Automatic Detection of Pathologies in the Voice by HOS Based Parameters

Jesus B. Alonso
Departamento de Señales y Comunicaciones, Universidad de Las Palmas de Gran Canaria, Campus de Tafira, 35017 Las Palmas de Gran Canaria, Spain
Email: jalonso@cibeles.teleco.ulpgc.es

José de Leon
Medical Doctor at the ENT Department, General Hospital G.C. Dr. Negrín, Las Palmas de Gran Canaria, Spain
Email: jleond@seorl.org

Itziar Alonso
Departamento de Ingeniería Telemática, Universidad de Las Palmas de Gran Canaria, Campus de Tafira, 35017 Las Palmas de Gran Canaria, Spain
Email: ialonso@dit.ulpgc.es

Miguel A. Ferrer
Departamento de Señales y Comunicaciones, Universidad de Las Palmas de Gran Canaria, Campus de Tafira, 35017 Las Palmas de Gran Canaria, Spain
Email: mferrer@dsc.ulpgc.es

Received 27 September 2001 and in revised form 3 October 2001

In the current panorama the conclusive identification of a laryngeal pathology relies inevitably on the observation of the vocal folds by means of laryngoscopical techniques. This inspection technique is inconvenient for a number of reasons, such as its high cost, the duration of the inspection, and, above all, the fact that it is an invasive technique. This paper looks into the possibility of measuring the quality of a voice starting from an audio recording. The existing parameters in current literature ("classic parameters") which allow quantifying the quality of a voice have been studied, and the parameters that present better results have been selected. Also, seven new high order statistics (HOS) based parameters are proposed to parameterize the voice signal. On the other hand, a software package has been developed which carries out the automatic detection of dysfunction in phonation. A success rate of 98.3% has been obtained by using both the classic and the HOS based proposed parameters.

Keywords and phrases: laryngeal pathology, voice quality, automatic detection of dysfunction, speech processing.

1. INTRODUCTION

Within the medical environment, and especially in the field of ENT, diverse techniques exist to assess the state of the voice of a patient. The most utilized techniques are based on the audition of the patient’s voice and the visual inspection of the vocal folds by means of laryngoscopical techniques. The technique based on the patient’s audition has the advantage of being a simple and straightforward technique to apply. However, it presents the inconvenience of being a subjective technique, that is, the result of the evaluation in some cases depends on the medical professional who carries out the assessment of the state of the vocal folds. Furthermore, the medical professional involved requires previous specific training. On the other hand, the technique based on the visual inspection of the vocal folds offers the advantage of being a very precise technique, but it is expensive, in some cases ineffective and time consuming, and it is above all an invasive technique which causes discomfort to patients and a natural reluctance to undergo the tests.

Within the medical environment it would be very useful to have tool that did not entail the inconveniences quoted above, a tool which would allow the documentation of the quality of a patient’s voice in a quantitative way. This tool would enable medical professionals to show in numerical values the evolution of a patient being treated for pathology in the vocal folds.
The objective of this work is the search for a technique that will allow the quantification of a speaker’s voice quality by means of an audio sample. This technique will allow us not only to show the evolution of the patient’s voice quality throughout the treatment, but it could also be applied in the field of preventive medicine in order to achieve early detection of laryngeal pathologies, as well as in the field of telemedicine. This paper examines and assesses the different parameters published in current literature (quoted here as “classic” parameters) which allow the quantification of voice quality. Furthermore, new HOS based voice parameters are presented in this work in order to compare healthy and pathological voice samples. These parameters are denominated: the interference value of the bicoherence index, the relative value of the high frequency energy of the one-dimensional bicoherence index, the relative value of the low frequency energy of one-dimensional bicoherence index and the value of their interference, the variation of the estimated noise by means of the module of the bispectrum, the variation of the noise obtained through the same technique taking into account in this case the phase information and the interference of the kurtosis.

A software package has also been developed which carries out the automatic detection of dysfunction in phonation. The voice signal is parametrized by means of the classic parameters (Hitter, Shimmer, Energy Balance, Spectral Distance) and the new proposed parameters, and then they are evaluated by a nonlinear classifier (Neural Network). The developed application has obtained a global rate of success of 98.3%.

2. CLASSIC PARAMETERS

In current literature there are diverse works about parameters which allow the quantification of voice quality, such as Boyanov and Hadjitodorov [1], Fröhlich et al. [2], Gavidia-Ceballos and Hansen [3], Guoxin et al. [4], and Feijoo et al. [5].

In this paper, we have considered that the best method to evaluate voice quality from a voice recording is to start from the analysis of a sustained voiced sound. In this study the five Spanish vowels have been used in a sustained way (a duration of about 2 seconds of phonation and a processing of 500 central milliseconds). They are assessed separately, as indicated by Koreman and Pützer [6].

A database was used to undertake the test, which consisted of 100 samples of voices of healthy speakers and 68 samples of voices of pathological speakers, which presented different laryngeal pathologies. One of the objective in the creation of the database was to have the most varied selection of pathologies. Examples of pathologies used in the database are: vocal folds nodule, sessile polyp, pedunculated polyp, Reinke’s edema, adult papilloma, extensive papilloma, vocal fold scar, hyperfunction, hyperfunction, vocal fold paralysis, and carcinoma. Also, pathological voice samples of different dysphonic degrees have been looked at. These voice samples were grouped into three levels according to their dysphonic degree: light, moderate, and severe.

