Optimal Size of Time Window in Nonlinear Features for Voice Quality Measurement

Jesús B. Alonso[1], Fernando Díaz-de-María[2], Carlos M. Travieso[1], and Miguel A. Ferrer[1]

[1] Dpto. de Señales y Comunicaciones, Universidad de Las Palmas de Gran Canaria, Campus de Tafira, 35017 - Las Palmas de Gran Canaria, Spain
{jalonso, ctravieso, mferrer}@dsc.ulpgc.es
http://www.gpds.ulpgc.es/index.htm
[2] Dpto. de Teoría de la Señal y Comunicaciones, Universidad Carlos III de Madrid, Avda. de la Universidad, 30, 28911 Leganes (Madrid), Spain
fdiaz@tsc.uc3m.es

Abstract. In this paper we propose the use of nonlinear speech features to improve the voice quality measurement. We have tested a couple of features from the Dynamical System Theory, namely: the Correlation Dimension and the largest Lyapunov Exponent. In particular, we have studied the optimal size of time window for this type of analysis in the field of the characterization of the voice quality. Two systems of automatic detection of laryngeal pathologies, one of them including these features, have been implemented with the purpose of validating the usefulness of the suggested nonlinear features. We obtain slight improvements with respect to a classical system.

1 Introduction

The medical community uses subjective techniques (evaluation of the voice quality by the specialist doctor's direct audition) or invasive methods (which allow the direct inspection of vocal folds thanks to the use of laryngoscopical techniques) for the evaluation and the diagnostic of voice pathologies. The voice quality measurement has received much attention during the last decade ([2] [3] [4] [5] are good examples). These systems allow us to quantify the voice quality effectively and to document the patient's evolution using objective measures. These techniques provide the ability to detect quickly and simply laryngeal pathologies; thus they can be applied in preventive medicine and telemedicine environments.

On the other hand, automatic laryngeal pathologies detection systems have been developed [6] [7] [8] [9]. In these works, different success rates are obtained in the classification between healthy voices and pathological voices, being evaluated each system with different data bases, since a data base of reference does not exist.

In [1], the authors proposed a classification system to distinguish healthy from pathologic voices using a Neuronal Networks (NN). In the feature extraction phase, diverse measures based on the High Order Statistics (HOS) were used in addition to a

M. Faundez-Zanuy et al. (Eds.): NOLISP 2005, LNAI 3817, pp. 206 – 218, 2005.

selection of classical voice quality measurements present in the current literature. These measurements of the voice quality based on the HOS achieve good results, but in exchange for a high computational cost.

In this work, the viability of the nonlinear dynamic-based speech analysis has been studied with the purpose of obtaining information on the voice signal nonlinear behavior. The tested nonlinear features are less computationally demanding than HOS-based ones. The viability of characterizing the voice signal by means of the Lyapunov Exponents has been already suggested in other works [10] [11] . In another paper [12] , the utility of the correlation dimension to detect the presence of laryngeal pathologies has also been proposed. However, different aspects of these measurements are explored, for example, the optimal size of the time window. Some preliminary results on this topic are presented in this work.

2 Nonlinear Dynamical System: The Embedding Theorem

The Chaos Theory can be used to gain a better understanding and interpretation of observed complex dynamical behaviour. Besides, It can give some advantages in predicting or controlling such time evolution 13.

Deterministic dynamical systems describe the time evolution of a system in some state space $\Gamma \subset \mathrm{R}^{\mathrm{d}}$. Such an evolution can be described case by ordinary differential equations:

$$\dot{x}(t) = F(x(t)) \tag{1}$$

or in discrete time $t = n\Delta t$ by maps of the form:

$$x_{n+1} = F(x_n) \tag{2}$$

Unfortunately, the actual state-vector only can be inferred for quite simple systems, and as anyone can imagine, the dynamical system underlying the speech production process is very complex. Nevertheless, as established by the "embedding theorem" 14, it is possible to reconstruct a state space equivalent to the original one. Furthermore, a state-space vector formed by time-delayed samples of the observation (in our case, the speech samples) could be an appropriate choice:

$$\mathbf{s_n} = [s(n), s(n-T), \ldots, s(n-(d-1)T)]^t \tag{3}$$

where $s(n)$ is the speech signal, d is the dimension of the state-space vector, T is a time delay and t means transpose.

