Revista Facultad de Ingeniería
Universidad de Antioquia
ISSN: 0120-6230
revista.ingenieria@udea.edu.co
Universidad de Antioquia
Colombia

Orozco-Arroyave, Juan Rafael; Vargas-Bonilla, Jesús Francisco; Vásquez-Correa, Juan Camilo; Castellanos-Domínguez, Cesar German; Nöth, Elmar
Automatic detection of hypernasal speech of children with cleft lip and palate from spanish vowels and words using classical measures and nonlinear analysis
Revista Facultad de Ingeniería Universidad de Antioquia, núm. 80, septiembre, 2016, pp. 109-123
Universidad de Antioquia
Medellín, Colombia

redin Revista Facultad de Ingeniería, Universidad de Antioquia, No. 80, pp. 109-123, 2016

Automatic detection of hypernasal speech of children with cleft lip and palate from spanish vowels and words using classical measures and nonlinear analysis

Detección automática de voz hipernasal de niños con labio y paladar hendido a partir de vocales y palabras del español usando medidas clásicas y análisis no lineal

Juan Rafael Orozco-Arroyave[1,2*], Jesús Francisco Vargas-Bonilla[1], Juan Camilo Vásquez-Correa[1], Cesar German Castellanos-Domínguez[3], Elmar Nöth[2]

[1]Facultad de Ingeniería, Universidad de Antioquia. Calle 67 # 53-108. A. A. 1226. Medellín, Colombia.
[2]Pattern Recognition Lab, Friedrich-Alexander University Erlangen-Nürnberg. Martensstraße 3. 91058. Erlangen, Germany.
[3]Facultad de Ingeniería y Arquitectura, Universidad Nacional de Colombia. Cra 27 # 64-60. Manizales, Colombia.

ARTICLE INFO
Received April 30, 2015
Accepted November 12, 2015

KEYWORDS
Automatic hypernasality detection, cleft lip and palate, perturbation measures, noise measures, nonlinear dynamics

Detección automática de hipernasalidad, labio y paladar hendido, medidas de perturbación, medidas de ruido, dinámica no lineal

ABSTRACT: This paper presents a system for the automatic detection of hypernasal speech signals based on the combination of two different characterization approaches applied to the five spanish vowels and two selected words. The first approach is based on classical features such as pitch period perturbations, noise measures, and *Mel-Frequency Cepstral Coefficients* (MFCC). The second approach is based on the *Non-Linear Dynamics* (NLD) analysis. The most relevant features are selected and sorted using two techniques: *Principal Components Analysis* (PCA) and *Sequential Forward Floating Selection* (SFFS). The decision about whether a voice record is hypernasal or healthy is taken using a *Soft Margin - Support Vector Machine* (SM-SVM). Experiments upon recordings of the five Spanish vowels and the words /coco/ and /gato/ are performed considering three different set of features: (1) the classical approach, (2) the NLD analysis, and (3) the combination of the classical and NLD measures. In general, the accuracies are higher and more stable when the classical and NLD features are combined, indicating that the NLD analysis is complementary to the classical approach.

RESUMEN: Este artículo presenta un sistema para la detección automática de señales de voz hipernasales basado en la combinación de dos diferentes esquemas de caracterización aplicados en las cinco vocales del español y dos palabras seleccionadas. El primer esquema está basado en características clásicas como perturbaciones del periodo fundamental, medidas de ruido y *coeficientes cepstrales en la frecuencia de Mel*. El segundo enfoque está basado en medidas de *dinámica no lineal*. Las características más relevantes son seleccionadas usando dos técnicas: *análisis de componentes principales y selección flotante hacia adelante secuencial*. La decisión acerca de si un registro de voz es hipernasal o sano es tomada usando una *máquina de soporte vectorial de margen suave*. Los experimentos consideran grabaciones de las cinco vocales del idioma español y las palabras /coco/ y /gato/, y se consideran, asimismo, tres conjuntos de características: (1) el enfoque clásico, (2) el análisis de dinámica no lineal y (3) la combinación de ambos esquemas. En general, los aciertos son mayores y más estables cuando las características clásicas y no lineales son combinadas, indicando que el análisis de dinámica no lineal se complementa con el esquema clásico.

1. Introduction

The automatic evaluation of pathological voices is an interesting field of research due to the possibility of performing objective assessments of speech quality without using invasive procedures. Most of the studies in this area have been focused on laryngeal pathologies [1], taking aside other pathologies such as hypernasality,

* Corresponding author: Juan Rafael Orozco Arroyave
e-mail: rafael.orozco@udea.edu.co
ISSN 0120-6230
e-ISSN 2422-2844
DOI: 10.17533/udea.redin.n80a12

which also affects the speech production. Hipernasality is a voice pathology that may cause a significant reduction in speech intelligibility [2] and it is mainly suffered by children that have born with a craniofacial malformation called *Cleft Lip and Palate* (CLP). Typically children with CLP also have velopharyngeal insufficiency which leads to a lack control of the pharyngeal velum during phonation. Due to velopharyngeal insufficiency, an excess of air coming out through nasal cavities is revealed, generating unintelligible speech [3].

Unlike laryngeal pathologies which affect the normal vibration of vocal folds, hypernasality is produced by an insufficiency in the velar movements changing the resonant properties of the nasal cavity [4]. People suffering from hypernasality should receive early and constant speech therapy to develop complete control of the pharyngeal velum in the shortest possible time. Such therapy contributes also to the easy and fast integration of patients to the society [3]. Currently, it is difficult to perform continuous evaluation of the speech quality in CLP patients due to the change of speech therapists during the treatment provided by the social care system in Colombia, generating different opinions in the diagnosis and the therapy for the same patient. The differences among diagnoses performed by the speech therapists appear due to their subjective evaluation, evidencing the need for computer aided systems to support the medical diagnosis from the objective information provided by the speech signals.

The automatic assessment of hypernasality in speech of CLP patients comprises a process with at least three main stages, (1) the automatic detection of hypernasality which allows the phoniatricians to make informed decisions regarding the speech and language therapy of the patients, (2) the automatic detection of hypernasality is the first step in the development of computer aided tools that assist the evaluation of the progress of the speech therapy, and (3) once the automatic detection of hypernasality is performed, the extent of hypernasality needs to be measured in order to objectively quantify the progress of the therapy. This paper is focused on the automatic detection of hypernasality in voice signals, thus it comprises a contribution in the first two stages of the aforementioned process.

Hypernasality has been classically evaluated using acoustic and spectral analysis. In [5] the authors used a modified group delay function for the automatic detection of hypernasality in the voice of children with non-repaired CLP. The database contained speech recordings from 33 children with non-repaired CLP and 30 healthy controls. The authors report accuracies of 100%, 88.78%, and 86.66% for the vowels /a/, /i/, and /u/ respectively. Considering the high accuracies reported in this study, the methodology is reproduced here for the sake of comparison and also to confirm the usefulness of the technique in CLP patients whose the velum is already repaired.

Automatic hypernasality detection can also be performed using voice quality measures such as Jitter, Shimmer, and noise [6]. Their use is motivated due to the fact that velopharyngeal insufficiency leads the patients to perform compensatory movements that modify acoustic and resonance properties of the vocal tract, producing glottal stops and general problems with glottal articulations [7].

Another kind of features commonly used for the characterization of speech pathologies is the *Mel-frequency Cepstral Coefficients* (MFCC), which can model irregular movements of the vocal tract [8]. These features have been used in the context of hypernasality to characterize the irregularities derived from compensatory movements in the vocal tract [6]. In [9] it is presented a set of pronunciation features and 12 MFCC to detect different articulation problems in CLP patients. The database used in that study included recordings sustained vowels and a total of 1916 German words. The authors reported accuracies of 71.1% with the sustained phonations and 75.8% with the isolated words.

On the other hand, bearing in mind the presence of nonlinearities in the vocal tract movements [10], in [11] the authors compare the accuracies obtained with sustained vowels modeled using acoustic features, with respect to the same phonations characterized with four *Non-Linear Dynamics* (NLD) features. The authors considered a database with 156 hypernasal and 110 healthy speakers, reporting accuracies of 93.86% for the acoustic analysis and 92.05% with the NLD features. This paper comprises a step forwards with respect to the study presented in [11]. Further to the analysis of vowels, in this paper we evaluate hypernasality in words; additionally, the acoustic features are merged with the NLD measures in the same representation space. This paper demonstrates that such a fusion improves the classification rate with respect to those obtained in the previous study. The results show that NLD and acoustic features provide complementary information to model hypernasal speech signals.

In this study, the use of perturbation, noise, and MFCC for the automatic evaluation of hypernasality in Spanish vowels and words is presented. The discriminant capability of complexity measures based on NLD such as *Correlation Dimension* (CD), *Largest Lyapunov Exponent* (LLE), *Hurst Exponent* (HE), and *Lempel-Ziv Complexity* (LZC) is also studied [12]. Finally, the two features sets are merged to analyze whether the information from both sets is complementary.