Different parameters from current literature were initially evaluated to carry out a selection of those that presented better results. The approach for the selection of the parameters based on the probability density functions of each one of the parameters for the population of samples of healthy voices and pathological voices. The parameters that allowed a bigger discrimination percentage between the two populations were selected.

It is necessary to keep in mind that there is no parameter which is completely conclusive in the detection of laryngeal pathologies. Laryngeal pathologies affect voice quality in different ways. For example, there are pathologies that present a great content of nonstationary noise in the high frequency components, whereas other pathologies are characterized by the uncertainty of the value of the pitch throughout the duration of the phonation of a sustained voiced sound. Classic parameters have consequently been divided into 5 groups: quantifying the variation in amplitude (shimmer), quantifying the presence of unvoiced frames, quantifying the absence of wealth spectral (Hitter), quantifying the presence of noise, and quantifying the regularity and periodicity of the waveform of a sustained voiced voice.

In the following sections, the parameters that presented better results are indicated and their classification are presented within the different aforementioned groups. Also, new parameters based on classic parameters are proposed.

2.1. Quantifying the variation in amplitude (shimmer)

Some pathology voice samples are characterized by the fact that they present a great variation of amplitude and waveform during the phonation of a sustained voiced sound. This phenomenon is due to the presence of fleshy growths in the vocal folds or to the decrease of the elastic properties of the vocal folds. The parameters that have been selected in this group are indicated next.

2.1.1 Variation in the highest value of each voice frame

The great variation in the maximum positive value of each voice frame in the time domain is a phenomenon observed in pathological voice samples. This parameter, also called “shimmer,” is determined by means of the variation in the maximum amplitude among adjacent frames.

Amplitude interference is defined by Boyanov and Hadjitodorov [1], Wallen and Hansen [7], Michaelis and Strube [8], and Fröhlich et al. [9] as

\[
AP = \left( \frac{1}{N-1} \right) \left( \frac{1}{S_{\text{max}}} \right) \sum_{i=1}^{N-1} \left| S_{A}(i+1) - S_{A}(i) \right|, \quad (1)
\]

where \(AP\) is the value of amplitude interference over a period of time, \(S_{A}\) is the value of the positive maximum pick of each voice frame, \(S_{\text{max}}\) is the maximum value of all \(S_{A}\), and \(N\) is the number of audio frame that is analyzed.

2.1.2 Variation in the mean quadratic value of each voice frame

The interference of the mean quadratic value of each frame is another parameter that quantifies the variation in amplitude
over a period of time, as discussed by Marasek [10]. The mean quadratic value of each frame is defined in the following way:

\[ VCM = \frac{1}{L} \cdot \frac{1}{L} \sum_{i=1}^{L} (x(i) - \mu_x)^2, \]  

(2)

where VCM is the mean quadratic value of one frame, \( x(n) \) is the voice frame that has a length equal to \( L \) samples, and \( \mu_x \) is the mean value of the frame.

The interference of the mean quadratic value is defined in a similar way to the interference of the maximum value.

### 2.1.3 Variation in the highest value of the short time cross correlation function of each voice frame

The “Short Time Cross Correlation Function” (STCCF) parameter, defined by Michaelis and Strube [8], quantifies the degree of similarity between the waveforms of two consecutive periods of pitch. This quantification of the degree of similarity is gotten by means of the cross correlation function of two consecutive pitch periods.

The mean value of the STCCF for a sustained voice sound is defined as “mean waveform matching coefficient” (MWC) and was defined by Michaelis and Strube [8] and Fröhlich et al. [11]. This proposed parameter is the interference of the maximum value of the STCCF in each frame.

### 2.2. Quantifying the presence of unvoiced frames

Some pathological voice samples present unvoiced frames due to the impossibility of carrying out a periodical glottal closure. This phenomenon is caused by the patients having significant fleshy growths in their vocal folds.

#### 2.2.1 Relationship between the number of unvoiced frames and the total number of frames of the sample voice

This parameter is calculated as the relationship between the number of unvoiced frames and the total number of frames during the phonation of a sustained vowel, as defined by Boyanov and Hadjitodorov [1]. The degree of sonority “DUV” is calculated by means of the following equation:

\[ DUV = \frac{N_{\text{unvoiced frames}}}{N_{\text{frames}}}, \]  

(3)

where \( N_{\text{unvoiced frames}} \) is the number of unvoiced frames and \( N_{\text{frames}} \) is the total number of frames.

#### 2.2.2 The unvoiced periodicity index of a sample voice

In this work, we have studied if the production of unvoiced frames during the phonation of a voiced sound is produced in a consecutive way or in a quasiperiodical way. As a result, a new parameter denominated “unvoiced periodicity index” (UPI) has been proposed, that quantifies the number of transitions: unvoiced-voiced frames or voiced-unvoiced frames with regard to the total frames (\( N_{\text{frames}} \))

\[ UPI = \frac{\text{number of transitions}}{N_{\text{frames}}}, \]  

(4)

#### 2.3. Quantifying the absence of spectral wealth (Hitter)

Some samples of pathological voices present an absence of spectral wealth due to the difficulty in performing a correct glottal closure and this fact produces in a pathological speaker a different frequency of pitch during a phonation of a sustained voiced sound.

