Expert Systems With Applications 82 (2017) 184–195

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Detection of different voice diseases based on the nonlinear characterization of speech signals

Carlos M. Travieso[a,*], Jesús B. Alonso[a], J.R. Orozco-Arroyave[b,c], J.F. Vargas-Bonilla[b], E. Nöth[c], Antonio G. Ravelo-García[a]

[a] Signals and Communications Department, Institute for Technological Development and Innovation in Communications, Las Palmas de Gran Canaria, Spain
[b] Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia
[c] Pattern Recognition Lab., Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany

ARTICLE INFO

Article history:
Received 13 September 2016
Revised 30 March 2017
Accepted 4 April 2017
Available online 8 April 2017

Keywords:
Nonlinear dynamic parameterization
Hidden Markov models
Laryngeal pathologies
Hypernasality
Disarhtria
Disphonia
Parkinson's disease
Speech signal

ABSTRACT

This work describes a novel methodology to characterize voice diseases by using nonlinear dynamics, considering different complexity measures that are mainly based on the analysis of the time delay embedded space. The feature space is represented with a DHMM and a further transformation of the DHMM states to a hyperdimensional space is performed. The discrimination between healthy and pathological speech signals is peformed by using a RBF-SVM which is trained following a K-fold cross-validation strategy. Results of around 99% of accuracy are obtained for three different voice disorders, disphonia due to laryngeal pathologies, hypernasality due to cleft lip and palate, and dysarthria due to Parkinson's disease.

1. Introduction

Most of the methods used by the medical community to evaluate and diagnose speech pathologies are either invasive ones, which require direct inspection of vocal folds (using laryngoscopical techniques such as fiberscope), or subjective ones in which voice quality is evaluated hearing the patient's speech, i.e. GRBAS and RBH methods (Hirano, 1981). Too, Voice Handicap Index (VHI) is another subjective method to analysis the voice quality (Jacobson et al., 1997). These techniques require trained expert doctors. The use of voice quality measures obtained from speech recordings allows to quantify voice quality and to follow the evolution of the patient. These measures are non-invasive, quick, automatic, and can improve classical techniques used in medicine.

In the last decades, several studies have provided objective measures of voice quality. The measures can be obtained from the voice signal in time, spectral or cepstral domains. There are many different measures used to evaluate the quality of speech and the most common in the literature include: fundamental frequency (Boyanov, Hadjitodorov, Teston, & Doskov, 1997), temporal variation of the fundamental frequency (jitter) (Demirhan, Unsal, Yilmaz, & Ertan, 2016; Kasuya, Endo, & Saliu, 1993), amplitude variation of the fundamental frequency (shimmer) (Demirhan et al., 2016; Kasuya et al., 1993), harmonics to noise ratio (HNR) (Yumoto, Gould, & Baer, 1982), low to high energy ratio (LHR) (Yunik & Boyanov, 1990), normalized noise energy (NNE) (Kasuya, Ogawa, Mashima, & Ebihara, 1986), glottal to noise excitation ratio (GNE) (Frohlich, Michaelis, & Strube, 1998), dynamic time warping and Itakura-Saito distorsion measure (Gu, Harris, Shrivastav, & Sapienza, 2005), among others. Using combinations of these measures, laryngeal pathologies detection systems have been developed obtaining different accuracies in the classification between healthy and pathological voices: 93.5% (Boyanov & Hadjitodorov, 1997), 85.8% (Wallen & Hansen, 1996), 75.2% (Uloza, Padervinskis, Uloziene, Saferis, & Verikas, 2015), 96.1% (Hadjitodorov & Mitev, 2002). It is not possible to compare the performance of these systems since most of them have been evaluated using different databases, moreover as reported in Saenz-Lechon, Godino-Llorente, Osma-Ruiz, and Gomez-Vilda (2006), the evaluation of the results is far from being robust and comparable.

* Corresponding author.
E-mail addresses: carlos.travieso@ulpgc.es, ctravieso@dsc.ulpgc.es (C.M. Travieso), jesus.alonso@ulpgc.es (J.B. Alonso), rafael.orozco@udea.edu.co (J.R. Orozco-Arroyave), francisco.vargas@udea.edu.co (J.F. Vargas-Bonilla), noeth@cs.fau.de (E. Nöth), antonio.ravelo@ulpgc.es (A.G. Ravelo-García).

http://dx.doi.org/10.1016/j.eswa.2017.04.012
0957-4174/

Nevertheless, most of the measures considered in the literature do not take into account nonlinearity in speech, although there are studies demonstrating that vocal fold vibration is highly influenced by nonlinearity in tissue and air movement (Titze, 1995). Other nonlinear phenomena that explain nonlinearty in pathological speech signals include abnormal vocal fold collision, increased pressure-flow in the glottis, and stress-strain curves of vocal fold tissue (Herzel, Berry, Titze, & Saleh, 1994). More evidence of nonlinear behavior of speech production can be found in Jiang, Zhang, and Stern (2001), Teager and Teager (1990) and Robertson, Zañartu, and Cook (2016).

Recent works consider this approach in order to reveal measures able to discriminate between healthy and pathological voices. Examples of this approach are based on high order statistics (HOS) (Steinecke & Herzel, 1995), AM-FM modeling of speech signal (Alonso, De Leon, Alonso, & Ferrer, 2001), and nonlinear operators (Wang, Yu, Yan, Wang, & Ng, 2016.; Little, McSharry, Roberts, Costello, & Moroz, 2007; Vaiciukynas, Verikas, Gelzinis, Bacauskiene, Minelga, & Hållander, 2015) .

Chaos theory, an area of nonlinear dynamics systems theory, applied to time series has been adopted as a new nonlinear approach to speech signal processing. The application of nonlinear techniques in speech signal processing so far are based on modeling or extraction of characteristics based on chaotic systems or/and with chaotical behaviour of dynamical systems (Lyapunov exponents, correlation dimension, etc.). The main chaotic characteristics studied are the Lyapunov exponents (Cairns, Hansen, & Kaiser, 1996; Chaitra, Mohan, & Dutt, 2013; Huang, Zhang, Calawerts, & Jiang, 2016) and dimensions of attractor, especially the correlation dimension. The correlation dimension has shown its capability to discriminate between healthy and pathological speech signals (Cairns, Hansen, & Kaiser, 1996; Calawerts, Lin, Sprott, & Jiang, 2016; Yu, Ouaknine, Revis, & Giovanni, 2001; Zhang & Jiang, 2003), and even distinguishing among different types of pathologies such as ataxic dysarthia and hyperkinetic extrapyramidal dysarthia (Calawerts et al., 2016).

In this work, three different databases are used to evaluate the speech of people with three different voice disorders: hypernasality due to cleft lip and palate (CLP), dysphonia due to organic diseases, i.e., laryngeal pathologies (LP), and dysarthria due to Parkinson's disease (PD).

The automatic detection of hypernasality started approximately in 1994, when Cairns, Hansen, and Riski (1994) proposed a characterization technique based on the Teager Energy Operator (TEO) and reported accuracies of about 98.8% considering 11 healthy and 11 simulated hypernasal speech recordings. The technique was also applied on consonant-vowel-consonant (CVC) words (hypernasal speech signals were also simulated) and the reported accuracy was about 93% (Cairns, Hansen, & Riski, 1996). Later, Vijayalakshmi and Reddy (2005) used the modified group delay functions to detect hypernasality in speech. The authors evaluated the speech from children with non-repaired CLP and reported accuracies of 100%, 88.7% and 86.66% for the Indian vowels /a/, /i/ and /u/, respectively. In Orozco-Arroyave, Arias-Londono, Vargas-Bonilla, and Nöth (2013), the authors worked with the five Spanish vowels and with the words */coco/* and */gato/*, and applied four nonlinear dynamics features to characterize hypernasal speech signals. The database used by the authors included recordings from 54 healthy children and 65 with repaired cleft lip and palate whose voice was labeled as hypernasal by a phoniatric expert. The set of nonlinear dynamics features considered by the authors included the largest Lyapunov exponent (*LLE*), Hurst exponent (*H*), Lempel-Ziv complexity (*LZC*), and correlation dimension (*CD*). The reported accuracies are around 92% for the Spanish vowels and 89% for the words.

The automatic detection of dysphonic speech signals using nonlinear dynamics techniques has been addressed in sustained vowels and in continuous speech. In Henríquez, Alonso, Ferrer, Travieso, Godino-Llorente, and Díaz-de-María (2009), sustained phonations of the vowel /a/ were evaluated using six nonlinear dynamics features including first minimum of the mutual information (FMMI), *CD*, first-order Renyi block entropy, second-order Renyi block entropy and Shannon entropy. The experiments presented by the authors were carried out on the *Massachusetts Eye & Ear Infirmary* (MEEI) database (KayPENTAX, 2005) and the reported accuracy was around 99.69%. With the same database, Vaziri, Almasganj, and Behroozmand (2010) reported accuracies of about 94.44% when both, the sustained vowel /a/ and continuous speech signals were evaluated using the *CD*. Additionally, in Arias-Londoño, Godino-Llorente, Sáenz-Lechón, Osma-Ruiz, and Castellanos-Domínguez (2010) nonlinear dynamics features and acoustics measures were merged to characterize a subset of the sustained phonations of the MEEI database (Parsa and Jamieson, 2000). The reported accuracy in the automatic classification of phonations from healthy and pathologic speakers was 98.23%. Further, in Godino-Llorente, Fraile, Sáenz-Lechón, Osma-Ruiz, and Gómez-Vilda (2009) the authors used the mel-frequency cepstral coefficients along with three noise measures to characterize continuous speech signals that are included in the same subset of the MEEI database described in Parsa and Jamieson (2000). The authors reported accuracies of 96.3% using a multi-layer perceptron neural network to decide whether a signal belongs to a healthy or pathologic speaker. Thereafter, in Orozco et al. (2012), *CD, H, LLE*, and *LZC* were considered to characterize continuous speech signals of the same subset of the MEEI database. The reported accuracy was 98.21% in the automatic classification of pathologic and healthy speech signals.

