Analysis of Kohonen's Neural Network with application to speech recognition

Carlos A. de Luna-Ortega [1, 2], Miguel Mora-González [2], Marco A. Álvarez-Medina[1], Julio C. Martínez Romo [3], Francisco J. Luna Rosas [3], Víctor E. Gómez-del Villar [1].

[1]Universidad Politécnica de Aguascalientes.
Av. Prol. Mahatma Gandhi Km.2, Col. San Fco. Del Arenal, Aguascalientes, Ags., C.P. 20280, México
alejandro_de_luna@hotmail.com,
{marco.alvarez, victor.gomez}@upa.edu.mx
[2]Universidad de Guadalajara, Centro Universitario de los Lagos.
Av. Enrique Díaz de León 1144, Col. Lomas del Valle, Lagos de Moreno, Jalisco, C.P. 47460, México
mmora@culagos.udg.mx
[3]Instituto Tecnológico de Aguascalientes.
Av. Adolfo López Mateos 1801 oriente, Fracc. Bonagens, Aguascalientes, Ags., C.P. 20256, México
jucemaro@yahoo.com, fjluna@ita.mx

Abstract. In this paper we present the use of Kohonen's Neural Network (or Self-Organizing Map - SOM) in an Automatic Speech Recognition (ASR) for isolated words in Spanish language with Mexican accent for a single speaker, the words that used indicated directionality, this application could be used in an automatic wheelchair. The corpus of this application uses four words, "adelante" (forward), "Atrás" (backward), "Izquierda" (left) and "Derecha" (rigth). Our algorithm proposes has structure in five steps: recording, filtering, begin-end detection, feature extraction and word recognition (Speech Recognition). The signal was filtering using a wavelet denoising algorithm. We propose a begin-end voiced algorithm with the use of a filters bank, this algorithm found it automatic in the signal recording. Then, we apply LPC algorithm for a feature extraction of each word that we use, after the coefficients (obtained with LPC), are introduce in SOM Network for search what word pronounce. Word recognition accuracy of 91% for average of four words**.**

Keywords. Kohonen's Network, SOM, Self-Organization Map, Wavelet Denoising, LPC, Automatic Speech Recognition.

1 Introduction

Some troubles of an Automatic Speech Recognition (ASR) are: 1) variation in an physiology conditions in a human being [1], [2], [3], [4], that exists because all people have different vocal registers, that registers depends of an age, gentle, and regional accents. All of them give place to origin that two persons can't pronounce the same word with the same characteristics. In particular, two pronunciations of one person can't be matched, in signal question, because exists variations when some person pronounces this words sequence [5], like to speed of pronounce, amplitude of each word, energy with that pronounce, emotional state, etc., we can say that one person can't pronounce the same word. 2) the noise that exists in a digital signal when recording, this trouble is annoying when use all kind of system for convert an analog signal like to the voice, this factor is crucial for an accuracy recognition in an ASR [6].

All kind of algorithm that was designed for an ASR try to solve this troubles, in one hand, search the best form for obtained an high accuracy for recognition, this is, obtain efficient and effective systems for each pronunciation, in an other hand, exist algorithms that remove the noise, always try to increase the accuracy percentage.

For solution of the first trouble exists some methodologies that searching the best solution for an ASR, between them to emphasize the use of algorithms for linear programming, like to the Dynamic Time Warping (DTW) [5], [7], [8], [9], the algorithms that uses statistics and probabilistic tools, like to Regression Linear Algorithm (RLA), Bayesian Networks (BN) [12], [13], Hidden Markov Model (HMM) [9], [14], [15], [16], [35], etc., in another hand, exist algorithms that uses an artificial intelligence like to genetic algorithms [17], and neural networks [18], [19], [20], inclusive the mixed of some algorithms and the build if of algorithms with majority models.

The Solution for the second trouble, typically was used two techniques, all of them try to recognize when the voice is present and when exist an unvoiced, the fist technique is using the zero crossing rate, and another using the energy for determining the presence of voice in signal that was recorded [21], [22]. The central objective is only processing the signal when exist voiced, and discriminate the typical noise.