#### 2.3.1 Variation of pitch energy cepstral measure

Power cepstrum is defined by Petropulu and Nikias [12] as

\[ p_x(n) = F^{-1} \{ \log (|X(\omega)|^2) \}, \]  

(5)

where \( \{ x(n) \} \) is the discrete sequence with length \( N \) of a voice frame, whose Fourier transform is \( X(\omega) \).

The well-known parameter “pitch energy cepstral measure” (PECM), defined by Boyanov and Hadjitodorov [1], evaluates the relationship between the cepstral energy concentrated on the pulse of the pitch and the total cepstral energy of each cepstral window, in the context of the real cepstrum. The PECM is calculated for each frame.

The PECM is expressed as follows:

\[ \text{PECM} = \frac{E_{\text{pulse of pitch}}}{E_{\text{total}}}, \]  

(6)

where

\[ E_{\text{pulse of pitch}} = \sum_{n=n_0}^{n_1} c(n), \quad E_{\text{total}} = \sum_{n=0}^{L} c(n), \]  

(7)

where \( c(n) \) is the cepstral coefficient, \( n_0 \) is the number of the sample after the last crossing by zero which is produced to the left of the pitch cepstral position, and \( n_1 \) is the number of the sample before the first crossing by zero which is produced to the right of the pitch cepstral position. \( L \) is the total number of cepstrum coefficients of each frame.

#### 2.3.2 Variation of the first harmonic value in the derived cepstrum domain

A cepstrum window that is spatial derived is denominated derived cepstrum, defined by Picone [13]. In this section the effect of the pitch has been studied in the context of the derived cepstrum domain. The derived cepstra are obtained by means of a differentiating ideal, which is implemented using a linear phase filter. This filter is defined in the following way:

\[ \hat{s}(n) = d \frac{d}{dn} s(n) \approx \sum_{m=-N_d}^{N_d} m \cdot s(n + m), \]  

(8)

where \( N_d \) has been used in this paper as equaling five samples.

The parameter “value of the first harmonic in the derived cepstrum domain” (Vald1) is defined as the maximum value that the derived cepstrum reaches around the frequency of the pitch (first harmonic). The new parameter that is proposed in this study is the interference of the “Vald1.”
2.3.3 Variation in the first/second harmonic relationship value within the derived cepstrum domain

The parameter “value of the quotient second/first harmonic in the derived cepstrum domain” (Vald2/d1) is defined as the relational quotient between the maximum value of the derived cepstrum around the second harmonic and the maximum value that the derived cepstrum reaches around the frequency of the pitch.

This new parameter proposed in this study is the interference of “Vald2/d1.”

2.4. Quantifying the presence of noise

The parameters that allow quantifying either the presence of stationary noise in certain regions of the spectrum of frequencies or the presence of nonstationary noise along the whole spectrum of frequencies have been included in this group of parameters. This phenomenon is characteristic of certain pathological voices and it is due to the impossibility for a pathological speaker to carry out a perfect glottal closure either permanently or over certain periods.

2.4.1 Energy spectral balance

There are some studies on the relationship between the physiology of vocal folds and acoustic measures, as discussed by Fröhlich et al. [2], Guoxin et al. [4], Hansen et al. [14], Rosa et al. [15], and Maraske [10]. There are certain acoustic measures that relate calculated parameters from the formants and from the energy of the spectrum which allow the detection of different dysfunctions in the vocal folds, as discussed by Fröhlich et al. [2].

Beside these measures, there are others which quantify the energy of the spectrum for different regions of the spectrum of frequencies, as discussed by Fröhlich et al. [2]. Although Boyanov and Hadjitodorov [1] distinguished only between high and low frequencies, the localization of the different bands of frequencies, used in this work, is the following one:

- Region 0: from 60 Hz to 400 Hz.
- Region 1: from 400 Hz to 2 KHz.
- Region 2: from 2 KHz to 5 KHz.
- Region 3: from 5 KHz to 8 KHz.
- Region 4: from 8 KHz to 11025 KHz.

The parameters that relate these energy bands are $L_3L_2$ and $L_4L_1$. The general expression for the previous parameters is

$$L_jL_i = \frac{\sum_{\omega=\text{region } j} |Y_{\text{frame}}(\omega)|^2}{\sum_{\omega=\text{region } i} |Y_{\text{frame}}(\omega)|^2},$$

where $|Y_{\text{frame}}(\omega)|^2$ is the power spectrum of a frame.

The parameter “relative spectral energy between the regions 4 and 1” ($L_4L_1$) is defined as the relationship between the energy of the power spectrum in the region between 8 KHz and 11 KHz and the energy of the power spectrum in the region between 60 Hz and 400 Hz. This is quantified in the following way:

$$L_4L_1 = \frac{\sum_{\omega=8 \text{ KHz–11 KHz}} |Y_{\text{frame}}(\omega)|^2}{\sum_{\omega=60 \text{ Hz–400 Hz}} |Y_{\text{frame}}(\omega)|^2}. $$

The parameter that is proposed is the value of “$L_4L_1$.”