For the case of speech of people with Parkinson's disease (PD), nonlinear characterization is an emerging topic that has captured the attention of many researchers around the world. In Tsanas, Little, McSharry, and Ramig (2010) the authors used four nonlinear dynamics features along with 13 acoustic measures for the automatic detection of PD. The set of nonlinear features included *CD*, recurrence period density entropy (*RPDE*), detrended fluctuation analysis (*DFA*) and pitch period entropy (*PPE*). The database used in the work includes sustained phonations of the English vowel /ah/, repeated six times by 31 patients with PD and 8 healthy controls (HC). The reported accuracy was 91.4%; *however, it is important to highlight that the authors were not aware of the speaker independence in their experiments, thus their results are a little bit optimistic*. When recordings of the same speaker are included in training and test it allows the system to "know" several characteristics of the target speaker before evaluating his/her condition. For instance, in the speaker verification field it is well known that the mel-frequency cepstral coefficients (MFCCs) describe particular characteristics of the speaker, thus if those features are included in train and in test, the decision of the system will be very biased. A similar problem with other features happens in pathological speech assessment.

Further, the discriminant capability of different nonlinear features to classify between speech of people with PD and HC was studied in Orozco-Arroyave, Vargas-Bonilla et al. (2013). The authors performed automatic classification of sustained phonations of the five Spanish vowels uttered by 20 speakers with PD and 20 HC and reported accuracies of up to 76%. Another study focused on the detection of Parkinson's disease is presented in Orozco-Arroyave, Hönig, Arias-Londoño, Vargas-Bonilla, and Nöth (2015) where the authors introduce a methodology based on the spectral and ceptral modeling of sustained vowels and words uttered by patients with Parkinson's patients. The authors report accuracies of up to 79% when combining all the measures and the five Spanish vowels. For the case of isolated words, the accuracies range between 74% and 87% depending on the word. The

database used in that study is the same as the database used in this paper; the only difference is that here we are not considering the isolated words.

Further to the reviewed literature, there is a recent study considering linear and nonlinear techniques to model different speech pathologies, (Orozco-Arroyave, Belalcázar-Bolaños et al., 2015). In that study, the authors modeled sustained vowel phonations, isolated words, and continuous speech signals (depending on the database). They considered speech recordings from three different pathologies: Parkinson's disease, laryngeal cancer, and cleft lip and palate. The recordings were modeled, among others, with nonlinear dynamics features extracted from the time-series. The recordings for studying laringeal pathologies consisted of sustained vowel phonations and when the nonlinear measures were used, the accuracy is between 72% and 78% depending on the analyzed vowel. For the case of hypernaslity detection, the authors used recordings of the Spanish vowels and report accuracies of up to 97% also using the nonlinear features. For the automatic detection of Parkinson's disease, the authors use nonlinear features and considered also sustained phonations of the Spanish vowels and reported accuracies ranging from 72% to 79% depending on the vowel.

The reviewed literature indicates that there are two aspects that are not covered in the literature so far: (1) the modeling of sustained phonations in Parkinson's disease using nonlinear features needs more work because the obtained accuracies discriminating between Parkinson's disease vs. heatlhy controls are below 90%, and (2) the use of nonlinear features to model pathological speech signals is still an open problem because the obtained accuracies are not in the same range of other state of the art methods like those based on the stability measures or the spectral-ceptral modeling.

The advances of expert systems allow automatize many applications, and too, it has been applied to the analysis of the voice. The sequential character of the voice has defined the type of expert systems applied and developed. Among them Dynamic Time Warping (DTW), Hidden Markov Models (HMM), Support Vector Machines (SVM), Gaussian Mixture Model (GMM), Artificial Neural Networks (ANN) are was and is very method, and later, Hidden Markov Models (HMM) were applied for the decision-making task of a human expert. Due to the a characteristic of complexity of the voice signal, the use of expert systems facilitates the different analysis of the voice.

The aim of this paper is to present a new methodology, for the automatic detection of pathological speech signals. The proposed approach is based on the methods presented in Orozco-Arroyave, Belalcázar-Bolaños et al. (2015) and includes a modification with a further transformation of the extracted features. The method is based on three stages, the first comprises the characterization of the speech recordings using a set of 10 nonlinear dynamics features, the second is the transformation of the feature space using a Discrete Hidden Markov Model (DHMM), and the third is the classification between healthy and pathological speech signals using a support vector machine (SVM). During the second stage of the proposed methodology, a kernel is built to improve the accuracy of the SVM. The robustness of this method is tested using three different databases, with different utterances and with speech recordings of three different diseases. A direct comparison, using the same feature sets, is performed by doing direct classification (without any further transformation of the feature space) using a radial basis function – SVM (RBF-SVM). According to the results, the proposed approach improves the baselines in all of the cases, indicating that the HMM-based transformation is suitable to perform the automatic detection of several voice diseases.

The rest of the paper is organized as follows: Section 2 describes the nonlinear parameterization. Section 3, includes details of the proposed transformation and the classification process. In Section 4 the databases and the experimental setting are described. In Section 5, the experiments, results, and discussions are presented. Finally, the conclusions derived from this work are presented in Section 6.

2. Nonlinear parameterization

The presence of nonlinearities in the vocal fold vibration was demonstrated in Titze, Baken, and Herzel (1993). Additionally, in Titze (1995) the authors performed a classification of voice signals according to their "level of periodicity". Depending on the voice impairment, this characterization is directly related to different phenomena in speech production like nonlinear pressure-flow in the glottis, nonlinear stress-strain of vocal fold tissues, and nonlinearities associated with vocal fold collision Jiang, Zhang, and McGilligan (2006). For instance, laryngeal pathologies produce abnormal vocal fold vibration, cleft lip and palate produces abnormal resonances in the vocal tract and problems in the movement of the velum, and Parkinson's disease produces problems to control most of the limbs and muscles involved in the speech production process, e.g., tongue, jaw, vocal folds, and velum.

The nonlinear parameterization is based on the reconstruction of the phase space, which describes topological features of the dynamics of the system that produces the speech signal, i.e., vocal tract. It is generated following the embedding process that consists of representing the solutions of the differential equations that describe the dynamics of the system in the phase space. The trajectories of such a representation form different figures called *attractors*. However, in real life the equations that describe real phenomena are unknown, thus the trajectories of the attractors have to be reconstructed following an indirect procedure based on the *embedding theorem* proposed by Takens (1981). It allows the reconstruction of diffeomorphic attractors, i.e., those that hold topological properties of the dynamical system. The set of points that form the attractor is given by $S[k] = \{x[k], x[k+\tau], x[k+2\tau, \ldots, x[k+(\vartheta-1)\tau]\}$, where $k = 1, 2, \ldots, l$ and $l = N - (\vartheta - 1)\tau$, N is the number of points in the voice signal, ϑ is the dimension of the embedding space and τ is the time delay that guarantees the minimum correlation among the sate variables. The dimension ϑ is estimated following the false neighbors method (Kennel, Brown, & Abarbanel, 1992) and τ is calculated with the FMMI method described in Fraser and Swinney (1986).

The set of nonlinear measures considered in this paper allow quantifying morphological features of the reconstructed attractors, i.e., how jumbled the trajectories are. According to Titze (1995), voice signals can be grouped according to their degree of impairment and such a grouping can be modeled through nonlinear analyses, thus the more impaired the voice, the more jumbled the trajectories of its reconstructed attractor. Fig 1 illustrates this idea showing a sustained phonation uttered by a healthy speaker and its corresponding attractor (left side) and a sustained phonation uttered by a Parkinson's patient and its corresponding attractor (right side).

In this paper, a set of ten nonlinear dynamics measures are calculated: four nonlinear dynamics features and six entropy measures. The set of nonlinear dynamics features includes *CD, LLE, H*, and *LZC*. The set of entropy measures includes approximate entropy (A_E), Gaussian kernel entropy (GA_E), sample entropy (S_E), Gaussian kernel sample entropy (GS_E), *RPDE*, and *DFA*. Further details of the process to calculate these features are included in the next subsections.

2.1. Nonlinear dynamics features

Correlation dimension (CD): it is a measure of the space dimensionality occupied by the points in the reconstructed state

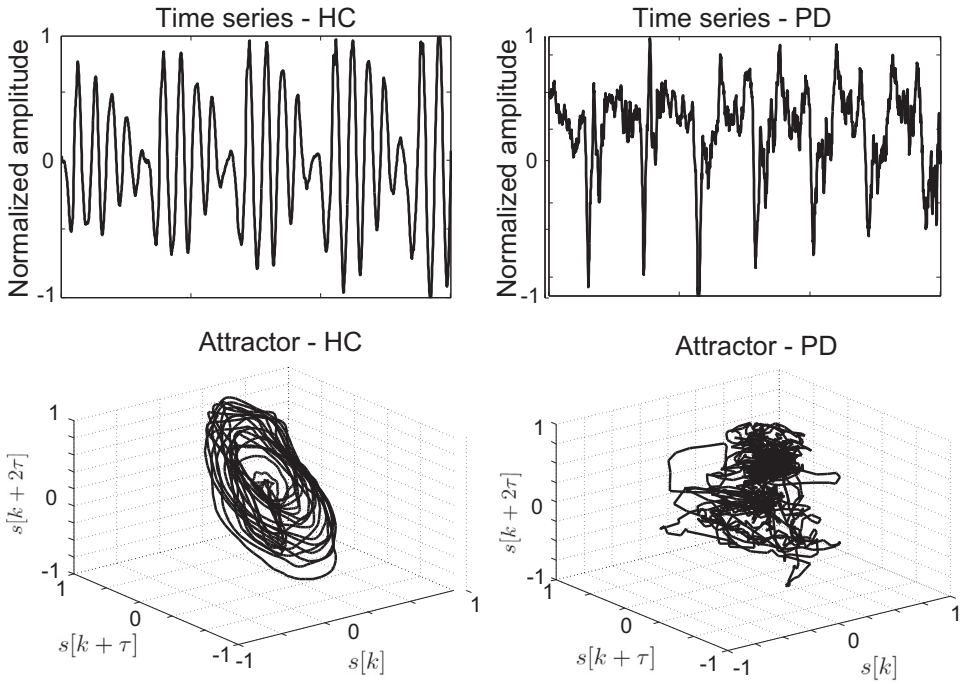

Fig. 1. Time series and the estimated attractors for health speaker (HC) and for Parkinson patient (PD).

space (attractor). *CD* is estimated according to the Takens' method (Takens, 1981). The process requires to calculate the correlation sum ($C(r)$) as in Eq. (1). $C(r)$ is the number of possible pair of points closer than a given distance r in a particular norm, and it can be interpreted as the probability of having pairs of points inside a sphere of radius r that is moving along to the trajectories of the attractor.

$$C(r) = \sum_{i=1}^{N} C_i^m(r) \tag{1}$$

Where:

$$C_i^m(r) = \frac{2}{N(N-1)} \sum_{j=i+1}^{N} \Theta\left(r - \left\| \overrightarrow{x_i} - \overrightarrow{x_j} \right\|\right) \tag{2}$$

N is the number of points in the state space, Θ is the Heaviside function, m is the dimension of the embedded space and $\| \cdot \|$ illustrates the norm defined in any consistent metric space.

