World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:14, No:1, 2020

Speaker Recognition Using LIRA Neural Networks

Nestor A. Garcia Fragoso, Tetyana Baydyk, Ernst Kussul

Open Science Index, Electrical and Computer Engineering Vol:14, No:1, 2020 publications.waset.org/10010982/pdf

***Abstract*—**This article contains information from our investigation in the field of voice recognition. For this purpose, we created a voice database that contains different phrases in two languages, English and Spanish, for men and women. As a classifier, the LIRA (Limited Receptive Area) grayscale neural classifier was selected. The LIRA grayscale neural classifier was developed for image recognition tasks and demonstrated good results. Therefore, we decided to develop a recognition system using this classifier for voice recognition. From a specific set of speakers, we can recognize the speaker's voice. For this purpose, the system uses spectrograms of the voice signals as input to the system, extracts the characteristics and identifies the speaker. The results are described and analyzed in this article. The classifier can be used for speaker identification in security system or smart buildings for different types of intelligent devices.

***Keywords*—**Extreme learning, LIRA neural classifier, speaker identification, voice recognition.

I. Introduction

THE problem of speaker recognition is not new. For many decades this theme has provoked the interest of scientists and engineers. Different systems of voice recognition have been developed. The results are satisfactory, but until now it is possible to improve these processes. Our task is to develop a method for speaker recognition through voice spectrograms and the neural classifiers application. In the present study, this involves the acquisition and transformation of data, the extraction of characteristics, and obtaining and processing the results.

The first stage of this study was connected with the design and creation of a voice database of different speakers. The database is not limited to the acquisition of voice signals from different people. In addition, it must meet different acquisition conditions and conversion requirements for use in the system. Below, we present the conditions and characteristics of voice database generation:

- All voice signals must comply with the same acquisition conditions.
- All voice signals must have the same coding format (the WAV format).
- All recordings must have a phonetic content representative of the language.
- The sentences must be the same and be said by each speaker.
- The bandwidth of the recordings must be unified.
- The same filters and the same number of filters will be used to obtain the spectrograms for each voice signal.

Additional research that we needed to conduct was as follows:

- Investigate different methodologies used in speech recognition.
- Develop and use resonant filters to obtain frequency characteristics of voice signals. In this stage, the conversion of an audio signal to an image or spectrogram is done.
- Create different support programs for the administration, manipulation and preprocessing of data.

The diagram in Fig. 1 shows the procedures used to conduct this study. Therefore, different methodologies used in speech recognition were investigated. For the creation of a human voice database, voice sample acquisitions were obtained from different speakers. We developed and used the resonant audio filters for the extraction of voice signal characteristics. In this stage, the conversion of an audio signal to an image or spectrogram is performed. Different support programs for the data administration, manipulation and preprocessing were created. We developed a LIRA neural classifier that can extract and classify the voice characteristics of each speaker [1]. This LIRA neural classifier is one of the extreme learning algorithms [2], [3]. Different experiments permitted us to evaluate the efficiency of the classifier. We describe all results obtained with voice recognition.

In the first stage of this study, the voices of a set of people were recorded. The samples of the voices were obtained using digital audio recordings. All samples were obtained under the same recording conditions: a unified bandwidth, the same coding, and a silent environment to avoid other sources of sound. In addition, each speaker was assigned a unique identification key, and the sex of each speaker was recorded. In this way, the voice database was structured.

For the recordings, 15 different phrases were used, each one said by each speaker. These phrases use the entire phonetic set of the Spanish and English languages. The objective of using this type of phrase is to have as much information as possible about the phonetic changes of the voice of each person. Once the digital audio voice signals were obtained from each speaker, 64 resonant filters were designed for the extraction of the natural frequencies of the human voice. Using the frequencies selected by the filters, the speech spectrogram was constructed. The spectrogram of each phrase was transformed into a 64 × 1000 pixel image that served as input to the LIRA neural classifier. The LIRA grayscale neural classifier is a neural network specializing in image classification [1].

This work was partly supported by project UNAM-DGAPA-IT102320.