The weight of the energy of each region is also analyzed with regard to the total energy of the spectrum. This is quantified in the following way:

$$L_{\text{region } i}L_t = \frac{\sum_{\omega=\text{region } i} |Y_{\text{frame}}(\omega)|^2}{\sum_{\omega=60 \text{ Hz–11 KHz}} |Y_{\text{frame}}(\omega)|^2}. $$

The parameter “relative spectral energy in region 1” ($L_1L_t$) is defined as the relationship between the energy of the power spectrum in the region between 400 Hz and 2 KHz and the total energy of the power spectrum. This is quantified in the following way:

$$L_1L_t = \frac{\sum_{\omega=400 \text{ Hz–2 KHz}} |Y_{\text{frame}}(\omega)|^2}{\sum_{\omega=60 \text{ Hz–11 KHz}} |Y_{\text{frame}}(\omega)|^2}. $$

The parameter that is proposed is the interference of the value of “$L_1L_t$.”

The parameter “relative spectral energy in region 2” ($L_2L_t$) is defined as the relationship between the energy of the spectrum of power in the region between 2 KHz and 4 KHz and the total energy of the spectrum of power. This is quantified in the following way:

$$L_2L_t = \frac{\sum_{\omega=2 \text{ KHz–4 KHz}} |Y_{\text{frame}}(\omega)|^2}{\sum_{\omega=60 \text{ Hz–11 KHz}} |Y_{\text{frame}}(\omega)|^2}. $$

The parameter that is proposed is the interference of the value of “$L_2L_t$.”

2.4.2 Spectral distance (metrics based on the spectral module)

The following distance measure, which is obtained from the spectrum module, is proposed

$$D_{\text{module}} = \sum_{\omega=0}^L \left| |T(\omega)| - |T_{\text{ref}}(\omega)| \right|, $$

where $D_{\text{module}}$ is the value of the distance measured between the modules, $|T(\omega)|$ is the spectrum module of the analyzed voice frame which is calculated by FFT, $|T_{\text{ref}}(\omega)|$ is the module of the previous frame, and $L$ is the FFT length.
2.4.3 **Spectral distance (metrics based on the spectral phase)**

Likewise, in the spectral domain, a new parameter is proposed which is defined by means of the following expression:

\[ D_{\text{phase}} = \sum_{\omega=0}^{L} \left| |\varphi_T(\omega)| - |\varphi_{T_{\text{ref}}}(\omega)| \right|, \quad (16) \]

where \( D_{\text{phase}} \) is the value of the phase distance measure, \( \varphi_T(\omega) \) is the spectrum phase vector of the voice frame, and \( \varphi_{T_{\text{ref}}}(\omega) \) is the spectrum phase vector of the previous frame.

2.5. **Quantifying the regularity and periodicity of the waveform of a sustained voiced voice**

There are studies indicating that the resonance characteristics of a voice production model can change quickly as much in amplitude as in frequency, and even over a period of pitch, as discussed by Hansen et al. [14] and Gavidia-Ceballos et al. [16]. By means of a nonlinear differential operator \( T_{\text{EO}} \) (the operator Teager), a modulation of amplitude (AM) and a modulation of frequency (FM) can be detected around a resonance (formant) of the vocal tract, as discussed by Hansen et al. [14], Gavidia-Ceballos et al. [16], Lu and Doerschuk [17], and Lu and Doerschuk [18]. The operator Teager is regarded as an estimator of high-resolution energy. The operator \( T_{\text{EO}} \) is used in the signal processing in several areas, as in, for example, the detection of a hypernasal voice, Cairns and Hansen [19], and the detection of stress by means of the voice as discussed by Zhou et al. [20].

The proposed parameters are: the value and variation in energy and the variation of the slope, of the envelope in the autocorrelation function of an AM modulating signal, obtained from a pass band filter around the first formant.

3. **NEW PARAMETERS**

The new parameters proposed in this work have been obtained from a study on the utility of the high-order statistical (HOS) applied to the detection of dysphony in phonation. Most of the parameters that are proposed were obtained directly or indirectly starting from the bispectrum of a voice frame.

3.1. **Quantifying the nonlinear behaviour**

In same healthy voice samples there is a greater presence of quadratic phase coupling. This phenomenon is quantified by means of the irregularity of the bicoherence index and the energy balance of the one-dimensional bicoherence index.

3.1.1 **Value of the bicoherence index interference (IB)**

The bicoherence index is defined by Mendel [21] and Nikias and Raghuveer [22] as

\[ b(\omega_1, \omega_2) = \frac{B(\omega_1, \omega_2)}{P(\omega_1)P(\omega_2)P(\omega_1 + \omega_2)}, \quad (17) \]

where \( B(\omega_1, \omega_2) \) is the bicoherence index, the dimension of the bicoherence index is \( \dim_x \) and \( \dim_y \), and \( N \) is the product between \( \dim_x \) and \( \dim_y \). In Figures 1 and 2 you can see a representation of the bicoherence index as a sample of a healthy voice and a sample of a pathological voice, respectively.

The ability of the IB parameter to discriminate between healthy and pathological voices can be attributed to the fact that in a healthy voice there is a greater presence of quadratic phase coupling, and this is because the healthy voice is characterized by a vocal tract which is more markedly non-linear than in pathological voices.