CD is defined in Eq. (3) for an infinity amount of data ($N \to \infty$) and for small values of r.

$$CD = \lim_{r \to 0} \frac{\partial \ln C(r)}{\partial \ln (r)} \tag{3}$$

According to Abarbanel (2012), a proper estimation of *CD* must guarantee that the embedding dimension complies the expression $m = 2CD + 1$.

Largest Lyapunov Exponent (LLE): this feature represents the average divergence rate of neighbor trajectories in the state space. The Lyapunov spectrum reflects the sensitivity of the system to the initial conditions. When a system has at least one positive Lyapunov exponent the trajectories will diverge exponentially. *LLE* indicates whether the divergence exists.

The divergence rate of neighbor trajectories in the state space is calculated according to the Rosenstein's method (Rosenstein, Collins, & De Luca, 1993). In this algorithm the nearest neighbors to every point in the trajectories must be estimated. A neighbor must fulfill a temporal separation greater than the "period" of the time series to be considered as a nearest neighbor. It can be stated that the separation of points in a trajectory can be described by the expression $d(t) = Ce^{\lambda_1 t}$, where λ_1 is the *LLE*, $d(t)$ is the average divergence taken at the time t, and C is a constant.

If we assume that the $j-th$ pair of nearest neighbors approximately diverge at a rate of λ_1, it is possible to obtain $\ln(d_j(i)) = \ln(C_j) + \lambda_1(i\Delta t)$, where λ_1 is the slope of the average line that appears when such expression is drawn on a logarithmic plane (Kantz & Schreiber, 2004).

Hurst Exponent (H): this parameter allows the analysis of long term dynamics of a dynamical system, stating the possible long term dependencies of different elements in a given time series.

The estimation of H from a time series $x(n)$ with $n = 1, 2, \ldots, N$, is based on the rank scaling method proposed in Hurst, Black, and Simaika (1965). The authors demonstrated that the relation between the variation rank of the signal (R), evaluated in a segment, and the standard deviation of the signal (S) is given by $\frac{R}{S} = cT^H$, where c is a scaling constant, T is the duration of the segment and H is the Hurst exponent. A value of $H = 0.5$ indicates a completely uncorrelated series (Brownian time series), meaning that there is no correlation between current and future points in the time series. A value of H in the range $0 < H < 0.5$ indicates an "anti-persistent behavior", which means that the trend of the time series will be the opposite of the current.

On the other hand, a value of H in the range $0.5 < H < 1$ indicates positive auto-correlation, i.e. the trend of the time series will remain.

Lempel-Ziv Complexity (LZC): the process to calculate this feature consists on finding the number of different "patterns" present in a given sequence. The algorithm only considers binary strings, thus for the practical case of speech signals, it is necessary to assign the value of 0 when the difference between two successive samples is negative, and 1 when the difference is positive or null. The estimation of *LZC* is based on the reconstruction of a sequence X by means of the copying and insertion of symbols in a new sequence.

The binary sequence $X = x_1, x_2, \ldots, x_n$ is analyzed from left to right, the first bit of the string is taken by default as the initial point. The variable S is defined to store the bits that have been inserted, i.e. at the beginning S only has x_1. The variable Q is defined to accumulate the bits that have been analyzed from left to right in the bit stream. On each iteration, the union of S and Q (de-

noted by *SQ*) is performed. When the sequence *Q* does not belong to the string $SQ\pi$, which is the result of eliminating the last bit in the stream *SQ*, the insertion of the bits in the subset of symbols finishes. The value of *LZC* will be the number of subsets used to represent the original signal (Kaspar & Schuster, 1987).

2.2. Entropy measures

Approximate entropy (A_E): entropy can measure the uncertainty of a random variable and the most common definition is the Shannon's entropy, which is expressed as

$$H(X) = -\sum_{x\in\chi} p(x)\log p(x) \tag{4}$$

Where X is a random variable with alphabet χ and its probability mass function is $p(x)$.

One stochastic process with a set of independent, but not identically distributed variables has joint entropy with growing rate that depends on the number of variables n and is given by:

$$H(X) = -\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} H(X_i) \tag{5}$$

For state spaces (attractors), their trajectories can be partitioned into hypercubes of volume ϵ^m (in an n-dimensional space) and observed at time intervals δ, defining the Kolmogorov-Sinai entropy (H_{KS}) as Costa, Goldberger, and Peng (2005):

$$H_{KS} = -\lim_{\substack{\delta\to 0\\ \epsilon\to 0\\ n\to\infty}}\frac{1}{n\delta}\sum_{k_1,\ldots,k_n} p(k_1,\ldots,k_n)\log p(k_1,\ldots,k_n) \tag{6}$$

Where $p(k_1,\ldots,k_n)$ is the joint probability of the state of the system to be in the hypercube k_1 at the time $t=\delta$, in the hypercube k_2 at the time $t=2\delta$, etc.

For stationary processes, it can be shown that $H_{KS} = \lim_{\delta\to 0}\lim_{\epsilon\to 0}\lim_{n\to\infty}(H_{n+1}-H_n)$, where $H_{n+1} = -\lim_{n\to\infty}\frac{1}{n+1}\sum_{i=1}^{n+1} H(X_i)$ and $H_n = H(X)$.

Note that it is not possible to compute entropy for $n\to\infty$; however, there are alternative methods to approximate H_{KS}, one of them is called *approximate entropy* (A_E), which is conceived to measure the average conditional information generated by diverging points on a trajectory in the state space (Costa et al., 2005). For fixed values of m and r, A_E is estimated as

$$A_E(m,r) = \lim_{N\to\infty}\left[\phi^{m+1}(r) - \phi^m(r)\right] \tag{7}$$

Where $\phi^m(r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1}\ln C_i^m(r)$, and $C_i^m(r)$ is defined in Eq. (2).

Sample entropy (S_E): note that the estimation of A_E depends on the length of the signal because each point of the attractor is compared to each other. To overcome this problem, the sample entropy was proposed as

$$S_E(m,r) = \lim_{N\to\infty}\left(-\ln\frac{\Gamma^{m+1}(r)}{\Gamma^m(r)}\right) \tag{8}$$

Note that the estimation of $\Gamma(r)$ does not include self-comparisons of points in the state space (Xu, Wang, & Wang, 2005).

Gaussian kernel approximate entropy (GA_E): this measure uses a Gaussian kernel function to give greater weights to nearby points. The process consists in replacing the Heaviside function in C_i^m in A_E with a Gaussian kernel to measure the distance between every point in the attractor (Costa et al., 2005). The function to calculate such distance is

$$d_G\left(\vec{x_i},\vec{x_j}\right) = \exp\left(-\frac{\left\|\vec{x_i}-\vec{x_j}\right\|}{10r^2}\right) \tag{9}$$

3. Kernel based on non-linear information

In this work, two classification methods are evaluated. The first one is a classic Hidden Markov Model (HMM), the second one uses a transformation based on a Discrete Hidden Markov Model (DHMM) previous to be classified with a Support Vector Machine (SVM). The details of both squemes are presented in this section, and Fig 2 depicts the experimental methodology. Thus, HMM have different role in each method, as classifier and as transformation block, respectively. Moreover, the effect of this proposal is compared versus the use of an isolated HMM.

A hidden Markov model (HMM) is a doubly stochastic process that can only be observed through another set of stochastic processes, which produce a sequence of observations. The two stochastic process of an HMM are: one hidden process associated with the probability of transition between states (not directly observable), and one observable process associated with the probability of obtaining each of the possible values at the output depending on the current state. In this paper, we are using a particular case of HMM which is called Discrete HMM, defined in Kirk (2014) and with the following set of characteristics:

- The number of states is N and the number of different observations is M.
- The transition probability matrix is A.
- The probability vector of the starting state is π.
- The probability matrix B defines the possible states at each of the observations.

The model used here is called "Left to Right" or "Bakis" HMM, which is particularly appropriate for the evaluation of sequences. As the voice signal can be seen as a sequence of values, it can be modeled using an HMM with a single direction. This fact provides the ability to keep a certain order with respect to the observations produced on the temporary distance among the more representative changes.

In the DHMM approach, the conventional technique for quantifying features is applied. For each input data, the quantifier takes the decision about which was the most convenient value from the information of the previous input vector. To avoid taking a software decision, a fixed decision on the value quantified was made. Multi-labeling was used in order to expand the possible values that the quantifier was going to acquire, so that the possible quantified values were controlled varying this parameter. Note that the number of labels in a DHMM is related to the number of symbols per state.