The rest of the paper is organized as follows: Section 2 presents the methodology addressed in the paper. The details of the estimated features, the strategies for the automatic selection of features and the classification process are also provided. Section 3 includes a description of the experiments and the obtained results; finally Section 4 presents the conclusions derived from this study.

2. Methods

Figure 1 depicts a block diagram with the steps of the methodology addressed in this study. The voice signal is first divided into frames to perform a short-time analysis. Subsequently, for each frame, the two different

characterization approaches are performed. The first one is based on perturbation, cepstral and noise features, which will be referred to *Classical Analysis of Voice* (CAV). The second one is based on the NLD analysis. The final feature vector per voice signal is composed by the estimations of mean and standard deviation of the values obtained per feature upon the frames.

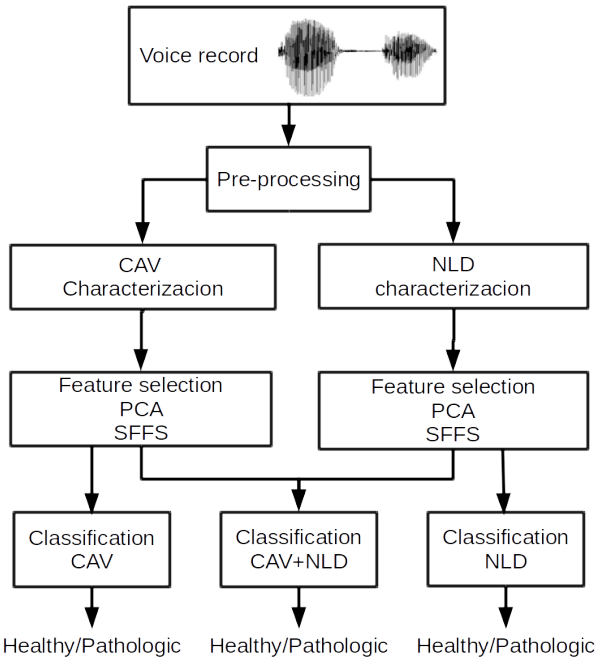


Figure 1 General scheme of the different approaches carried out for the automatic evaluation of hypernasality in voice

In order to find an optimal representation of the voice recordings in the features space and to avoid redundant information, automatic feature selection is performed by means of two different approaches, *Principal Component Analysis* (PCA) and *Sequential Forward Floating Selection* (SFFS). Three different feature vectors are considered: CAV, NLD, and CAV+ NLD.

The decision whether a recording comes from a CLP or a healthy speaker is taken with a *Soft-Margin - Support Vector Machine* (SVM). The results are presented in terms of accuracy, sensitivity, specificity, and area under the *Receiving Operating Characteristic* (ROC) curve, which are commonly used for the evaluation of medical systems [13].

2.1. Feature estimation

Perturbation, cepstral, and noise features - CAV

The set of features described in this section includes the most common acoustic features used for the modeling of pathological voices.

Jitter and Shimmer. These features are calculated to analyze the stability in time and amplitude of the fundamental voice period. Jitter indicates variations on the frequency vibration of the vocal folds due to the lack of control of the vocal fold muscles, while Shimmer represents reduction of glottal resistance and the possible presence of mass lesions in the vocal folds [14]. The variations on the tone of hypernasal voices due the velopharyngeal insufficiency have been already evaluated in adults [15], finding that there are significant correlations between the nasality scores and the hypernasality ratings. Perturbation measures have also been tested in children with phonological disorders, which allow finding problems associated to improper or incomplete closure of the vocal folds [14]. In this study, the estimation of shimmer and jitter has been performed according to the methods evaluated in [16].

Noise measures. Four different noise features are included to model the noise that appears in phonations uttered by CLP patients when producing involuntary movements in articulators e.g., tongue, lips, and jaw, and in the entire vocal tract to compensate their velopharyngeal incompetence [7].

The first measure corresponds to *Harmonics to Noise Ratio* (HNR). This feature quantifies the relationship between the harmonics structure energy and the additive noise present on the signal produced by the pathology [17]. HNR is calculated according to the procedure presented in [18]. The voice signal $\boldsymbol{x(n)}$ is divided in $\boldsymbol{L}$ time intervals, $\boldsymbol{x_i(m)}$, then the signal is averaged as is shown in Eq. (1).

$$x_A(n) = \sum_{i=1}^{L} \frac{x_i(m)}{L} \tag{1}$$

The energy of the harmonic component of the signal is defined according to Eq. (2), where T is the duration of each time interval.

$$\boldsymbol{H}(n) = L \sum_{n=0}^{T} x_A^2\,(n) \tag{2}$$

The noise wave in each interval is $\boldsymbol{x_i(n) - x_A(n)}$, where $\boldsymbol{x_i(n)}$ is the voice signal in the $\boldsymbol{i-th}$ time interval. Then, the energy of the noise component of $\boldsymbol{x(n)}$ is defined using Eq. (3). The relation $\frac{\boldsymbol{H}}{\boldsymbol{N}}$ is HNR.

$$\boldsymbol{N}(n) = \sum_{i=1}^{L} \sum_{n=0}^{T} |x_i(n) - x_A(n)|^2 \tag{3}$$

The HNR can also be calculated in the cepstral domain, producing the feature called *Cepstral Harmonics to Noise Ratio* (CHNR), which provides more accurate estimation of noise levels on different spectral components [19, 20]. The method followed in this study is the presented in [20]. In this case, the voice signal is windowed, and each window is transformed according to Eq. (4) to calculate the cepstral

version. Where $\hat{x}_i$ are the frames of the voice signal $x(n)$, and the terms FFT and IFFT refer to the Fast Fourier Transform and its inverse, respectively.

$$C_{\hat{x}_i}(n) = IFFT[\log|FFT(\hat{x}_i)|] \quad (4)$$

Higher components on the cepstrum are related to the harmonic content on the original signal. These are zeroed by means of the liftering process, and the result is Fourier transformed to provide the noise spectrum N_i [20]. The harmonic content in the signal is found as $H_i = log|FFT(\hat{x}_i)| - N_i$, and to calculate the noise component in the signal, the values of N_i are corrected subtracting the value of the successive minimum between successive harmonics $\boldsymbol{m_{Hi}}$. Finally, CHNR is calculated according to Eq. (5).

$$CHNR_{dB} = 10\log\left(\frac{|FFT(\hat{x}_i)|}{N_i - \boldsymbol{m_{Hi}}}\right) \quad (5)$$

Another commonly used noise measure is the *Normalized Noise Energy* (NNE) which is calculated following the algorithm presented in [21], finding the relation between the energy of noise and the total energy of the signal. The detailed procedure is as follows: the voice signal is divided into L frames $x_i(n)$. Every frame has two components, the first one related to the harmonic $s_i(n)$ and the other one to the additive noise $w_i(n)$, thus $x_i(n) = s_i(n) + w_i(n)$, $n = 0, 1, 2, \ldots, M-1$, where M is the number of voice samples per frame.

The N-points FFT of the signals $x_i(n)$, $s_i(n)$ and $w_i(n)$, must be calculated, obtaining the sequences $X_i(k)$, $S_i(k)$ and $W_i(k)$, respectively. With $N \geq M$ and considering that $x_i(n) = s_i(n) = w_i(n) = 0$ for $L < n < N$, then $X_i(k) = S_i(k) + W_i(k), k = 0,1,2,\ldots,N-1$. If $|\widehat{W}_i(k)|^2$ is defined as the estimated value of $|W_i(k)|^2$, then NNE will be defined by Eq. (6).

$$NNE_{dB} = 10\log\left(\frac{\frac{1}{L}\sum_{k=N_L}^{N_H}\sum_{i=1}^{L}|\widehat{W}_i(k)|^2}{\frac{1}{L}\sum_{k=N_L}^{N_H}\sum_{i=1}^{L}|X_i(k)|^2}\right) \quad (6)$$

Where $N_L = \lceil N f_L T_s \rceil$, $N_H = \lceil N f_H T_s \rceil$; f_L and f_H are respectively the lower and higher frequencies of the bands where the energy of the noise is evaluated. The brackets $\lceil\ \rceil$ indicate that the value inside must be rounded up, and T_s is the sample period. The estimated value of the noise spectrum $|\widehat{W}_i(k)|$ is formed by the intervals with the lowest amplitude in $X_i(k)$. According to the procedure presented in [21], $|\widehat{W}_i(k)|^2$ can be estimated as follows: the spectra $S_i(k)$ and $W_i(k)$ can be expressed in polar coordinates as $S_i(k) = |S_i(k)|e^{j\theta(k)}$ and $W_i(k) = |W_i(k)|e^{j\theta(k)}$, then $|X_i(k)|$ is given by Eq. (7)

$$|X_i(k)|^2 = |S_i(k)|^2 + |W_i(k)|^2 + 2|S_i(k)|\,|W_i(k)|\cos[\theta(k) - \varphi(k)] \quad (7)$$