3.2. **Study of the one-dimensional bicoherence index**

The possibility of carrying out the calculation of the one-dimensional version of the bicoherence index has been studied, in the same way as it has been carried out in the case of the bispectrum by Elgar and Guza [23], Moreno et al.
Figure 2: Representation of the bicoherence index showing a sample of a pathological voice.

[24], and Moreno and Rutllán [25]. Following them, the one-dimensional bicoherence index is defined as:

\[ b_{\text{one-dimensional}}(\omega) = \sum_{\omega_i=0}^{\text{dimy}-1} b(\omega, \omega_i), \]  (21)

where \( b(\omega_1, \omega_2) \) is the bicoherence index.

A phenomenon that is observed when representing a one-dimensional version of the bicoherence index \( b_{\text{one-dimensional}}(\omega) \) is the presence of higher spectral components in low frequencies in pathological, as opposed to healthy, voice samples. This phenomenon can be observed in Figure 3, where the one-dimensional bicoherence index is shown for a sample of both a healthy and a pathological voice.

![Figure 3: One-dimensional bicoherence index of a sample of healthy voice and a sample of pathological voice.](image)

Dividing the one-dimensional bicoherence index in two regions on the boundary around the frequency of 5300 Hz, a superior and an inferior area of a one-dimensional bicoherence index can be distinguished. The parameters studied for the one-dimensional bicoherence index are the following ones: high and low frequency energy relative to the total energy. These energies are defined in (22) and (23), respectively,

\[ \text{HighEnergy} = \frac{\sum_{\omega_i=0}^{\omega_{\text{max}}} b_{\text{one-dimensional}}(\omega)}{\sum_{\omega} b_{\text{one-dimensional}}(\omega)} = \frac{\sum_{\omega_i=0}^{\omega_{\text{max}}} b(\omega, \omega_i)}{\sum_{\omega} b(\omega, \omega_i)}, \]  (22)

\[ \text{LowEnergy} = \frac{\sum_{\omega_i=0}^{\omega_{\text{threshold}}} b_{\text{one-dimensional}}(\omega)}{\sum_{\omega} b_{\text{one-dimensional}}(\omega)} = \frac{\sum_{\omega_i=0}^{\omega_{\text{threshold}}} b(\omega, \omega_i)}{\sum_{\omega} b(\omega, \omega_i)}. \]  (23)

The parameters that are proposed for the detection of dysfunctions in phonation are the following ones:

1. Value of the High Frequency Energy of the one-dimensional bicoherence index (HFEUBI): it is the value that the parameter “HighEnergy” takes for each voice frame (equation (22)).

2. Value of the Low Frequency Energy of the one-dimensional bicoherence index (LFEUBI): it is the value that the parameter “LowEnergy” takes for each voice frame (equation (23)).

3. Interference of the Value of the Low Frequency Energy of the one-dimensional bicoherence index (ILFEUBI): this parameter quantifies the variation over time of the value of the low frequency energy of the one-dimensional bicoherence index that each voice frame takes (equation (24)). The value of the interference is calculated in the following way:

\[ \text{ILFEUBI} = \left( \frac{1}{N-1} \right) \left( \frac{1}{\text{LFEUBI}_{\text{max}}} \right) \times \sum_{i=1}^{N-1} | \text{LFEUBI}(i+1) - \text{LFEUBI}(i) |, \]  (24)

where \( N \) is the number of voice frames of the sample that is evaluated.

The ability of the above parameters to discriminate between a healthy and a pathological voice can be attributed to a greater number of quadratic phase coupling taking place in the low frequencies, as opposed to the high frequencies. This could indicate that a model of healthy vocal tract presents a more marked nonlinear characteristic in low frequencies, and that it is also variable over time.

3.3. Quantifying glottal noise

In some pathological voice samples there is a greater presence of nonstationary noise because of a lack of a correct glottal closure. The nonstationary noise can be detected by the spectral subtraction of the spectrum calculated using the bispectrum from the spectrum calculated using the FFT.

In this paper, we have experimented with the possibility of separating the harmonic and the noise components of the power spectrum by means of the bispectrum, using the technique put forward by Sundaramoorthy et al. [26]. It was presupposed that most of the noise present in the power spectrum of the voice is distributed symmetrically, as suggested by
Automatic detection of pathologies in the Voice by HOS based parameters

by means of the FFT and Spectrum where estimation of the noise is obtained as follows: into account in this case the phase information (MPN). The of the noise obtained through the same technique, taking means of the bispectrum module (MN) and the interference in this section are: the interference of the estimated noise by the same voice frame obtained by means of the bispectrum. This phenomenon is shown in Figure 4.