Finally, the reconstructed state-space vector dynamic, $\mathbf{s_{n+1}} = \mathrm{F}(\mathbf{s_n})$, can be learned through either local or global models, which in turn will be polynomial mappings, neural networks, etc.

2.1 Correlation Dimension

The correlation dimension D_2 gives an idea of the complexity of the dynamics. A more complex system has a higher dimension, which means that more state variables

are needed to describe its dynamics. The correlation dimension of a random noise is not bounded while the correlation dimension of a deterministic system yields a finite value. The correlation dimension can be obtained as follows:

$$D_2 = \lim_{r\to 0} \lim_{N\to\infty} \frac{\log C(N,r)}{\log r} \tag{4}$$

with *C(N,r)* being,

$$C(N,r) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i\neq j}}^{N} \theta(r - \left\| X_i - X_j \right\|) \tag{5}$$

where *r* is the radius around X_i and $\theta(x)$ is the step function. Equation [4] converges very slowly as *r* tends to zero. To circumvent this problem, the local slope can be estimated:

$$D_2 = \frac{d \log C(N,r)}{d \log r} \cong \lim_{\Delta \log r \to 0} \frac{\Delta \log C(N,r)}{\Delta \log r} \tag{6}$$

When the length N is significantly large, D_2 will converge with the increase of the embedding dimension, *m*.

2.2 The Largest Lyapunov Exponent

Chaotic behaviour arises from the exponential growth of infinitesimal perturbations. This exponential instability is characterized by the Lyapunov exponents. Lyapunov exponents are invariant under smooth transformations and are thus independent of the measurement function or the embedding procedure.

The largest Lyapunov exponent can be determined without the explicit construction of a model for the time series. It considers the representation of the time series as a trajectory in the embedding space, and assume that you observe a very close return $s_{n'}$ to a previously visited point s_n. Then one can consider the distance $\Delta_0 = s_n - s_{n'}$ as an small perturbation, which should grow exponentially in time. Its evolution can be followed from the time series: $\Delta_l = s_{n+l} - s_{n'+l}$. If one finds that $|\Delta_l| \approx \Delta_o e^{\lambda l}$, λ is the largest Lyapunov exponent.

3 New Voice Disorder Parameterisation

In the current literature, some works suggest the viability of characterizing the voice signal by means of the Lyapunov Exponents (for example in synthesis of phonemes 10 and 11), and characterizing the voice disorder signal by means of the correlation dimension 12.

For the study of the presence of laryngeal pathologies based on the voice recording, it is very common to use recordings of sustained vowels.

In this work, we have studied which is the optimal size of the time window for the nonlinear analysis (Correlation Dimension and the Largest Lyapunov Exponent), with the purpose of deciding whether a vowel utterance comes from a healthy or a patho-

logical voice. Different sizes of time window has been studied: 10, 30, 50, 100, 150, 300, 500 ms or the whole vowel utterance and pitch-synchronous segments of 3, 5, 7, 10 pitch periods.

In the case of obtaining multiple frames for each vowel, the following parameters have been extracted for each feature (Correlation Dimension and the Largest Lyapunov Exponent):

- The mean value of the feature P for the different frames $\{T_i\}_N$:

$$M_P = \frac{1}{N}\sum_{i=1}^{N} P_i \tag{7}$$

- Variation of the value of the feature along the time:

$$V_P = \frac{1}{(N-1)}\frac{1}{\max\{P_i\}}\sum_{i=2}^{N}\left|P_i - P_{i-1}\right| \tag{8}$$

3.1 Voice Database

The voice signals used in this study were digitalized with a sample frequency of 22050 Hz and 16 bits per sample. The speaker's voice was recorded with a conventional sound card and a basic microphone. The database consists of 100 voices of healthy speakers and 68 voices of pathological speakers. Each sample of the database is composed by the five Spanish vowels (‘a’, ‘e’, ‘i’, ‘o’ and ‘u’) pronounced in a sustained way by the speakers during approximately two seconds for each vowel. In case of pathological speakers there are situations of vocal folds without lesion (hypofunction, hyperfunction, vocal fold paralysis,...) and vocal folds with lesion (carcinoma, vocal folds nodule, sessile polyp, pedunculated polyp, Reninke’s edema, adult papiloma,...). The database has been created contemplating different disphonia levels: “light pathological voice”, “moderate pathological voice” and “severe pathological voice.”

