UDC 654.14:681.3.06:517.4
Original scientific paper
Received: 18.09.2007.

Recognition of speaking in a phone channel with a neural network backpropagation

Luis A. Cruz-Beltrán, Marco Antonio Acevedo-Mosqueda and **Jose Luis López-Bonilla**
Seccion de Estudios de Postgrado e Investigation, Escuela Superior de Ingenieria Mecanica y Electrica,
Instituto Politécnico Nacional, Edif. Z-4, 3er. Piso, Col. Lindavista, C.P. 07738, MEXICO D.F.
e-mails: lcruzb06@ipn.mx; macevedo@ipn.mx; jlopezb@ipn.mx

SUMMARY

*In this paper we propose an algorithm for identification of speaking in the phone channel. The investigation is based on the behavior of the Artificial Neural Network (ANN) Backpropagation. The analysed voice is captured with format *.wav. We introduce the procedure for the speaking identification in the phone channel which offered the best results in several tests that we carry out with our method.*

***Key words**: artificial neural networks, wavelets, backpropagation, CLP, wav.*

1. INTRODUCTION

The verification or people's identification in a phone channel, using their voice pattern, it is the base for the realization of this article which exhibits a system that uses the voice pattern's characteristics, Code of Linear Prediction (CLP) and Artificial Neural Network (ANN) to be able to solve, for example, a juridical case efficiently in which the accused's phone recording is mixed as test of the case and it is needed a system that authenticates and corroborates if indeed the voice into the recording belongs to the inculpated that is in a penal process.

The system consists of three stages. The first of them refers to the capture of the voice signal, after is made a preprocessing of the same, in which is carried out the extraction of the characteristics of the voice by means of Wavelets and the coefficients of CLP. In the third stage is carried out the speaker's recognition using one ANN Backpropagation.

2. ARTIFICIAL NEURAL NETWORKS

The ANN are systems information's processing whose structure and operation were inspired by the biological neural networks [1]. In all models of ANN four basic elements are present:

1) A set of connections, weights or synapses that determine the behavior of the neuron, which can be excitatory, exhibit a positive signal (positive connections) and the inhibiting has a negative signal (negative connections);
2) A function that takes charge of adding all the inputs multiplied by their corresponding weights;
3) An activation function that can be linear or nonlinear which limits the amplitude of the output from the neuron;
4) An external gain that determines the threshold of activation of the neuron.

Ever since the psychologist Frank Rosenblatt in 1957 [1] introduced the model of perceptron of a single layer, the ANN became a powerful tool to solve several

types of problems related with the classification, functional estimation and optimization of pattern recognition.

The proposed model is described with Eq. (1), where $x_{p1}\ldots x_{pj}$, are the inputs units, $w_{j1}\ldots w_{ji}$ are the weights of the ANN, b_i is the gain or threshold of activation, N_{pj} is the product of the weights with respect to the input, f is the function of activation of the ANN, and finally y_{pj} is the output of the ANN, having the form:

$$y_{pj} = f(N_{pj} = \sum_{i=1}^{m} x_{pi} w_{ji} + b_i),\ m \in IR,\ m < \infty. \tag{1}$$

2.1. Code of linear prediction

A great part of the applications related with the treatment of the speech, are based on the analysis of codes of linear prediction (CLP), since it is able to extract the linguistic information and to eliminate the corresponding to the particular person. The linear prediction models human vocal zone as an answer to the infinite impulse that produces the voice signal.

The term linear prediction refers to the method to predict or to approach a sample of a signal in the domain of time $s[n]$ based on several previous samples $s[n\text{-}1]$, $s[n\text{-}2]$, $s[n\text{-}M]$:

$$s[n] \approx \hat{s}[n] = -\sum_{i=1}^{M} a_i s[n-i], \tag{2}$$

where $\hat{s}[n]$ is called signal sample, and a_i, $i=1,2\ldots M$ are the predictor or coefficients of CLP. A small number of coefficients of CLP a_1, $a_2\ldots a_M$ can be used to represent a signal efficiently $s[n]$, Ref. [2]. The values $a_1, a_2\ldots a_M$ are the base for the realization of this work participating because they help us to model the parameters of the voice of each one of the speakers that are used in this proposed system.