Since $s_i(n)$ is assumed to be the periodic component of $x_i(n)$, then $|S_i(k)|$, will contribute to the harmonic structure in $|X_i(k)|$. It is possible to state that in the frequency bands where there is no evidence of harmonic structure, the signal components will be due to the noise. In this case the estimate of the noise is given using Eq. (8), where D_j is the set of points that correspond to the $j-th$ interval of the spectrum where the harmonic component is minimum.

$$|\widehat{W_i^2}(k)| = |X_i(k)|^2,\ k \in D_j \quad (8)$$

The last noise measure used in this work is the *Glottal to Noise Excitation Ratio* (GNE), which was introduced in [22] to quantify the amount of vocal excitation due to vocal fold vibration versus the amount of excitation due to turbulent noise in the vocal tract. The procedure begins re-sampling the voice signal at 10KHz then, it is necessary to find the glottal pulses of the voice signal. This finding is performed by applying an inverse linear filtering over voiced intervals of 30 ms. Next, different band-pass filters are applied using Hanning windows. The number, location, and bandwidth were calculated for real voices in [23], finding that the optimal value of the band width for the band-pass filters is 1KHz. Band steps of 300 Hz must also be applied. Lastly, the Hilbert envelopes are calculated for each filtered voice interval $x_i(n)$ whose duration is given by the glottal pulses found in the inverse filtering process. The maximum value of autocorrelation sequence of such envelopes is the GNE.

Mel-Cepstral coeffcients. 11 MFCC are calculated considering their capability to model both the vocal folds and the vocal tract [8]. MFCC can be estimated using a parametric approach derived from *Linear Prediction Coefficients* (LPC), or using a non-parametric FFT-based approach. However, FFT-based MFCC typically encode more information from excitation; while LPC based MFCC remove it, as is demonstrated in [24]. FFT based MFCC has been considered suitable for our purpose because in presence of voice disorders they show the inherent ability to model either an irregular movement of the vocal folds, or a lack of closure induced by the compensatory movements in the vocal tract due to the velopharyngeal incompetence.

The procedure to calculate the MFCC begins with the windowing of the signal using a Hamming window. Then it is calculated the FFT to obtain the power spectrum of the signal. Subsequently, a filter bank in Mel scale is created to obtain a higher resolution in lower frequencies. Finally, the log-energy of the output signals from each filter is calculated, and the *Discrete Cosine Transform* (DCT) is applied.

Nonlinear dynamics analysis

NLD methods such as Poincaré maps, fractal dimension, Kolmogorov entropy, correlation dimension and Lyapunov exponents have been used for the analysis of irregular or chaotic activities in voice signals [25]. In order to understand the NLD analysis, the concept of phase space, also known as state space will be introduced.

Phase space is a multidimensional representation which allows the study of topological or qualitative features of the voice production system. For a time series, the state space can be reconstructed by applying the embedding theorem originally proposed in [26]. This theorem allows the reconstruction of difeomorphic attractors i.e., those that hold topological properties of the system. The state space S is defined according to Eq. (9), where n is the number of points in the time series $s[n]$, ϑ and τ are the dimension and embedding delay, respectively.

$$S[n] = \{s[n], s[n+\tau], s[n+2\tau], \ldots, s[n+(\vartheta-1)\tau]\} \quad (9)$$

The embedding dimension is found by applying the false neighbor's method, which is based on the assumption of a minimum embedding dimension ϑ_0 to reconstruct the topological properties of the attractor of the time series s_i. Surrounding each embedded point, there will be a set of neighbors which number will depend on the neighborhood size. If you suppose that the series s_i will be embedded in a space R^{ϑ} such that $\vartheta < \vartheta_0$. The topological features of the attractor will be destroyed because the new space shall be a projection of the previous one R^{ϑ_0}, thus there will be points projected in wrong neighborhoods which originally belong to other spaces with higher dimensionality. These points are known as false neighbors.

For the estimation of the time delay τ, the *First Minimum of Mutual Information* (FMMI) method is applied. This estimation consists in finding the first minimum in the mutual information of the signal $x(n)$, which is defined by Eq. (10). Where $P(x_n, X_{n+T})$ is the probability to observe x_n and x_{n+T} at the same time, and $P(x_n)$ is the probability to observe x_n. $I(T)$ is an information measure of x_n when x_{n+T} is observed and the value of the delay can be found when the value $T = \tau$ is the first local minimum of $I(T)$.

$$I(T) = \sum_{n=1}^{N} P(x_n, X_{n+T}) \log_2 \frac{P(x_n, x_{n+T})}{P(n)P(x_{n+T})} \quad (10)$$

With the aim of illustrating the concept of complexity based on the reconstructed attractor, Figure 2 shows the attractor for a sinusoid. Note that its form is a perfect circle, indicating that its complexity is low or null.
When a more complex phenomenon is considered, such as

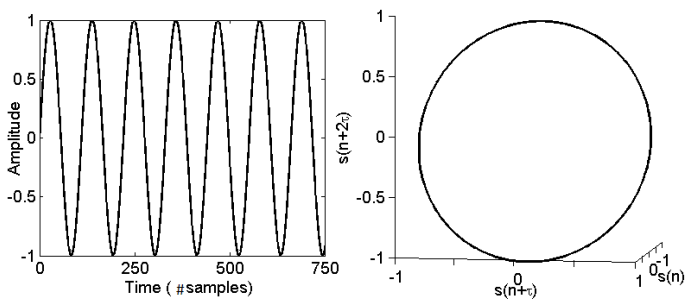


Figure 2 Sinusoid signal and its attractor

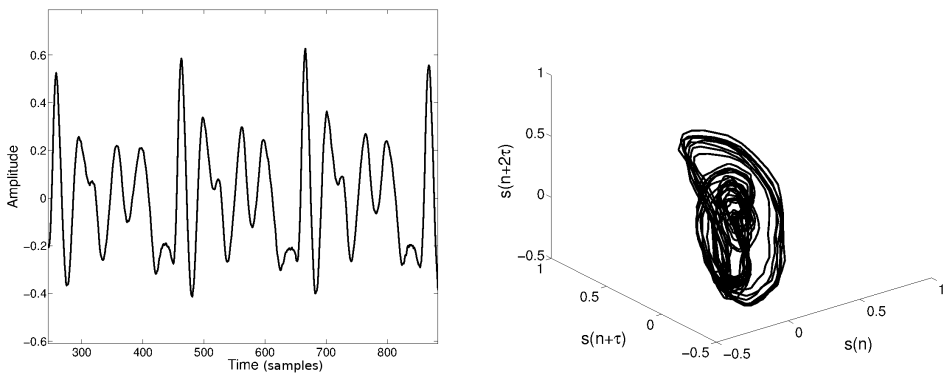
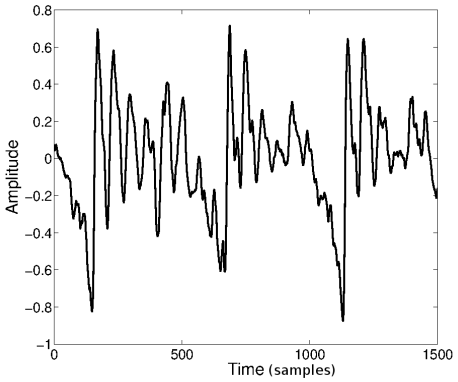


Figure 3 Healthy voice and its attractor



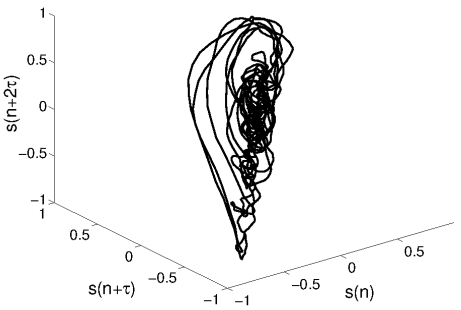


Figure 4 Hypernasal voice and its attractor

the voice production, the obtained attractor shows more irregular forms. Additionally, according to [25], when the pathology level is higher, the associated attractor is more complex, i.e. more irregular. Figures 3 and 4 show the attractors for healthy and hypernasal voices.

There are works that demonstrate the existence of NLD in the voice production process and analyze its capability in the automatic detection of pathologies [27-29]. The suitability of those features to detect hypernasality in sustained vowels is demonstrated in [11]. Although there are other NLD features in the state of the art, in this paper we want to study the suitability of the same set of features to model hypernasality in words. Additionally, the fusion of this set of features with acoustic ones is also evaluated.