3.2 Evaluation of the Parameterization

The attractor dimension has been fixed to 2 since the result obtained does not justify the increment of the time consuming, and the delay, *T*, has been estimated to 8 samples.

A one-second interval, located in the centre of the utterance, has been studied. This alteration has been carried out with the purpose of eliminating the beginning and end of the phonation, because it presents a transitory character. This modification has been implemented except when the whole vowel is used.

Four different attributes have been studied:

- *Atrib1*: Mean value of the Correlation Dimension.
- *Atrib2*: Time Variation of the Correlation Dimension values.
- *Atrib3*: Mean value of the Largest Lyapunov Exponent.
- *Atrib4*: Time Variation of the Largest Lyapunov Exponent values.

A neural network has been used to evaluate the benefits of the different attributes in the environment of the automatic pathologies detection. Each attribute has separately been evaluated using neural network Multilayer feedforward with 2 hidden layers, with Backpropagation train algorithm. Different sizes of asynchronous time window have been evaluated of using like evaluation function the success rate in the classification. Each attribute has been evaluated separately, differentiating between the five vowels. The different sizes of asynchronous time window are: 10, 30, 50, 100, 150, 300, 500 milliseconds and the whole vowel utterance ('full' in the figures). The result is showed in the Figures 1, 2, 3 and 4.

The asynchronous time window has a disadvantage: because the vibration frequency of the vocal folds (picth) of the women is greater than the men, for a certain temporal window is obtained different number of periods between men and women. In order to be able to make the parameterization process independent of the pitch frequency, it is possible dividing the vowel in pitch-synchronous segments of 3, 5, 7, 10 pitch periods (To). The result is showed in the Figures 5, 6, 7 and 8.

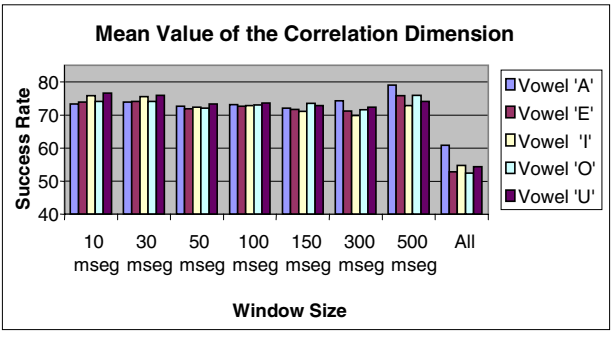

Fig. 1. Results of the study about size of asynchronous time window for "mean value of the Correlation Dimension"

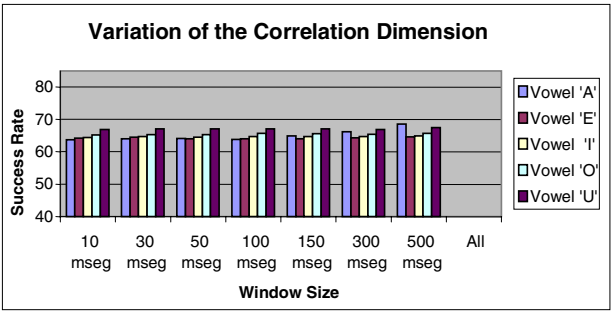

Fig. 2. Results of the study about size of asynchronous time window for "Time Variation of the Correlation Dimension values"

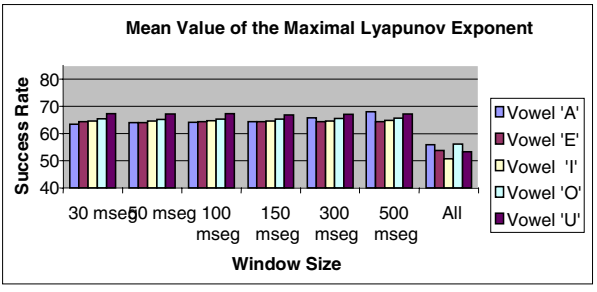

Fig. 3. Results of the study about size of asynchronous time window for "mean value of the Maximal Lyapunov Exponent"

