Voice Disorders Identification Using Multilayer Neural Network

Lotfi Salhi, Talbi Mourad, and Adnene Cherif
Signal Processing Laboratory Sciences Faculty of Tunis, University Tunis ElManar, Tunisia

Abstract: *In this paper we present a new method for voice disorders classification based on multilayer neural network. The processing algorithm is based on a hybrid technique which uses the wavelets energy coefficients as input of the multilayer neural network. The training step uses a speech database of several pathological and normal voices collected from the national hospital "Rabta - Tunis" and was conducted in a supervised mode for discrimination of normal and pathology voices and in a second step classification between neural and vocal pathologies (Parkinson, Alzheimer, laryngeal, dyslexia…). Several simulation results will be presented in function of the disease and will be compared with the clinical diagnosis in order to have an objective evaluation of the developed tool.*

Keywords: *Speech processing, pathological voices, classification, wavelet transform, neural networks, energy.*

Received August 27, 2008; accepted December 28, 2008

1. Introduction

Pathological voice recognition has been received a great attention from researchers in the last decade. Speech processing has proved to be an excellent tool for voice disorder detection. Among the most interesting recent works are those concerned with Parkinson's Disease (PD), Multiple Sclerosis (MS) and other diseases which belong to a class of neuro-degenerative diseases that affect patients speech, motor, and cognitive capabilities [3, 11]. The speech production is a complex motor act that implies a big number of muscles, of physiological variables and a neurological control implying different cortical and under cortical regions.

We distinguish three systems contributing to the production of the speech: the respiratory system, the laryngeal system and the supra-laryngeal system (the articulators) [14, 15]. The nervous system also controls the prosody. This one schematically covers the variations of height (intonation, melody), the variations of intensity (accentuation) and the temporal progress (pauses, debit, and rhythm).

The analysis of the voice disorder stays essentially clinic [12]. The instrumental measures are spilled little in practice clinic. The most used are the acoustic and aerodynamic measures [13]. The speech analysis is complex and has been disregarded for a long time. A difficulty result is to analyze in the literature the different treatment effect (medical or surgical). Indeed, a many studies don't return specific analysis of the speech. On the other hand, we attend confusion between the modifications of the motivity orofacial and the speech quality that remains the clinic objective [1].

Features assessment of a voice disorder is that the disorder carries on a patient's capacity to communicate are a crucial step to conceive a program of its management. A process of the prosperous assessment allows the pathologist of the speech to diagnose the voice disorder, determine the relative efficiency of several treatment approaches and formulate a prognosis [5]. Physicians often use invasive techniques like endoscopy to diagnose the symptoms of vocal fold disorders. However, it is possible to identify disorders using certain features of speech signals [12, 13].

Different classic techniques are used to extract the vocal parameters and so to make the classification of the pathological voices such as pitch and formant detection: PDA (cepstrum, FFT, spectrogram…).

2. Disorder Identification by Pitch and Formants Analysis

2.1. Speech Processing Algorithm

Figure 1 illustrates the algorithm steps for speech processing.

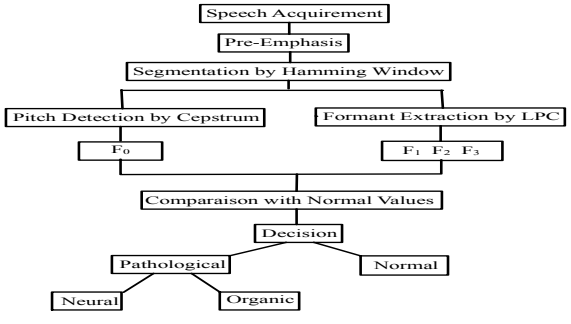
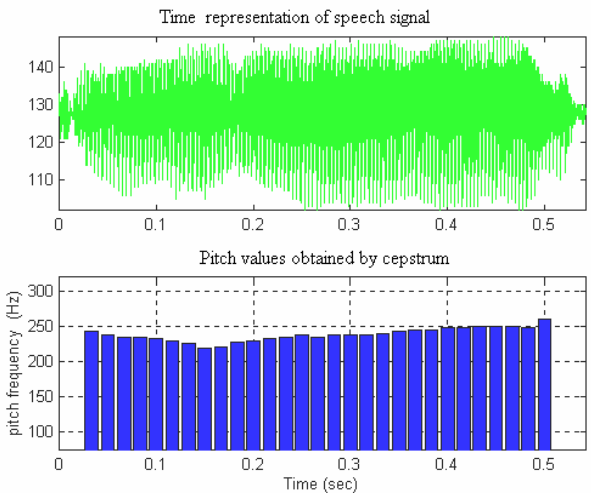