The different complexity measures which have been implemented for the automatic detection of hypernasality in Spanish vowels and words will be described in the following sections.

Correlation dimension (CD): To describe CD, it is necessary to introduce the concept of "correlation sum" in the state space $C_\vartheta(\varepsilon)$, which can quantify the number of points x_i that are correlated with the others inside an sphere with radius ε. Intuitively, this sum can be interpreted as the probability to have pairs of points in a trajectory of the attractor inside the same sphere of radius ε. It is possible to define an expression for the correlation sum according to Eq. (11), where $\Theta(z)$ is a Heaviside step function and $\|\vec{x}_i - \vec{x}_j\|$ is the Euclidean distance between every pair of points inside the sphere of radius ε.

$$C_\vartheta(\varepsilon) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \Theta(\varepsilon - \|\vec{x}_i - \vec{x}_j\|) \tag{11}$$

In [30], the authors demonstrated that $C_\vartheta(\varepsilon)$ represents a volume measure, thus the correlation dimension is been defined by Eq. (12).

$$CD = \lim_{\varepsilon\to 0} \frac{\log C_\vartheta(\varepsilon)}{\log \varepsilon} \tag{12}$$

CD is estimated according to the slope of the curve $\log C_\vartheta(\varepsilon)$ vs $\log \varepsilon$ after a linear regression for small values of ε. A proper estimation of CD must guarantee that the embedding dimension complies with the expression $\vartheta = 2CD + 1$ [31].

Largest Lyapunov exponent (LLE): This feature represents the average divergence rate of the neighbor trajectories in the state space; due to its robustness to noisy and short term signals, its estimation in this study has been developed according to the algorithm proposed in [32].

After the estimation of ϑ and τ, the nearer neighbors to every point in different trajectories are found. Nearer neighbors must have a temporal separation greater than the temporal period of the time series. Considering every pair of neighbors on each trajectory as the representation of the initial conditions of the phenomenon, LLE is estimated as the average separation rate of the nearer neighbors in the embedding space.

According to the Oseledec's theorem [33], the points on a trajectory in the state space can be represented by the expression $d(t) = Ce^{\lambda_1 t}$, where λ_1 is the LLE, $d(t)$, is the average divergence take nat the time t, and C, is a normalization constant. Assuming that the $j-th$ pair of nearer neighbors diverge approximately at a rate of λ_1, it is possible to obtain the expression $\ln d_j(i) = \ln C_j + \lambda_1(i\Delta t)$, where λ_1, is the slope of the average line that appears when such expression is drawn on a logarithmic plane.

Hurst Exponent HE: This feature allows analyzing the long term dynamics of a system, stating the possible long term dependencies of the different elements in a given time series.

The estimation of HE for a time series $x(n)$ with $n = 1, 2, \ldots, N$, is based on the rank scaling method, proposed in [34]. Hurst 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 $\frac{R}{S} = cT^{HE}$, where c is a scaling constant, T is the duration of the segment and HE is the Hurst exponent.

Lempel-Ziv Complexity (LZC): This feature can be used for the estimation of the complexity in a time series. The computation consists in finding the number of different "patterns" in a binary sequence [35]. The binary sequence is formed according to the difference between consecutive samples in the time series, i.e., if the difference is negative it is assigned a 0 to the sequence, in the other case, it is assigned a 1. The estimation of the LZC is based on the reconstruction of a sequence X according to the copy and insertion of symbols in a new sequence. Let's consider the binary string $X = x_1, x_2, \ldots x_n$. 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 contains x_1. The variable Q is defined for the accumulation of bits that have been analyzed from left to right in the string. On each iteration, the union of S and Q (denoted by SQ) is generated. Also, the string $SQ\pi$ is formed by the subtraction of the last bit in the stream SQ. If the sequence Q does not belong to $SQ\pi$, the insertion of the bits in the subset of symbols is finished. The value of LZC will be the number of subsets used for the representation of the original signal [35]. It ranges from 0 for deterministic sequences, to 1 for random sequences.

2.2. Automatic feature selection

The aim of automatic feature selection is to find out the m most relevant characteristics of the original feature space $\boldsymbol{X} \in R^{n \times p}$ (n: number of observations, p: number of original features), which allows building the subspace representation $\boldsymbol{Y} \in R^{n \times m}, (m < p)$. The features contained in $\boldsymbol{Y}$ reduce the redundant information and the computational load in classification stages. In this study, two different algorithms for feature selection have been considered. The first one is based on PCA in order to have a subset of features that holds the maximum variance of the original features space. The second selection technique is based on a heuristic floating search, whose aim is to find a subset of features that better discriminates the studied classes: healthy and hypernasal.

Selection based on Principal Component Analysis

PCA is a statistical technique that allows finding a low-dimensional representation of the original features space, searching for directions with greater variance to subsequently project the data [36]. The PCA algorithm can be summarized as follows:

Let $\boldsymbol{X} \in R^{n \times p}$ the input data matrix with sample objects $(\boldsymbol{x_i} : i = 1, \ldots, n)$. Then, the estimated covariance matrix $\boldsymbol{S}$ is calculated. As $\boldsymbol{S}$ is a symmetric matrix, it is possible to calculate a new orthogonal basis given by its m largest eigenvectors $(m < p)$. Finally, $\boldsymbol{X}$ can be linearly mapped to a lower-dimension space $\boldsymbol{Y} \in R^{n \times m}$, with output vectors $(\boldsymbol{y_i} : i = 1, \ldots, n)$, so that $\boldsymbol{Y} = \boldsymbol{AX}$, being $\boldsymbol{A} \in R^{m \times n}$ a transformation matrix with columns equal to the eigenvectors $\boldsymbol{v_j}$ that conforms the orthogonal basis found by PCA.

Although, PCA is commonly used as a feature extraction method, it can also be useful to properly select a relevant subset of original features that better represents the studied process [37]. In this sense, given a set of features $(\xi_k : k = 1, \ldots, p)$ corresponding to each column of $\boldsymbol{X}$, we can analyze the relevance of each ξ_k for finding $\boldsymbol{Y}$. More precisely, we can identify the relevance of ξ_k looking at the vector $\rho = [\rho_1 \, \rho_2 \ldots \rho_p]^T$, defined according to Eq. (13).

$$\rho = \sum_{j=1}^{m} |\lambda_j \boldsymbol{v_j}| \tag{13}$$

Therefore, the main assumption is that the largest values of ρ_k points out to the best input attributes, since they exhibit higher overall correlations with principal components. Finally, for the redundancy elimination process, the features with a linear correlation greater than 80% are removed from the analysis.

Selection based on Sequential Forward Floating Selection

This method is a heuristic algorithm that finds the best subset of features of the original set through the inclusion and exclusion of features. In this procedure after each forward step, a number of backward steps are applied as long as the resulting subsets are better than the previous [38]. Sorting features according to its discriminant capacity is necessary to get stable and consistent results, which is reflected in the overall performance of the system. The algorithm for the SFFS is as follows (Figure 5) [39]:

Where J is the criterion for evaluating the goodness of a particular subset of features. In this study J corresponds to the detection rate provided by a 1-Nearest Neighbor classifier. With this technique, dimensionality reduction is also achieved.

Input:

$$\boldsymbol{X} = \{x_i | i = 1, \ldots, n\}$$

Output:

$$\boldsymbol{Y_m} = \{y_i | i = 1, \ldots, m,\ y_i \in \boldsymbol{X}\}, m = 0, 1, \ldots, n$$

Initialization:

$$\boldsymbol{Y_0} := 0;\ m := 0$$

Step 1 (Inclusion)

$$y^+ := \arg \max_{y \in (\mathbf{X} - \mathbf{Y_m})} J(\boldsymbol{Y_m} + y)$$

The most significant feature with respect to Y_m

$$\boldsymbol{Y_m} + 1 := \boldsymbol{Y_m} + y^+;\ m := m + 1$$

Step 2 (Conditional Exclusion)

$$y^- := \arg \max_{y \in \mathbf{Y_m}} J(\boldsymbol{Y_m} - y)$$

The least significant feature in $\boldsymbol{Y_m}$

if $J(\boldsymbol{Y_m} - \{y^-\})) > J(\boldsymbol{Y_m} - 1)$ then

$Y_{m-1} := \boldsymbol{Y_m} - y^-;\ m := m - 1$

go to step 2

else

go to step 1

Figure 5 Algorithm to compute the SFFS

2.3. Classification

Given the subsets of features $\boldsymbol{Y_{PCA}}$ or $\boldsymbol{Y_{SFFS}}$, a SM-SVM classifier is trained using a radial basis Gaussian kernel with band-width $\boldsymbol{\sigma}$. The SVM are supervised learning models based on the concept of decision hyperplanes. The SVM perform the classification task by constructing a set of hyperplanes that separates different class labels. The main aim of a SVM is to maximize the separation between classes finding a hyperplane that has the largest distance to the nearest training data point of any class by the concept of support vectors.