The majority of the ASR systems using windowing with a Hamming window, this windowing useful for the stationary of the signal [23], [24], [25], generated that the signal has a quasi-stationary and work it with the classical techniques for the stationary signals for the feature extraction. This windowing represents more computer-time and recourses for an ASR.

We propose an adaptation of an algorithm of ASR without windowing and pre-emphasis, use the Linear Predictive Coding (LPC) for feature extraction for the pronunciation of each word without the window of 20ms that is frequently use in all kind of algorithm for ASR, likewise we include two algorithms: 1) for denoising and begin-end voice detection, respectively; 2) for speech recognition uses a Self Organizing Map network (SOM), with a single speaker, isolated word and the corpus pronounce of words in Mexican Spanish language. The words to recognition are "adelante" (forward), "atrás" (backward), "izquierda" (left) and "derecha" (right), this application could be used in an automatic wheelchair. These studies of the corpus try to define the variations that could be exist with the accent of the different regions of Mexico, in the pronunciation, and their relationship with accuracy of ASR.

The rest of the paper is organized as follows. The methodologies for filtering, begin-end voiced detection and SOM algorithm was presented in Section 2. In Section 3, is presented the experimental setup with the proposed algorithms. The experimental results are given in Section 4, and finally we conclude with the conclusion Section.

2 Methodology

In this section we present the methodologies using in the implementation of our ASR, adaptation with our corpus. Our algorithm proposes do not use a windowing and do not have pre-emphasis, and have structure in five steps: recording, filtering, begin-end detection, feature extraction with LPC and word recognition with SOM algorithm.

2.1 Filtering with wavelet denoisinig

Most of the electrical signals presented different kinds of noise (could be randomize, pop, Gaussian, etc.) [26], [27]. The noise modifies the original signal. A simple model of this phenomenon is the adding with the original signal, this could be represents for

$$f = s + r, \tag{1}$$

where f, s and r are the contaminated signal, the original signal and the noise, respectively. The method of denoising with wavelet transform consists in assumed that f is the same that the transmitted signal with noise, and supposed the next conditions:

1. To recover the energy of s is covered, in a higher percentage, for the values of the Wavelet Transform (WT) and this values presents magnitudes higher that the threshold $T > 0$.

2. All the noise signal values of the WT presents magnitudes below that the noise threshold T_r, to fulfill with $T_r < T$.

The steps of the algorithm for removing the noise use the method of "Wavelet Threshold". These steps are:

1. Select the T value for a threshold, where:

$$WT(f) = \begin{cases} 0, & WT(f) < T \\ WT(f), & WT(f) \geq T \end{cases} \tag{2}$$

2. When obtained the Inverse Wavelet Transform (IWT) has an approximated of the signal without noise [26], this is $IWT(WT)$.

In [28] describe the calculus of the threshold for the Wavelet coefficients with empirical method using a universal threshold:

$$T = \sqrt{2\sigma^2 \log N}, \tag{3}$$

where σ^2 is the variance of the original values and N is the simple size.

The results of the wavelet decomposition presents an approximation of the coefficients, see Figure 1, with multiresolution filter bank with 4 Daubechies coefficients of the WT for the analysis stage. The stage of the signal WT reconstruction and the post-processing, can see in Figure 2, with two levels and filter bank with 4 Daubechies coefficients. This analysis of multiple resolution enable the capacity for analysis the signal with different kind of frequency bands; with this, can see it every transient stage in the time domain and the frequency domain [29].

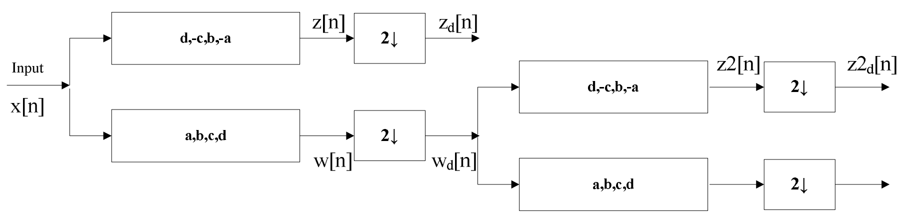

Figure 1. Two levels resolution of the filter bank for the decomposition signal and 4 Daubechies coefficients [29].

