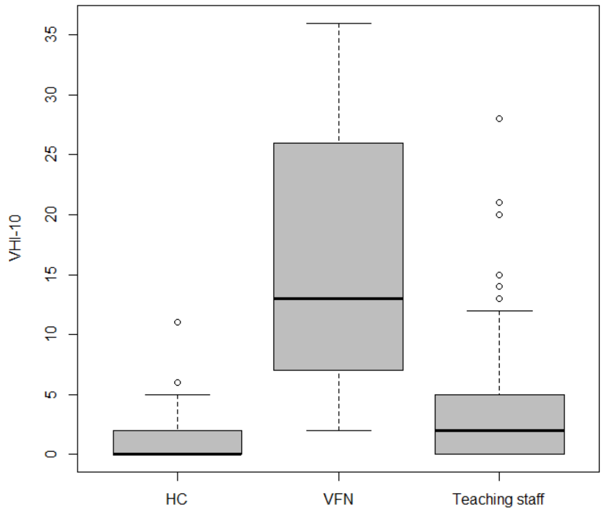

Figure 2. Box plots of VHI-10 for VFN, HC, and teaching staff.

The application of the RF classifier to the VFN and HC groups involved a learning process to recognize patterns of people suffering from vocal fold nodules and of healthy subjects. This process yielded the classifier parameters that were then applied to the teaching staff predictors (SampleEn, D2, PPE, GNE, and sex) to estimate the probability of their having vocal fold nodules or other pathologies with similar dysphonic speech. The result was that 60.8% (73 out of 120) of the lecturers were assigned to the low-risk group, 29.2% (35 out of 120) to the medium-risk group, and 10% (12 out of 120) to the high-risk group. In the questionnaire, 17 lecturers recognized the existence of voice problems, whereas 103 did not. Table 3 presents the contingency table resulting from crossing the variable indicating the recognized voice status with the assigned risk group.

There was a statistically significant association between the two variables (p<0.001), indicative of a relationship of dependence (Cramér's V=0.487).

			Risk group			
			Low	Medium	High	Total
Voice problem	No	Count	72	25	6	103
		% by row	69.9%	24.3%	5.8%	100%
	Yes	Count	1	10	6	17
		% by row	5.9%	58.8%	35.3%	100%
	Total	Count	73	35	12	120
		% by row	60.8%	29.2%	10.0%	100%

Table 3. Contingency table displaying the bivariate frequency distribution of the recognized voice status and the risk group assigned.

As with any diagnostic test, this automatic assignation to risk groups may yield misclassifications. Up to this point, only information on health status, sex, and acoustic features had been used for the lecturers. Then, a decision-making process was undertaken by the occupational medicine doctor to refer the lecturers to one of the three prevention levels. This process was based on the lecturers' medical history and a follow-up of some but not all of them.

Almost 70% of the lecturers who had not recognized that they had any voice problem were classified in the low-risk group, followed by 24.3% who were classified as medium risk. The actions involved in primary and secondary prevention were not as costly as those in tertiary prevention. Indeed, those in primary prevention were of almost zero cost. For this reason, all the lecturers assigned to the low-risk and medium-risk groups were assigned to primary and secondary prevention, respectively. One must note that even in the case that any subject of the medium-risk group had been assigned only to primary prevention there would not have been any health consequence. All teaching staff should have some proficiency in vocal techniques to control tone and volume, and only funding constraints prevented secondary prevention being made available in a generalized form. With respect to the six lecturers (3 men and 3 women) who did not recognize problems in their voice but were assigned by the model to the high-risk group, none presented any vocal problem in the voice recordings, but two had some risk factors (seniority, heavy smokers, high VHI-10 scores, and some previous episodes of vocal problems). These six were followed up by an otorhinolaryngologist to evaluate whether or not any voice disorder was indeed present or emerging, but no clinical signs were observed. The two with risk factors were therefore assigned to secondary prevention, and the other four to primary prevention.