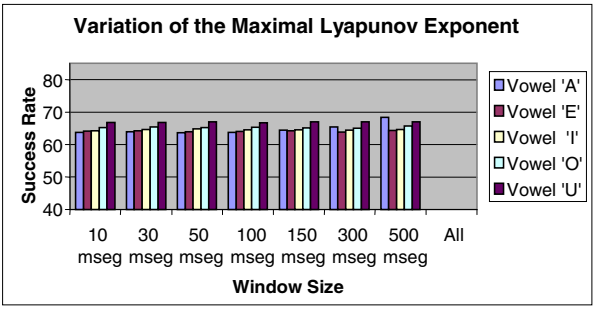

Fig. 4. Results of the study about size of asynchronous time window for "Time Variation of the Maximal Lyapunov Exponent values"

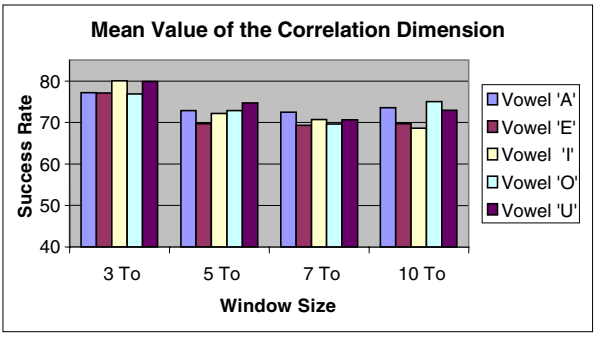

Fig. 5. Results of the study about size of synchronous time window for "mean value of the Correlation Dimension"

To sum up, it is observed better results dividing the vowel in pitch-synchronous segments of 3 pitch periods. It is also observed better results for the attribute "mean value of the Correlation Dimension", during the individual evaluation of the parameter.

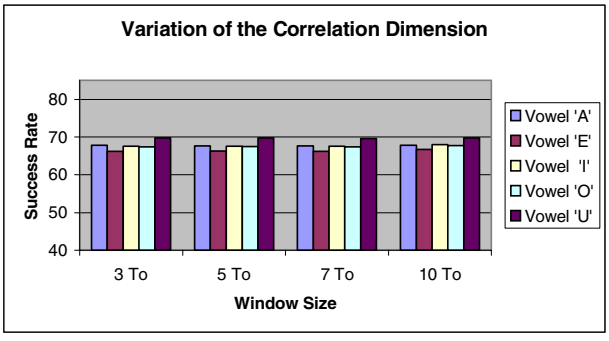

Fig. 6. Results of the study about size of synchronous time window for "Time Variation of the Correlation Dimension values"

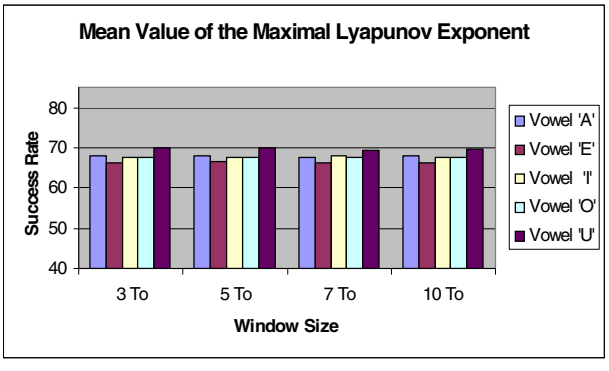

Fig. 7. Results of the study about size of synchronous time window for "mean value of the Maximal Lyapunov Exponent"