According to the above, the parameters that are proposed in this section are: the interference of the estimated noise by means of the bispectrum module (MN) and the interference of the noise obtained through the same technique, taking into account in this case the phase information (MPN). The estimation of the noise is obtained as follows:

$$\text{Noise}_{\text{module}} = \sum (|\text{Spectrum}_{\text{FFT}}| - |\text{Spectrum}_{\text{Bispectrum}}|)^2,$$

$$\text{MN} = \left(\frac{1}{N-1}\right) \left(\frac{1}{\text{Noise}_{\text{module}}_{\text{max}}} \right) \times \sum_{i=1}^{N-1} \text{Noise}_{\text{module}}(i + 1) - \text{Noise}_{\text{module}}(i),$$

$$\text{Noise}_{\text{module-phase}} = \sum (\text{Spectrum}_{\text{FFT}} - \text{Spectrum}_{\text{Bispectrum}})^2,$$

$$\text{MPN} = \left(\frac{1}{N-1}\right) \left(\frac{1}{\text{Noise}_{\text{module-phase}}_{\text{max}}} \right) \times \sum_{i=1}^{N-1} \text{Noise}_{\text{module-phase}}(i + 1) - \text{Noise}_{\text{module-phase}}(i),$$

where $\text{Spectrum}_{\text{FFT}}$ is the spectrum of a voice frame obtained by means of the FFT and $\text{Spectrum}_{\text{Bispectrum}}$ is the spectrum of the same voice frame obtained by means of the bispectrum.

### 3.4. Quantifying statistical behaviour

Some pathological voice samples present longer variation in their statistics, this phenomenon is quantified by mean of the Kurtosis. The Kurtosis is defined as

$$\text{Kurtosis} = E\left(\frac{(X - \mu)^4}{\sigma^4}\right).$$

The last proposed parameter is the interference of the value that each voice frame takes of the Kurtosis, shown in the following equation:

$$IK = \left(\frac{1}{N-1}\right) \left(\frac{1}{\text{Kurtosis}_{\text{max}}} \right) \times \sum_{i=1}^{N-1} |\text{Kurtosis}(i + 1) - \text{Kurtosis}(i)|.$$

### 4. Pathologies Detector Diagram

In a general way, voice recognition allows the identification of predetermined classification units from a registered voice signal. The type of classification units will depend on the type of voice recognition. There are different types of voice recognition, examples of which are digits recognition, which allows us to identify different digits, or language recognition, which allows the identification of the speaker’s language.

In this work, a software package has been developed which allows carrying out the automatic detection of laryngeal pathologies. This software consists of a voice recognition system which allows identifying healthy voice samples and pathological voice samples as units of classification. In the developed software package, the speaker must pronounce the five Spanish vowels in a sustained way during approximately 2 seconds.

The general diagram of an automatic detection of laryngeal pathologies is shown in Figure 5.

In the following section, the most outstanding characteristics in each block are commented.

#### 4.1. Voice acquisition

The block “voice acquisition” has the purpose of digitizing the acoustic signal of a speaker’s voice. In the digitization process a frequency of sampling of 22050 Hz and 16 bits has been used per sample. Besides digitizing the audio signal, in this block a discrimination has been carried out between the part of the audio signal that corresponds to voice and the part of the audio signal that corresponds to noise, using the algorithm proposed by Boll [28]. This process is necessary because in most of the voice recognition applications, the signal that corresponds to audio among different voice blows, or embedded at the beginning and end of the voice blow, belongs to noise and its study is not useful. Also in this block, we have carried out a segmentation of the voice signal in diverse frames. For this reason we have used a Hamming window of a variable length of frame of three pitch periods (in the case of it being an unvoiced audio, the length of the frame is 30 milliseconds).
4.2. Parametrization

In the block “parametrization” we have carried out a quantification of the different frames. This quantification will allow us to discover differential characteristics among the different classification units. In this block, each voice frame is turned into a parameter vector.

In the parametrization phase, both classic parameters and the new HOS based parameters have been used. To evaluate the parameter vector in the classifier, it is necessary to carry out a normalization of the parameter vector, because each parameter quantifies different magnitudes.

4.3. Classifier

The HOS based parameters proposed together with the classic parameters (hitter, shimmer, energy balance, spectral distance) have been studied simultaneously by means of a non-linear classifier (neural network), for the automatic detection of dysfunctions in phonation.

A database of vowels consisting of 100 healthy speakers and 68 pathological speakers has been used to evaluate the system. Each speaker was asked to say the five Spanish sustained vowels. In the case of pathological speakers there are cases of vocal folds without lesion (hypofunction, hyperfunction, vocal fold paralysis,...) and vocal folds with lesion (carcinoma, vocal folds nodule, sessile polyp, pedunculated polyp, Reinke’s edema, adult papilloma,...).

The dependence of the parameters on the analyzed vowel has been taken into account, as pointed out by Koreman and Pützer [6]. Consequently, a “vowel classifier” has been used for each vowel, such as is shown in Figure 6. First of all, each “vowel classifier” emits an estimation dependent on whether the analyzed vowel is related to a “healthy vowel” or to “pathological vowel.” Secondly, the results of the different vowel classifiers are evaluated by means of an “output logic.” The output logic will indicate that the voice sample corresponds to a “pathological voice” if two or more vowels are classified as “pathological vowels,” whereas the voice sample will be classified as a “healthy voice” if only one vowel or none of them are classified as “pathological vowels.”

Each vowel classifier evaluates one by one each voice frame of a certain vowel. 500 milliseconds of each vowel is enough for the analysis, as pointed out by Koreman and Pützer [6]. Each voice frame is evaluated in two neural networks, and an assessment is emitted: “healthy frame” or “pathological frame.” If 70% or more of the frames correspond to healthy frame, the analyzed vowel will be labeled as a
“healthy vowel,” otherwise it will be labeled as a “pathological vowel.” The scheme of a vowel classifier is shown in Figure 7.