Most of the lecturers who had recognized that they had voice problems were classified in the medium-risk group (58.8%), followed by those classified in the high-risk group (35.3%). Only one was assigned to the low-risk group (5.9%). This subject had dysphonia from chronic rhinopharyngitis, so that she should have been put into the high-risk group. She was assigned to tertiary prevention. The other subjects at medium and high risk were also evaluated by an otorhinolaryngologist to refer them to the proper prevention level. With respect to the medium-risk group (7 men and 3 women), they all had some relatively slight voice affectation such as dysphonia or throat irritation, with several episodes during the previous few years. They were therefore assigned to the secondary prevention level. With respect to the high-risk group (5 men and 1 woman), different situations were found. One had a simple throat irritation of no relevance, and had experienced no previous episode of voice problems. This subject was therefore assigned to primary prevention. Three had chronic dysphonia with repeated episodes during the previous few years, and the two remaining subjects certainly had vocal fold nodules, so these five were assigned to tertiary prevention.

Once the prevention levels had been established, the statistical association between the recognized voice status and the assigned prevention level was studied. Table 4 presents the contingency table. Again, a statistically significant association was found between the two variables ($p<0.001$), with there being a clear dependence relationship (Cramér's V = 0.630).

			Prevention level			
			Primary	Secondary	Tertiary	Total
Voice problem	No	Count % by row	76 73.8%	27 26.2%	0 0.0%	103 100%
	Yes	Count % by row	1 5.9%	11 64.7%	5 29.4%	17 100%
	Total	Count % by row	77 64.2%	38 31.7%	5 4.1%	120 100%

Table 4. Contingency table displaying the bivariate frequency distribution of the recognized voice status and the assigned prevention level.

5. DISCUSSION

In this paper, we have shown that it is possible to develop and apply a preventive voice care program for lecturers by assigning them to risk groups based on features automatically extracted from voice recordings. The objective of the vocal health program that has been implemented in the University of Extremadura is risk identification followed by the rapid application of preventive or corrective measures. This program makes it possible to avoid, or at least reduce, the exposure of lecturers to known causes of voice problems by means of appropriate intervention, as well as to avoid initiation of a voice disorder or to apply the corresponding measures in the event that the disorder was already present. This is coherent with experts' recommendations and regulatory legislation [6, 23, 24].

With the proposed approach, support for risk assessment is enhanced by means of automatic extraction of objective parameters and their subsequent analysis. To the best of the authors' knowledge, this is the first time that a preventive voice care program based on acoustic features has been applied to university lecturers. In the following paragraphs we shall discuss some relevant aspects related to the proposal and the experiment.

In this program we have focused on vocal fold nodules because they constitute one of the commonest vocal pathologies among teaching staff. Urrutikoetxea et al. [21] studied 1046 teachers and found 218 subjects with organic lesions, 94 of which were nodules, 39 Reinke's edema, 24 hypertrophy of vestibular folds, 19 polyps, and the others less frequent lesions. This means that 43.12% of the organic lesions were vocal fold nodules. In the same sense, Tavares [22] observed that 63.16% of the organic lesions found in a study of 80 teachers were vocal fold nodules. Perez et al. [37] identified some factors potentially predisposing for vocal fold nodules, and emphasized that phonotraumatic lesions, especially vocal nodules, can be diagnosed even for student teachers, reinforcing the importance of preventive vocal guidance. Besides the high prevalence of vocal fold nodules, other voice disorders, such as vocal fold polyps, cysts and other causes of functional dysphonia, share some characteristics with vocal fold nodules that can be detected by means of acoustic features. Finally, vocal fold nodules are the only voice disorder recognized as an occupational disease for voice professionals in Spain (Royal Decree 1299/2006). These were the main reasons for considering this disorder. Nevertheless, adding other voice pathologies would be of interest. To this end, another experiment adding groups of people with other pathologies could be considered in a multi-class problem. Such an increase in the number of categories to classify would, however, reduce the accuracy of the resulting classifier. Even so, an extensive study should be performed to find acoustic features that are capable of simultaneously discriminating various pathologies since this is still an open research line.