To achieve a more robust machine, the number of support vectors are also optimized with respect to the accuracy in the training process; with this optimization, the over fitting is avoided, thus the implemented SVM can generalize the process and exhibit good results in the classification stage [40].

3. Experiments and results

3.1. Database

The database used in this study was provided by *Grupo de Procesamiento y Reconocimiento de Señales* - (GPRS) from the Universidad Nacional de Colombia, branch Manizales. It contains recordings from 65 children with repaired CLP and 54 healthy ones; the ages of the children range from 5 to 15. All of the patients with CLP were evaluated by phoniatric experts and labeled as hypernasal. The recordings were performed in a quiet room, using an omnidirectional microphone, a professional audio-card, and with a sampling frequency of 44.1 KHz with 16 resolution-bits.

The Spanish vowels were uttered twice by every child, thus for each vowel in the database a total of 130 recordings from hypernasal voices and 108 from healthy ones were captured. For the case of the Spanish words */coco/* and */gato/*, each word was recorded once, forming a total of 65 hypernasal recordings and 54 healthy ones. The features considered in this paper are calculated through all of the utterance, i.e., no prior segmentation of voiced and unvoiced frames, is performed. This approach allows analyzing particular consonant sounds and articulation movements that are performed to produce the words */coco/* and */gato/*. These words were selected because both have occlusive and velar characteristics that are useful to show problems in the velopharyngeal movement. Phonemes */k/* and */g/*, which are present in the words */coco/* and */gato/*, are consonants that force the velum occlusion to produce proper phonations. The inclusion of these two phonemes allows the evaluation of the velar movement in children with repaired CLP. This information is important for the speech therapy experts because velopharyngeal problems can be evidenced in phonations with weak obstruent consonants [41].

3.2. Experimental setup

The features described in the section 2.1 are calculated following a short-time strategy. CAV features are implemented considering frames with 40 ms length, and NLD features are implemented using frames of 55 *ms* length, as in [29]. The frame length defined for NLD feature assures the number of points required for a successful reconstruction of the embedded attractor, which has been established at around 10^{CD} [42, 43].

After having built both feature vectors, mean values and standard deviations are calculated for each one, building the input spaces $\boldsymbol{X_{CAV}} \in R^{238\times34}$, $\boldsymbol{X_{NLD}} \in R^{238\times8}$, and $\boldsymbol{X_{CAV+NLD}} \in R^{238\times42}$ for vowels, and $\boldsymbol{X_{CAV}} \in R^{119\times34}$, $\boldsymbol{X_{NLD}} \in R^{119\times8}$, and $\boldsymbol{X_{CAV+NLD}} \in R^{119\times42}$ for words.

To find the features that better represent the phenomenon, the feature selection is performed repeating 10 times cross-validations with 10 folds; for a total of 100 vectors with the selected features on each technique (PCA and SFFS). In the case of SFFS, the best feature vector is obtained after counting and sorting the times that each feature is selected as relevant on each fold; only those that appear as relevant on every fold will be considered in the final vector. For vectors selected as relevant using the PCA-based selection technique, since each feature is associated to a relevance weight, the final vector will be the result of sorting the features considering the sum of their weights on each fold.

Once the most relevant features are chosen with each selection strategy, classification is performed using a SVM with a radial basis Gaussian Kernel. The parameters of the classifier: regularization trade-off (C) and the standard deviation of the Kernel's machine (σ), were optimized also using a 10- fold cross-validation strategy, where the feature set is divided into 10 groups or folds of approximately equal size. The first fold is used as test set for the SVM, and the remaining 9 folds are grouped and used as train set for the SVM. The procedure is repeated 10 times, in each time a different fold is used as test set.

The methodology that the authors proposed in [13] has been used for the evaluation of the system, thus results are presented in terms of the overall accuracy of the system and also with specificity and sensitivity to indicate the probability of a healthy register to be correctly detected, and the probability of a pathological signal to be correctly classified, respectively.

3.3. Results and discussion

Low frequency analysis

For the sake of comparison, the experiments presented in [5] were reproduced. The authors consider features based on *Modified Group Delay Functions* (MGDF) to model the low frequency region (around 250 Hz) of registers from non-repaired CLP patients. Our database contains registers from repaired CLP patients, it is interesting to compare the performance of the pointed out method with the one we are proposing in this work.

The sampling frequency of the recordings is 44.1 kHz, in order to get comparisons to the methodology described in [5], for the experiments based on the MGDF, the first step was to down sample the recordings to 8000 Hz. Table 1 summarizes the results obtained by applying the cited methodology upon the recordings of the five Spanish vowels of our dataset.

Table 1 Results with MGDF applied over voice recordings of children with repaired CLP

Vowel	/a/	/e/	/i/	/o/	/u/
Accuracy (%)	77.6	75.7	58.2	75.1	57.6

Notwithstanding, the good results reported in [5], this methodology is not suitable to assess the voice of children with repaired CLP.

Experiments with CAV features

Table 2 shows the results obtained with the features grouped in CAV (perturbation, noise and MFCC) for the automatic detection of hypernasality in the five Spanish vowels and in the words /*coco*/ and /*gato*/. The results include accuracy rates, specificity, sensitivity and the *area under the ROC curve* (AUC) when the techniques for feature selection are used (SFFS and PCA-based) and when no feature selection technique is applied. The last two rows in the table also show the results obtained when the most relevant features are combined for vowels and words separately. Note that the best result is reached out using vowel /i/, and the combination of all vowels. For Tables 2, 3 and 4, the results with the highest AUC values are highlighted in bold-face.

Table 3 shows the results obtained in the automatic detection of hypernasality in the Spanish vowels and in the words /*coco*/ and /*gato*/. using NLD features. The accuracies obtained by combining the most relevant NLD features for vowels and words are also indicated in the last two rows. Note that in this case, again the overall performance of the system increases when the feature space includes information from all the best subsets of features per vowel and words, and that the best individual accuracy is reached out with the vowel /i/. Note also that the results for NLD features are similar to those obtained in CAV; it indicates that nonlinear analysis provide relevant information that should be considered as an alternative to complement classical techniques for the evaluation of hypernasality.

Table 4 shows the results of the combination of features related to NLD and CAV for the five Spanish vowels, and the two words, in order to verify if the NLD analysis can be used to improve the overall accuracy of the system.

Note that the results have been improved when the features of both domains are combined. Also, the confidence intervals of the results are narrower when both spaces are joined, indicating that combining CAV and NLD features not only increases the accuracy of the system but also improves its stability, which is in accordance with results obtained for different pathologies [29].

When the five vowels are considered and the NLD and CAV features are merged, an accuracy of 95.4% is obtained. This is the highest accuracy obtained in all of the experiments, and it represents an improvement in the absolute and relative errors of about 2.2% and 32%, respectively, with respect to the results obtained with only NLD features. For CAV features the reduction in the absolute and relative errors is about 1% and 18%, respectively.

For the case of the combination of features from words /*coco*/ and /*gato*/, the best results are also obtained for the fusion of NLD and CAV features. The highest accuracy is 93.3%, which means improvements in the absolute and relative errors of about 3% and 32%, respectively, with respect to the results obtained with only NLD features. Comparing the result of the fusion with respect to those obtained with CAV features, the reduction in the absolute and relative errors is about 2% and 24%, respectively. In terms of AUC, there is an absolute improvement of 10%.

Figures 6 and 7 show the ROC curves with the best results obtained with the Spanish vowels and the words /*coco*/ and /*gato*/. Note that for both cases (vowels and words) the AUC values present a significant increment when the CAV and NLD are merged.