3. NETWORK BACKPROPAGATION

In 1986, Rumelhart, Hinton and Williams, based on another works formalized a method so that a neural network learned the association that exists between the input patterns to same and the corresponding classes, using more levels of neurons than those that Rosenblatt employed to develop the Perceptron.

This new method is known as Backpropagation (retropropagation of error) that is a type of network of supervised learning, which uses a propagation-adaptation cycle of two phases.

Once applied a training pattern to the input of the network, this propagates from the first layer through the subsequent layers of the network, until generating an output, which is compared with the output wished and an error signal is calculated for each one of the outputs, this is propagated as well backwards, beginning from the output layer, towards all the layers of the network until arriving at the input layer, with the purpose of updating the weights of connection of each neuron, to make that the network converges to a state that allows it to classify all the training pattern correctly.

The general structure is shown in Figure 1.

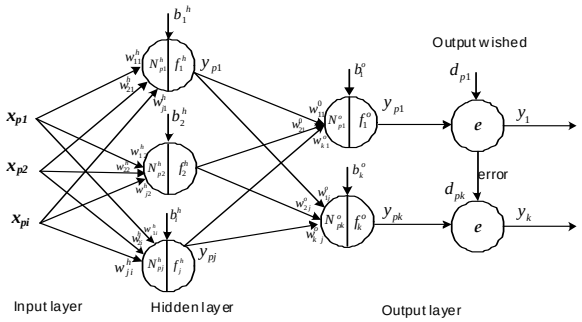

Fig. 1 Model of the ANN Backpropagation

3.1 Training algorithm of the network

Next, the algorithm used for the training of the ANN Backpropagation is presented, Refs. [1, 3, 4]:

1) To initialize the weights of the network (w) with small random values;
2) To carry out the steps (3-6) while the stop condition is false;
3) A pattern of inputs is presented, (x_{p1}, x_{p2},…, x_{pi}) and specific wished output that it should generate the network (d_{p1}, d_{p2},…,d_{pk});
4) Calculate output actuality of the network. The inputs to the network are presented as well as calculating the output that presents each layer until arriving to the output layer ($y_1, y_2,\ldots, y_k$). The steps are the following:
   a. The net inputs are determined for the hidden neurons coming from the input neurons:

$$N_{pj}^{h} = \sum_{i=1}^{m} w_{ji}^{h} x_{pi} + b_i^{h} \tag{3}$$

   where h refers to the magnitudes of the hidden layer; the subindex p, to the vectorial p-th of training, and j to j-th hidden neuron; b is bias (it acts like an input more).

   b. The activation function is applied to each one of the input of the hidden neuron to obtain its respective output:

$$y_{pj} = f_j^{h}(N_{pj}^{h} = \sum_{i=1}^{m} w_{ji}^{h} x_{pi} + b_i^{h}) \tag{4}$$

   c. The same calculations are carried out to obtain the respective outputs from neurons of the output layer:

$$N_{pk}^{o} = \sum_{j=1}^{n} w_{kj}^{o} y_{pj} + b_{k}^{o} \quad (5)$$

$$y_{pk} = f_{k}^{o}(N_{pk}^{o} = \sum_{j=1}^{m} w_{kj}^{o} y_{pj} + b_{k}^{o}) \quad (6)$$

where *o* refers to the magnitudes of the output layer;

5) Determination of error terms for all the neurons:
   a. To compute the error term of each output neuron (wished output–output obtained):

$$e = (d_{pk} - y_{pk}) \quad (7)$$

   b. Obtaining the delta (product of the error with the derived activation function and with regard to the weight of the network):

$$\delta_{pk}^{o} = e * f_{k}^{o'}(N_{pk}^{o}) \quad (8)$$

6) Updating the weights. The algorithm recursive of the descending gradient is used, beginning with the output neurons and working back until arriving to the input layer:
   a. For the weights of the neurons of the output layer:

$$w_{kj}^{o}(t+1) = w_{kj}^{o}(t) + \Delta w_{kj}^{o}(t+1);$$

$$\Delta w_{kj}^{o}(t+1) = miu\delta_{pk}^{o} y_{pj} \quad (9)$$

   b. For the weights of the neurons of the hidden layer:

$$w_{ji}^{h}(t+1) = w_{ji}^{h}(t) + \Delta w_{ji}^{h}(t+1);$$

$$\Delta w_{ji}^{h}(t+1) = miu\delta_{pj}^{h} x_{pi} \quad (10)$$

7) The unemployment condition is completed (error is less or achieved reached number of iterations).