Another issue of interest is the sample size. In this experiment, 50 subjects (25 healthy, and 25 suffering from vocal fold nodules) were recruited. This seems to have been a reasonable sample size given the results that were obtained. The number of subjects in the VFN and HC groups will be increased as the preventive voice care program advances. The two groups were chosen to match in sexes and approximately in age. In general, more women have vocal problems than men [30, 38]. According to the Spanish National Institute of Statistics, in 2017, there were 12 times more women with vocal fold nodules than men. The percentage of women who had sick leave for this reason was 51.59% of those affected by vocal fold nodules, compared to 63.63% in the case of men. Vocal fold nodules are reported to be rare in adult men [38]. Clearly, there is a sex imbalance in the appearance of vocal fold nodules as was reflected in the present sample. The recruitment process for the VFN group was performed at the SP Hospital in accordance with the availability of the patients. In particular, a random assignment was not possible. The HC group was then matched so that both the VFN and HC groups consisted of 21 women and 4 men. In the proposed preventive vocal health program, sex is included as a predictor variable in the classifier.

The lack of public databases with a reasonable sample size of recordings of pathological and healthy voices made it necessary to specifically build a new database. Most have been very limited in the number of participants, and their recordings are not public due to non-disclosure agreements. The language also can be of influence. Even more important are the recording conditions. Sometimes the voice recordings were made in conditions which were so controlled that they can not be compared with others made under more realistic situations. For example, the MEEI database recordings (Massachusetts Eye and Ear Infirmary, [39]) were made in an acoustic chamber, with these conditions being very difficult to reproduce in everyday situations. Henríquez et al. [28] highlighted the differences, finding an accuracy of 99.69% with the MEEI database which was much greater than that of other results reported in the scientific literature. The database used in the present experiment was formed with recordings made in two quiet (but not acoustically prepared) rooms at the SP Hospital and at the University of Extremadura. The two rooms had similar levels of environmental noise (below 40 dB). The recording device is also very important. All the recordings were made using the same equipment. This equipment and the personnel in charge ensured that the voice recordings were of high quality.

Once the voice recordings had been properly performed, the feature extraction and selection processes were considered. A set of linear and nonlinear features were considered first. They contained the classical perturbation measures such as jitter and shimmer, a measure related to the glottal closure, and seven nonlinear features including various entropy-based measures. These last measures are appropriate for Type 2 (with strong modulations or bifurcations) and Type 3 (irregular and aperiodic) signals obtained from patients with a moderate-to-severe pathology [18]. Using a large number of features may cause problems (multicollinearity or over-fitting) when estimating the parameters in classification models, leading them to fail to generalize well for new data. The interpretation of the results also becomes more difficult. For this reason, the dimensionality of the feature space was reduced based on Fisher's discriminant ratio [9] and the parsimony principle applied to this context [14], i.e., that the number of features should be kept as low as possible while maintaining the same prediction accuracy. In some applications, it is desirable to trade off accuracy against complexity (number of features), but this was not the case in the present situation.

Some of the best classifiers, including ensemble models [33], were tested with these data. After an extensive analysis, the four features having the greatest Fisher's discriminant ratios (SampleEn, D2, PPE, and GNE) and sex provided the best results with the RF classifier. The RF is a powerful classification model based on decision trees. It provides good results in many different contexts. Fernández-Delgado et al. [40] evaluated 179 classifiers with 121 datasets from the UCI Machine Learning repository (http://archive.ics.uci.edu/ml/index.php), leading to the conclusion that the classifiers most likely to be the best are versions of RF, the best of which achieved 94.1% of the maximum accuracy, and surpassed 90% accuracy in 84.3% of the datasets. The present results are consistent with those findings in that we found an RF classifier with the selected predictors to give the best model, having an accuracy of 92% and an AUC of 94%. Adding more features to the set of predictors did not improve the accuracy with any classifier tested, but simply made the process more complex and computationally costlier.

