CLASSIFICATION OF VOICE PATHOLOGIES BASED ON NEURAL NETWORKS AND PREDICTABILITY MEASURES

Lyvia Regina Biagi Silva, * Arthur Hirata Bertachi, * Helder Luiz Schmitt, * Paulo Rogério Scalassara, * Alessandro Goedtel*

*Federal University of Technology - Paraná, Cornélio Procópio, Brazil

Email: lybiagi@hotmail.com, arthurltda@hotmail.com, helderschmitt@gmail.com, prscalassara@utfpr.edu.br, agoedtel@utfpr.edu.br

Abstract— This work presents a proposal of the application of artificial neural networks to voice signals for classification of larynx pathologies. We use an alternative set of patterns based on predictability measures estimated from relative entropy of wavelet-packet components. We tested a radial basis function neural network in order to classify voice signals from people of three groups: with healthy larynx, with Reinke's edema, and with nodule on the vocal folds. Using the proposed method, the signals were correctly classified, even the pathological ones, which was not achieved by other studies with the same database. Experimental results are presented to validate the methodology.

Keywords— Voice processing, larynx pathologies, relative entropy, neural networks

Resumo— Este trabalho apresenta uma proposta de aplicação de redes neurais artificiais à sinais de voz para classificação de patologias na laringe. Utiliza-se um conjunto de padrões baseados em medidas de previsibilidade estimadas a partir da entropia relativa de componentes de decomposição *wavelet-packet*. Implementa-se uma rede de função de base radial neural para classificar sinais de voz de pessoas de três grupos: com laringe saudável, com edema de Reinke e com nódulo nas pregas vocais. Utilizando o método proposto, os sinais foram classificados corretamente, até mesmo entre patologias, resultado não obtido em outros estudos com o mesmo banco de dados. Resultados experimentais são apresentados para validar a proposta.

Palavras-chave— Processamento de voz, patologias da laringe, entropia relativa, redes neurais

1 Introduction

Voice evaluation is a subject of constant study in the speech field. The most common form of evaluation is the use of the human ear as an inspection instrument (perceptual evaluation). Some tests can be used in conjunction with this assessment to diagnose larynx disorders.

According to Davis (1979), larynx pathologies usually produce asymmetrical changes in the vocal folds, this results in different types of vibrations. The most common larynx pathology test is the laryngoscopy, which is classified as a visualization method that aims to make an especific diagnosis of the disease, evaluate the stage of evolution, monitor and perform the prognosis (Behlau, 2008).

Automatic methods for voice analysis has shown great progress recently. Numerous techniques for digital signal processing have been developed, which resulted in many tools for the analysis and classification of voice signals (Rabiner and Schafer, 2007).

Studies related to acoustic analysis are usually based on the frequency of vocal fold vibration and the volume of air that escapes through the glotis during speech (Rosa et al., 2000). Thus, the analysis of voice signals can aid the diagnosis of diseases of the larynx. Because it is a noninvasive method, signal analysis has many advantages over other methods, which also allows the construction of an automatic diagnosis system.

One application involvig the use of intelligent systems, according to Silva et al. (2010), is the classification of speech patterns. Espinosa et al. (2000) presents an application of neural networks to classify healthy and pathological speech signals from acoustic parameters extracted from speech samples. In Voigt et al. (2010), healthy and pathological voices classification was obtained from vibration patterns analysis of the vocal folds. Scalassara et al. (2011) uses wavelet packet decomposition (WPD) to obtain a detailed description of the voice signals and presents a method to classify the signals in accordance with its pathological condition. This method is based on a predictability measure of signals reconstructed from selected components of the wavelet packet decomposition. In that work, the tools are tested by using voice signals from people with healthy and pathological larynx with a fuzzy c-means classifier. But the pathological signals were not separable.

The voice signals used in this work were obtained from the database of the Laboratory of Signal Processing of the School of Engineering of São Carlos at the University of São Paulo, Brazil and it is the same used in Scalassara et al. (2011), Rosa et al. (2000) and Santos and Scalassara (2012).

This work aims to implement a neural network classifier to better separate healthy and pathological voice signals. To do this, we established the input signals of the neural network from the sample database. We compare the results obtained with Scalassara et al. (2011).