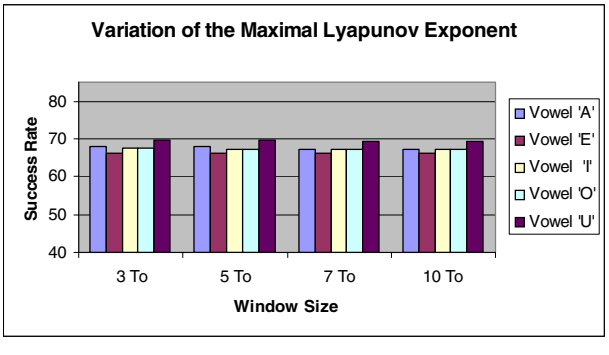

Fig. 8. Results of the study about size of synchronous time window for "Time Variation of the Maximal Lyapunov Exponent values"

4 Detector Model

The voice automatic classification system allows us to discriminate healthy voices from pathological ones. It is based on a pattern recognition model.

These systems are typically structured in three steps, namely: “Voice Acquisition”, “Parameterization” and “Classification”. The proposed automatic laryngeal pathologies recognition system follows this structure, illustrated in figure 9. Firstly, it captures the speaker's voice using a sound card and a microphone. The parameterization step uses parameters presented in [1], where a combination of a selection of parameters exposed in the current literature with new parameters based on Higher Order Statistics (HOS) was exposed. Finally, a net of classifiers based on Neural Networks (NN) is used to classify between healthy and pathological voices.

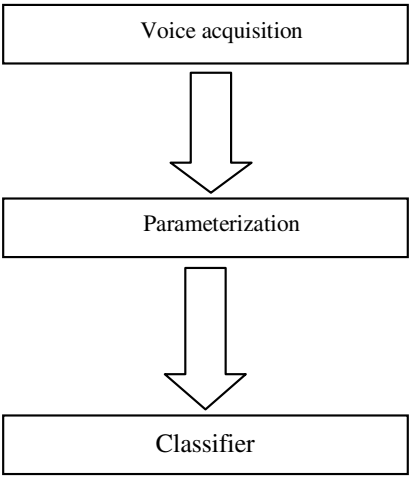

Fig. 9. Pattern recognition model

4.1 Parameterization

The parameterization step uses the same parameters that has been used in [1], where the authors made a selection of 17 parameters for the laryngeal pathologies classification (in the rest of the paper “classic parameters”), among multiple voice quality characterization parameters well-known in the literature.

4.1.1 Classic Parameters

There is no parameter which is completely conclusive in the detection of laryngeal pathologies, because each pathology affects the voice in a different way. For example, there are pathologies that present a great content of non-stationary noise in the high frequency components. On the other hand, other pathologies are characterised by the

Table 1. Classic characteristics

Group of characteristics	Name of the attributes
Quantifying the variation in amplitude (shimmer)	- Variation in the mean quadratic value of each voice frame - Variation in the highest value of the short time cross correlation function of each voice frame
Quantifying the presence of unvoiced frames	- Relationship between the number of unvoiced frames and the total number of frames of the sample voice - The unvoiced periodicity index of a sample voice
Quantifying the absence of wealth spectral (Hitter)	- Variation of pitch energy cepstral measure - Variation in the first harmonic value in the derived cepstrum domain - Variation in the first/second harmonic relationship value within the derived cepstrum domain
Quantifying the presence of noise	- Energy spectral balances - Spectral distance (based on the spectral module) - Spectral distance (based on the spectral phase)
Quantifying the regularity an periodicity of the waveform of a sustained voiced voice	- Value an variation in energy of the slope of the envelope in the autocorrelation function of an AM modulated signal - Variation of the slope of the envelope in the auto-correlation function of an AM modulated signal

uncertainty of the pitch value throughout the duration of the phonation of a sustained voiced sound. This is why classical characteristics have been divided into five groups depending on the physical phenomenon that each parameter quantifies: 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 an periodicity of the waveform of a sustained voiced voice. All the classic characteristics used are shown in Table 1.

4.1.2 New Nonlinear Parameters

In this work the possibility of using nonlinear features with the purpose of detecting the presence of laryngeal pathologies has been explored. The four measures proposed will be used: mean value and time variation of the Correlation Dimension and mean value and time variation of the Maximal Lyapunov Exponent values.

4.2 Classifier

The proposed system is based on the use of a net of classifiers, where each one discriminates frames of a certain vowel. Combinational logic has been added to evaluate the success rate of each classifier.