The selected feature subset consisted of GNE and three nonlinear measures, with no conventional perturbation measures being selected. GNE compares the amount of voice excitation caused by vocal-fold oscillations with excitation by turbulent noise [26]. This feature correlates with the presence of incomplete glottal closure which generates air turbulence among the vocal folds, perceived as noise. The presence of a benign lesion characterized by an increase of mass in the vocal folds, as in the case of nodules, polyps, or cysts, may cause insufficient glottal closure, leading to air leakage. Vocal fold nodules caused by excessive vocal load or voice misuse are typically located superficially on the vibrating free edge of the fold, and thus their presence typically constrains the approximation of the vocal folds during the closed phase of the cycle producing incomplete glottal closure. Due to incomplete vocal fold closure, it is also difficult for a person with vocal fold nodules to sustain stable pitch. PPE is a nonlinear measure of fundamental frequency variation. Its great advantage in comparison with classical perturbation measures (jitter variants) is that it can distinguish natural healthy variations of fundamental frequency coming from a natural smooth vibrato from dysphonic variations produced by a voice disorder. It was introduced by Little et al. [29] for Parkinson's disease detection, since this neurological disease also produces impaired control of stabilized pitch. Finally, the feature selection process also included D2 and SampleEn, two complementary nonlinear features that characterize the complexity of a dynamic system under two different perspectives: by measuring the dimensionality of the underlying dynamics (D2) or the unpredictability within the time series (SampleEn).

Once the RF classifier had learnt from the selected predictors of the subjects in the VFN and HC groups, the model parameters were applied to the teaching staff to predict the probability of their having vocal fold nodules or other pathologies with similar dysphonic speech. The assignation to the three risk groups provided a triage method that was later reviewed by an occupational medicine doctor and an otorhinolaryngologist. The potential users of primary, secondary, or tertiary prevention are, respectively, subjects assigned to low-, medium-, or high-risk groups. However, the automatic assignation method made some mistakes. Specifically, 6 of the 103 lecturers who reported no voice problems should not have been classified as high risk, and 3 of the 17 lecturers who recognized the existence of voice problems were not assigned to the right risk group. This meant that 7.5% of the assignations had to be corrected by the occupational medicine doctor in charge. A significant association was found between the existence of voice problems and the prevention level. Therefore, based on the results obtained in this experiment, we consider the methodological approach that was applied to be promising.

The proposed way for the lecturers to be assigned to risk groups is non-invasive and low-cost. Their voices were recorded with inexpensive equipment (computer, sound card, and microphone), and the group assignation was performed with our own software specifically developed for this task. The costliest part of the preventive program was that of the actions taken in the tertiary prevention level. This included visits to the otorhinolaryngologist and personalized monitoring by a speech therapist. The primary prevention program is almost free of cost since the course offered is taught to large groups, and, furthermore, could even be posted online since the concepts are easy.

The secondary prevention program is aimed at small groups of up to 10 people. No specific cost analysis has been done since the actions were integrated into the Program of Disease Prevention and Health Promotion that covers many other occupational risks, so that the exact economic cost of the prevention program itself is difficult to quantify. With the present experiment, it is not possible to know exactly how many subjects would have taken sick leave if they had not attended the corresponding level of prevention. It is known, however, that, on average, five days of sick leave can cover the economic cost of the full treatment necessary for a subject in tertiary prevention for a month. Therefore, without taking personal welfare into account which can not be quantified financially, it is expected that the prevention program will be economically profitable.

6. CONCLUSION

University lecturers' increased risk of developing voice disorders due to their continued and sustained use of the voice makes it necessary to plan and develop a preventive voice care program to identify those risks, and provide the rapid application of preventive or corrective measures. The proposed decision-support diagnostic tool is based on features automatically extracted from voice recordings and their subsequent processing. The proposed method provided promising results that are expected to improve as the preventive voice care program progresses. The different prevention levels seek the dissemination of health education that stimulates the promotion and maintenance of vocal health, the early identification of asymptomatic subjects, and the application of corrective measures when the disease is present. To the best of the authors' knowledge, this is the first preventive voice care program that has been applied to faculty staff based on acoustic features.