For the evaluation of the operating capacity of the automatic recognition system, 80% of the voice samples have been used for the training phase of the neural network, and the 20% of the remaining samples have been used for the evaluation phase. A global rate of success of 94.4% has been obtained by the use of the classic parameters, whereas a global rate of success of 98.3% has been obtained by using both the classic parameters and the new proposed parameters based on HOS.

5. SUMMARY

In this paper, certain general characteristics of pathological voices have been observed which allow their identification. These characteristics are the following ones:

1. “Some pathological voices present a short time greater variation in certain acoustic characteristics during the phonation of a sustained voice sound.” Examples of these acoustic parameters are the frequency of vibration of the vocal folds, the maximum amplitude of a voice frame, the waveform of the signal and the energy of the signal concentrated on the harmonic spectral components. This phenomenon is due to some type of growth or irregularity on the vocal folds.

2. “There can be an abnormal presence of spectral noise during the phonation of a sustained voice sound, which can be located in certain regions of the spectrum in a stationary way, or which can have a non stationary character and present short time variations.” This phenomenon is due to the impossibility of a speaker to carry out a correct glottal closure.

3. “The samples of healthy voices show a greater presence of a non-linear behaviour, while the samples of pathological voices show a bigger variability of their statistical over a period of time.” This phenomenon could be motivated by a deterioration or disappearance of the mucosae that coats the surface of the vocal folds, and especially in the area of the pleats.

In the context of laryngeal illnesses and within the world of medicine, quantifying the quality of a voice is of great interest. This is justified by different reasons: it shows the temporary evolution of the quality of a patient’s voice during the application of a treatment for a laryngeal illness, it contributes positively to medical-legal documentation, and it also helps to diagnose dysphonic degree in an objective way.

Applying automatic classification techniques is another utility of the numerical values of voice quality. This possibility is attractive for the early detection of illnesses (preventive medicine) and its use in telemedicine.

ACKNOWLEDGEMENTS

This work contributes to the automatic detection of voice pathologies with seven new parameters based on the HOS. They show an improvement on the results obtained using the classic parameters of about 4%.

This work was funded by the Spanish government (TIC 2000-0586).

REFERENCES

[1] B. Boyanov and S. Hadjitodorov, “Acoustic analysis of pathological voices: a voice analysis system for the screening of laryngeal diseases,” IEEE Engineering in Medical and Biology, vol. 16, no. 4, pp. 74–82, July/August 1997.

[2] M. Fröhlich, D. Michaelis, and H. W. Strube, “Acoustic “breathiness measures” in the description of pathological voices,” in Proc. ICASSP-98, vol. 2, pp. 937–940, Seattle, WA, May 1998.

[3] L. Gavidia-Ceballos and J. H. L. Hansen, “Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection,” IEEE Trans. on Biomedical Engineering, vol. 43, no. 4, pp. 373–383, 1996.

[4] K. Guoxin, L. Jiurui, Y. Quiang, and C. Guoying, “A method of spectra-analysis for abnormal voice signals,” in 11th Annual Inter. Conf. of the IEEE Engineering in Medicine and Biology Society, November 1989.

[5] S. Feijo, C. Hernández, A. Carollo, R. C. Hermida, and E. Molides, “Acoustic evaluation of glottal cancer based on short-term stability measures,” in 11th Annual Inter. Conf. of the IEEE Engineering in Medicine and Biology Society, November 1989.

[6] J. Koreman and M. Pützer, “Finding correlates of vocal fold adduction deficiencies,” in PHONUS, W. J. Barry and J. Koreman, Eds., vol. 3, pp. 155–178, Institute of Phonetics, University of the Saarland, Saarbrücken, 1997.

[7] E. J. Wallen and J. H. L. Hansen, “A screening test for speech pathology assessment using objective quality measures,” in Inter. Conf. on Spoken Language Processing, vol. 2, pp. 776–779, Philadelphia, PA, October 1996.

[8] M. Michaelis and H. W. Strube, “Empirical study to test the independence of different acoustic voice parameters on a large voice database,” in Proceedings of Eurospeech’95, J. M. Pardo et al., Eds., vol. 3, pp. 1891–1893, 1995.

[9] M. Michaelis, D. Fröhlich and H. W. Strube, “Selection and combination of acoustic features for the description of pathologic voices,” J. Acoust. Soc. Am., vol. 103, no. 3, pp. 1628–1639, 1998.

[10] K. Marasek, “EGG and voice quality,” http://www.ims.uni-stuttgart.de/phonetik/EGG/rrmsst1.htm, 1997.

[11] M. Fröhlich, D. Michaelis, H. Strube, and E. Kruse, “Acoustic voice quality description: case studies for different regions of the hoarseness diagram,” in Advances in Quantitative Laryngoscopy using Motion-, Image- and Signal Analysis, 2nd “Round Table”, T. Wittenberg et al., Eds., http://www.physik3.gwdg.de/ miche/papers.html, pp. 143–150, Erlangen, Germany, 1997.

[12] A. P. Petropulu and C. L. Nikias, “The complex cepstrum and bicepstrum: analytic performance evaluation in the presence of gaussian noise,” IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 38, no. 7, pp. 1246–1256, 1990.