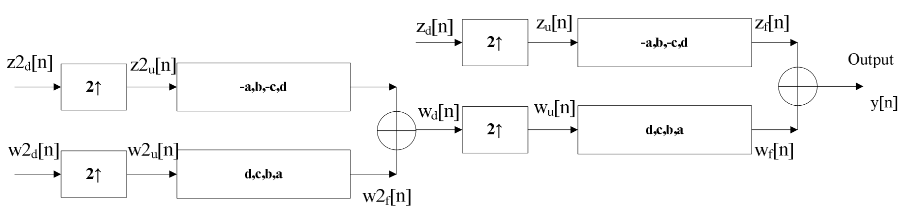

Figure 2. Two levels resolution of the filter bank for the reconstruction signal and 4 Daubechies coefficients [29].

In the Figures 1 and 2, has that n is the simple rate, x[n] is the input signal, a, b, c and d are the coefficients of each filter, z[n], z2[n], w[n], w2[n] are the output signals, f is the filtering stage, 2↓ is a *downsampler*, 2↑ is an *upsampler*, z_d[n] is the output of the first high pass filter, w_d[n] is the output of the first low pass filter, $z2_d$[n] is the output of the second high pass filter, $w2_d$[n] is the output of the second low pass filter, and $z2_u$[n], z_u[n], $w2_u$[n] and w_u[n] are the outputs of the *upsamplers*, $z2_f$[n], z_f[n], $w2_f$[n], are w_f[n] are the outputs of the high pass and low pass filters for the reconstruction signal, and y[n] is the output signal.

2.2 Begin-End Pronunciation

For the search of begin and end of each word pronounced, was used the next block diagram (see Figure 3) [30]. First, the algorithm used a Band Pass Filter (BPF) with a cascade composition of two filters, Low Pass Filter (LPF) and High Pass Filter (HPF), this function obtained the zone with more agglomeration of the signal without noise. After apply a derivative filter, where was find the zero crossing. Then, was calculated the energy of the signal and applied an adaptative threshold that found it begins and end of each pronunciation.

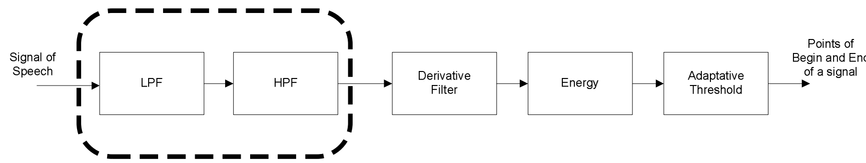

Figure 3. Block Diagram of the Search begin-end word pronounce algorithm.

2.3 SOM Algorithm

The SOM Neural Network as an ordered, no linear, smooth mapping of high-dimensional input data manifolds onto the elements of a regular, low-dimensional array. The input is definable as a vector $x = \{x_1, x_2, ..., x_n\}^T \in R^n$ and the output as a vector $m_i = [m_{i1}, m_{i2}, ..., m_{in}]^T \in R^n$. Assuming a measure distance between x and m_i, denoted for $d(x, m_i)$, the image of an input vector x on the SOM array is defined as the array elements m_c that has best matches with x, this index [31] is defined by the next equation

$$c = \arg\min_i \{d(x, m_i)\}. \tag{4}$$

For common selection of the distance *d,* the Euclidean distances are defined by [32]

$$d(x, m_i) = \sqrt{\sum_{k=1}^{n} (x_k - m_{ik})^2} \tag{5}$$

In general, if exists dismatches of x and m_i when obtained d, found it the first winner neuron m_c.

The SOM is a special type of competitive learning network that defines a spatial neighborhood for each output unit. During competitive learning, all weight vectors associated with the winner and its neighboring units are updated.

In the next lines presents the steps of the learning SOM algorithm [33].