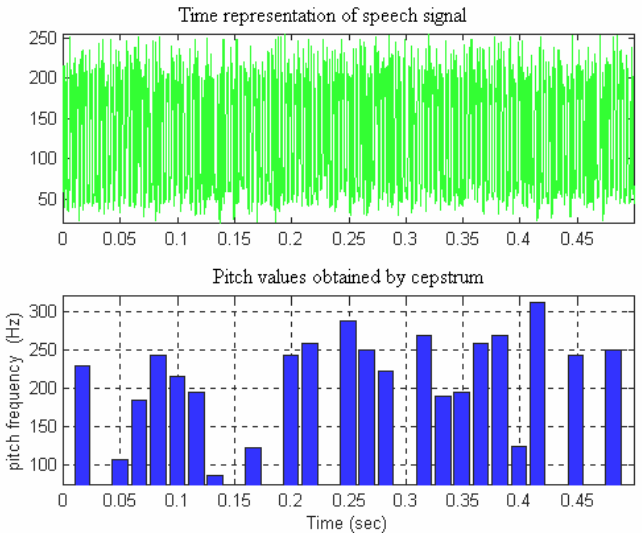
Figure 1. Speech processing algorithm.

2.2. Pitch and Formants Analysis Results

Figure 2 illustrates the pitch variation by application of the cepstrum method analysis of normal and pathological female sounds (32 years). The high distortion and the variation of the pitch around the expected value (250 Hz) demonstrate a state of the glottic signal anomaly, resulting of a laryngeal pathology [2, 7, 8].



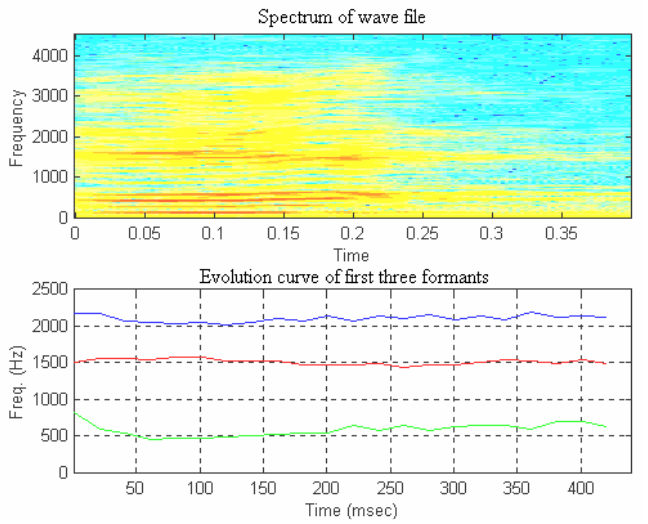
a. Normal female voice.



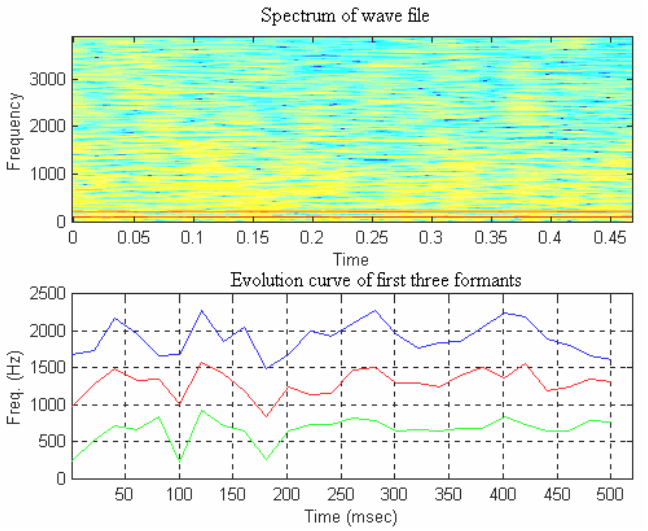
b. Pathological female voice.

Figure 2. Pitch evolution.

Linear Predictive Coding (LPC) method applied on the same voices allows evolution to extract the formants. By comparison to the normal values as shown in Figure 3, the high variations of the formants F_1, F_2 and F_3 of the pathological male sound confirm the conclusions, [2, 7, 8]. Although these methods can help us to distinguish a pathological voice but they remain subjective methods that don't give any quantification values to take the decision. For it, we try to improve this idea by new methods that use wavelet transformed and neural networks.



a. Normal male voice.



b. Pathological male voice.

Figure 3. Formants evolution.

3. New Approach for Voice Pathology Classification

This work presents a development of the basic idea presented in [10]. This paper propose a technique that uses wavelet analysis to extract a feature vector from speech samples, which is used as input to a Multilayer Neural Network (MNN) classifier. Wavelet analysis provides a two dimensional pattern of wavelet coefficients. The energy content of wavelet coefficients at various level of scaling is used to formulate a feature vector of speech sample. Attempt is made to use this feature vector as a diagnostic tool to identify pathological disorders in the voice. A three layer feed forward network with sigmoid activation is used for classification. Generalized Back Propagation Algorithm (BPA) is used for training of the network.

3.1. Algorithm of the Hybrid Method

Figure 4 illustrates the algorithm of the hybrid method.