This paper is organized as follows: Section 2 presents the voice signals and Section 3 shows

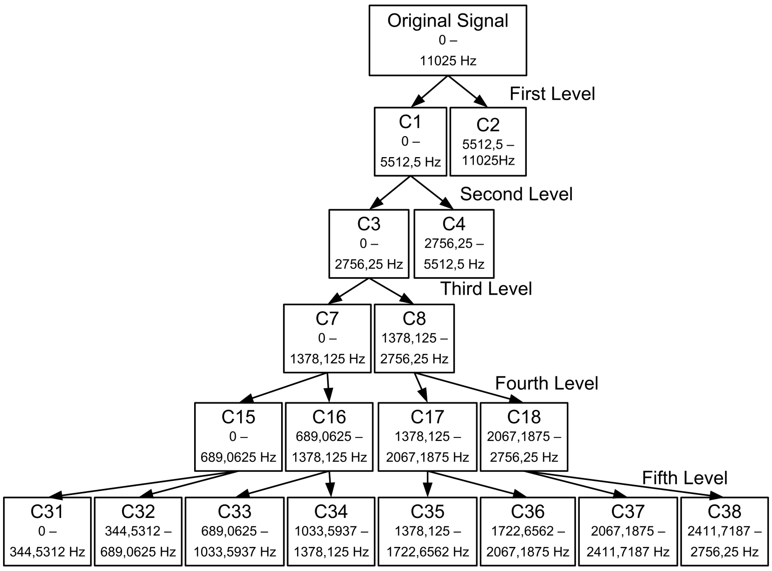

Figure 1: Diagram of the fifth level wavelet packet decomposition showing the relevant components of this study.

how we obtained the vectors used as inputs to the neural networks. Section 4 presents the topologies of neural networks used in this work. Section 5 shows the classification of the voice signals: at first, we distinguish healthy and pathological signals, and subsequently we define which pathology it presents. Finally, section 6 presents the conclusions of this paper.

2 Voice Signals

The database consists of 48 voice signals of adults between 18 and 60 years. The signals were divided into three groups of people: with healthy voice (without pathology in the larynx), with nodule on the vocal folds and with Reinke's edema. Each group consists of 16 signals, which are recordings of Brazilian Portuguese phoneme /a/ for nearly 5 s. The acquisition was performed by the standards presented in Rosa et al. (2000). The sampling frequency was 22,050 Hz and the sampling quantization level was 16 bits. For the analysis presented in this paper, only one second of the most stationary part of each signal was selected and used (Scalassara et al., 2011; Scalassara et al., 2009).

3 Feature Vector

In this section, we explain how to obtain the feature vectors used as the inputs of the neural networks. We explain some details of the WPD and the predictability measures applied to the reconstructed signals.

3.1 Wavelet Packet Decomposition

Wavelet analysis is based on the decomposition of a signal in shifted and scaled versions of a single function called Wavelet, which allows different resolutions of time and frequency. Differently of the Fourier transform, which transform a signal into a sum of sines of different frequencies (Mallat, 1999). Due to the non-stationarity of the voice signals, the wavelet analysis becomes more suitable for these signals (Scalassara et al., 2011). We chose the Daubechies family, as proposed in Guido et al. (2006), with filter order 20.

According to Diniz et al. (2010) and Scalassara et al. (2011), a description of WPD of a given signal $x[n]$, with N samples, consists of a j-level decomposition of the signal into projections of itself on two variable functions: scaling (ϕ) and wavelet (ψ).

These functions scaling and wavelet can be defined using shifted and scaled versions of themselves and also by a pair of decomposition lowpass (h_n) and highpass (g_n) filters, according to Equations (1) and (2).

$$\phi\left[n\right] = \sum_{k} h_n \phi\left[2n-k\right] \qquad (1)$$

$$\psi\left[n\right] = \sum_{k} g_n \psi\left[2n-k\right] \qquad (2)$$

The computation is obtained by convolution of the signal with the pair of filters. The signal is decomposed into two parts, separating the components of low and high frequency. For each level, the signals are downsampled by 2. In the WPD, both components are decomposed again each level (Mallat, 1999).