1. Initialize weights with random numbers; set initial learning rated and neighborhood.
2. Present a pattern **x** and evaluate the network outputs.
3. Select the unit (c_i, c_j) with the minimum output:

   $$\left| x - w_{c_i c_j} \right\| = \min_{ij} \left\| x - w_{ij} \right|.$$
4. Update all weights according the following learning rule:

   $$w_{ij}(t+1) = \begin{cases} w_{ij}(t) + \alpha(t)[x(t) - w_{ij}(t)], & \mathit{if}\ (i,j) \in N_{c_i c_j}(t) \\ w_{ij}(t), & \mathit{otherwise} \end{cases},$$

   where $N_{c_i c_j}(t)$ is the neighborhood of the unit time (c_i, c_j) and $\alpha(t)$ is the learning rate.
5. Decrease the value of $\alpha(t)$ and shrink the neighborhood $N_{c_i c_j}(t)$.
6. Repeat steps 2 through 5 until the change in weight values is less that a pre-specified threshold or a maximum number of interactions are reached.

3 Experimental Set-up

In the Figure 4 see the block diagram of the experimental set-up and the next lines explain all the stages were constituted.

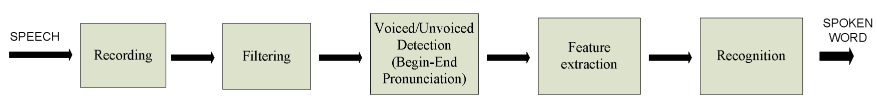

Figure 4. Block diagram of the experimental set-up.

3.1 Recording

The recording of word pronunciation was performed with a Laptop Computer XPS Dell with a microphone integrated in the computer. Was realized a consecutive series of pronunciations with the same word during two minutes. This signal was obtained with variations in speed, frequency and intensity. All this was successful for a single speaker and with the software Simulink® of Matlab®.

The time recording was established in 2 minutes with a sample rate of 8kHz, because with this time we obtained 50 samples of each pronounce. This number of samples was considering enough for our experimental, 35 samples for training and 15 samples to recognize test. These words were pronounced in controlled environment and with characteristic noise generated for the lighting lamps, sometimes with spontaneous sound outside of the experiment, moreover exists an offset in the signal.

3.2 Filtering

For Filtering was realized with wavelet denoising using a multiresolution filter of 12 levels with 4 Daubechies coefficients (a=-0.4830, b=0.8365, c=-0.2241, d=-0.1294). The Daubechie wavelet was selected because this type of wavelet presented the best results with the noise filtering, likewise a comparison was made with different levels of the filter bank, was obtained that over of 12 levels presents a greater reduction of noise and obtained a better signal, we used 12 levels because is the minimum in computer time with the best results.

This algorithm generated one change of amplitudes of the signal and a little distortion of pronounce, but this effect did not affected in the information that was implemented in the algorithm, because this considerations was used in the learning and recognizing for the word pronunciation.

3.3 Voiced/Unvoiced Detection (beginning and end pronunciation)

With the filter bank described in section 2.2 was realized the search begin and end of each pronunciation, in the Figure 5 presents the results for the implementation of our algorithm. The algorithm returned the index where the word begin and end of the pronunciation; this algorithm detected the beginning and end in a large sequence of pronunciation, and is important say that all beginning and end of the pronunciation, that was find it, presents different length in time, whereby was necessary to apply an alignment.

It should be mentioned that the derivative filter was used, is based it on the spline wavelet of order 4, which allows us to obtain the signal derivative function; this gain detected the zero crossings of the signal and removed the offset that occurs during the recording of sound.