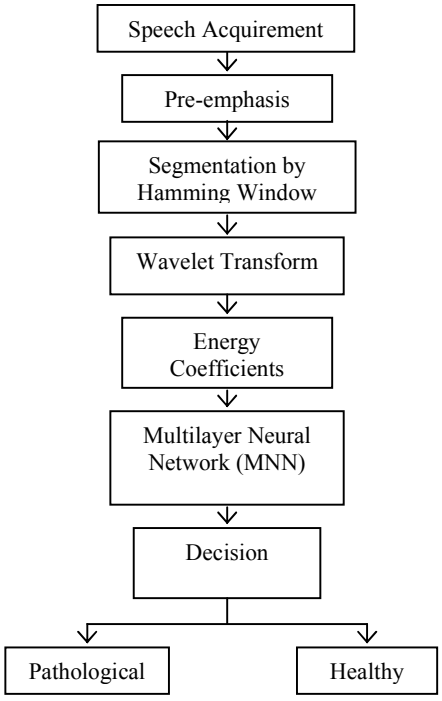


Figure 4. Hybrid method algorithm.

3.2. Wavelet Transforms Analysis

The wavelet transform can be viewed as transforming the signal from the time domain to the wavelet domain. This new domain contains more complicated basis functions called wavelets, mother wavelets or analysing wavelets. A wavelet prototype function at a scale *s* and a spatial displacement *u* is defined as [9]:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi(\frac{t-u}{s}) \quad (u \in IR,\ s \in IR_+^*) \qquad (1)$$

This localisation feature, along with wavelets localisation of frequency, makes many functions and operators using wavelets sparse when transformed into the wavelet domain. This sparseness, in turn results in a number of useful applications such as data compression and detecting features in signals.

3.3. Continuous Wavelet Transforms

The Continuous Wavelet Transform (CWT) is used to decompose a signal into wavelets, small oscillations that are highly localized in time. Whereas the Fourier transform decomposes a signal into infinite length sinus and cosines, effectively losing all time-localization information, the CWT's basis functions are scaled and shifted versions of the time-localized mother wavelet. The CWT is used to construct a time-frequency representation of a signal that offers very good time and frequency localization.

The CWT is an excellent tool for mapping the changing properties of non-stationary signals. The CWT is also an ideal tool for determining whether or not a signal is stationary in a global sense. When a signal is judged non-stationary, the CWT can be used to identify stationary sections of the data stream.

Specifically, a Wavelet Transform function f(t) Є L^2(R) (defines space of square integrated functions) can be represented as:

$$\begin{aligned} W(f)(u,s) &= \int_{-\infty}^{+\infty} f(t)\psi_{u,s}^*(t)dt \\ &= \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\psi^*(\frac{t-u}{s})dt \end{aligned} \qquad (2)$$

The factor of scale includes an aspect transfer at a time in the time brought by the term u, but also an aspect dilation at a time in time and in amplitude brought by the terms s and s et $\sqrt{s}$. The dilation in amplitude permits to preserve a constant norm for all elements of the basis (wavelet energy). The most important criteria for the choice of a wavelet is to present, for it and its Fourier transformed of the possible weakest oscillations; it is what will permit to assure a good temporal and frequency resolution.

3.4. Discrete Wavelet Transform

Discrete Wavelet Transform (DWT), which is based on sub-band coding, is found to yield a fast computation of Wavelet Transform. It is easy to implement and reduces the computation time and resources required. DWT involves choosing scales and positions based on powers of two, so called dyadic scales and positions. The mother wavelet is rescaled or dilated by powers of two and translated by integers.

In CWT, the signals are analyzed using a set of basis functions which relate to each other by simple scaling and translation. In the case of DWT, a time scale representation of the digital signal is obtained using digital filtering techniques. The signal to be analyzed is passed through filters with different cut off frequencies at different scales.

The criteria for selecting a proper mother wavelet is to have a wavelet function with enough number of vanishing moments in order to represent the salient features of the disturbance. At the same time, this wavelet should provide sharp cut-off frequencies. Furthermore, the selected mother wavelet should be orthonormal. Daubechies 40 shows the sharper cut off frequency compared with the others and hence the leakage energy between different resolution levels is reduced. The number of vanishing moments of db40 wavelet is large, and hence it gives a meaningful wavelet spectrum of the analyzed signal. The daubechies wavelet db40 is selected as a good choice because of its high performance in an informal listening test [4].

Wavelet that one often uses in the setting of the treatment of the discreet dimensional mono signal is the wavelet of daubechies. For fast wavelets

transformed DWT, the functions are defined by a game of indications that one designates under the appellation "coefficients of the filters in wavelets" [9]. The daubechies wavelet with compact support is functions to *p* hopeless moments, their regularity increases with *p*. Figure 5 shows the variation of the analysis function of the daubechies wavelet (db40) and its corresponding spectrum.