DHMM algorithms should be generalized to be adjusted to the multi-labeling output ({vk} k = 1,..,C), where C is the size of the vector values codebook, in order to generate the output vector $(\{w(x_t,v_k)\}k=1,..,C)$. Therefore, for a given state j of the DHMM, the probability that a vector xt is observed in the instant t can be written as:

$$b_j(x_t) = \sum_{k=1}^{C} w(x_t,v_k)b_j(k) \tag{10}$$

where $b_j(k)$ is the output discrete probability associated with the value v_k and the state j.

3.1. HMM transformation

The Fisher evaluator transforms the DHMM states to a hyper-dimensional space (Travieso, Ticay-Rivas, Briceño, del Pozo-Baños, & Alonso, 2014). Only the gradients of the transmitted DHMM's probabilities are considered, as described in the equation: $U_X = \nabla logP(X|\lambda)$

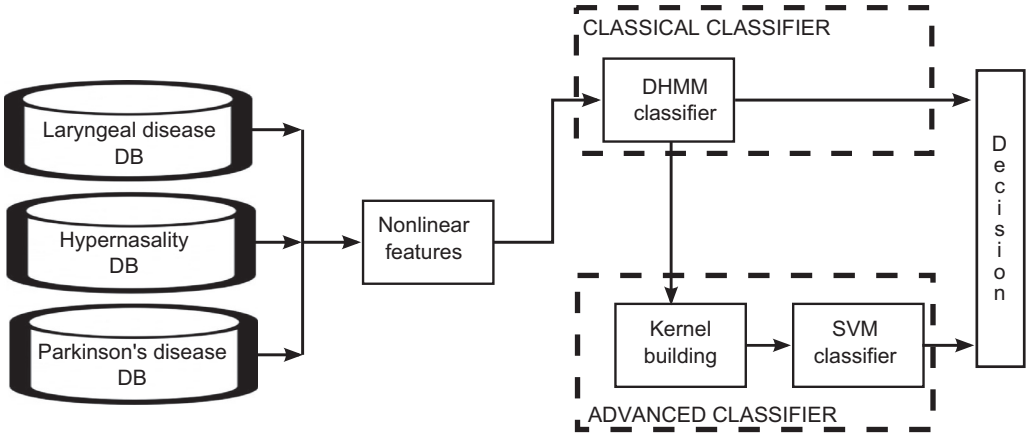

Fig. 2. Proposed experimental methodology.

1. The probability of transmitting a residual x (from the observed chain $X=(x_1,...,x_n)$) from the alphabet, while it is in the state $s \in \{s_1, ..., s_n\}$, is defined by $P(x|s,\theta) = \theta_{x|s}$.
   a. Note that:
   $$\theta_{x|s} = P(x|s,\lambda) = b_s(x) \tag{11}$$
   b. With the property:
   $$\forall s \quad \sum_x \theta_{x|s} = \sum_x P(x|s,\lambda) = \sum_x b_s(x) = 1 \tag{12}$$
   By the implementation of the DHMM
      i. Where $b_s = P(x = v_k|s)\ 1 \le k \le L$ is defined with vk the labels of the DHMM quantization.
2. The probability of transition from the state s to the state s' is defined as: $P(s'|s,\tau) = \tau_{s'|s}$.
   a. Noting that:
   $$\tau_{s'|s} = P(s'|s) = a_{s's}. \tag{13}$$
   To simplify, a unique initial state s_0 is assumed, i.e. $\pi_{s_0} = 1$. Note that, in terms of the derivation ∂X; where x is characterized by $\theta_{x|s} = b_s(x)$ (depending on the descriptors x), π_{s_i} are constants (*).
3. The defined λ DHMM assigns a probability for each sequence $X = (x_1, x_2, ..., x_T)$ given by:
$$P(X|\theta,\tau) = P(X|\lambda) = \sum_{s_1,...,s_m} \prod_i P(x_i|s_i,\lambda) P(s_i|s_{i-1},\tau)$$
$$= \sum_{s_1,...,s_m} \prod_i \theta_{x_i|s_i} \tau_{s_i|s_{i-1}}. \tag{14}$$
4. And according to (1.a) and (2.a):
$$P(X|\theta,\tau) = P(X|\lambda) = \sum_{s_1,...,s_n} \prod_i b_{s_i}(x_i) a_{s_i,s_{i-1}}$$
$$= \sum_{s_1,...,s_n} b_{s_1}(x_1) a_{s_1 s_2} b_{s_2}(x_2) ... a_{s_{n-1} s_n} b_{s_n}(x_n). \tag{15}$$
   where the sum is applied over all possible states' sequences.

The interest falls in the derivatives of $\log P(X|\theta,\tau) = \log P(X|\lambda)$ with respect to the emission probabilities $\theta_{x|s} = b_s(x)$ as commented in (*), as they are the components of the evaluator vector U_X.

By (12), the vectors $\theta_{x|s}$ are linked by the fact that the sum must be 1 for any fixed state s. In order to be able to implement independent derivations, an independent description must be implemented. To achieve this, the terms $\theta_{x|s}$ must be written in terms of a set of independent parameters:

5. $\theta_{x|s} = \frac{\theta_{x,s}}{\sum_{x'} \theta_{x',s}}$, with the values $\theta_{x,s}$ such that:
$$\left(\sum_{x'} \theta_{x',s} = 1 \right) \Rightarrow \theta_{x,s} = \theta_{x|s \cdot (*)} \tag{16}$$

Therefore, the HMM kernel (HMMK) can be defined as:

$$\frac{\delta}{\delta P(x,q)} \log P(x/q,\lambda) = \frac{\xi(x,q)}{b_q(x)} - \xi(q) \tag{17}$$

where $\xi(x,q)$ represents the number of times that the model is located in a state q during the generation of a sequence emitting a certain symbol x, and $\xi(q)$ represents the number of times that the model has been in q during the process of sequence generation (Travieso et al., 2014). These values were directly obtained from the forward backward algorithm applied to the DHMM by (Travieso et al., 2014). The application of this score U_X to the SVM is given by the following expression, using the technique of the natural gradient:

$$U_X = \nabla_{P(x,q)} log(P(x/q,\lambda)) \tag{18}$$

where U_X defines the direction of maximum slope of the logarithm of the probability of having a certain symbol in a given state.

Therefore, the proposed HMM Kernel (HMMK) is defined as the calculation of the natural distance between the scores of two sequences X and Y:

$$D^2(X,Y) = \frac{1}{2}(U_X - U_Y)^T F^{-1} (U_X - U_Y) \tag{19}$$

where F is the HMM information matrix. Note that Eq. (19) is equivalent to the covariance matrix of vectors U_X and U_Y.

Finally, the final decision will be made by Support Vector Machine (SVM) (Chandorkar, Mall, Lauwers, Suykens, & De Moor, 2015). SVM is based on a bi-class system, in other words only two classes are considered. In particular, for this present work, we have worked with 2 classes, pathological and control classes (Chandorkar et al., 2015). A linear and RBF kernels have been used with SVM.

4. Experimental methodology

4.1. Datasets

Cleft Lip and Palate (CLP): this database contains recordings of the five Spanish vowels uttered by 65 children with repaired CLP and 54 healthy controls. The age of the children ranged from 5 to 15 and the voices of the CLP speakers were labeled as hypernasal

by a phoniatric expert. Two repetitions of the sustained phonation of the five Spanish vowels are considered. The recordings were captured in noise-controlled conditions, using a sound-proof booth, with a sampling rate of 44,100 Hz and 16 bit-resolution. The ethics committee of Universidad de Caldas in Manizales (Colombia) approved this dataset. A written informed consent was given by next of kin/caregiver of children for their clinical records to be used in this study. This dataset is non-public.

Laryngeal pathologies (LP): 72 recordings of the "rainbow passage" which are part of the MEEI public database are considered. 36 speakers are patients with a variety of voice impairments such as organic and traumatic disorders, and the other 36 speakers are healthy.

Besides the continuous speech signals, sustained phonations of the vowel /a/ produced by the same group of speakers were also considered. The age of the patients ranges from 21 to 51 years and the age of the control group ranges from 22 to 52 years. The speech samples were captured using a condenser microphone, in a sound-proof booth, and the distance between the microphone and the speakers was 15 cm. The recordings were downsampled from 44,100 Hz to 25,000 Hz using the CSL system model 4300 (Yunik & Boyanov, 1990), and 16 bit-resolution. Note that this database was recorded in noise controlled conditions and using a professional setting, i.e., professional microphone, sound-card, high quality cables, and a sound-proof booth.

Parkinson's disease (PD): This database contains speech recordings of 50 patients with PD and 50 HC sampled at 44,100 Hz with 16 bit-resolution. As in the previous databases, the recordings were captured in a sound-proof booth using a professional microphone and a sound card. All of the speakers are balanced by gender and age i.e. the age of the 25 male patients ranges from 33 to 77 and the age of the 25 female patients ranges from 44 to 75. For the case of HC, the age of the 25 men ranges from 31 to 86 and the age of the 25 women ranges from 43 to 76. All of the patients were diagnosed by neurologist experts; the values of their evaluation according to the UPDRS and Hoehn & Yahr scales are 36.7 ± 18.7 and 2.3 ± 0.8, respectively. The speakers pronounced different speech tasks during the recording session including, sustained phonations of the Spanish vowels, rapid repetition of the syllables /pa-ta-ka/, i.e., Diadochokinetic evaluation, read texts, monologue, and others. In this paper we only consider the sustained phonations in order to study the impact of the further HMM-based transformation introduced here on the classification results.

The speech samples were recorded with the patients in ON-state, i.e. no more than 3 h after the morning medication. The ethics committee of Clínica Noel in Medellín (Colombia) approved this dataset. A written informed consent was given by participants for their clinical records to be used in this study.

Further details of this database can be found in Orozco-Arroyave, Arias-Londoño, Vargas-Bonilla, González-Rátiva, and Nöth (2014). In this study only the sustained vowel phonations are used because we wanted to analyze the specific effect of the DHMM transformation that is applied before the classification stage. This dataset is distributed only for research purposes and upon request.[1]

NOTE: in all of the databases considered here, only one phonation per speaker is considered, thus they are not mixed between the train and test subsets.