N.A. Garcia Fragoso is with the National Autonomous University of Mexico (UNAM), Mexico city, Mexico (e-mail: dooegato@live.com).

T. Baydyk is with the Instituto de Ciencias Aplicadas y Tecnología, UNAM, Mexico city, Mexico (corresponding author, phone: 52 55 56228602; e-mail: t.baydyk@icat.unam.mx).

E. Kussul is with the Instituto de Ciencias Aplicadas y Tecnología, UNAM, Mexico city, Mexico (e-mail: ernst.kussul@icat.unam.mx).

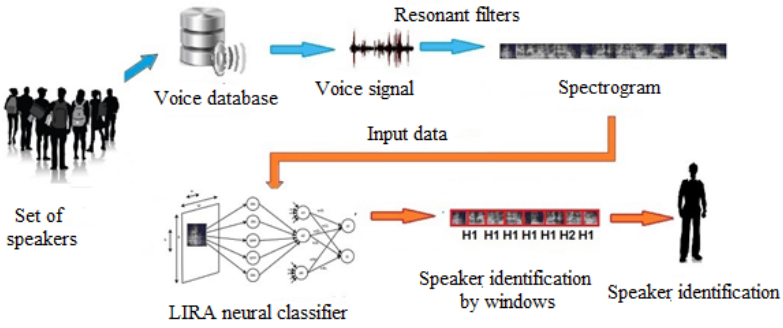


Fig. 1 Speaker recognition scheme

The system is implemented in MATLAB, which is a very powerful tool for the solution of numeric matrix algorithms.

In Section II we describe the LIRA neural classifier. Section III contains the voice recognition description. In Section IV we provide a database description and the procedure for its generating. We provide examples of phrases and the description of resonant filters that give us the spectrograms. We use these spectrograms as input for our recognition system. In Section V we describe different experiments with our system and the results. Section VI contains the conclusions.

II. LIRA Neural Classifier

A. LIRA Structure

The LIRA grayscale neural classifier consists of 4 layers: an input layer (*S*), an intermediate layer (*I*), an association layer (*A*) and an output layer (*R*). The input layer has pixels from windows randomly selected from the original image. The values of the pixels are within the range of [0 255]. The *S* layer is connected to the *A* layer by the *I* layer by means of neuron connections and neuron activation processes. Every pixel corresponds to the neuron and pixel brightness corresponds to neuron excitation.

The *I* layer is formed by ON neurons and OFF neurons, and each of these neurons has an associated excitation threshold that determines the activation of an *A* layer neuron. For ON neurons, the input value of the neuron must be greater than the excitation threshold to activate the neuron. For OFF neurons, the input value must be lower to excite the neuron. The neurons of the *A* layer will only be activated if all the ON neurons and the OFF neurons of the *I* layer of a window are activated.

Finally, all neurons in the *A* layer are connected to all neurons in the *R* layer and have a weight associated with them. For each neuron in the *R* layer, which also corresponds to a class, the weights of the active neurons in the *A* layer are added, and the highest value is selected as the winning class.

Before testing, the classifier must be trained. Hebb's rule is used for the training process.

B. LIRA Characteristics

The specifications of the LIRA neural classifier as implemented are as follows: 64000 neurons in the *A* layer, window sizes of 20 × 20 pixels, 3 ON neurons, 2 OFF neurons, and a spectrogram size of 64 × 1000 pixels. Each spectrogram is segmented into 64 × 100 pixel images with an overlap of 50%. We selected 15 spectrograms per speaker and from them 12 spectrograms for training and 3 for recognition.

III. Voice Recognition

The voice treatment process includes a wide range of functions and tasks [4]. The voice is a complex signal and is the result of several transformations occurring at different levels: semantic, linguistic, articulator and acoustic. The differences in these transformations appear as differences in the acoustic properties of the speech signal. Differences related to the speaker are the result of the anatomical combination inherent in the vocal tract and the learned language habits. For speaker recognition, all these differences can be used to discriminate between speakers. For almost a century, the ability of humans to recognize voices has been studied [5]. This task became even more ambitious after the development of digital computers and parallel processing. The goal of speaker recognition changed to automatic speaker recognition in a manner more similar to how people do it [6].