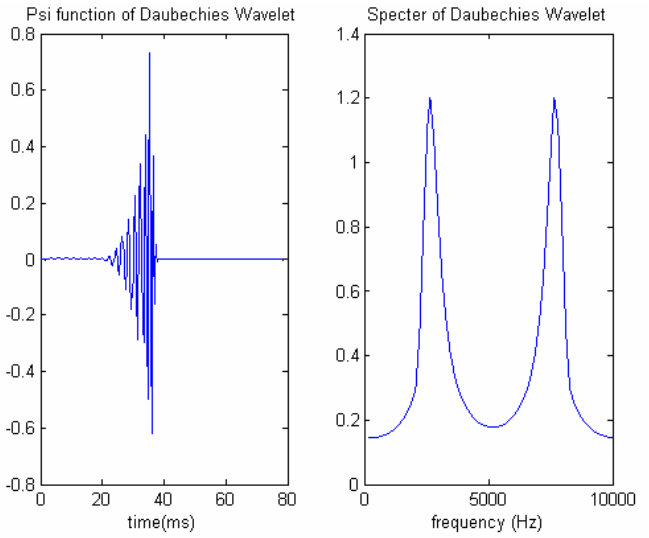


Figure 5. Wavelet of daubechies to 40 hopeless moments and its spectrum.

3.5. DWT and Filter Banks

Starting with a discrete input signal vector x[n], the first stage of the FWT algorithm decomposes the signal into two sets of coefficients. These are the approximation coefficients cA1 (low frequency information) and the detail coefficients cD1 (high frequency information). The DWT is computed by successive low pass and high-pass filtering of the discrete time-domain signal as shown in Figure 6.

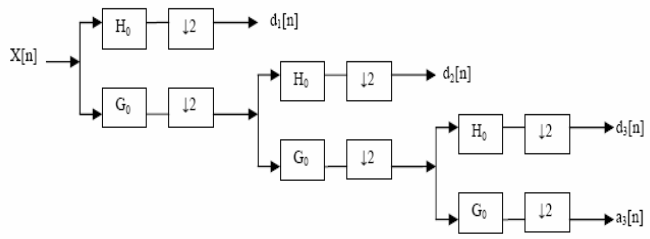


Figure 6. Three level wavelet decomposition tree.

This is called the Mallat algorithm or Mallat tree decomposition. Its significance is in the manner it connects the continuous-time multiresolution to discrete time filters. In the figure, the signal is denoted by the sequence x[*n*], where n is an integer. The low pass filter is denoted by G_0 while the high pass filter is denoted by H_0. At each level, the high pass filter produces detail information d[*n*], while the low pass filter associated with scaling function produces coarse approximations, a[*n*].

3.6. Neural Networks

Neural networks were chosen as a method of pattern matching for many main reasons. First the Matlab software has a fantastic implementation of several different types of neural networks in its neural network toolbox. The big advantage of the neural networks resides in their automatic training capacity, what permits to solve some problems without requiring to the complex rule writing, while being tolerant to the errors [6].

Neural networks consist of several simple parallel computational units called neurons. These units form a neural network that resembles a biological nervous system. The functioning of a neural network is greatly determined by the ways in which its units connect to other units.

A neuron as shown in Figure 7 is an information processing unit, which is an essential part of a neural network. A neuron consists of three main elements: synapses (links), a linear combiner, and an activation function. Each synapse (link) contains a weight factor. Input p(*i*), which is connected to neurone *k*, is multiplied by synaptic weight w(*k*, *i*).

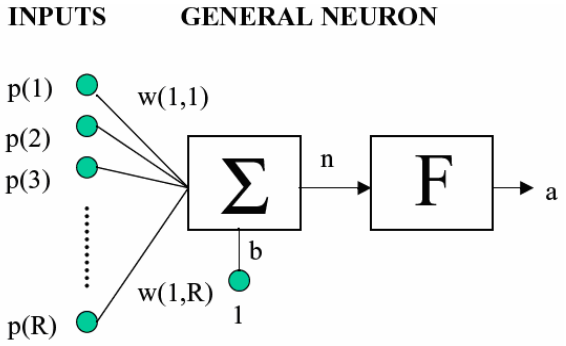


Figure 7. Model of an artificial neuron.

The linear combiner adds the neurone's weighted inputs together and the activation function *F* limits the neuron's output. The figure also shows the bias factor *b*. Hence, the output of a neuron depends on its inputs and its activation function. There are different types of activation functions that can be used in Matlab. The most commonly used activation functions are hard limit, linear, or sigmoid functions. Naturally, one can also construct her own activation function.

A neural network consists of one or more layers. The neurons are arranged into layers so that the input vector values are fed to the first layer, the output from the first layer is fed to the next and so on, until the output layer is reached. Normally the neurons are completely connected in between layers, so each neuron in layer is connected to every neuron in the next layer.

Each layer has a weight matrix *W*, a bias vector *b*, and an output vector a as shown in Figure 8. The number of neurons usually varies between each layer. In Figure 8, the number of inputs is *R*, and the number

of neurons in the first layer is S_1, while in the second layer it is S_2, also the same for other layers. The layers, which are situated between the inputs and the output layer, are called hidden layers. Thus, Figure 8 shows two hidden layers [6].