4.2. Experimental settings

Preprocessing: the speech recordings are divided into frames of 55 ms with an overlap of 50%. The length of these frames is chosen to guarantee that the number of points of each frame is around 1500. The analysis of the number of points was introduced in Arias-Londoño, Godino-Llorente, Sáenz-Lechón, Osma-Ruiz, and Castellanos-Domínguez (2011) to assure an appropriate number of "cycles" for the embedding process. Besides, with this windows size and from "real condition versus the noise" point of view, some features as Entropy and Hurst parameters model that noise, which is intrinsic to laryngeal pathologies; and finally, to use this number of data sequence is adequate.

Experiments: the samples are divided into training and test sets applying the $K-$Folds and Hold-Out cross-validation techniques (Shao, Er, & Wang, 2015). The system is trained and tested with totally different samples. In particular, the experiments are performed with $K = \{2, 3, 4, 5, 10\}$. For instance, when a Hold-Out cross-validation strategy is addressed, 50% of the data is considered for training and the remaining 50% for testing ($K = 2$). The composition of the data per each folder is proporcional to the number of healthy and pathological subjects. It is worth to mention that the training and testing sets were computed individually for each class. The experiments were repeated 10 times.

Three different classification approaches are tested here. The first one is based on direct classification, i.e., without any further transformation, using RBF-SVM, the second one consists on using a HMM-based model directly applied upon the feature vectors, and the third one comprises the proposed approach, which consits on the DHMM states to a hyperdimensional space by using the Fisher evaluator. The first experiment is peformed following a 5-fold cross-validation strategy, while the other two experiments considered 5 folds in the beginning while varying the number of states, and finally, when the best accuracy was found, the number of HMM states was fixed and the number of folds was varied.

Table 1
Results obtained for the three pathologies using a RBF-SVM classifier trained following a 5-fold cross-validation strategy.

Database	Corpus	RBF-SVM
LP	Spontaneous speech	$91.41\% \pm 1.01$
LP	A	$91.71\% \pm 1.65$
CLP	A	$90.56\% \pm 1.76$
CLP	E	$87.89\% \pm 3.71$
CLP	I	$89.86\% \pm 3.03$
CLP	O	$86.34\% \pm 3.45$
CLP	U	$83.80\% \pm 2.59$
PD	A	$72.49\% \pm 2.68$
PD	E	$71.58\% \pm 5.29$
PD	I	$76.81\% \pm 3.77$
PD	O	$70.23\% \pm 3.71$
PD	U	$73.36\% \pm 3.96$

5. Experimental results

The three databases considered in this study: CLP, LP, and PD, are tested on similar conditions and using different classification approaches. The baseline is stated by a direct classification, i.e., without any further transformation, using a RBF-SVM. The results are displayed in Table 1. The accuracy of classifying speakers with laryngeal pathologies is around 90%, while the accuracies for CLP and PD are around 85% and 70%, respectively. These results suggest that a feature set based only on nonlinear measures is, to some extent, able to model nonlinearities in speech production like abnormal vocal fold vibration, nonlinear preassure-flow in the glottis, stress-strain in the vocal fold tissues, and others; however, it is not suitable to accurately discriminate between healthy and pathological speech signals. There are two approaches that can be addressed here: to include more features to model

[1] The interested researcher may write an email to rafael.orozco@udea.edu.co.

Table 2
Results obtained for the three pathologies using 5-fold cross-validation.

Database	Corpus	# states	HMM	Linear SVM Kernel	RBF SVM Kernel	Υ^*
LP	Spontaneous speech	5	92.45% ± 2.59	98.41% ± 1.24	99.91% ± 0.26	2×10^{-5}
LP	Spontaneous speech	10	93.55% ± 3.24	96.87% ± 3.14	99.95% ± 0.26	2×10^{-5}
LP	Spontaneous speech	15	93.24% ± 3.21	96.65% ± 3.77	99.87% ± 0.39	4×10^{-5}
LP	Spontaneous speech	20	93.21% ± 2.56	96.73% ± 3.42	99.74% ± 0.46	4×10^{-5}
LP	A	5	80.69% ± 5.25	97.91% ± 2.99	97.04% ± 5.89	2×10^{-7}
LP	A	10	81.35% ± 5.46	99.12% ± 0.86	98.76% ± 2.69	2×10^{-7}
LP	A	15	83.07% ± 3.75	99.37% ± 0.71	99.68% ± 0.41	2×10^{-7}
LP	A	20	83.60% ± 2.51	99.41% ± 0.75	99.68% ± 0.41	2×10^{-7}
CLP	A	5	86.82% ± 3.39	98.83% ± 2.01	99.74% ± 0.42	4×10^{-5}
CLP	A	10	88.95% ± 1.64	99.55% ± 0.48	99.80% ± 0.41	8×10^{-5}
CLP	A	15	87.85% ± 2.20	99.48% ± 0.19	99.67% ± 0.30	4×10^{-5}
CLP	A	20	87.73% ± 2.30	99.48% ± 0.54	99.80% ± 2.29	4×10^{-5}
CLP	E	5	87.66% ± 1.83	99.35% ± 0.67	99.61% ± 0.87	6×10^{-5}
CLP	E	10	87.66% ± 1.79	99.55% ± 0.63	99.93% ± 0.19	8×10^{-5}
CLP	E	15	87.66% ± 2.05	99.48% ± 0.61	99.80% ± 0.29	6×10^{-5}
CLP	E	20	87.92% ± 0.70	99.42% ± 0.41	99.54% ± 0.56	6×10^{-5}
CLP	I	5	88.89% ± 1.83	99.03% ± 1.30	99.80% ± 0.40	2×10^{-5}
CLP	I	10	88.63% ± 1.20	99.81% ± 0.29	99.94% ± 0.19	2×10^{-5}
CLP	I	15	87.40% ± 2.58	99.94% ± 0.19	100% ± 0	2×10^{-5}
CLP	I	20	89.15% ± 2.34	99.68% ± 0.31	100% ± 0	2×10^{-5}
CLP	O	5	83.12% ± 1.79	99.35% ± 0.87	99.67% ± 0.42	6×10^{-5}
CLP	O	10	82.04% ± 3.07	99.41% ± 0.45	99.80% ± 0.25	6×10^{-5}
CLP	O	15	83.27% ± 3.01	99.74% ± 0.31	99.87% ± 0.26	6×10^{-5}
CLP	O	20	82.56% ± 4.09	99.42% ± 0.41	99.55% ± 0.48	6×10^{-5}
CLP	U	5	81.13% ± 4.28	98.90% ± 0.79	98.97% ± 0.76	2×10^{-5}
CLP	U	10	81.21% ± 1.88	99.35% ± 0.45	99.35% ± 0.45	2×10^{-5}
CLP	U	15	81.65% ± 3.29	99.35% ± 0.45	99.42% ± 0.45	2×10^{-5}
CLP	U	20	81.97% ± 2.18	99.48% ± 0.19	99.48% ± 0.19	2×10^{-5}
CLP	U	25	80.68% ± 6.24	99.68% ± 0.31	99.68% ± 0.31	2×10^{-5}
CLP	U	30	84.88% ± 1.35	99.48% ± 0.35	99.48% ± 0.35	2×10^{-5}
PD	A	5	61.53% ± 6.33	99.39% ± 0.40	98.24% ± 1.24	1×10^{-7}
PD	A	10	62.73% ± 4.89	98.43% ± 1.83	97.40% ± 1.35	8×10^{-8}
PD	A	15	62.41% ± 5.69	99.40% ± 0.59	99.12% ± 0.94	6×10^{-8}
PD	A	20	66.62% ± 2.79	98.84% ± 1.14	98.70% ± 1.05	1×10^{-7}
PD	E	5	58.84% ± 4.49	99.16% ± 0.77	98.01% ± 1.66	3×10^{-7}
PD	E	10	60.18% ± 2.89	99.39% ± 0.55	98.88% ± 0.65	3×10^{-7}
PD	E	15	60.74% ± 3.61	98.84% ± 1.53	97.59% ± 2.11	3×10^{-7}
PD	E	20	60.83% ± 3.17	98.98% ± 1.36	98.05% ± 2.38	3×10^{-7}
PD	E	25	61.99% ± 3.59	99.67% ± 0.54	99.30% ± 0.55	3×10^{-7}
PD	E	30	59.67% ± 3.84	99.63% ± 0.32	98.93% ± 1.65	3×10^{-7}
PD	I	5	57.54% ± 4.28	98.70% ± 0.67	98.51% ± 0.77	3×10^{-7}
PD	I	10	57.26% ± 3.71	99.44% ± 0.51	99.40% ± 0.36	1×10^{-7}
PD	I	15	55.93% ± 3.10	99.44% ± 0.46	99.68% ± 0.34	1×10^{-7}
PD	I	20	55.51% ± 2.83	99.31% ± 0.62	99.35% ± 0.55	1×10^{-7}
PD	O	5	61.25% ± 2.52	98.05% ± 1.35	98.51% ± 1.20	3×10^{-7}
PD	O	10	61.12% ± 4.75	98.37% ± 0.96	98.88% ± 0.81	3×10^{-7}
PD	O	15	62.73% ± 6.22	98.28% ± 1.30	98.70% ± 0.99	3×10^{-7}
PD	O	20	62.17% ± 2.06	99.16% ± 0.72	99.30% ± 0.69	3×10^{-7}
PD	O	25	60.83% ± 2.92	98.61% ± 0.83	99.12% ± 0.67	3×10^{-7}
PD	U	5	60.60% ± 6.16	98.75% ± 1.18	94.25% ± 2.21	2×10^{-2}
PD	U	10	60.18% ± 3.05	99.33% ± 0.61	98.47% ± 1.38	2×10^{-7}
PD	U	15	61.81% ± 2.84	99.44% ± 0.42	99.03% ± 1.41	1×10^{-7}
PD	U	20	62.45% ± 2.58	99.40% ± 0.98	99.03% ± 1.41	8×10^{-8}
PD	U	25	63.28% ± 2.77	98.89% ± 1.42	98.56% ± 1.50	4×10^{-8}