As the WPD results in a more detailed analysis of signals, it was choosen for this study and, as in Scalassara et al. (2011), we select only the first eight components of the fifth level as a basis for creating the feature vector, this is because

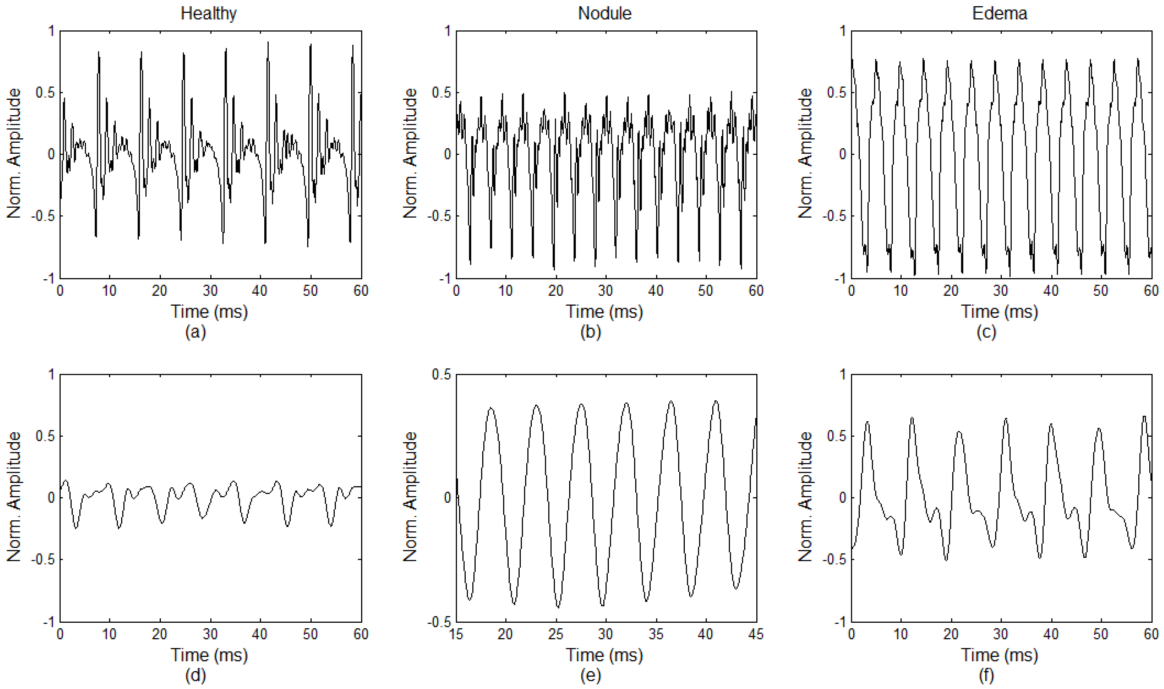

Figure 2: Example of the fifth level of the wavelet packet decomposition of the voice signals. (a)(b)(c): 60 ms of the original voice signal. (d)(e)(f): reconstructed signal using only the component C31 of each voice.

the frequency range covers almost all the relevant harmonic information of the vowel ‘a’ (Lieberman and Blumstein, 1988). Figure 1 shows the components of the WPD that are relevant to this study and their approximate frequency range.

After the decomposition, the reconstructed signal is obtained by using each of the eight components of the fifth level of decomposition (C31 to C38). The reconstruction of a signal using its wavelet components is the inverse process of the decomposition. If all the components are used in this process, the original signal is obtained. This process is performed by upsampling by 2, followed by convolution with the reconstruction filters, which are created with the decomposition filters, h and g (Diniz et al., 2010).

Due to the upsampling, the reconstructed signals have the same size of the original signal. For the creation of the feature data, only one component is used at a time to obtain the reconstructed signal. Each reconstructed signal is considered as an estimation (prediction) of the original voice signal. Figure 2 presents an example of this process, showing, in the first line, the three voice signals (healthy, with nodule and with Reike’s edema) and the second line the reconstruction of the C31 component of each voice.

The predictability measure applied to the reconstructed signal is the Predictive Power (Scalassara et al., 2009).

3.2 Predictive Power

The feature vector used as input of the classifier was obtained by the computation of the Predictive Power (PP) of each reconstructed signal.

The PP is based on entropy estimations and it was presented in Schneider and Griffies (1999). In accordance with Scalassara et al. (2009), the PP is obtained using the relative entropy between the original voice signal and its prediction error. The relative entropy between two signals, can be treated as a measure of distance between two probability distributions (PDF). According to Cover and Thomas (2006), the relative entropy is calculated by the Equation (3).

$$D_{p||q} = \sum_{i=1}^{k} p\left(i\right) \log_2 \frac{p\left(i\right)}{q\left(i\right)} \qquad (3)$$

where $D_{p||q}$ is the relative entropy and $p(i)$ and $q(i)$ are two probability distributions and k is the number of points of the probability function.