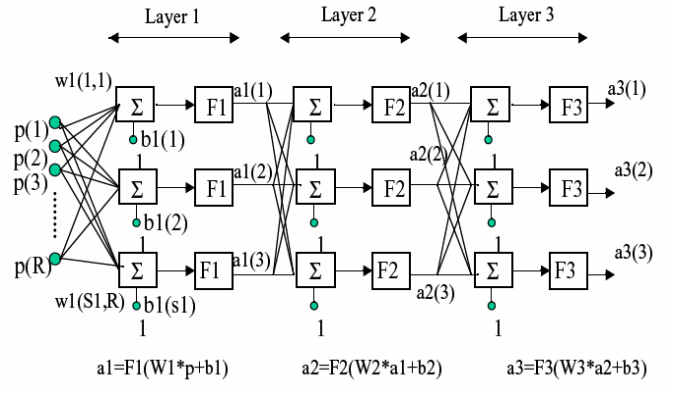


Figure 8. Multi layer neural network structure.

A multi layer neural network can be used to implement an arbitrary Boolean operation, it can be used in formulating boundary surfaces in classification problems, and it can realise almost any arbitrary non linear function.

A neural network is trained by giving a target output to a certain input group, in which case the term supervised learning is used. Alternatively, a network can be trained through self guidance, which means that the network parameters adapt according to the input. In both cases, the free parameters in the network, weights and biases, adapt according to the measured data. The training can be gradual (incremental training), which means that the weights and biases are adapted every time that a new training example is fed to the network, or it can be done in batches (batch training), in which case the parameters are not adapted until all the examples have been fed.

The BPA is often used in the training of multi-layer neural networks. Backpropagation consists of two stages: forward pass and backward pass through the network. In the forward pass, the input is conveyed layer by layer all the way to the output neuron, which produces the true output of the network. In the backward pass, an error signal is produced by deducting the desired output from the actual output. This error signal is conveyed backwards through the network, layer by layer, simultaneously modifying the values of the network weights, thus bringing the actual output closer to the desired output.

In practice, an activation function is used to form a gradient into the network's weight factor space, and the weights are modified into the opposite direction from the gradient. Thus, the activation function must be constant and derivable.

If sigmoid activation functions are used in the output layer, the outputs of the network are limited to a small range. Also, if a linear activation function is used in the output layer, the network outputs can have any real number values.

4. Simulation Results

The new method used in this survey is said hybrid since it takes as a basis on a wavelet transformed followed by a neural network. The input vector for the neural network is constituted by the normalized coefficients energy corresponding to the coefficients of wavelet transformed of input speech signal. The input signal is the voice recorded of a speaker who can be healthy (normal) or pathological. The pathological voice data base has been prepared with help of the G.E laboratory of the UCLA-LosAngeles University and the RABTA hospital of Tunis.

4.1. Wavelet Coefficients

We apply discreet and continuous wavelet transform coefficients on the same word pronounced by two speakers that have the same sex and same age. One of these speakers is healthy whereas the other sulphur of Alzheimer's illness. The simulation result of different absolute coefficients is given in Figures 9 and 10. We notice a clean difference between wavelet coefficient evolutions of the two different signals. This analysis method can also provide a visual pattern, which can be of considerable help in diagnostics.

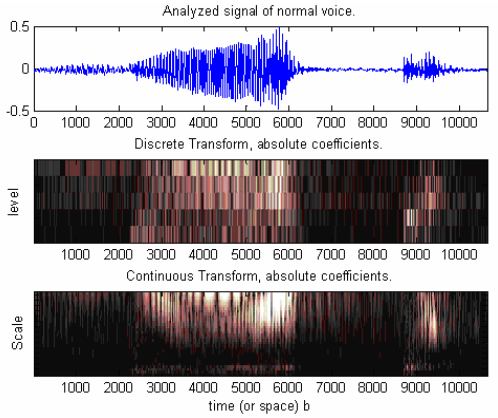


Figure 9. Wavelet analysis of normal voice.

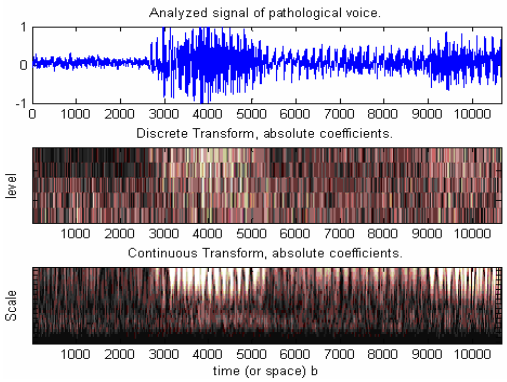


Figure 10. Wavelet analysis of pathological voice.

4.2. Proposed System

The Matlab7.0 platform is used for implementation of the neural network formed of three layers, one of input, one of output and a hidden layer as shown in Figure 11. The input layer is formed of the same neurons number that corresponds to the components of the input vector. The input is the feature vector obtained from wavelet decomposition. The hidden layer contains fifteen neurons and the output layer contains only one neuron to give the decision pathological or normal as shown in Figure 11.