The structure of the proposed classifier is similar to the one proposed in [1], and represented in figure 10. Five NN–based vowel classifiers have been used to discriminate between healthy and pathological vowels, one for each vowel. The inputs of each vowel classifier are the feature vectors of the sequence of frames in which the

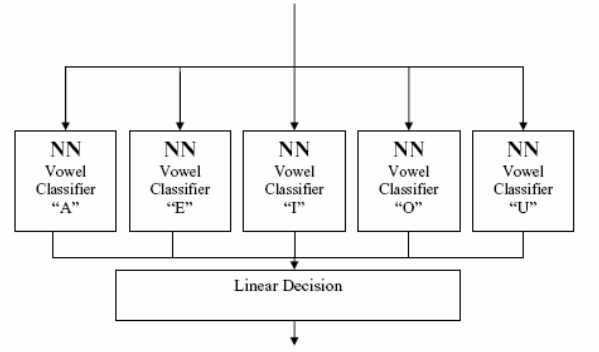

Fig. 10. Classification System Scheme

vowel corresponding to this classifier has been divided. The length of each frame is a three pitch period, for a voiced sound, or 30 ms if it is an unvoiced sound (typical of the pathological voices). Only the 500 central milliseconds of the vowel have been considered to avoid considering the frames that exhibit non-stationary behaviour at the beginning and end of each vowel.

The dependence of the parameters on the analysed vowel has been taken into account, as pointed out by Jacques Koreman and Manfred Pützer [15]. Consequently, a "vowel classifier" has been used for each vowel, such as is shown in figure 10.

First of all, each "vowel classifier" emits an estimation dependent on whether the analysed 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".

In each "vowel classifier", the different voice frame are 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 analysed vowel will be labelled as a "healthy vowel", otherwise it will be labelled as a "pathological vowel." The scheme of a vowel classifier is shown in figure 11. In this study, normalized data (zero-mean and variance one) have been used.

The characteristics of the Neural Network are described in the table 2.

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".

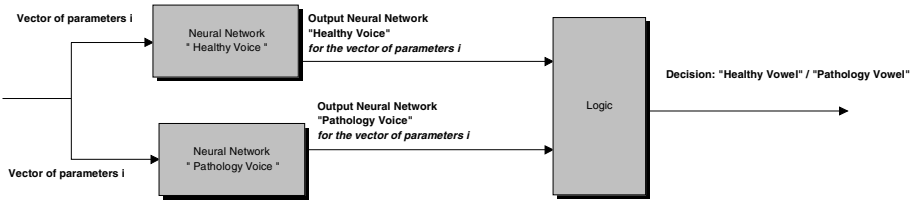

Fig. 11. Classification System Scheme

Table 2. Characteristics of the Neural Network

	Characteristic	Neural Network "Healthy Voice"	Neural Network "Pathological Voice"
Topology	**Number of Layer**	2	2
	Number of inputs	Number of parameters	Number of parameters
	Number of neurons in the first layer	40	40
	Number of outputs	50	50
Training	**Maximum Threshold of absolute error**	0.01	0.01
	Maximum Threshold of relative error	0.015	0.015
	Maximum number of epochs	10000	10000
	Training method	Back-propagation	Back-propagation
	No linear function	Hidden layer: "tansig"	Hidden layer: "tansig"
		Output layer: "purelin"	Output layer: "purelin"

5 Results

Two systems have been compared using the same data base. The first one, only works with the "classic parameters", while the second one uses both "classic" and "nonlinear" parameters, obtaining the results displayed in Tables 2 and 3.

A global success rate of 91,77% is obtained using "classic parameters", whereas a global success rate of 92,76% using "classic parameters" and "nonlinear" parameters. These results show the utility of new parameters.

Table 3. Success Rate using "Classic characteristics"

		Input	
		Healthy Voice	Pathological Voice
Output	Healthy Voice	95.65 %	12.10 %
	Pathological Voice	4.35 %	87.90 %

Table 4. Success Rate using "Classic characteristics + New parameters"

		Input	
		Healthy Voice	Pathological Voice
Output	Healthy Voice	96.12 %	10.60 %
	Pathological Voice	3.88 %	89.40 %

6 Conclusions

In this work, the possibility of using nonlinear features to improve the performance of an automatic detector of laryngeal pathologies has been explored.