The PP can be described by the Equation (4):

$$PP = 1 - e^{-D_{p||q}} \qquad (4)$$

If the relative entropy grows, it is because the probability distribution function (PDF) of the voice signal and its prediction error become more different and the previsibility is high. The values of $D_{p||q}$ vary between 0 and infinity, thus the values of the PP vary between 0 and 1.

Figure 3 shows the PP of all the components of the fifth level of the decomposition (C31 a C62)

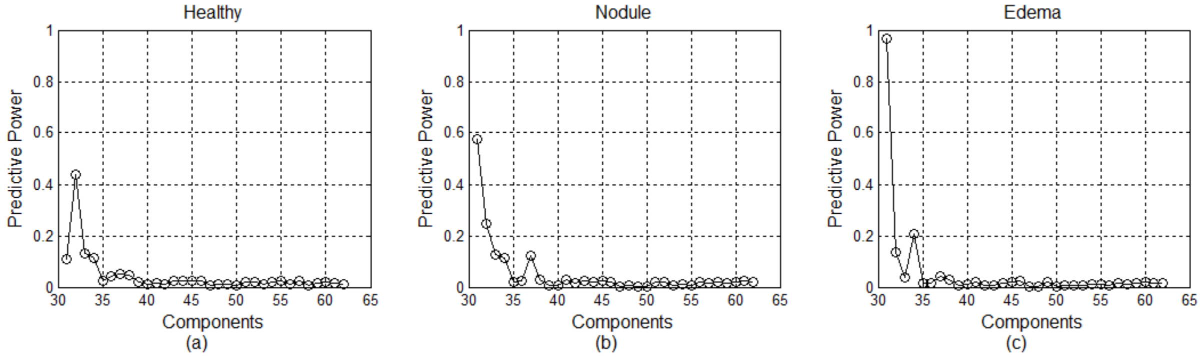

Figure 3: Predictive Power of the voice signals: (a) healthy voice, (b) voice with nodule in the vocal folds, (c) voice with Reinke's edema.

of the voices of Figure 2 (healthy voice, with nodule and with Reinke's edema). The greater predictabilities are presented by components C31 to C38, that is why these components are used as the inputs of the neural network classifier .

4 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models inspired by biological nervous system. These systems have the ability of acquiring and maintaining knowledge (based on information), and can be defined as a set of processing units, characterized by artificial neurons, which are interconnected (artificial synapses) (Silva et al., 2010).

The structure of ANNs was developed from known biological nervous systems models and from models of the human brain itself. The computational elements, or processing units, called artificial neurons (Figure 4), are simplified models of biological neurons (Silva et al., 2010). The operation of the artificial neuron can be summarized by:

- Presenting a set of input values ($\mathsf{x}_1 ... \mathsf{x}_n$);
- Multiplying each neuron input by its synaptic weight ($\mathsf{w}_1 ... \mathsf{w}_n$);
- Obtaining the activation potential ($\mathsf{u} = \sum_{\mathsf{i}=1}^{\mathsf{n}} \mathsf{w}_\mathsf{i} \cdot \mathsf{x}_\mathsf{i} - \theta$), where θ is the bias;
- Applying the activation function (g);
- Compiling the output ($\mathsf{y} = \mathsf{g}(\mathsf{u})$).

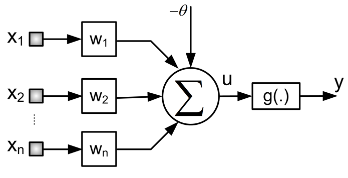

Figure 4: Artificial neuron.

In accordance with Haykin (1999), the maximum computational power of a neural network is extracted through its parallel distributed structure and its ability to learn and generalize.

Three neural networks common topologies used as classifiers are: Multilayer Perceptron (MLP), Radial-Basis Functions networks (RBF) and Kohonen self-organizing networks. In this work, we used a only the RBF network topology.

The typical structure of the RBF network is composed of three layers: input layer, only one intermediate layer (**h**), in which activation functions are Gaussian type and the output layer (**o**), which activation functions are linear (Haykin, 1999). RBF network belongs to feedforward architecture, however the training strategy of the RBF consists of two very distinct stages (Silva et al., 2010). Figure 5 illustrates the RBF network.