4. ALGORITHM

The Figure 2, shows the proposed system, which consists of three stages: the stage of the capture of the signal of voice of the phone channel, the stage of preprocessing of the signal and finally the stage of the speaker's verification using the characteristics extracted in the first two stages.

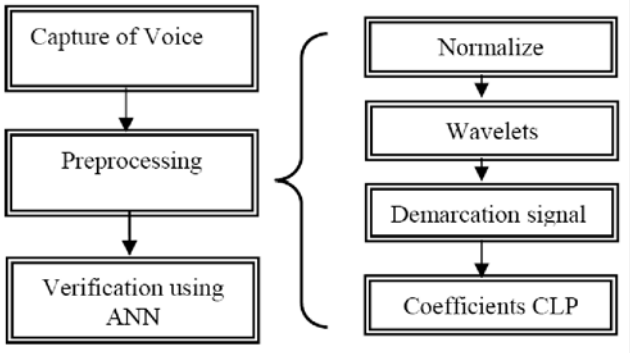

Fig. 2 Proposed system

In the Preprocessing of the voice signal, the objective is to condition the input signal so that this can be processed by the ANN, first we normalize the signal of audio between the values of [*-1,1*], later by means of the Wavelets, as it is seen in Figure 3, takes the "a" corresponding to the low frequencies from the voice signal where the biggest quantity in energy of the same is located, eliminating the "b" that corresponds to the high frequencies since it is where it finds the biggest quantity of noise of the signal. Obtaining, in this way, a signal of voice compacted and filtered with respect to the original. Then the voice signal is clipped at the beginning and end of the conversation, for, finally, to extract the coefficients of CLP of the signal, that will serve for the design of the training patterns of the ANN.

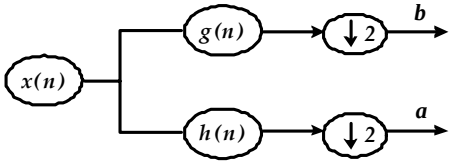

Fig. 3 Structure of the Wavelet

VOICE CAPTURE

For the realization of the capture of the voice we suggest to register 5 times word "Zoological" with five different people, each one of them recorded the same phrase with different emotional states. The speakers were Luis, Orlando, Alejandro, Diana and Leydi of 23, 29, 33, 5 and 22 years old, respectively.

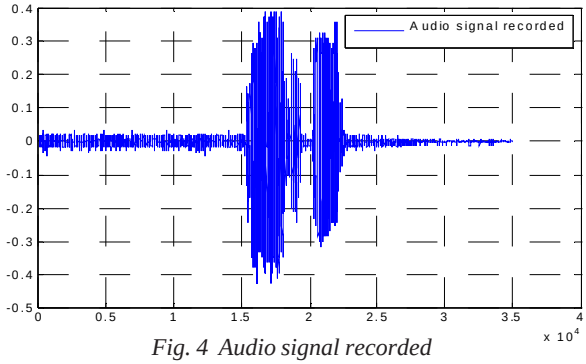

Fig. 4 Audio signal recorded

The word was chosen word "Zoological" because it contains most of the formants of the voice, great quantity of spectral characteristics of the same. The procedure was carried out in the following way: settled the Software Mercury in a PC Dell, was connected to the phone line via MODEM, then each one of the 5 speakers carried out a phone call to the number of the house where it was connected the PC to the phone line and via the software Mercury the pronounced word "Zoological" could be recorded in the PC from the phone house. This process was carried out 5 times for each one of the 5 speakers, obtaining in this way 25 files of audio that were turned format *.wav by their versatility

of handling with software Matlab, each one of the 25 files has the characteristics shown in Table 1.