The proposed method can be extended to other groups, such as telemarketers, singers, actors, presenters, coaches…, who use the voice as a work tool. Although vocal fold nodules constitute the commonest voice disorder among teaching staff, other voice disorders could be considered and a similar experiment could be developed. Finally, implementation of this approach in smartphones could provide clinical voice professionals and users considerable advantages in both health and economic terms. Nevertheless, this is currently far from realistic application, since different recording devices and environmental noise have major effects on the acoustic features that need to be measured. A robustness analysis of these acoustic features is an open line of research.

ACKNOWLEDGEMENTS

We are grateful to Antonio Moreno, Gloria Grajales, and Esther de la O for their help in the experimental part of this study. We also thank the volunteers from the SP Hospital and the University of Extremadura. This research was supported by projects MTM2014-56949-C3-3-R and MTM2017-86875-C3-2-R (MINECO), and projects IB16054 and GR15106 (Junta de Extremadura/European Regional Development Funds, EU).

REFERENCES

[1] Sataloff RT. Professional Voice: The Science and Art of Clinical Care. San Diego: Singular Publishing Group; 1997.

[2] Przysiezny PE, Przysiezny LTS. Work-related voice disorder. *Brazilian Journal of Otorhinolaryngol*. 2015;81(2):202-211.

[3] Sliwinska-Kowalska M, Niebudek-Bogusz E, Fiszer M, et al. The prevalence and risk factors for occupational voice disorders in teachers. *Folia Phoniatr Logop.* 2006; 58(2):85–101.

[4] Roy N, Merrill RM, Thibeault S, et al. Prevalence of Voice Disorders in Teachers and the General Population. *J Speech Lang Hear Res*. 2004;47(2):281-293.

[5] Martins RHG, Pereira ERBN, Hidalgo CB, et al. Voice Disorders in Teachers. A Review. *J Voice*. 2014;28(6):716-724.

[6] Kyriakou K, Petinou K, Phinikettos I. Risk Factors for Voice Disorders in University Professors in Cyprus. *J Voice.* 2018 (In press).

[7] Higgins KP. The Prevalence of Voice Disorders in University Teaching Faculty. Electronic Theses and Dissertations 286. 2006.

[8] Korn GP, De Lima Pontes AA, Abranches D, et al. Hoarseness and Risk Factors in University Teachers. *J Voice*. 2015;29(4):518.e21-518.e28.

[9] Al-nasheri A, Muhammad G, Alsulaiman M, et al. An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification. *J Voice*. 2017;31(1):113.e9–113.e18.

[10] Baghai-Ravary L, Beet SW. Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders. Springer Briefs in Electrical and Computer Engineering, Springer, New York, 2013.

[11] Calle-Alonso F, Pérez CJ, Arias-Nicolás JP, et al. Computer-aided diagnosis system: A Bayesian hybrid classification method. *Comput Methods Programs Biomed*. 2013;112(1):104-113.

[12] Baken RJ, Orlikoff RF. Clinical Measurement of Speech and Voice. Singular Thomson Learning, San Diego, 2000.

[13] Bishop MC. Pattern Recognition and Machine Learning, Springer, New York, 2010.

[14] Tsanas A, Gómez-Vilda P. Novel robust decision support tool assisting early diagnosis of pathological voices using acoustic analysis of sustained vowels. Multidisciplinary Conference of Users of Voice, Speech and Singing, pp. 3-12, Las Palmas de Gran Canaria, 2013.

[15] Forero LA, Kohler M, Vellasco MRB, et al. Analysis and classification of voice pathologies using glottal signal parameters. *J Voice*. 2016;30(5):549-556.

[16] Dehqan A, Scherer RC. Acoustic Analysis of Voice: Iranian Teachers. *J Voice*. 2013;27(5):655.e17–655.e21.