Automatic speaker recognition systems (ARS) are used to verify and identify a person. ARS allows automated control of voice services for tasks such as banking transactions or the control of confidential information.

In recent decades, there have been great advances in applications of speaker recognition for industry, laboratories and universities [5], [7]-[12].

IV. Digital Signal

In this section, we present the characteristics that define the audio, the characteristics of the human voice and the mechanism that makes possible the identification of people.

A digital audio file is a data structure that allows one to store acoustic information from the outside world or from a sound synthesizer. The quantity and quality of the information contained in these files depends on the coding standards used for their creation, which are the sampling frequency, quantization level, number of channels, filters, etc.

The spectrum is the information of the frequencies that a certain sound contains and their respective amplitudes. The spectrum is obtained by calculating the energy that contributes

Open Science Index, Electrical and Computer Engineering Vol:14, No:1, 2020 publications.waset.org/10010982/pdf

each frequency to the total sound [13], [14]. The representation of a sound spectrum can be accomplished using a table that relates the number of harmonics to the amplitude or using a graph that relates the time interval to the range of frequencies present in the sound. This graph is called a spectrogram and constitutes a fundamental tool in acoustic analysis.

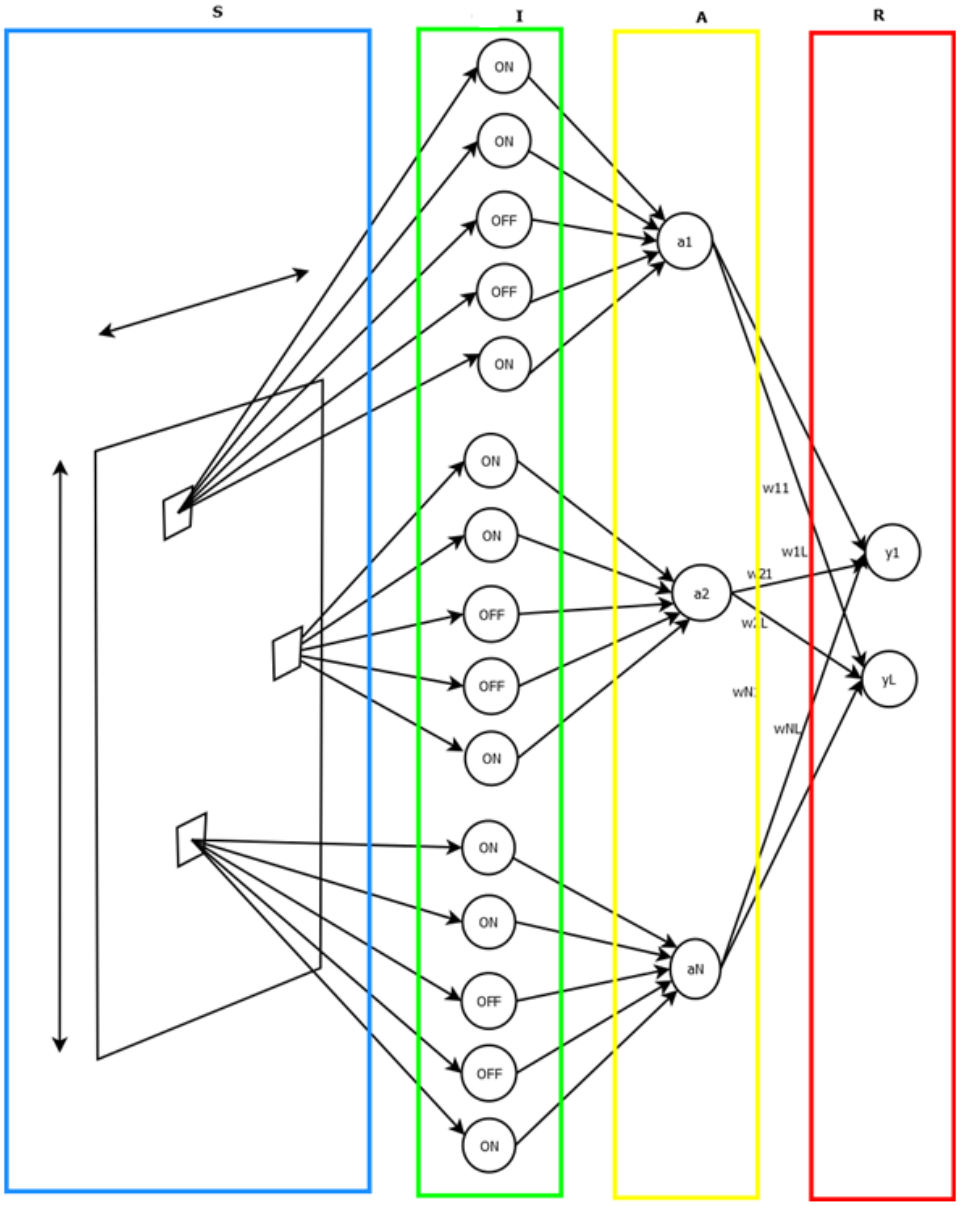


Fig. 2 LIRA neural classifier

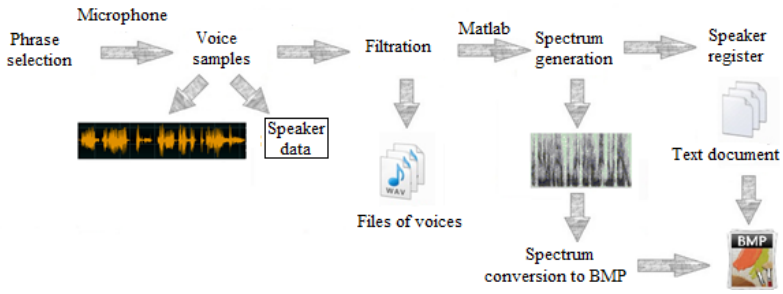


Fig. 3 Voice database creation

Open Science Index, Electrical and Computer Engineering Vol:14, No:1, 2020 publications.waset.org/10010982/pdf