Table 1 Characteristic of each voice file

SPEED OF TRANSMISSION	*128 KBPS*
Size of sound sample	*16 bits*
Channel Type	*Monophonic*
Speed of sound sample	*11 Khz.*
Format of audio	**.wav*

It is necessary to stand out that a speed of sample of sound of *11 Khz* was used, with the purpose of fulfilling the criterion of Nyquist that is bigger or equal to 2 times the sampling frequency, in this case the sampling frequency belongs to the phone channel that is approximately *4 Khz*.

PREPROCESSING

The stage of preprocessing of the voice signal consists of the following steps which are observed in the Figure 2.

Normalization

The normalization consists of adjusting all the parameters to a single scale so that to the moment of being used by the ANN do not cause problems of stability, in this case the used scale is given by the parameters of the activation function of the ANN, that is a tangent bipolar sigmoid and works with values of [*-1,1*], therefore each one of the 25 archives was normalized to this scale, as it is observed in Eq. (11), where the data that are wanted to normalize are within vector $x(i)$, with $i=1,\ldots n$. The procedure to continue is the following:

a) Compute the mean u and the standard deviation σ of the vector $x(i)$;
b) The data are normalized according to the relation:

$$xnor(i) = \frac{x(i) - \mu}{\sigma} \qquad (11)$$

c) The maximum and the minimum of the vectorial $xnor(i)$ are calculated, it is divided by that of greater absolute value and the normalized data fall inside the interval [*-1,1*].

The results are illustrated in Figure 5.

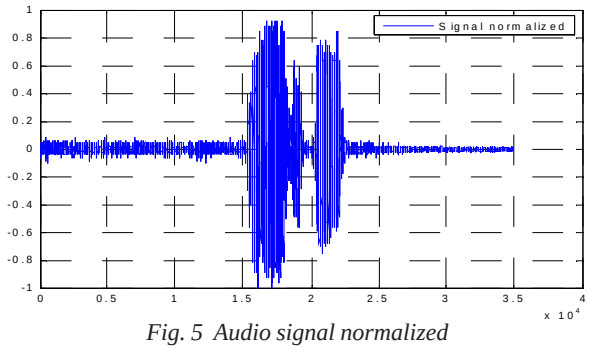

Fig. 5 Audio signal normalized

Wavelets

All signals of voice in nature are affected by noise, and the signal of voice of the phone channel is not the exception. For such reason the Wavelets are used to reduce this effect, therefore it suggests to use three types of Wavelets which are Haar, Coiflet and Daubechies, observing that the best of them is the Daubechies according to Table 2, due to the percentage of energy compression for each one of the twenty-five files is much bigger in all the cases with respect to the another two types of Wavelets, which turns out ideal to compact the energy of the audio signal present in the low frequencies "a" of the Wavelet and as well eliminating the noise stored in high frequencies "b" of the Wavelets.

Table 2 Energy compression with Wavelets

PERCENTAGE OF ENERGY COMPRESSION WITH WAVELETS			
Name files	*Haar*	*Coiflet*	*Daubechies*
Lb1	*80.27*	*89.45*	*202.35*
Lb2	*74.94*	*80.72*	*200.45*
Lb3	*83.13*	*93.21*	*212.00*
Lb4	*83.47*	*87.48*	*201.15*
Lb5	*80.37*	*87.45*	*198.56*
Aa1	*98.01*	*101.13*	*114.64*
Aa2	*98.76*	*105.47*	*122.73*
Aa3	*97.72*	*102.40*	*120.44*
Aa4	*98.07*	*101.88*	*116.63*
Aa5	*96.33*	*99.70*	*122.55*
Le1	*91.79*	*96.67*	*141.86*
Le2	*91.99*	*96.89*	*139.70*
Le3	*86.53*	*94.78*	*173.36*
Le4	*92.33*	*97.01*	*134.93*
Le5	*93.97*	*97.87*	*126.85*
Oc1	*95.84*	*98.71*	*117.98*
Oc2	*96.02*	*98.71*	*118.01*
Oc3	*96.31*	*98.78*	*116.50*
Oc4	*96.24*	*98.73*	*117.51*
Oc5	*96.31*	*98.78*	*116.50*
Dd1	*54.98*	*64.53*	*194.51*
Dd2	*61.79*	*76.92*	*272.13*
Dd3	*62.99*	*75.49*	*243.80*
Dd4	*84.90*	*94.76*	*210.59*
Dd5	*63.18*	*76.50*	*315.09*

Demarcation of the signal

Once all the audio signals have been compacted and filtered by means of the wavelets, we eliminate the time samples that alone contain "silences" different to each speaker's acoustic characteristics, in general these are at the beginning and end of the files of audio. The result of this process is exhibited in Figure 6.