Table 2 Classification results for vowels and words using perturbation, noise and cepstral features, where SM:Selection Method, WS: Without Selection, NF: Number of Features

Vowel	SM	NF	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
/a/	WS	34	87.8±7.5	94.1±7.3	79.7±13.4	0.93
	PCA	23	88.3±7.3	92.6±6.2	83.5±12.2	0.94
	SFFS	26	86.5±6.4	91.2±7.1	81.3±13.4	0.93
/e/	WS	34	92.5±5.5	94.3±6.4	92.2±8.6	0.95
	PCA	29	90.4±6.2	93.7±8.6	85.5±8.7	0.96
	SFFS	24	93.3±2.2	93.1±6.8	93.4±6.2	0.96
/i/	WS	34	92.9±6.8	95.5±6.7	90.3±12.1	0.97
	PCA	32	91.1±3.8	96.5±3.8	84.6±11.2	0.96
	SFFS	23	93.2±3.1	93.7±6.2	92.7±5.9	0.97
/o/	WS	34	91.2±8.7	90.9±12.0	92.0±8.2	0.92
	PCA	31	91.2±6.0	90.0±7.8	92.6±7.2	0.94
	SFFS	25	89.9±4.0	90.0±8.4	89.9±11.0	0.94
/u/	WS	34	87.4±7.1	90.2±7.2	84.4±13.4	0.91
	PCA	29	86.6±6.5	84.4±13.2	89.2±8.7	0.89
	SFFS	23	86.6±8.6	86.8±11.1	87.0±10.7	0.91
/coco/	WS	34	87.4±9.1	83.1±18.2	90.9±9.7	0.94
	PCA	29	87.4±9.0	87.5±14.7	84.7±14.5	0.90
	SFFS	22	86.6±8.0	90.7±10.6	83.5±14.8	0.89
/gato/	WS	34	90.8±6.1	91.5±9.7	89.8±11.2	0.94
	PCA	31	90.8±10.7	91.8±13.0	90.2±21.0	0.93
	SFFS	23	88.3±8.0	87.8±9.7	85.5±20.9	0.87
All Vowels	WS	170	94.6±2.9	93.0±6.4	96.5±4.6	0.96
	PCA	144	93.7±3.6	94.7±5.4	93.1±7.1	0.97
	SFFS	121	92.9±4.4	92.4±4.42	93.2±6.5	0.95
All Words	WS	68	90.8±10.7	91.6±13.8	90.7±16.2	0.94
	PCA	60	91.6±8.8	93.4±11.7	91.4±11.5	0.92
	SFFS	45	86.7±10.5	87.5±13.3	82.5±16.4	0.90

* The results are presented in terms of mean value ± standard deviation

Table 3 Classification results for vowels and words using non-linear dynamics analysis, where SM: Selection Method, WS: Without Selection, NF: Number of Features

Vowel	SM	NF	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
/a/	WS	8	85.4±7.9	83.64±9.5	87.5±11.7	0.91
	PCA	8	84.8±9.0	81.70±13.1	88.4±13.2	0.90
	SFFS	8	85.3±6.6	83.73±8.1	87.0±8.9	0.91
/e/	WS	8	89.0±7.2	89.14±8.8	89.7±8.7	0.94
	PCA	6	89.1±7.0	87.39±10.3	91.2±11.5	0.93
	SFFS	7	88.6±5.4	87.20±8.1	90.2±8.4	0.92
/i/	WS	8	88.7±7.4	85.77±8.3	91.6±10.8	0.92
	PCA	6	87.9±7.5	86.24±15.1	89.2±13.6	0.88
	SFFS	7	90.4±6.2	88.90±8.7	92.1±6.1	0.93
/o/	WS	8	88.2±8.0	88.21±13.2	88.2±11.7	0.93
	PCA	8	87.8±5.3	86.81±11.0	87.9±8.9	0.94
	SFFS	7	87.8±7.2	87.84±8.7	88.8±10.7	0.93
/u/	WS	8	85.7±5.6	86.37±7.6	84.0±10.7	0.92
	PCA	8	88.2±4.3	89.30±4.3	85.5±11.0	0.92
	SFFS	7	88.2±4.8	90.11±8.4	86.6±7.4	0.91

/coco/	WS	8	84.8±9.5	82.84±12.9	89.5±11.5	0.87
	PCA	6	82.5±8.3	76.80±11.8	87.8±16.3	0.86
	SFFS	7	85.6±7.3	79.50±19.04	90.1±10.9	0.84
/gato/	WS	8	90.0±8.6	86.87±17.1	90.0±11.9	0.92
	PCA	6	89.0±8.1	87.64±12.7	91.3±12.5	0.90
	SFFS	6	88.3±13.7	84.56±20.1	95.7±7.1	0.92
All Vowels	WS	40	93.2±5.6	93.16±4.1	89.8±12.5	0.97
	PCA	36	92.8±5.9	97.02±3.9	88.2±10.2	0.96
	SFFS	36	91.6±4.9	94.7±5.0	88.0±13.2	0.97
All Words	WS	16	89.0±8.0	91.0±12.1	90.5±11.5	0.92
	PCA	12	89.9±6.8	91.6±11.0	89.7±14.3	0.92
	SFFS	13	90.6±9.6	88.7±15.9	91.3±16.4	0.94

* The results are presented in terms of mean value ± standard deviation

Table 4 Classification results for vowels and words using a combination of perturbation, noise, cepstral and non-linear dynamics features, where SM: Selection Method, WS: Without Selection, NF: Number of Features

Vowel	SM	NF	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
/a/	WS	42	91.2±4.20	88.4±5.5	94.6±4.8	0.94
	PCA	31	91.6±3.38	93.3±5.0	90.0±8.3	0.95
	SFFS	27	92.9±5.28	92.8±7.3	93.3±7.5	0.97
/e/	WS	42	93.7±5.63	94.3±6.4	93.2±7.2	0.96
	PCA	35	94.6±3.95	94.3±6.8	93.6±7.4	0.96
	SFFS	27	94.1±4.91	93.9±6.1	94.8±6.0	0.97
/i/	WS	42	93.3±4.87	93.6±5.4	93.9±8.5	0.97
	PCA	37	92.0±6.40	96.2±4.1	87.6±12.6	0.95
	SFFS	31	94.6±5.60	96.5±4.6	92.2±10.5	0.98
/o/	WS	42	94.5±4.51	93.5±5.0	94.8±7.6	0.96
	PCA	39	94.1±4.01	92.4±5.0	96.1±5.5	0.94
	SFFS	28	93.7±3.65	92.7±7.6	94.5±6.2	0.95
/u/	WS	42	90.8±7.0	91.6±11.8	91.3±7.3	0.91
	PCA	37	90.7±6.2	88.5±9.8	92.3±8.2	0.91
	SFFS	22	91.2±6.6	91.5±9.2	91.8±6.5	0.95
/coco/	WS	42	87.3±9.9	84.7±16.6	93.2±9.4	0.91
	PCA	35	88.3±8.1	87.6±14.3	92.1±14.5	0.93
	SFFS	22	90.8±4.7	95.2±8.0	81.6±15.0	0.92
/gato/	WS	42	92.5±9.2	93.6±8.8	91.7±14.7	0.94
	PCA	37	94.2±6.9	97.5±5.4	89.2±14.5	0.95
	SFFS	21	94.2±7.9	94.6±8.9	93.7±10.7	0.95
All Vowels	WS	210	94.9±3.9	95.8±4.9	94.4±6.7	0.98
	PCA	179	95.4±3.7	95.5±5.3	95.9±6.9	0.98
	SFFS	135	95.0±4.7	94.2±8.6	95.6±4.9	0.98
All Words	WS	84	91.7±5.6	91.4±10.0	93.5±10.9	0.94
	PCA	72	91.6±8.8	93.0±9.7	91.7±11.5	0.95
	SFFS	43	93.3±6.6	95.1±8.1	91.9±11.6	0.95

* The results are presented in terms of mean value ± standard deviation

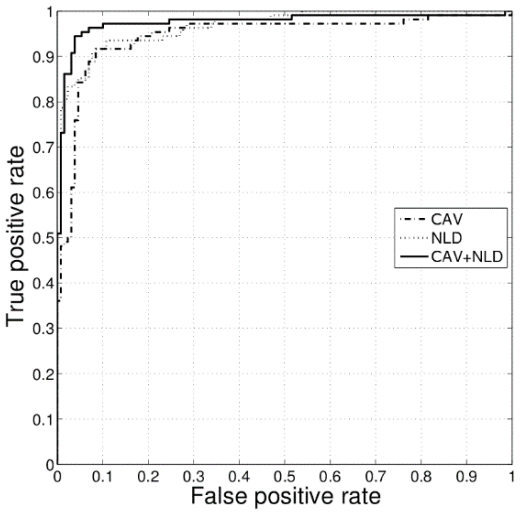


Figure 6 ROC curves for the evaluation of the five Spanish vowels in the same representation space considering CAV, NLD analysis and its combination

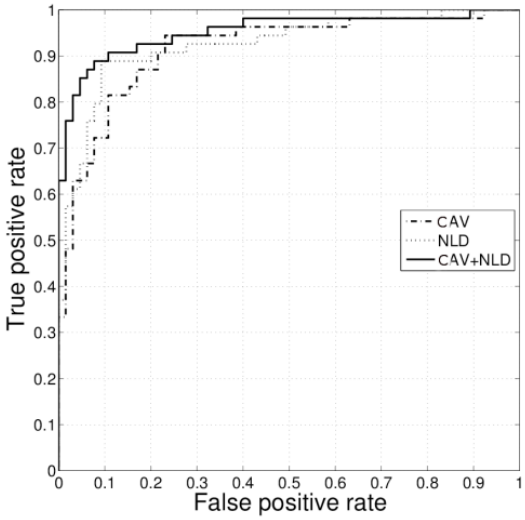


Figure 7 ROC curves for the evaluation of the words */coco/* and */gato/* in the same representation space considering CAV, NLD analysis and its combination