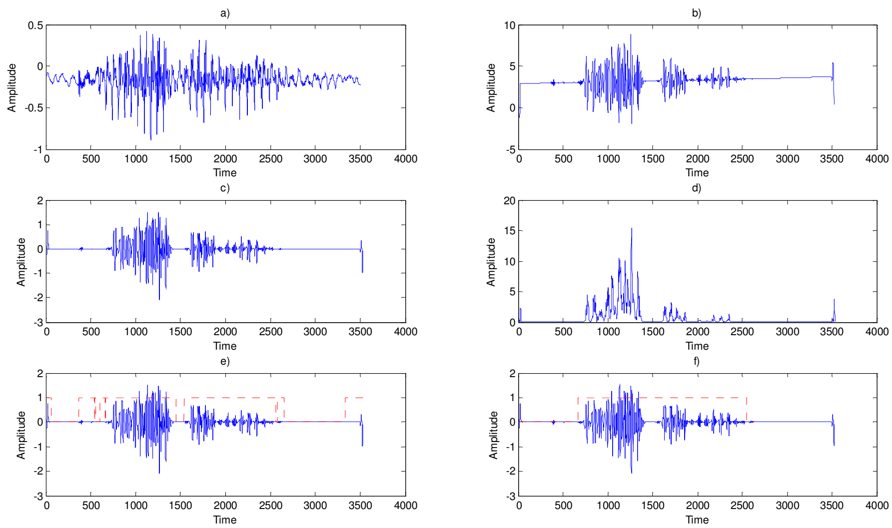

Figure 5. a) Original Pronunciation, b) Pronunciation after Filtering with Denoising and Band Pass applied, c) Pronunciation after of derivative filter, d) Energy of the pronunciation after c), e) Points of Voiced/Unvoiced Detection of the pronunciation, and f) Beginning and End Detection of the Pronunciation.

3.4 Recognition

SOM Training, Learning and Simulation.

The SOM network was created in Matlab® with an array of 25X5 of classification neurons, using a hexagon topology, calculating the neighbors with Euclidean distances algorithm. The inputs of the net was two, first, the word to used for training, learning or recognizing, and second, the average vector of the word pronounced (this vector was obtained with 35 samples of each words and calculated the average of their coefficients and); this vector change according to the pronunciation to try to recognized, for example if the pronunciation was "izquierda", the learning was made with the "izquierda" average vector, along with the other pronunciations.

The 35 samples of each pronunciation used for training with the average vector of each word obtained, after this, the network obtained a vector with the characteristics weights of each word trained. Finally, we obtained 4 vectors with characteristics weights (vectors of "adelante", "atrás", "izquierda" and "derecha").

Our algorithm was trained with 10,000 epochs, because Kohonen [34] in his paper, provides this epochs as sufficient training for applications of speech recognition.

Recognition Network

The algorithm used for recognition from the data obtained from the SOM network is performed by the sequence of steps outlined below:

1 The word to recognize is taken and is introduced to the neuronal network next to the media vectors of the four words.
2 The simulation of the neuronal network is carried out with the word to recognize and each one of the media vectors.
3 It is obtained with each simulation, a distance comparison with each one of the four words, using the method of Euclidean distances.
4 A comparison among the four distances obtained is carried out, and the distance with smaller value represents the similar pattern that defines that is the same word.

5 Experimental Results

The proposed algorithm did not use a pre-emphasis and a Hamming window, with this; we could avoid the use of fragments and worked with full signal, by obtaining the LPC coefficients. When this considerations, in Table 1 presents the results was obtained with the SOM network. The accuracy of the algorithm was 91% (average of 4 pronounces), in each pronounce exist some variations in recognition rate, because it involves factors such as length, duration and amplitude in each pronunciation.

Pronounce	**Percentage of Recognition**
"Adelante" (forward)	**87%**
"Atrás" (backward)	**86%**
"Izquierda" (Left)	**100%**
"Derecha" (right)	**90%**
Average	**91%**

Table 1. Percentage of Recognition (each word pronunciation)

6 Conclusions

We implemented a filtering stage to reduce noise that will support future application of the implementation of the algorithm in real time, and its possible relevance in an embedded system.

The search for the beginning and end gave us a reduction in computation time, because we don't have to process the silent before and after the word pronounced, whit this, we could handle some files with several pronunciations without problem for training and recognition.

The LPC coefficients us favorably reduced the amount of data to work and thus could reduce the time, resulting in faster training and recognition.