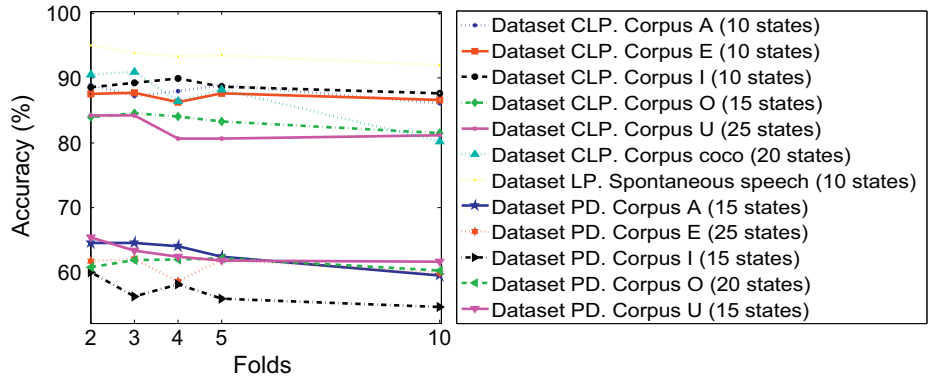

Fig. 3. Accuracy obtained with HMM classifier for three datasets for different k-folds cross-validation and number of states with optimal performance.

Table 3
Results obtained for the three datasets varing the number of folds in the validation.

Database	Corpus	k-fold	# states	HMM	linear SVM Kernel	RBF SVM Kernel	Gamma
LP	Spont. Speech	10	10	91.96% ± 2.92	94.54% ± 4.11	99.76% ± 0.52	2×10^{-5}
LP	Spont. Speech	5	10	93.55% ± 3.24	96.87% ± 3.14	99.95% ± 0.26	2×10^{-5}
LP	Spont. Speech	4	10	93.34% ± 1.93	98.90% ± 1.67	99.95% ± 0.19	2×10^{-5}
LP	Spont. Speech	3	10	93.86% ± 2.65	99.46% ± 1.26	99.95% ± 0.15	2×10^{-5}
LP	Spont. Speech	2	10	95.08% ± 3.07	99.57% ± 0.99	99.97% ± 0.08	2×10^{-5}
LP	A	10	15	77.07% ± 5.15	99.11% ± 0.92	99.48% ± 0.71	2×10^{-7}
LP	A	5	15	83.07% ± 3.75	99.37% ± 0.71	99.68% ± 0.41	2×10^{-7}
LP	A	4	15	85.61% ± 4.14	99.22% ± 0.70	99.59% ± 0.50	2×10^{-7}
LP	A	3	15	85.24% ± 4.26	99.37% ± 0.71	99.62% ± 0.46	2×10^{-7}
LP	A	2	15	89.47% ± 3.47	99.45% ± 0.64	99.68% ± 0.41	2×10^{-7}
CLP	A	10	10	86.14% ± 2.92	97.65% ± 1.59	99.26% ± 1.52	2×10^{-5}
CLP	A	5	10	88.95% ± 1.64	99.55% ± 0.48	99.80% ± 0.41	4×10^{-5}
CLP	A	4	10	88.00% ± 1.49	99.77% ± 0.47	99.85% ± 0.29	4×10^{-5}
CLP	A	3	10	87.24% ± 2.71	99.48% ± 0.78	99.65% ± 0.41	4×10^{-5}
CLP	A	2	10	88.89% ± 2.07	99.59% ± 0.59	99.80% ± 0.40	4×10^{-5}
CLP	E	10	10	86.65% ± 0.79	98.85% ± 0.76	99.25% ± 0.78	2×10^{-5}
CLP	E	5	10	87.66% ± 1.83	99.35% ± 0.67	99.61% ± 0.87	2×10^{-5}
CLP	E	4	10	86.29% ± 1.77	99.33% ± 0.58	99.70% ± 0.35	2×10^{-5}
CLP	E	3	10	87.76% ± 2.27	99.65% ± 0.69	99.91% ± 0.26	2×10^{-5}
CLP	E	2	10	87.55% ± 1.68	99.90% ± 0.31	100% ± 0	6×10^{-5}
CLP	I	10	10	87.62% ± 2.17	99.59% ± 0.22	99.71% ± 0.27	2×10^{-5}
CLP	I	5	10	88.63% ± 1.20	99.81% ± 0.29	99.94% ± 0.19	2×10^{-5}
CLP	I	4	10	89.93% ± 2.07	99.78% ± 0.33	100% ± 0	6×10^{-5}
CLP	I	3	10	89.24% ± 2.21	99.65% ± 0.41	100% ± 0	2×10^{-5}
CLP	I	2	10	88.58% ± 3.56	99.58% ± 0.49	100% ± 0	2×10^{-5}
CLP	O	10	15	81.50% ± 1.35	99.37% ± 0.62	99.54% ± 0.60	6×10^{-5}
CLP	O	5	15	83.27% ± 3.01	99.75% ± 0.35	99.79% ± 0.40	6×10^{-5}
CLP	O	4	15	84.07% ± 2.79	99.74% ± 0.31	99.87% ± 0.26	6×10^{-5}
CLP	O	3	15	84.59% ± 3.57	99.79% ± 0.40	99.92% ± 0.22	6×10^{-5}
CLP	O	2	15	84.01% ± 2.82	99.83% ± 0.32	100% ± 0	6×10^{-5}
CLP	U	10	25	81.16% ± 3.27	99.65% ± 0.25	99.65% ± 0.25	2×10^{-5}
CLP	U	5	25	80.68% ± 6.24	99.68% ± 0.31	99.68% ± 0.31	2×10^{-5}
CLP	U	4	25	80.67% ± 2.97	99.63% ± 0.35	99.63% ± 0.35	2×10^{-5}
CLP	U	3	25	84.20% ± 4.07	99.48% ± 0.39	99.48% ± 0.39	2×10^{-5}
CLP	U	2	25	84.26% ± 2.96	99.59% ± 0.59	99.59% ± 0.59	2×10^{-5}
PD	A	10	15	59.51% ± 7.56	98.14% ± 2.25	97.86% ± 1.48	3×10^{-7}
PD	A	5	15	62.41% ± 5.69	99.40% ± 0.59	99.12% ± 0.94	3×10^{-7}
PD	A	4	15	64.02% ± 2.87	99.47% ± 0.60	99.42% ± 0.71	3×10^{-7}
PD	A	3	15	64.57% ± 4.28	99.57% ± 0.46	99.44% ± 0.73	3×10^{-7}
PD	A	2	15	64.52% ± 6.05	99.48% ± 0.87	99.26% ± 0.84	3×10^{-7}
PD	E	10	25	59.91% ± 4.33	99.13% ± 0.99	98.47% ± 1.22	3×10^{-7}
PD	E	5	25	61.99% ± 3.59	99.67% ± 0.54	99.30% ± 0.55	3×10^{-7}
PD	E	4	25	58.57% ± 3.49	99.68% ± 0.63	98.67% ± 1.36	3×10^{-7}
PD	E	3	25	62.09% ± 3.28	99.87% ± 0.25	99.51% ± 0.70	3×10^{-7}
PD	E	2	25	61.70% ± 2.16	99.63% ± 0.48	99.26% ± 0.22	1×10^{-7}
PD	I	10	15	54.67% ± 2.88	99.05% ± 0.87	99.18% ± 0.43	1×10^{-7}
PD	I	5	15	55.93% ± 3.10	99.44% ± 0.51	99.68% ± 0.34	1×10^{-7}
PD	I	4	15	58.15% ± 3.89	99.52% ± 0.48	99.42% ± 0.42	1×10^{-7}
PD	I	3	15	56.23% ± 3.24	99.81% ± 0.28	99.81% ± 0.28	1×10^{-7}
PD	I	2	15	60.00% ± 4.48	99.70% ± 0.35	99.70% ± 0.35	1×10^{-7}
PD	O	10	20	60.28% ± 2.61	98.14% ± 1.56	98.68% ± 1.05	3×10^{-7}
PD	O	5	20	62.17% ± 2.06	99.16% ± 0.72	99.30% ± 0.69	3×10^{-7}
PD	O	4	20	61.95% ± 5.03	99.36% ± 0.71	99.68% ± 0.41	2×10^{-7}
PD	O	3	20	61.85% ± 6.60	99.44% ± 0.73	99.69% ± 0.56	2×10^{-7}
PD	O	2	20	60.81% ± 5.49	99.63% ± 0.35	99.77% ± 0.33	2×10^{-7}
PD	U	10	15	61.60% ± 4.18	97.50% ± 1.62	98.06% ± 1.39	3×10^{-7}
PD	U	5	15	61.81% ± 2.84	99.44% ± 0.42	99.03% ± 1.41	1×10^{-7}
PD	U	4	15	62.40% ± 4.84	99.64% ± 0.49	98.94% ± 1.73	1×10^{-7}
PD	U	3	15	63.34% ± 4.15	99.72% ± 0.59	99.38% ± 1.09	1×10^{-7}
PD	U	2	15	65.37% ± 3.93	99.78% ± 0.33	98.59% ± 1.26	1×10^{-7}

other phenomena, e.g., spectral wealth and stability/periodicity, or to perform an additional transformation prior to the classification stage. The first approach was already addressed in Orozco-Arroyave, Hönig et al. (2015), thus we decided to address the secon approach by transforming the DHMM states to a hyperdimensional space.