Open Science Index, Electrical and Computer Engineering Vol:14, No:1, 2020 publications.waset.org/10010982/pdf

Jean Baptist Fourier was the first mathematician to study harmonics. He discovered that a signal, however complex, can be decomposed into an algebraic sum of harmonic sinusoidal signals obtained from an original signal [15].

A. Database of Voices

The database for speech recognition consists of a set of audio files in WAV format, its corresponding spectrogram as a BMP image and an identifier of the speaker to which it belongs. The database has been made following the process presented in Fig. 3.

B. Phrase Selection

The phrases constitute a list of 15 pangrams. Pangrams are designed phrases that contain all the letters of the alphabet at least once. In this way, it can be ensured that all the phonetic sounds that can be produced by the voice will be generated in a single statement.

The language was considered with the aim of generalizing the identification of a speaker regardless of the language that is spoken: The set of sentences consists of 10 statements in Spanish and 5 in English:

1. Whisky bueno: excitad mi frágil pequeña vejez.
2. El viejo señor Gómez pedía queso, kiwi y habas, pero le ha tocado un saxofón.
3. Aquel biógrafo se zampo un extraño sándwich de vodka y ajo.
4. El veloz murciélago hindú comía feliz cardillo y kiwi. La cigüeña tocaba el saxofón detrás del palenque de paja.
5. Hoy bajo su valor la wulfenita, extraño molibdato que se cotiza por kilogramo.
6. El pingüino Wenceslao hizo kilómetros bajo exhaustiva lluvia y frío, añoraba a su querido cachorro
7. Tengo un libro de papiroflexia sobre las hazañas y aventuras de Don Quijote de la Mancha en Kuwait.
8. Le gustaba cenar un exquisito sándwich de jamón con zumo de piña y vodka fría.
9. Queda gazpacho, fibra, látex, jamón, kiwi y viña.
10. Manchaba una y otra la pequeña hoja de fax con kiwi y grasa.