The implementation of SOM neural network provides training in a non supervised pattern giving a wide variety of classification of pronunciation of one word, thereby gaining an important feature that gives us greater flexibility in the variance of the pronunciation of words, and this type of network it offers the advantage of not having to assign output values for training and the speed of trains gives greater flexibility to be able to calculate weights of each keyword with 10,000 training times in short periods of time. Also how the network fits the pattern gives pronunciations to get better results in every test performed. With the above was achieved by 91% of general recognition of the four words.

The Automatic Speech Recognition was presented an 9% error rate, is quite acceptable, because if we pronounce 10 words, our system recognize 9, it may be mentioned that the word "izquierda" was obtained 0% error rate, because exist enormous difference in the word with the rest of the corpus managed.

References

1. Tebelskis J. PhD Thesis, Speech Recognition using Neural Networks, School of Computer Science Carniege Mellon University. Pennsylvania (1995).
2. Merlo, G., Fernández, V., Caram D., Priegue, R.: Reconocimiento de voz mediante una Red Neuronal de Kohonen. In: Proceedings of CACIC, pp. 1--7. CACIC Press, Buenos Aires(1997).
3. Campbell J.: Speaker Recognition: A Tutorial. Proceedings of the IEEE, Vol. 85, No. 9. pp. 1437-1462(1997).
4. Benzeghiba M., De Mori R., Deroo O., *et. al.*: Automatic speech recognition and speech variability: A review. Speech Communication. Vol. 49, pp. 763-786 (2007).
5. De Luna-Ortega C.A., Mora-González M. and Martínez-Romo J.C.: Reconocimiento de voz con Redes Neuronales, DTW y Modelos Ocultos de Markov. Conciencia Tecnológica. Vol. 32, pp. 13-17 (2006).
6. Obaidat M.S., Lee C., Sadoun B., Nelson D.: Estimation of pitch period of speech signal using a new dyadic wavelet algorithm. Information Sciences. Vol. 119. pp. 21-39, (1999).
7. Furui, Z.: Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech. IEEE Transactions on acoustics, speech, and signal processing, vol. ASSP-34(1), pp. 52--59. (1986)
8. Irwin M.J.: **A** Digit Pipelined Dynamic Time Warp Processor. IEEE Transactions On Acoustics, Speech, And Signal Processing, Vol. 36, No. 9, pp.1412-1422 (1988).
9. De Wachter M., Matton M., Demuynch K., Wambacq P. Cools R.: Template-Based Continuous Speech Recognition. IEEE Transactions On Audio, Speech, And Language Processing, Vol. 15, No. 4, pp. 1377-1390 (2007).
10. Cox, S. "Speaker Adaptation in Speech Recognition Using Linear Regression