[17] Remacle A, Garnier M, Gerber S, et al. Vocal Change Patterns During a Teaching Day: Inter- and Intra-Subject Variability. *J Voice*. 2018;32(1):57-63.

[18] Titze IR. Summary Statement: Workshop on Acoustic Voice Analysis. National Center for Voice and Speech, National Center for Voice and Speech, Denver, pp 26-30, 1995.

[19] Little MA, McSharry PE, Roberts SJ, et al. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. *BioMedical Engineering OnLine*. 2007;6-23.

[20] Boltežar L, Šereg Bahar M. Voice Disorders in Occupations with Vocal Load in Slovenia. *Slovenian Journal of Public Health*. 2014;53(4):304-310.

[21] Urrutikoetxea A, Ispizua A, Matellanes F. Vocal pathology in techers: a videolaryngostroboscopic study in 1046 teachers. *Rev Laryngol Otol Rhinol*. 1995;116(4):255–262.

[22] Tavares EL, Martins RH. Vocal evaluation in teachers with or without symptoms. *J Voice*. 2007;21(4):407–414.

[23] Santana MC, Goulart BN, Chiari BM. Voice disorders in teachers: critical review on the worker's health surveillance practice. *Jornal da Sociedade Brasileira de Fonoaudiologia*. 2012;24(3):288-295.

[24] De Oliveira Bastos PRH, Hermes EC. Effectiveness of the Teacher's Vocal Health Program (TVHP) in the Municipal Education Network of Campo Grande, MS. *J Voice*. 2017 (in press).

[25] Lages Moselli LD, Assunção AA, Mesquita de Medeiros, A. Absenteeism due to voice disorders in teachers: literature review, 2005-2015, *Distúrbios da Comunicação*. 2017;29(3):579-587.

[26] Michaelis D, Gramss T, Strube HW. Glottal-to-noise excitation ratio-a new measure for describing pathological voices, *Acta Acust united Ac.*1997;83(7):700-706.

[27] Bravi A, Longtin A, and Seely AJE. Review and classification of variability analysis techniques with clinical applications. *BioMedical Engineering OnLine.* 2011;10(1):90.

[28] Henríquez P, Alonso JB, Ferrer MA, et al. Characterization of Healthy and Pathological Voice Through Measures Based on Nonlinear Dynamics. *IEEE Speech Audio Process.* 2009;17(6):1186-1195.

[29] Little MA, McSharry PE, Hunter, EJ, et al. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Trans Biomed Eng.* 2009;56:1015–1022.

[30] Russell A, Oates J, Greenwood KM. Prevalence of voice problems in teachers. *J Voice.* 1998;12:467–479.

[31] Tsanas A. Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning. PhD Thesis. University of Oxford, St. Cross College, 2012.

[32] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[33] Kuhn M. Caret: classification and regression training. Astrophysics Source Code Library, 2015.

[34] Refaeilzadeh P, Tang L, Liu H. Cross-validation. In Encyclopedia of database systems, pp. 532-538, Springer 2009.

[35] Pizolato RA, Rehder MIBC, de Castro Meneghim, et al. Impact on quality of life in teachers after educational actions for prevention of voice disorders: a longitudinal study, *Health Qual Life Outcomes*. 2013;11(1):28.

[36] Rosen CA, Lee AS, Osborne J, et al. Development and validation of the Voice Handicap Index10, *Laryngoscope*. 2010;114(9):1549-1556.

[37] Perez Fernandez CA, Preciado Lopez J. Vocal fold nodules. Risk factors in teachers. A case control study design. *Acta Otorrinolaringol Esp*. 2003;54: 253–260.

[38] Dejonckere, PH. Occupational Voice: Care and Cure. Kluger Publications, 2001.

[39] Kay Elemetrics Corporation. Massachusetts Eye and Ear Infirmary voice disorders database. Version 1.03, Lincoln Park, NJ, USA, 1994.

[40] Fernández-Delgado M, Cernadas E, Barro S, et al. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* 2014;15(1):3133-3181.