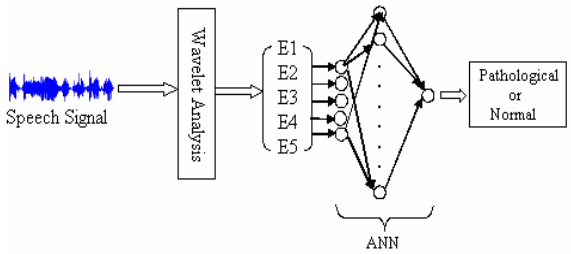


Figure 11. Voices classification model.

4.3. Neural Network Design

The classic cycle of a neural network development can be separate in seven stages:

- The collection of data.
- The analysis of the data.
- The separation of the data bases.
- The choice of a neural network.
- The formatting of the data.
- The training.
- The testing.

In our survey we use a MNN with only one layer hidden between the input layer and the output layer. Every neuron of the hidden layer is connected to the neurons of the input layer and those of the output layer and there is not a connection between the cells of a same layer. The activation functions used in this type of network are the doorstep or sigmoid functions.

This network follows a supervised training according to the rule of errors correction. The training type used for this network is the supervised fashion. To every well stocked input an answer corresponds waited at the output. So the network is going to alter until it finds the good output.

The speech data base used for the training and for the system validation is formed by normal and pathological voices containing about ten words of every type (words for normal and pathological voices). The data base implies several pathological voices as laryngeal, neurological (Parkinson, Alzheimer, dyslexia, dysphonic…) to come from different mixed speakers (men and women). Thus, the results will be compared to those gotten by the classic methods as the cepstrum, LPC and the spectrogram methods for the extraction of the pitch and formants.

4.4. Feature Extraction

The speech is a highly random signal, and then the classic parameter instability as the pitch, jitter and formants can be common for the two types of voice (pathological and normal). So that, the classification will be efficient and effective one chose to use the normalized energies correspondents to the wavelet transformed coefficients as part of input feature vector for the neural network [10]. A Filter Bank is used to extract the wavelet coefficients.

The energy of every level is normalized against total energy content in the signal.

$$E_N(i) = \frac{E_i}{E_T} \quad (3)$$

where i = 1, 2 …

E_T : Total energy across all the levels.

$$E_T = \sum_i E_i \quad (4)$$

E_i : Energy at each level.

The lower band scale presents a more dominant periodicity than the in the higher band scale. This periodicity is decreasing in the pathological speech but it is very consistent in the normal speech [10].

5. Neural Networks Results

We use 60 words on the total, pronounced by different speakers of which 30 are normal and the other present pathologies of vocal or neurological origin. For the training, we use 40 words (20 normal and 20 pathological). After training, the network will be tested with 20 words different from those used for the training (10 normal and 10 pathological).

In order to obtain optimal result, we vary at every time the number of the energy coefficients to the input of the neural network. This procedure requires a variation of the choice of the wavelet filter bank.

5.1. Three Energy Coefficients

We use a wavelet filter bank to extract a three wavelet coefficient then we calculate their corresponding energy.

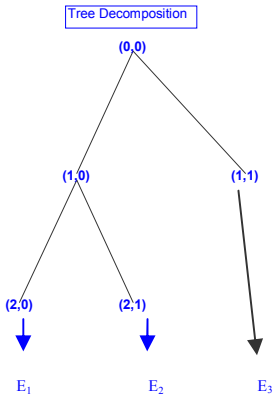


Figure 12. Three energy coefficients extraction.

The obtained training curve is given in Figure 13.

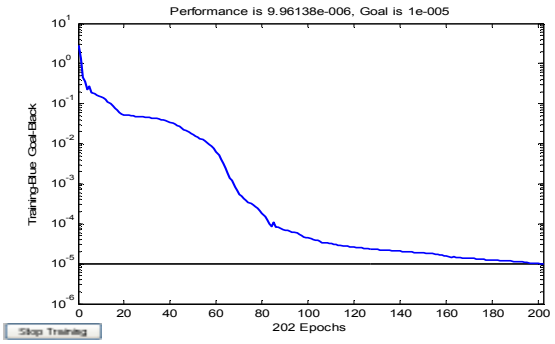


Figure 13. Training curve with three coefficients.

The results of neural network training and testing are regrouped in Table 1.

Table 1. Neural network results with three coefficients.

Pronounced Word	Normal	Pathological
Training Number	20	20
Test Number	10	10
Correct Classification	9	8
Rate of Classification	90 %	80%

5.2. Five Energy Coefficients

We use a wavelet filter bank to extract a five wavelet coefficient then we calculate their corresponding energy.