Techniques". Electronics Letters. Vol.28, No.22, pp. 2093-2094 (1994).
11. de Luna-Ortega C.A., Mora-González M., Martínez Romo J.C., Luna Rosas F.J., Álvarez-Medina M.A.: Reconocedor de Palabras por medio de coeficiente R2. No publicado.
12. Zweig G., Russell S.: Speech Recognition with Dynamic Bayesian Networks, AAAI-98 Proceedings, (1998). https://www.aaai.org/Papers/AAAI/1998/AAAI98-024.pdf
13. Nefian A., Liang L., Pi X., Lui X., Murphy K.: Dynamic Bayesian Networks for Audio-Visual Speech Recognition. EURASIP Journal on Applied Signal Processing, No. 11, pp. 1-15 (2002).
14. Rabiner, L.R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". Proceedings of the IEEE. Vol. 77, No.2, pp. 257-286 (1989).
15. Kinjo, T., Funaki, K.,"On HMM Speech Recognition Based on Complex Speech Analysis". Proc. IECON 2006 – 32nd Annual Conference on IEEE Industrial Electronics. No.1, pp. 3477-3480 (2006).
16. Oropeza R., Suárez G., "Algoritmos y Métodos para el Reconocimiento de Voz en Español Mediante Sílabas", Computación y Sistemas, Vol. 9, No. 3, pp. 270-286 (2006).
17. Romo, J.C., Rosas, F.J., Mora-González, M.: Combining Genetic Algorithms and FLDR for Real-Time Voice Command Recognition. In: Proceedings of the 2008 Seventh Mexican international Conference on Artificial intelligence, pp. 163--169. IEEE Computer Society, México (2008)
18. Orozco García, J., Reyes-García, C.A.: Clasificación de Llanto del Bebé Utilizando una Red Neural de Gradiente Conjugado Escalado. In: MICAI/TAIA 2002, pp 203-213. SMIA Press, Mérida (2002)
19. Ceccarelli, M., Hounsou, J.T.: Sequence recognition with radial basis function networks: experiments with spoken digits. Neurocomputing, 11(1), pp. 75--88 (1986).
20. Pardo J.D., Castro J.A. et. al., "Reconocimiento Automático del Habla utilizando la transformada de Fourier y Redes Neuronales", Revista Colombiana de Física. Vol. 38. No. 4, pp. 1595-1598. (2006).
21. Gökhun T.S., Özer H.: Voice Activity Detection in Nonstationary Noise, IEEE Transactions on speech and audio processing, Vol. 8, No. 4, pp. 478-482, (2000).
22. Hwai-Tsu H., Chu Y., Chih-Hang L.: Usefulness of the comb filtering output for Voiced/Unvoiced classification and Pitch Detection. 2009 International Conference on Signal Processing Systems. IEEE Computer Society. pp. 135-139. (2009).
23. Calvo Arias, R. Tesis Ingeniería. (2002). Reconocimiento de voz. Instituto Tecnológico de Costa Rica. Costa Rica. http://bibliodigital.itcr.ac.cr:8080/dspace/bitstream/2238/117/1/BJFIE200269.pdf
24. Juang B.H., Rabiner L.R., Wilpon J.G., "On the Use of Bandpass Liftering in Speech Recognition". IEEE Transactions On Acoustics, Speech, And Signal Processing. ASSP- Vol. 35, No.7, pp. 947-954 (1987).
25. San Martín S.C., Carrillo A.R., "Implementación de un Reconocedor de Palabras Aisladas dependiente del Locutor". Revista Facultad de Ingeniería U.T.A., Vol.12, No.1, pp. 9-14 (2004).
26. Walker, J. S., A primer on wavelets and their scientific applications. Second Edition. 26-87. Chapman & Hall/CRC. Florida, 2008.
27. Mora-González M., Casillas-Rodríguez F.J., Muñoz-Maciel J., Martínez Romo J.C., Luna Rosas F.J., de Luna-Ortega C.A., Gómez Rosas G., Peña Lecona G.,"Reducción de ruido digital en señales ECG utilizando filtraje por convolución". Investigación y Ciencia. Vol.16. No.040, pp. 26-32 (2008).
28. Thillard M.,"Wavelets in Soft Computing". First Edition. 28-29. World Scientific. Singapore, 2001.

29. Weeks M., "Digital Signal Processing using Matlab an Wavelets". First Edution. 315-316. Infinity Science Press LLC. Massachusetts (2007).
30. Álvarez-Medina M.A.. Master Thesis: "Análisis Y Reconocimiento De Patrones". Instituto Tecnológico de Aguascalientes, Aguascalientes, México (2005).
31. Kohonen T., "Self-Organization Maps", 3rd. Edition, Springer, New York, (2001)
32. Brugger D., Bodgan M, Rosenstiel W.: Automatic Cluster Detection in Kohone's SOM, IEEE Transactions on Neural Networks, Vol. 19, No. 3, pp. 442-459. (2008)
33. Jain, A. K., Mao, J., and Mohiuddin, K. M. 1996. Artificial Neural Networks: A Tutorial. Computer vol. 29, 3, 31-44. (1996)
34. Kohonen T.: The Self-Organizing Map. Proceedings of the IEEE, Vol. 78, No. 9, pp. 1464-1480. (1990).
35. De Luna-Ortega, C.A. Master Thesis: "Reconocimiento de voz con Redes Neuronales y Modelos Ocultos de Markov". Instituto Tecnológico de Aguascalientes. México, (2005).