Two features have been tested: Correlation Dimension and the Largest Lyapunov Exponent. In particular, the system works with their mean value and variation.

An experimental study aiming at selecting the best size for the time window of the nonlinear analysis has been conducted, concluding that the best option is using a pitch-synchronous window containing three periods.

Finally, the results of the classification system including the mean value and variation of the correlation dimension are slightly better than those achieved by the system using only the classic parameters.

Though the improvement is slight, we consider it an encouraging result, since the research is currently in the first stages. Further work is necessary in diverse directions.

Acknowledgement

This work was partially supported by the Spanish government (TIC2003-08956-C02-02).

References

1. J.B Alonso, J. de Leon, I. Alonso, M.A. Ferrer: Automatic detection of pathologies in the voice by HOS based parameters, *Proc. EUROASIP Journal on Applied Signal Processing*, (2001), vol.1, 275-284.
2. M. Fröhlich, D.Michaelis, H.W. Srube: Acoustic 'Breathiness Measures' in the description of Pathologic Voices, *Porc. ICASSP-98*, Seattle, WA, (1998), vol.2, 937-940.
3. L.Gavidia, J. Hansen: Direct Speech Feature Estimation Using an Iteractive EM Algorithm for Vocal Fold Pathology Detection, *Porc. IEEE Transactions on Biomedical Engineering,* (1996), vol.43, no.4, 373-383.
4. S. Feijoo, C. Hernandez, A. Carollo, R.C Hermida, E.Moldes: Acoustic Evaluation of glottal cancer based on short-term stability measures, *Porc. IEEE Engineering in Medicine & Biology Society 11th Annual Internatioonal Conference*, (1989), vol. 2, 675-676.
5. B.Boyanov, S.Hadjitodorov Ivanov: Analysis of voiced speech by means of bispectrum, *Electronics Letters*, (1991), vol. 27, no. 24, 2267-2268.
6. B. Boyanov, S.Hadjitodorov: Acoustic analysis of pathological voices: a voice analysis system for screening of laryngeal diseases, *Proc. IEEE Engineering in Medical and Biology*, (1997), vol. 16, no. 4, 74-82.
7. J.H.L Hansen, L. Gavidia, F. James: A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment, *Proc. IEEE Transactions on Biomedical Engineering* (1998), vol. 45, no. 3, 300-313.
8. M.O. Rosa, J.C. Pereria, C.P.L.F. Carvalho: Evaluation of neural classifier using statistics methods for identification of laryngeal pathologies, (1998), *Proc..Neural Networks*, vol.1, 220-225.
9. E. J. Wallen, J.H.L Hansen: A Screening test for speech pathology assessment using objective quality measures, *Proc. International Conference on Spoken Language Processing*, (1996), Philadelphia, PA, vol. 2, 776-779.

10. Michael Banbrook, Stephen McLaughlin, Iain Mann: Speech Characterization and Synthesis by Nonlinear Methods, *Proc.IEEE Transactions on Speech and Audio Processing*, January (1999), vol. 7, no.1.
11. V Pitsikalis, I Kokkinos, P Maragos: Nonlinear Analysis of Speech Signals: Generalized Dimensions and Lyapunov Exponents, *Proc EUROSPEECH-2003*,Geneva,(2003), 817-820.
12. J.J. Jiang,Yu Zhang: Nonlinear dynamic analysis of speech from pathological subjects, *Proc IEEE Electronics Letters*, March (2002), vol.38, no.6.
13. R. Hegger, H. Kantz, T. Schreiber: *Practical implementation of nonlinear time series methods: The TISEAN package*, CHAOS 9, 413, (1999)
14. E. Ott: Chaos in Dynamical Systems: Cambridge: Cambridge University Press, (1993).
15. Jacques Koreman and Manfred Pützer: Finding Correlates of Vocal Fold Adduction Deficiencies, *Phonus 3*, Institute of Phonetics, University of the Saarland, (1997), pp. 155-178