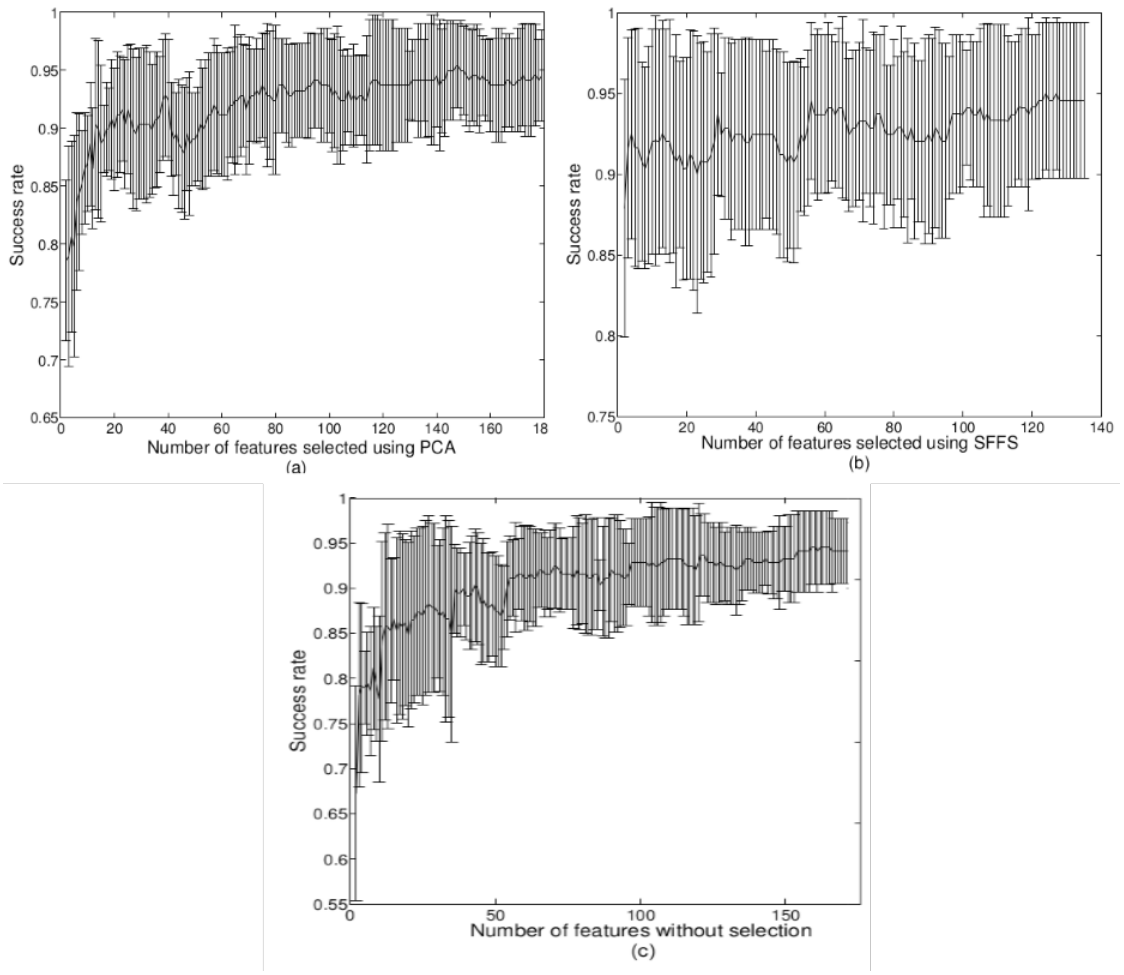


Figure 8 Improvement of the accuracies in vowels while the number of features is increased

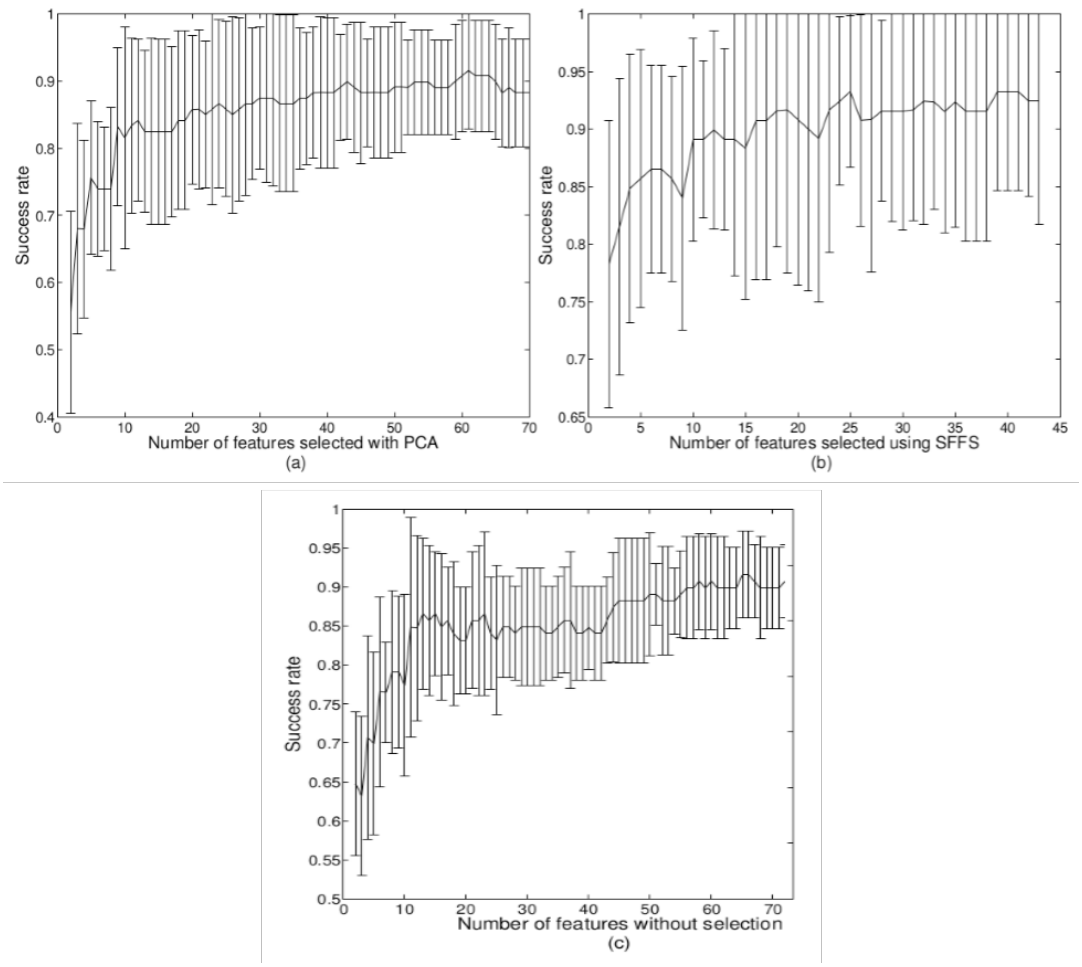


Figure 9 Improvement of accuracy in words while the number of features is increased

Figure 8 shows the improvement in the accuracy of the system when more features are considered to model the five Spanish vowels merged in the same representation space. Parts (a), (b), and (c) in Figure 8 are included to compare the results obtained with PCA, SFFS, and without feature selection, respectively. Figure 8 part (b) indicates that when SFFS selection technique is applied the number of required features is smaller than in the other cases. Note also that the accuracy obtained when using 130 features is the same as when using 60 features. On the other hand, it can also be observed that the results are more stable when PCA is applied than with SFFS.

Figure 9 shows the improvement of accuracies when more features are considered to model the words */coco/* and */gato/* and merged in the same space. Parts (a), (b), and (c) in Figure 9 include results with PCA, SFFS, and without feature selection technique, respectively. Note that the highest accuracies are obtained when no feature selection technique is applied.

The results indicate that the methodology proposed is suitable to detect Hypernasal speech for children with repaired cleft lip and palate, improving the results obtained using other methodologies such as the presented in [5]. Also, the presented results are higher than the obtained in previous works using the same database [11, 44].

4. Conclusions

Two different approaches for the characterization of healthy and hypernasal speech signals are presented. The first one is based on the classical analysis of voice (CAV), which includes pitch perturbation measures, noise measures, and 11 MFCC. The second approach is included to model the nonlinear behavior in speech and considers four complexity measures, correlation dimension, largest Lyapunov exponent, Hurst exponent and Lempel-Ziv complexity. The results show that nonlinear analysis provides complementary information to model hypernasal speech signals. The combination of both sets of features into a single feature space shows higher accuracies than applying each set of features separately.