The experiments considering the basic HMM approach and the further transformation are performed into two stages. The first one consisted on training the models following a 5-fold cross-validation strategy while varying the number of states of the Markov models. The results obtained in this stage are displayed in the fourth column of Table 2. Note that when using the basic HMM approach for CLP and LP diseases, the accuracies are around 90% while for PD are around 60%, which is in the same range of the baseline introduced in Table 1. The results obtained with CLP recordings are similar when vowels or words are tested. Note also that the highest accuracies are obtained when the number of the HMM states in the range of 10 to 20.

In order to improve the accuracy of the system, an additional step between the characterization and the classification stages is

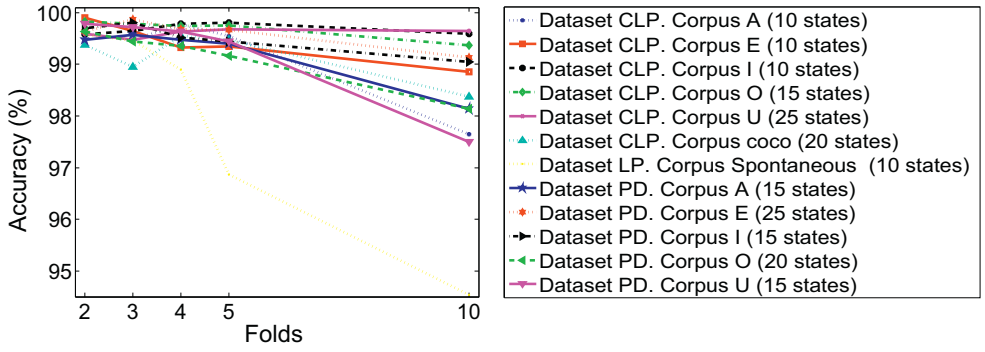

Fig 4. Accuracy obtained with the DHMM transformation and linear SVM for three datasets, different k-fold cross-validation and number of states with optimal performance.

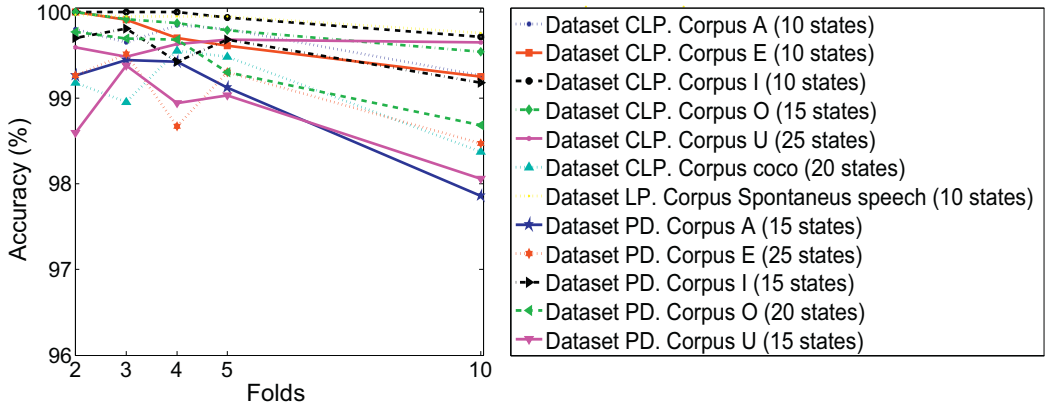

Fig 5. Accuracy obtained with the DHMM transformation and RBF SVM for three datasets, different k-fold cross-validation and number of states with optimal performance.

introduced. The experimental conditions are the same as in the previous experiments, i.e., same number and distribution of folds. The accuracies obtained when the DHMM-based transformation is applied are shown in the last three columns of Table 2. Note that two different kernels are used for the SVM, linear and RBF, γ is the bandwith of the RBF kernel. The results show that the further transformation of the feature space improve the accuracies in all of the cases; however, the stability and generalization capability of the method need to be evaluated by varying the distribution of the folds, i.e., the number of speakers in the train and test sets.

Table 3 shows the results obtained when the number of states is kept constant accoding to the highest accuarcy per speech task obtained in the previous experiments. The diference in the accuracy between K = 2 and K = 10 folds is very small (aproximately one percentage point), thus the results are stable when the number of training samples is varied. This result indicates that the nonlinear parameters are invariant and have a robust behaviour when the number of training samples is changed.

In general, the results of the direct classification based on the RBF-SMV and the HMM are relatively good for CLP and LP; however, for PD the results are not satisfactory. In order to improve the reults, it was necessary to perform additional transformations to the representation space. When the DHMM-based transformation is used the results improve not only in accuracy but in robustness and consistency, i.e., similar results are obtained along the three voice diseases considered here.

Fig 3 depicts the evolution of the recognition rate for the DHMM and number of states with maximum performance. Figs 4 and 5 show the accuracy of the classifier based on DHMM transformation with a linear and RBF SVM, respectively.

In general, the number of training samples of each fold is independent on the results, indicating that the nonlinear parameters with the further DHMM-based transformation are robust and stable detecting voice diseases. The prosposed approach is tested by using two different classifiers (linear SVM and RBF-SVM) and the results are similar.

The use of DHMM+SVM improves the accuracy compared to the direct use of a RBF-SVM and the HMM-based classifiers. This improvement can be explained due to the expansion of the dimensionality of the feature space. This increase in the dimension is improving the discriminant capability of the features. The results indicate that the combination of nonlinear parameters and the classifier applied to voice pathology detection is a good alternative. Our approach improves the results obtained in the state-of-the-art and seems to be the best option to detect different the voice diseases like hypernasality due to cleft lip and palate, dysphonia due to laryngeal pathologies and dysarthria due to Parkinson's disease.

This work has been developed on Matlab r2012a, using a personal computer with a Core i5 processor, 4 GB of RAM Memory, and 500GB HDD. Under these conditions, a recording with a duration of 3 s takes 122 s to be modeled with the no-linear parameters. The transformation performed upon the parameters spends 194 s per utterance and its classification with an RBF-SVM classifier takes 9.3 s. The timing is similar among the corpuses considered in this study.

6. Conclusions

In this work we propose a novel methodology for an automatic detection of voice diseases. The method consits on the transformation of DHMM states to a hyperdimensional space by using the Fisher evaluator. Afther such a transformation, the classification is peformed by using a RBF-SVM which is trained following a K-fold

cross-validation strategy. Results of around 99% of accuracy are obtained for three different voice disease datasets.

Linear methods or the combination of linear and nonlinear methods have been used in the state-of-the-art to detect voice pathologies in sustained phonations or in continous speech signals. For this approach, three datasets with different diseases were used: clef lip and palate which produces hypernasality in speech, Parkinson's disease which produces dysarthric speech, and laryngeal pathologies which produce disphonia. The results indicate that the proposed approach seems to be suitable and robust to detect all of these pathologies.

The proposed approach is compared with respect to other classification methods typically used in the state of the art. Particularly, a direct classification, i.e., without any further transformation, is peformed with an RBF-SVM and also with an HMM-based classfier. Both approaches showed reduced accuracies compare to those obtained with the DHMM-based transformation.

The main drawback of the methodology introduced here is the acoustic conditions of the speech recordings. The experiments presented here are based on speech samples captured in noise-controlled conditions, and according to our preliminary experiments with recordings captured in non-controlled conditions, the accuracy decreases. It is necessary to address further research on this topic in order to state which are the optimal recording conditions such that allow the application of the proposed methodology on different noise environments.

References

Abarbanel, H. (2012). *Analysis of observed chaotic data*. Springer Science & Business Media.

Alonso, J. B., De Leon, J., Alonso, I., & Ferrer, M. A. (2001). Automatic detection of pathologies in the voice by HOS based parameters. *EURASIP Journal on Applied Signal Processing, 4*, 275–284.

Arias-Londoño, J. D., Godino-Llorente, J. I., Sáenz-Lechón, N., Osma-Ruiz, V., & Castellanos-Domínguez, G. (2011). Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients. *IEEE Transactions on Biomedical Engineering, 58*(2), 370–379.

Arias-Londoño, J. D., Godino-Llorente, J. I., Sáenz-Lechón, N., Osma-Ruiz, V., & Castellanos-Domínguez, G. (2010). An improved method for voice pathology detection by means of a HMM-based feature space transformation. *Pattern Recognition, 43*(9), 3100–3112.

Boyanov, B., & Hadjitodorov, S. (1997). Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases. *IEEE Engineering in Medicine and Biology Magazine, 16*(4), 74–82.

Boyanov, B., Hadjitodorov, S., Teston, B., & Doskov, D. (1997). Robust hybrid pitch detector for pathologic voice analysis. In *Larynx: 97* (pp. 55–58). International Speech Communication Association.

Cairns, D. A., Hansen, J. H., & Kaiser, J. F. (1996). Recent advances in hypernasal speech detection using the nonlinear teager energy operator. In *Spoken language, 1996. ICSLP 96. Proceedings fourth international conference on: Vol. 2* (pp. 780–783). IEEE.

Cairns, D. A., Hansen, J. H., & Riski, J. E. (1994). Detection of hypernasal speech using a nonlinear operator. In *Engineering in medicine and biology society, 1994. Engineering advances: New opportunities for biomedical engineers. Proceedings of the 16th annual international conference of the IEEE* (pp. 253–254). IEEE.

Cairns, D. A., Hansen, J. H., & Riski, J. E. (1996). A noninvasive technique for detecting hypernasal speech using a nonlinear operator. *IEEE Transactions on Biomedical Engineering, 43*(1), 35.

Calawerts, W. M., Lin, L., Sprott, J. C., & Jiang, J. J. (2016). Using rate of divergence as an objective measure to differentiate between voice signal types based on the amount of disorder in the signal. *Journal of Voice*.

Chaitra, N., Mohan, D. M., & Dutt, D. N. (2013). Nonlinear dynamical analysis of speech signals. In *Proceedings of international conference on VLSI, communication, advanced devices, signals & systems and networking (VCASAN-2013)* (pp. 343–351). India: Springer.