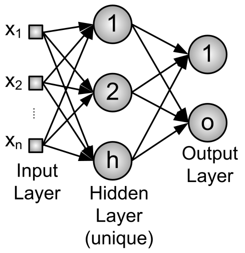

Figure 5: Illustration of the RBF network.

In the first training stage, which comprises the step the adjusting the synaptic weights of neurons in the hidden layer, a method of unsupervised learning is adopted (k-means), which is dependent on the characteristics of the input data. The second training stage consists in the adjustment of the synaptic weights of the output layer neurons, with the back-propagation algorithm.

5 Classification of the Voice Signals

First, the feature vectors were created through the computation of the Predictive Power using the reconstructed signals with the components C31 to C38 of the WPD. These data are used as the inputs of the neural networks.

The algorithm for voice signal classification is described by the flow chart shown in Figure 6

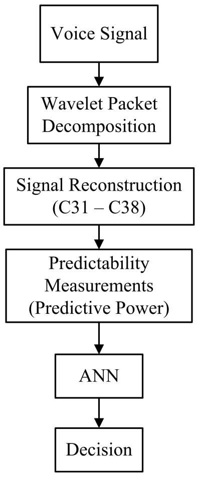
Voice Signal
Wavelet Packet Decomposition
Signal Reconstruction (C31 – C38)
Predictability Measurements (Predictive Power)
ANN
Decision

Figure 6: Algorithm for voice signals classification.

Considering the 48 original voice signals, 16 of each group, 33 were used to train the neural networks (11 of each group), and 15 for validation. Extreme values of samples (upper and lower values) were used to train the ANNs. We performed several tests varying the parameters of the ANNs empirically.

The topology of ANN used to classify the voice signals in three patterns has its parameters described in the Table 1. The results are presented in the bottom of the table.

Table 1: Parameters of the RBF network used to classify the voice signals in three patterns. Results of the voice signals classification in three patterns.

Architecture	**RBF**
Accuracy	10^{-7}
Neurons Hidden Layer	20
Neurons Output Layer	3
Learning Rate	0.01
Momentum Constant	0
Results	
Number of Epochs	151276
MSE	0.161
Classification Performance (%)	73.33

MSE - Mean-squared error

The classification performance was 73.33%. Considering 15 test samples, the network classified correctly 11 signals, and 4 nodule signals were classified as edema.

In a new approach, considering the voice signals in only two patterns (healthy and pathologic voices), we implement a second algorithm to classify the patterns, as it can be seen in Figure 7.

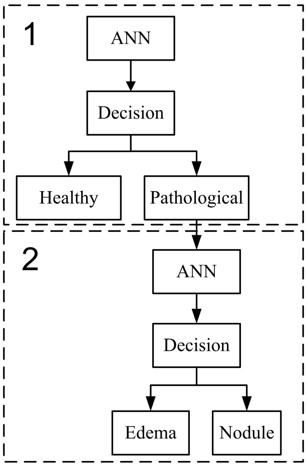

Figure 7: Stage 1: Classification of the signal between healthy and pathological. Stage 2: Classification of the pathologies (nodule or Reinke's edema).

In stage 1, the RBF network is used to classify the signals between healthy or pathological. In stage 2, only the pathological signals are used, so then the pathology is classified (nodule or Reinke's edema). Table 2 presents the parameters and the results of the RBF networks implemented in stage 1 and stage 2.

Table 2: Parameters of the RBF networks used in stage 1 and stage 2 of Figure 7. Results of the voice signals classification.

	Stage 1 Healthy/ Pathological	**Stage 2 Nodule/ Edema**
Architecture	RBF 1	RBF 2
Accuracy	10^{-5}	10^{-8}
Neurons Hidden Layer	2	8
Neurons Output Layer	1	1
Learning Rate	0.01	0.001
Momentum Constant	0	0.9
Results		
Number of Epochs	93	92219
MSE	0.0154	0.0051
Classification Performance (%)	100	100

MSE - Mean-squared error

Both RBF networks (RBF 1 and RBF 2) showed classification performance of 100% in the classification between healthy and pathological signals and in classification between nodule and edema.