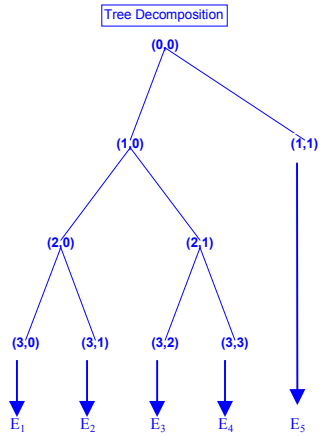


Figure 14. Training curve with five coefficients.

The results of neural network training and testing are regrouped in Table 2.

Table 2. Neural network results with five coefficients.

Pronounced Word	Normal	Pathological
Training Number	20	20
Test Number	10	10
Correct Classification	10	9
Rate of Classification	100 %	90 %

The obtained training curve is given in Figure 15.

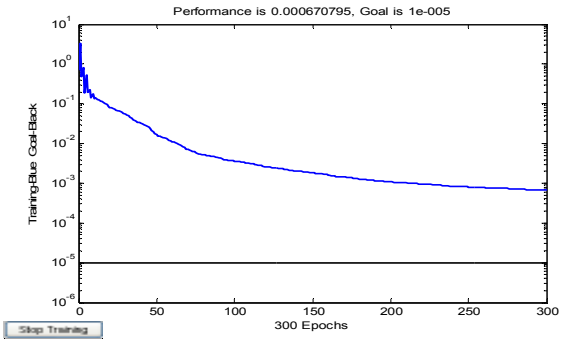


Figure 15. Training with five coefficients.

5.3. Seven Energy Coefficients

We use a wavelet filter bank to extract a seven wavelet coefficient then we calculate their corresponding energy.

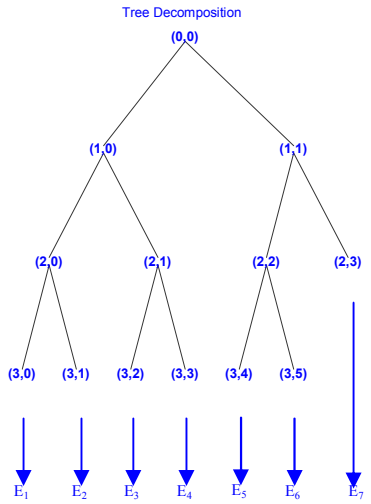


Figure 16. Training curve with five coefficients.

The obtained training curve is given in Figure 17.

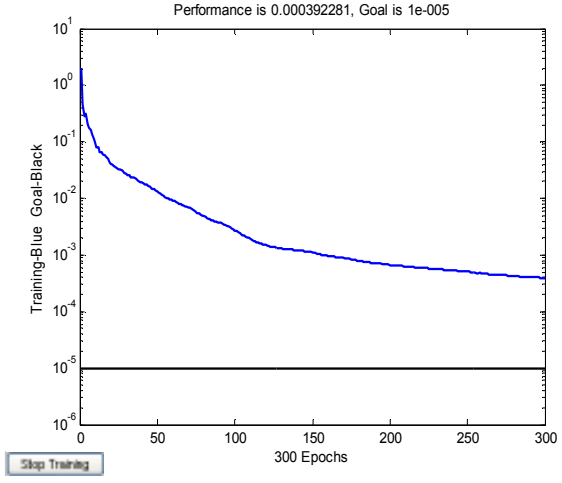


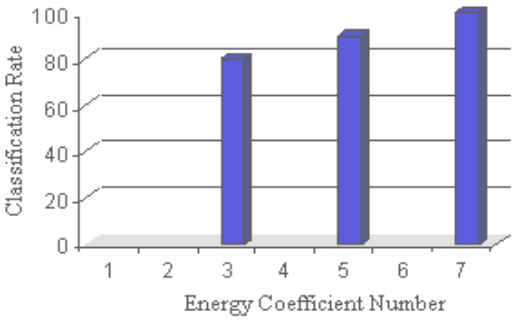
Figure 17. Training with seven coefficients.

The results of neural network training and testing are regrouped in Table 3.

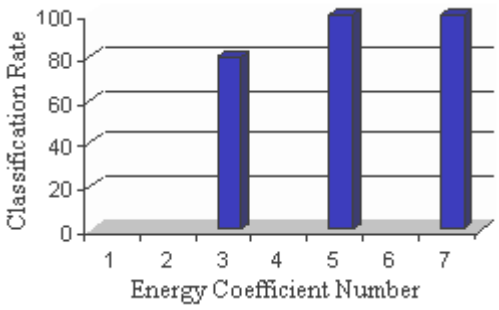
Table 3. Neural network results with seven coefficients.

Pronounced Word	Normal	Pathological
Training Number	20	20
Test Number	10	10
Correct Classification	10	10
Rate of Classification	100 %	100 %

Then we can summarize these different results in the following diagrams as shown in Figure 18.



a. Pathological voices.



b. Normal voices.

Figure 18. Classification by MNN.

6. Conclusions

The goal of this work is to conceive a tool of help to the clinicians in the Tunisian hospitals. This tool allow to follow up of patients who suffer from illness of vocal and neurological origin.