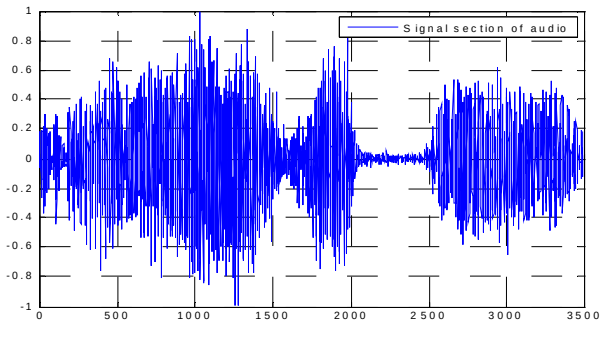

Fig. 6 Signal section of audio

Extraction of coefficients CLP

Due to the properties mentioned in the Section 2.1 of coefficients of CLP and in particular to that are able to model with great approach the linguistic information and the human vocal zone, in this work were used the coefficients a_i described in Eq. (2) forming this way a matrix of *25x25* elements that correspond to the extraction of 25 coefficients for each one of the 25 files by the speakers, creating in this way the training pattern of the ANN. The graph of the extraction of the coefficients of a speaker is observed in Figure 7.

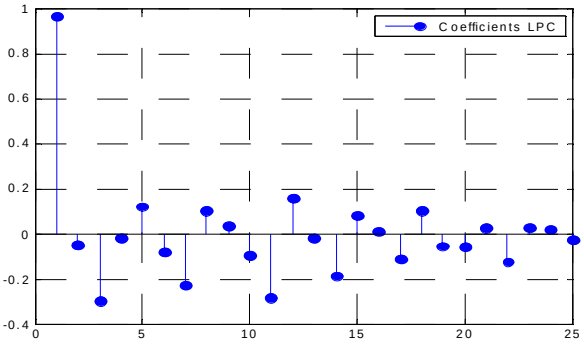

Fig. 7 Coefficients of CLP

DESIGN OF THE ARTIFICIAL NEURAL NETWORK

This stage consists of two parts, the first of them is the training of the ANN, which is carried out with the purpose of modifying the weights of the network in each one of the layers, so that the output wished by the user with the output obtained by the network before the presentation of a certain input pattern. The second part consists of a phase of validation of the network in front of any input pattern that is presented to it. An architecture Backpropagation was used with three layers, the input layer, hidden layer and the of output layer.

Phase of training

For the correct performance of this phase the following parameters were used:

1. Neurons of the input layer *=25*;
2. Neurons of the hidden layer *=21*;
3. Neurons of the output layer *=5*;
4. Number of trainings *=25*;
5. Number of epochs *=700*;
6. Weights of the input layer and the hidden layer. (Values within a rank of ([2.4 -2.4]) / Neurons of input), Ref. [5];
7. Training pattern;
8. Output Wished;
9. Average required quadratic error *=0.005*;
10. Learning-rate parameters *=0.009, 0.05, 0.02*.

Under these parameters and basing on the Section 3 where it is explained the operation of the ANN in details trained to the same, once trained the ANN with the number of proposed training is evaluated to generate and to keep the weights from the hidden layer and the output layer already trained to be used in the next stage.

Phase of evaluation

The kept weights obtained for the hidden layer and of output layer are opened of the training process, the points (1-4, 6 and 7) of the phase of training are defined, evaluates the network with a single training pattern which is the objective to identify inside our ANN, if the training pattern is the ANN it identifies it with one of the possible speakers employees in the training according to the characteristics of the values of its weights, but is not inside the speakers employees in the training of the network an error message it is emitted indicating that the person has not been able to be identified.

5. OBTAINED RESULTS

The Table 3 shows the results obtained in this investigation with that which we observe that our results are enough ideal because we obtain an effectiveness of the *100%*.