6 Conclusions

This study aimed to classify voice signals of three groups of individuals, with healthy larynx, with nodule in the vocal folds and Reinke's edema. To do this, the original signals were decomposed using the Wavelet Packet Transform, and we used predictive measures applied to the relevant components of this decomposed signals to create the inputs to one topology of classifier neural network: RBF.

In the first approach, we tried to classify three groups of signals directly, and the classification performance was of 73.33%. In a second strategy, we first aimed to classify first the signals as healthy or pathological, and after this, identify which pathology is present. In stage 1, the neural network presented a great classification performance: 100%. In stage of definition of the pathology (stage 2), the RBF neural network classified correctly the pathologies present in all samples studied (100% correct).

The results obtained in Scalassara et al. (2011), the authors showed that the pathological voice signals were not separable with their method (fuzzy c-means), our methodology, which consists first in the separation of the healthy and pathological signals, and then in the definition of the pathology, using ANNs, was effective in classifying all the signals.

Acknowledgments

This work is supported by the Araucária Foundation for the Support of the Scientific and Technological Development of the State of Paraná, Brazil (Processes Number 06/56093-3 and 338/2012), and the National Counsel for Technological and Scientific Development - CNPq (Processes number 474290/2008-5, 473576/2011-2, 552269/2011-5) and Capes-DS scolarships.

References

Behlau, M. (2008). *Voz: o livro do especialista*, Vol. 1, Revinter.

Cover, T. and Thomas, J. (2006). *Elements of Information Theory*, John Wiley & Sons.

Davis, S. B. (1979). Acoustic characteristics of normal and pathological voices, *in* N. J. Lass (ed.), *Speech and language: advances in basic research and practice*, Academic Publishers, New York, pp. 271–314.

Diniz, P. S. R., da Silva, E. A. B. and Netto, S. L. (2010). *Digital Signal Processing: Systems Analysis and Design*, Cambridge University Press.

Espinosa, C. H., Redondo, M. F., Vilda, P. G., Llorente, J. I. G. and Navarro, S. A. (2000). Diagnosis of vocal and voice disorders by the speech signal, *IEEE-INNS-ENNS International Joint Conference on Neural Networks*, Vol. 4, pp. 253–258.

Guido, R. C., Slaets, J. F. W., Köberle, R., Almeida, L. O. B. and Pereira, J. C. (2006). A new technique to construct a wavelet transform matching a specified signal with applications to digital, real time, spike, and overlap pattern recognition, *Digit. Signal Process.* **16**(1): 24–44.

Haykin, S. S. (1999). *Neural Networks: A Comprehensive Foundation*, 2 edn, Prentice Hall International.

Lieberman, P. and Blumstein, S. E. (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics*, Cambridge Studies in Speech Science and Communication, Cambridge University Press.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing*, Academic Press.

Rabiner, L. and Schafer, R. (2007). Introduction to digital speech processing, *Foundations and Trends in Technology* **1**(1/2): 1–194.

Rosa, M. d., Pereira, J. C. and Grellet, M. (2000). Adaptive estimation of residue signal for voice pathology diagnosis, *IEEE Transactions on Biomedical Engineering* **47**(1): 96–104.

Santos, L. A. and Scalassara, P. R. (2012). Análise de sinais de voz utilizando entropia relativa e estimador por janela de parzen, *XIX Congresso Brasileiro de Automatica*, p. 6.

Scalassara, P. R., Guido, R. C., Maciel, C. D. and Simpson, D. M. (2011). Fuzzy c-means clustering of voice signais based on predictability measurements of wavelet components, *X SBAI - Simpósio Brasileiro de Automação Inteligente* pp. 75–80.

Scalassara, P. R., Maciel, C. D. and Pereira, J. C. (2009). Predictability analysis of voice signals: analyzing healthy and pathological samples, *IEEE Engineering in Medicine and Biology Magazine* **28**: 30–34.

Schneider, T. and Griffies, S. M. (1999). A conceptual framework for predictability studies, *Journal of Climate* .

Silva, I. N., Spatti, D. H. and Flauzino, R. A. (2010). *Redes Neurais Artificiais para engenharia e ciências aplicadas - curso prático*, ARTLIBER, São Paulo.

Voigt, D., Döllinger, M., Braunschweig, T., Yang, A., Eysholdt, U. and Lohscheller, J. (2010). Classification of functional voice disorders based on phonovibrograms, *Artificial Intelligence in Medicine* **49**(1): 51 – 59.