Chandorkar, M., Mall, R., Lauwers, O., Suykens, J. A., & De Moor, B. (2015). Fixed–size least squares support vector machines: Scala implementation for large scale classification. In *Computational intelligence, 2015 IEEE symposium series on* (pp. 522–528). IEEE.

Costa, M., Goldberger, A. L., & Peng, C. K. (2005). Multiscale entropy analysis of biological signals. *Physical Review E, 71*(2), 021906.

Demirhan, E., Unsal, E. M., Yilmaz, C., & Ertan, E. (2016). Acoustic voice analysis of young Turkish speakers. *Journal of Voice, 30*(3), 378.e21.

Fraser, A. M., & Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information. *Physical Review A, 33*(2), 1134.

Frohlich, M., Michaelis, D., & Strube, H. W. (1998). Acoustic "breathiness measures" in the description of pathologic voices. In *Acoustics, speech and signal processing, 1998. Proceedings of the 1998 IEEE international conference on: Vol. 2* (pp. 937–940). IEEE.

Godino-Llorente, J. I., Fraile, R., Sáenz-Lechón, N., Osma-Ruiz, V., & Gómez-Vilda, P. (2009). Automatic detection of voice impairments from text-dependent running speech. *Biomedical Signal Processing and Control, 4*(3), 176–182.

Gu, L., Harris, J. G., Shrivastav, R., & Sapienza, C. (2005). Disordered speech assessment using automatic methods based on quantitative measures. *EURASIP Journal on Advances in Signal Processing, 2005*(9), 1–10.

Hadjitodorov, S., & Mitev, P. (2002). A computer system for acoustic analysis of pathological voices and laryngeal diseases screening. *Medical engineering & physics, 24*(6), 419–429.

Henríquez, P., Alonso, J. B., Ferrer, M. A., Travieso, C. M., Godino-Llorente, J. I., & Díaz-de-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.

Herzel, H., Berry, D., Titze, I. R., & Saleh, M. (1994). Analysis of vocal disorders with methods from nonlinear dynamics. *Journal of Speech, Language, and Hearing Research, 37*(5), 1008–1019.

Hirano, M. (1981). *Clinical examination of voice*: Vol. 5. Springer.

Huang, N., Zhang, Y., Calawerts, W., & Jiang, J. J. (2016). Optimized nonlinear dynamic analysis of pathologic voices with laryngeal paralysis based on the minimum embedding dimension. *Journal of Voice*.

Hurst, H. E., Black, R. P., & Simaika, Y. M. (1965). *Long-term storage: An experimental study*. Constable.

Jiang, J. J., Zhang, Y., & McGilligan, C. (2006). Chaos in voice, from modeling to measurement. *Journal of Voice, 20*(1), 2–17.

Jiang, J. J., Zhang, Y., & Stern, J. (2001). Modeling of chaotic vibrations in symmetric vocal folds. *The Journal of the Acoustical Society of America, 110*(4), 2120–2128.

Jacobson, B. H., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., Benninger, M. S., et al. (1997). The Voice Handicap Index (VHI) Development and Validation. *American Journal of Speech-Language Pathology, 6*(3), 66–70.

Kantz, H., & Schreiber, T. (2004). *Nonlinear time series analysis*: Vol. 7. Cambridge University Press.

Kaspar, F., & Schuster, H. G. (1987). Easily calculable measure for the complexity of spatiotemporal patterns. *Physical Review A, 36*(2), 842.

Kasuya, H., Endo, Y., & Saliu, S. (1993). Novel acoustic measurements of jitter and shimmer characteristics from pathological voice. *Third European conference on speech communication and technology*.

Kasuya, H., Ogawa, S., Mashima, K., & Ebihara, S. (1986). Normalized noise energy as an acoustic measure to evaluate pathologic voice. *The Journal of the Acoustical Society of America, 80*(5), 1329–1334.

KayPENTAX. (1996–2005). *Kay elemetrics disordered voice database, model 4337*. Lincoln Park, NJ, USA: Kay Elemetrics.

Kennel, M. B., Brown, R., & Abarbanel, H. D. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A, 45*(6), 3403.

Kirk, M. (2014). *Thoughtful machine learning: A test-driven approach*. O'Reilly Media, Inc.

Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A., & Moroz, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine, 6*(1), 1.

Orozco, J. R., Vargas, J. F., Alonso, J. B., Ferrer, M. A., Travieso, C. M., & Henríquez, P. (2012). Voice pathology detection in continuous speech using nonlinear dynamics. In *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on* (pp. 1030–1033). IEEE.

Orozco-Arroyave, J. R., Arias-Londono, J. D., Vargas-Bonilla, J. F., & Nöth, E. (2013). Analysis of speech from people with Parkinson's disease through nonlinear dynamics. In *International conference on nonlinear speech processing* (pp. 112–119). BerlinHeidelberg: Springer.

Orozco-Arroyave, J. R., Vargas-Bonilla, J. F., Arias-Londoño, J. D., Murillo-Rendón, S., Castellanos-Domínguez, G., & Garcés, J. F. (2013). Nonlinear dynamics for hypernasality detection in spanish vowels and words. *Cognitive Computation, 5*(4), 448–457.

Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., & Nöth, E. (2015). Spectral and cepstral analyses for Parkinson's disease detection in spanish vowels and words. *Expert Systems, 32*(6), 688–697.

Orozco-Arroyave, J. R., Belalcázar-Bolaños, E. A., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Skodda, S., & Rusz, J. (2015). Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases. *IEEE Journal of Biomedical and Health Informatics, 19*(6), 1820–1828.

Orozco-Arroyave, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J. F., González-Rátiva, M. C., & Nöth, E. (2014). New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. *Proceedings of the 9th LREC*, 342–347.

Parsa, V., & Jamieson, D. G. (2000). Identification of pathological voices using glottal noise measures. *Journal of Speech, Language, and Hearing Research, 43*(2), 469–485.

Robertson, D., Zañartu, M., & Cook, D. (2016). Comprehensive, population-based sensitivity analysis of a two-mass vocal fold model. *PloS One, 11*(2), e0148309.

Rosenstein, M. T., Collins, J. J., & De Luca, C. J. (1993). A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena, 65*(1-2), 117–134.

Saenz-Lechon, N., Godino-Llorente, J. I., Osma-Ruiz, V., & Gomez-Vilda, P. (2006). Methodological issues in the development of automatic systems for voice pathology detection. *Biomedical Signal Processing and Control, 1*(2), 120–128.

Shao, Z., Er, M. J., & Wang, N. (2015). An effective semi-cross-validation model selection method for extreme learning machine with ridge regression. *Neurocomputing, 151*, 933–942.

Steinecke, I., & Herzel, H. (1995). Bifurcations in an asymmetric vocal-fold model. *The Journal of the Acoustical Society of America, 97*(3), 1874–1884.

Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical systems and turbulence* (pp. 366–381). Berlin Heidelberg: Warwick 1980.

Teager, H. M., & Teager, S. M. (1990). Evidence for nonlinear sound production mechanisms in the vocal tract. In *Speech production and speech modelling* (pp. 241–261). Netherlands: SpringerSpringer.

Titze, I. R. (1995). *Workshop on acoustic voice analysis: Summary statement*. National Center for Voice and Speech.

Titze, I. R., Baken, R., & Herzel, H. (1993). Evidence of chaos in vocal fold vibration. In I. R. Titze (Ed.), *Vocal fold physiology: New frontiers in basic science* (pp. 143–188). San Diego, CA: Singular Publishing Group.

Travieso, C. M., Ticay-Rivas, J. R., Briceño, J. C., del Pozo-Baños, M., & Alonso, J. B. (2014). Hand shape identification on multirange images. *Information Sciences, 275*, 45–56.

Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE transactions on Biomedical Engineering, 57*(4), 884–893.

Uloza, V., Padervinskis, E., Uloziene, I., Saferis, V., & Verikas, A. (2015). Combined use of standard and throat microphones for measurement of acoustic voice parameters and voice categorization. *Journal of Voice, 29*(5), 552–559.

Vaiciukynas, E., Verikas, A., Gelzinis, A., Bacauskiene, M., Minelga, J., & Hållander, M. (2015). Fusing voice and query data for non-invasive detection of laryngeal disorders. *Expert Systems With Applications, 42*(22), 8445–8453.

Vaziri, G., Almasganj, F., & Behroozmand, R. (2010). Pathological assessment of patients' speech signals using nonlinear dynamical analysis. *Computers in biology and medicine, 40*(1), 54–63.

Vijayalakshmi, P., & Reddy, M. R. (2005). The analysis on band-limited hypernasal speech using group delay based formant extraction technique. In *INTERSPEECH* (pp. 665–668).

Wallen, E. J., & Hansen, J. H. (1996). A screening test for speech pathology assessment using objective quality measures. In *Spoken language, 1996. ICSLP 96. Proceedings., fourth international conference on: Vol. 2* (pp. 776–779). IEEE.

Wang, Z., Yu, P., Yan, N., Wang, L., & Ng, M. L. (2016). Automatic assessment of pathological voice quality using multidimensional acoustic analysis based on the GRBAS scale. *Journal of Signal Processing Systems, 82*(2), 241–251.

Xu, L. S., Wang, K. Q., & Wang, L. (2005). Gaussian kernel approximate entropy algorithm for analyzing irregularity of time-series. *2005 international conference on machine learning and cybernetics*.

Yu, P., Ouaknine, M., Revis, J., & Giovanni, A. (2001). Objective voice analysis for dysphonic patients: A multiparametric protocol including acoustic and aerodynamic measurements. *Journal of Voice, 15*(4), 529–542.

Yumoto, E., Gould, W. J., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America, 71*(6), 1544–1550.

Yunik, M., & Boyanov, B. (1990). Method for evaluation of the noise-to-harmonic–component ratios in pathological and normal voices. *Acustica, 70*(1), 89–91.

Zhang, Y., & Jiang, J. J. (2003). Nonlinear dynamic analysis in signal typing of pathological human voices. *Electronics Letters, 39*(13), 1021–1023.