We presented in this paper a material and software interface of numeric treatment of the patient's vocal signal based on neural networks. Result of the MNN classifier gives the correct classification. The classification rate is between 80% and 100%. We have demonstrated in this study, a feature vector based on wavelet coefficients is useful for classification of normal and pathological speech data. At a preliminary level, the speech data is classified into two classes normal or pathological. The MNN with BPA used as a classifier has been proved to be more efficient and more precise than the time-frequency analysis method. The MNN classifier represents a low cost, accurate, and automatic tool for pathological voice classification using wavelet coefficients normalized energy. It is presented in this paper as diagnostic tools to aid the physician and clinician in the analysis of speech disease. Therefore, future work will be focused on the specific recognition of illness type that causes the speech pathology.

This work has to be validated on a larger speech pathology database to increase the result reliability.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable suggestions and comments. The authors also express their thanks to Professor Saber ABID at ENSET of Tunis for his critical reading, comments, and help.

References

[1] Boyanov B. and Hadjitodorov S., "Acoustic Analysis of Pathological Voices: A Voice Analysis System for Screening of Laryngeal Diseases," *in Proceedings of IEEE Engineering in Medical and Biology*, vol. 16, no. 4, pp. 74-82. 1997.

[2] Cherif A., "Detection and Formant Extraction of Arabic Speech Processing," *Computer Journal of Applied Acoustics*, vol. 6, no. 3, pp. 55-58, 2001.

[3] Davis B., "Acoustic Characteristics of Normal and Pathological Voices," *in Proceedings of Speech and Language: Advances in Basic Research and Practice*, Orland, pp. 271-335, 1979.

[4] Gaouda A., Salama M., Chikhani A., and Sultan M., "Application of Wavelet Analysis for Monitoring Dynamic Performance in Industrial Plants," *in Proceedings North American Power Symposium*, Laramie, pp. 154-159, 1997.

[5] Jiang J. and Zhang Y., "Nonlinear Dynamic Analysis of Speech from Pathological Subjects," *in Proceedings of IEEE Electronics Letters*, vol. 38, no. 6, pp. 142-146, 2002.

[6] Kortelainen J. and Noponen K., "Neural Networks," *Technical Document*, University de Sherbrooke, 2005.

[7] Lotfi S. and Adnene C., "A Speech Processing Interface for Analysis of Pathological Voice," *in Proceedings of Communication Technologies from Theory to Applications Conference*, Damascus, pp. 250-254, 2006.

[8] Lotfi S., Haythem B., and Adnene C., "Interface d'Analyse Vocale a l'Identification de Certaines Pathologies d'Origine Neurologique et Vocale," *in Proceedings of JTM Conference*, Tunis, pp. 66-69, 2007.

[9] Mallat S., "A Theory for Multiresolution Signal Decomposition: Wavelet Representation," *Computer Journal of IEEE Transaction Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674-693, 1989.

[10] Nayak J. and Bhat S., "Classification and Analysis of Speech Abnormalities," *Computer Journal of ITBM-RBM*, vol. 26, no. 5, pp. 319-327, 2005.

[11] Parsa V. and Jamieson G., "Interactions between Speech Coders and Disordered Speech," *Computer Journal of Speech Communication*, vol. 40, no. 7, pp. 365-385, 2003.

[12] Plant F., Kessler H., Cheetham B., and Earis J., "Speech Monitoring of Infective Laryngitis," *in Proceedings of International Conference on Spoken Language Processing*, Philadelphia, pp. 749-752, 1996.

[13] Viera N., McInnes R., and Jack A., "Robust F_0 and Jitter Estimation in the Pathological Voices," *in Proceedings of International Conference on Spoken Language Processing*, Philadelphia, pp. 745-748, 1996.

[14] Wang J. and Cheolwoo J., "Performance of Gaussian Mixture Model as a Classifier for Pathological Voice," *in Proceedings of the ASST in Auckland*, Australian, pp. 165-169, 2006.

[15] Yu P., Ouaknine M., Revis J., and Giovanni A., "Objective Voice Analysis for Dysphonic Patients: A Multiparametric Protocol Including Acoustic and Aerodynamic Measurements," *Computer Journal of Voice*, vol. 15, no. 4, pp. 529-542, 2001.

Lotfi Salhi received the Diploma of Master degree in automatic and signal processing in 2004 from the National School of Engineers of Tunis with concentration in digital speech signal processing (cochlear filter). Currently, he is registered for Doctorate degree in electrical engineering (signal processing specialty) at the Faculty of Sciences of Tunis.

Talbi Mourad received his Bachelor degree in math at Sciences Faculty of Tunis and he obtained his Master in automatic and signal processing in 2004 at the National School of Engineers of Tunis. He is a researcher member of the Signal Processing Laboratory in Sciences Faculty of Tunis and preparing his Doctorate Thesis.

Adnene Cherif received his engineering Diploma in 1988 from the Engineering Faculty of Tunis and his PhD in electrical engineering and electronics in 1997. Currently, he is a professor at the Science Faculty of Tunis, responsible for the Automatic and Signal Processing Laboratory.