The results indicate that the relative errors can be reduced in up to 32%, when the Spanish vowels or the words */coco/* and */gato/* are modeled with the fusion of NLD and CAV features.

The results indicate that the proposed approach is suitable to detect Hypernansality in children with repaired cleft lip and palate, which contributes in two stages for the assessment of speech therapy of the patients. First, the detection allows the phoniatricians to make informed decisions regarding the speech and language therapy of the patients, and second it is the first step for the development of computer aided tools that assist the evaluation of the progress of the speech therapy. The third issue related with the measurement of the degree of Hypernasality must be evaluated in future studies.

Additional experiments should be performed using speech recordings of different languages to evaluate the suitability of this methodology in different languages and databases. Further work must be carried out to extrapolate the proposed methodology for the evaluation of running speech.

5. Acknowledgments

Juan Rafael Orozco-Arroyave is under grants of "Convocatoria 528 para estudios de doctorado en Colombia, generación del bicentenario, 2011" funded by COLCIENCIAS. This work was also financed by COLCIENCIAS, project # 111556933858. The authors thank CODI, "estrategia de sostenibilidad 2014-2015 from Universidad de Antioquia" for the support to develop this work.

6. References

1. J. Arias, J. Godino, N. Sáenz, V. Osma and G. J. Arias, J. Godino, N. Sáenz, V. Osma and G. Castellanos, "An improved method for voice pathology detection by means of a HMM-based feature space transformation", *Pattern Recognition*, vol. 43, no. 9, pp. 3100-3112, 2010.
2. T. Yun, W. Ching and L. Guo, "Voice low tone to high tone ratio, nasalance, and nasality ratings in connected speech of native Mandarin speakers: a pilot study", *The Cleft Lip and Palate Journal*, vol. 49, no. 4, pp. 437-446, 2012.
3. A. Kummer, *Cleft palate and craniofacial anomalies: effects on speech and resonance*, 2nd ed. Cincinnati, USA: Cengage Learning, 2007.
4. A. Kummer and L. Lee, "Evaluation and Treatment of Resonance Disorders", *Language, Speech, and Hearing Services in Schools*, vol. 27, pp. 271-281, 1996.
5. P. Vijayalakshmi, M. Reddy and D. O'Shaughnessy, "Acoustic analysis and detection of hypernasality using a group delay function", *IEEE Transactions on Biomedical Engineering,* vol. 54, no. 4, pp. 621-629, 2007.
6. L. He *et al.*, "Automatic evaluation of hypernasality based on a cleft palate speech database", *Journal of medical systems*, vol. 39, no. 5, pp. 1-7, 2015.
7. K. Golding, "Therapy techniques for cleft palate speech and related disorders", 1st ed. New York, USA: Singular Thomson Learning, 2001.
8. J. Godino, P. Gómez and M. Blanco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters", *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 10, pp. 1943-1953, 2006.
9. A. Maier, F. Hönig, C. Hacker, M. Schuster and E. Nöth, "Automatic evaluation of characteristic speech disorders in children with cleft lip and palate", in *9th Annual Conference of the International Speech Communication Association* (INTERSPEECH), Brisbane, Australia, 2008, pp. 1757-1760.
10. A. Giovanni *et al.*, "Nonlinear behavior of vocal fold vibration: the role of coupling between the vocal folds", *Journal of Voice*, vol. 13, no. 4, pp. 465-476, 1999.
11. J. Orozco *et al.*, "Automatic selection of acoustic and non-linear dynamic features in voice signals for hypernasality detection", in *12th Annual Conference of the International Speech Communication Association* (INTERSPEECH), Florence, Italy, 2011, pp. 529-532.
12. H. Kantz and T. Schreiber, *Nonlinear time series analysis*, 2nd ed. Cambridge, UK: Cambridge University Press, 2004.
13. N. Sáenz, J. Godino, V. Osma and P. Gómez, "Methodological issues in the development of automatic systems for voice pathology detection", *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 120-128, 2006.
14. H. Wertzner, S. Schreiber and L. Amaro, "Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders", *Rev. Bras. Otorrinolaringol.*, vol. 71, no. 5, pp. 582-588, 2005.
15. L. Guo, W. Ching and S. Fu, "Evaluation of hypernasality in vowels using voice low tone to high tone ratio", *Cleft Palate Craniofacial Journal*, vol. 46, no. 1, pp. 47-52, 2009.
16. B. Boyanov and S. Hadjitodorov, "Acoustic analysis of pathological voices: A voice analysis system for the screening of laryngeal diseases", *IEEE Engineering in Medicine and Biology*, vol. 16, no. 4, pp. 74-82, 1997.
17. K. Shama, A. Krishna and N. Cholayya N, "Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology", *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1-9, 2007.
18. E. Yumoto, W. Gould and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness", *Journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544-1550, 1982.
19. G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals", *Journal of Speech, Language and Hearing Research*, vol. 36, no. 2, pp. 254-266, 1993.
20. P. Murphy and O. Akande, "Cepstrum-based Harmonics to Noise Ratio Measurement in voiced speech", *Lecture Notes in Artificial Intelligence*, vol. 3445, pp. 199-218, 2005.
21. H. Kasuya, S. Ogawa, K. Mashima and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice", *Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1329-1334, 1986.
22. D. Michaelis, T. Gramss and H. Strube, "Glottal-to-Noise Excitation Ratio - A new measure for describing pathological voices", *Acta Acust. united Ac.*, vol. 83, no. 4, pp. 700-706, 1997.

23. J. Godino *et al.*, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders", *Journal of Voice*, vol. 24, no. 1, pp. 47-56, 2010.
24. S. Bou and J. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress", *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 429-442, 2000.
25. J. Jiang, Y. Zhang and C. McGilligan, "Chaos in voice, from modeling to measurement", *Journal of Voice*, vol. 20, no. 1, pp. 2-17, 2006.
26. F. Takens, "Detecting strange attractors in turbulence", *Lecture Notes in Mathematics*, vol. 898, pp. 366-381, 1981.
27. P. Henriquez *et al.*, "Characterization of healthy and pathological voice through measures based on nonlinear dynamics", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1186-1195, 2009.
28. A. Shaheen, N. Roy and J. Jiang, "Nonlinear dynamic analysis of disordered voice: the relationship between the correlation dimension (D2) and pre-/post-treatment change in perceived dysphonia severity", *Journal of Voice*, vol. 24, no. 3, pp. 285-293, 2010.
29. J. Arias, J. Godino, N. Sáenz, V. Osma and G. Castellanos, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients", *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 370-379, 2011.
30. P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors", *Physica D*, vol. 9, no. 1-2, pp. 189-208, 1983.
31. H. Abarbanel, *Analysis of observed chaotic data*, 1st ed. New York, USA: Institute of Nonlinear Science, 1999.
32. M. Rosenstein, J. Collins and C. De Luca, "A practical method for calculating largest Lyapunov exponents from small data sets", *Physica D*, vol. 65, no. 1-2, pp. 117-134, 1993.
33. V. Oseledec, "A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems", *Transactions of Moscow Mathematic Society,* vol. 19, pp. 197-231, 1968.
34. H. Hurst, R. Black and Y. Simaika, *Long-term storage: an experimental study*, 1st ed. London, UK: Constable, 1965.
35. F. Kaspar and H. Shuster, "Easily calculable measure for complexity of spatiotemporal patterns", *A Physical Review*, vol. 36, no. 2, pp. 842-848, 1987.
36. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, USA: Springer, 2002.
37. R. Bro and A. Smilde, "Principal component analysis", *Analytical Methods*, vol. 6, no. 9, pp. 2812-2831, 2014.
38. P. Pudil, J. Novovicova and J. Kittler, "Floating search methods in feature selection", *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119-1125, 1994.
39. P. Pudil, F. Ferri, J. Novovicova and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions", in 12th IAPR *International Conference on Pattern Recognition*, Jerusalem, Israel, 1994, pp. 279-283.
40. B. Scholköpf and A. Smola, *Learning with Kernels*, 1st ed. Cambridge, USA: The MIT Press, 2002.
41. D. Kuehn and K. Moller, "Speech and language issues in the cleft palate population: the state of the art", *Cleft Palate-Craniofacial Journal*, vol. 37, no. 4, pp. 1-35, 2000.
42. R. Carvajal, N. Wessel, M. Vallverdú, P. Caminal and A. Voss, "Correlation dimension analysis of heart rate variability in patients with dilated cardiomyopathy", *Computer Methods and Programs in Biomedicine*, vol. 78, no. 2, pp. 133-140, 2005.
43. M. Ding, C. Grebogi, E. Ott, T. Sauer and J. Yorke, "Estimating correlation dimension from chaotic time series: when does plateau occur?", *Physica D*, vol. 9, no. 3-4, pp. 404-424, 1993.