[13] J. Picone, “Signal modeling techniques in speech recognition,” Proceedings of the IEEE, vol. 81, no. 9, pp. 1215–1247, 1993.

[14] J. H. L. Hansen, L. Gavidia-Ceballos, and J. F. Kaiser, “A non-linear based speech feature analysis method with application to vocal fold pathology assessment,” IEEE Trans. on Biomedical Engineering, vol. 45, no. 3, pp. 306–313, 1998.

[15] M. Rosa, J. C. Pereira, and A. C. P. L. F. Carvalho, “Evaluation of neural classifiers using statistic methods for identification of laryngeal pathologies,” in Proceedings of the Vth Brazilian Symposium on Neural Networks, vol. 1, pp. 220–225, Belo Horizonte, MG, Brazil, December 1998.

[16] L. Gavidia-Ceballos, J. H. L. Hansen, and J. F. Kaiser, “Vocal fold pathology assessment using AM autocorrelation analysis of the teager energy operator,” in Inter. Conf. on Spoken Language Processing, vol. 2, pp. 757–760, Philadelphia, PA, October 1996.
[17] S. Lu and P. C. Doerschuk, “Demodulators for AM-FM models of speech signals: a comparison,” in Proceedings of the IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing, vol. 1, pp. 263–266, Atlanta, GA, 1996.

[18] S. Lu and P. C. Doerschuk, “Nonlinear modeling and processing of speech based on sums of AM-FM formant models,” IEEE Trans. Signal Processing, vol. 44, no. 4, pp. 773–782, 1996.

[19] D. A. Cairns, J. H. L. Hansen, and J. E. Riski, “A noninvasive technique for detecting hypernasal speech using a nonlinear operator,” IEEE Trans. on Biomedical Engineering, vol. 43, no. 1, pp. 35–45, 1996.

[20] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, “Classification of speech under stress based on features derived from the nonlinear teager energy operator,” in IEEE Inter. Conf. on Acoustics, Speech, Signal Processing, vol. 1, pp. 549–552, Seattle, WA, May 1998.

[21] J. M. Mendel, “Tutorial on higher order statistics (spectra) in signal processing and system theory: theoretical results and some applications,” Proceedings of the IEEE, vol. 79, no. 3, pp. 278–305, 1991.

[22] C. L. Nikias and M. R. Raghuveer, “Bispectrum estimation: a digital signal processing framework,” Proceedings of the IEEE, vol. 75, no. 7, pp. 869–891, 1987.

[23] S. Elgar and R. Guza, “Statistics of bicoherence,” IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 36, no. 10, pp. 1667–1668, 1988.

[24] A. Moreno, S. Törtola, J. Vidal, and J. A. R. Fonollosa, “New host-based parameter estimation methods for speech recognition in noisy environments,” in IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 429–432, Detroit, USA, May 1995.

[25] A. Moreno and M. Rutllán, “Integrated polispectrum on recognition,” in IEEE Inter. Conf. on Spoken Language Processing, vol. 2, pp. 1281–1284, Philadelphia, October 1996.

[26] G. Sundaramoorthy, M. R. Raghuveer, and S. A. Dianat, “Bispectral reconstruction of signals in noise: amplitude recovery issues,” IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 38, no. 7, pp. 1297–1306, 1990.

[27] L. Gavidia-Ceballos, J. H. L. Hansen, and J. E. Riski, “A noninvasive speech processing method for analysis of vocal fold cancer based on an iterative EM algorithm,” in 16th Annual Inter. Conf. of the IEEE Engineering in Medicine and Biology Society, vol. 2, pp. 1308–1309, 1996.

[28] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113–120, 1979.

Jesus B. Alonso was born in 1971 in Gran Canaria, Spain. He received the M.Sc. degree in 2000 in Telecommunication engineering from Universidad de Las Palmas de Gran Canaria (ULPGC), Spain. Currently he is a Ph.D. student in the Department of Signal and Communications of ULPGC. His interests include signal processing in bio-computing, and recognition systems based on neural networks.

José de Leon was born in Las Palmas de G.C., Canary Island, Spain in 1962. He received the M.Sc. degree in 1985 in Salamanca University, Spain, and the Ph.D. degree in Las Palmas University in 1997. He is an ENT Medical Dr. of ENT Department of Gran Canaria General Hospital since 1993. His current research interest include singing voice and voice disorders and their management where has presented a lot of works in several national and international conferences. He is now working in signal processing in bio-computing joint to ETSI of Telecommunication in Las Palmas University.

Itziar Alonso was born in 1968 in Gran Canaria, Spain. She received the Ph.D. degree in Telecommunication Engineering from University of Las Palmas de Gran Canaria, Spain. She is currently lecturing at Department of Telematic Engineering in the University of Las Palmas de Gran Canaria, Spain. Her areas of research include classification applications, neural networks and hybrid systems formed by NN and HMM.

Miguel A. Ferrer was born in Valencia, Spain, in 1965. He received the M.Sc. degree in 1988 and the Ph.D. degree in 1994, both in Telecommunication engineering, from the Universidad Politècnica de Madrid, Spain. Since 1990 he has been with Universidad de Las Palmas de Gran Canaria, Spain, where he is currently a professor. His current research interest include signal processing in bio-computing, and recognition systems based on neural networks and hidden Markov models.