Continuing with our tests, to the moment of evaluate the ANN with the voice files without these have passed through the stage of the preprocessing the results obtained in effectiveness they diminish from the *100* to *96%*.

Graphing the variations of the learning-rate we obtain different values from error for the training process of the ANN, that are shown in Figure 8. From the graph we observed that the best values for the construction of the ANN are those of the heaviest line (alfa = *0.009, 0.05, 0.02*) since with them we obtain the minimum errors in the ANN.

Table 3 Results

TESTS WITH DIFFERENT EMOTIONAL STATES					
	Alejandro	*Leydi*	*Orlando*	*Diana*	*Luis*
Speakers	*Aa1= identified*	*Le1= identified*	*Oc1= identified*	*Dd1= identified*	*Lb1= identified*
	Aa2= identified	*Le2= identified*	*Oc2= identified*	*Dd2= identified*	*Lb2= identified*
	Aa3= identified	*Le3= identified*	*Oc3= identified*	*Dd3= identified*	*Lb3= identified*
	Aa4= identified	*Le4= identified*	*Oc4= identified*	*D4= identified*	*Lb4= identified*
	Aa5= identified	*Le5= identified*	*Oc5= identified*	*Dd5= identified*	*Lb5= identified*
Recog nition	*100%*	*100%*	*100%*	*100%*	*100%*
				Effectiveness =	*100%*

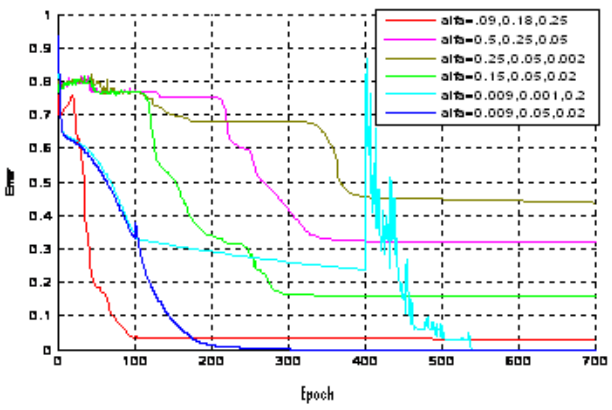

Fig. 8 Graph of the error

6. CONCLUSIONS

The system proposed in this work has a good operation since for its application we obtain recognition of the speakers of the 100% as it occurs to notice in Table 3 where the obtained results have been presented.

It is possible to emphasize that the procedure for the obtaining of the values of the parameters used in the design of the ANN Backpropagation does not exist as such very defined, however, the values proposed in this work were obtained on approval and error giving us account of how the learning-rate and the correct election of the initial weights influences a lot in the obtained result.

The proposed system presents a structure easy to develop and its mathematical complexity is minimum, reason which it can have diverse applications in the field of the identification and the speaker's verification.

7. REFERENCES

[1] R. José and H. Martinez, *Redes Neuronales Artificiales, Fundamentos, Modelos y Aplicaciones*, Alfa Omega, México, 2000.

[2] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, Cambridge University Press, 2001.

[3] S. Haykin, *Neural Networks*, Prentice Hall, New Jersey, 1999.

[4] L. Fausett, *Fundamentals of Neural Networks, Architectures, Algorithms and Applications*, Prentice Hall, New Jersey, 1995.

[5] B.M. del Rio and A.S. Molina, *Redes Neuronales y Sistemas Borrosos*, Ra-Ma, Madrid, 2001.

PREPOZNAVANJE GOVORA U TELEFONSKOM KANALU S NEURONSKOM MREŽOM POVRATNE PROPAGACIJE

SAŽETAK

*U ovom se radu preporuča algoritam za identifikaciju govora u telefonskom kanalu. Istraživanje je zasnovano na ponašanju umjetne neuralne mreže (ANN) povratne propagacije. Analizirani glas je zapisan u *.wav formatu. Predstavit će se procedura za identifikaciju govora u telefonskom kanalu koja je dala najbolje rezultate u nekoliko testova obrađenih našom metodom.*

***Ključne riječi**: umjetne neuralne mreže, impulsi, povratna propagacija, CLP, wav.*