In English:

1. Sphinx of black quartz, judge my vow.
2. The five boxing wizards jump quickly.
3. Pack my box with five dozen liquor jugs.
4. A quick Brown fox jumps over the lazy dog.
5. Sexy zebras just prowl and vie for quick hot mating.

C. Voice Samples

For the process of obtaining voice samples, 20 speakers were used. A recording was obtained for each speaker reading each of the 15 sentences. This is a single session per speaker. The recordings were made using a conventional microphone.

The following considerations were made during the sampling:

- The recordings of all sentences per speaker were made in a single session.
- The recordings were made in a room isolated from noise and without reverberation.
- The WAV audio format was used to save the samples.
- The limited bandwidth for each recording was 200 Hz to 3500 Hz.
- Sampling frequency is 16000 Hz.
- Resolution is 16 bits.

The information obtained from the speakers contains numeric code that allows identifying the speaker, numerical code that allows identifying the recording, male or female, date of the session when the samples were obtained.

For this work a total of 13 speakers were used, of which 8 were men and 5 were women. Each speaker made a recording by phrase, giving a total of 195 recordings and 195 spectrograms.

D. Resonant Filter for Obtaining of the Spectrogram

Different methods are used for time-frequency decompositions, for example, Fourier Spectrum and Wavelet Packet Decomposition [15], [16].

An analog model of discrete resonant filters was used in this work. The following mathematical formula was used to construct the model.

$$\sin(t) = \sin(t + 3/\pi) + \sin(t - 3/\pi). \quad (1)$$

If we generalize (1) for a function $y(t)$ and it is evaluated in 0 and 1, we can simplify the mathematical equation and obtain a representation as a function $y(t)$.

$$y(t+1) = y(t) - y(t-1), \quad (2)$$

where $t = 1, 2, 3,...$ and all the numbers in this sequence will be part of a sine wave with a period equal to 2π (Fig. 4).

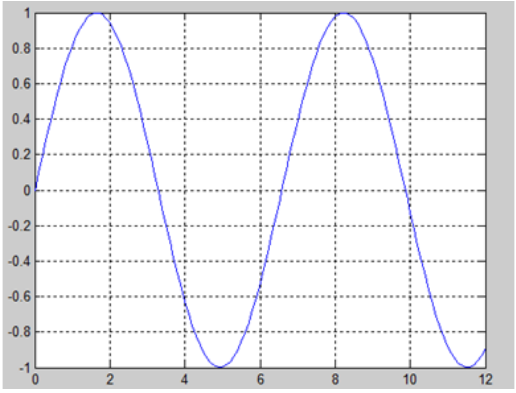


Fig. 4 Function $y(t)$ with period 2π

If we add a quality parameter λ to the function, and add the input function to the right of the expression, the filter is similar to (3):

$$y(t+1) = (y(t) - y(t-1))(1-\lambda) + x(t) \quad (3)$$

We applied the transformation Z to the analogical model (3). To obtain the digital model we use the following transformation function (4):

$$Y_Z = \left(Y - YZ^{-1}\right)(1-\lambda) + X \tag{4}$$

The resulting function is:

$$H = \frac{Y}{X} = \frac{1}{Z - \left(1 - Z^{-1}\right)(1-\lambda)} \tag{5}$$

The window works with 128 coefficients. To obtain these values we evaluate the transformation function with $Z = e^{\frac{\pi n}{128}}$, for n = 0, 1, 2, …, 127. Its spectral representation is similar to a resonant Gaussian filter (Fig. 5). The quality parameter directly affects the Gaussian acuity. Fig. 5 compares the spectrum for λ = 0, 0.5 and 1.

Many characteristics of the voice are hidden in high frequencies, but these frequencies have small amplitudes. In addition, they are important for the recognition of the person, because they are the main peculiarities for each person. To represent high frequencies, λ ≈ 1 is used and more time periods are necessary. The periods correspond to the development of resonance and depend on the quality parameter.

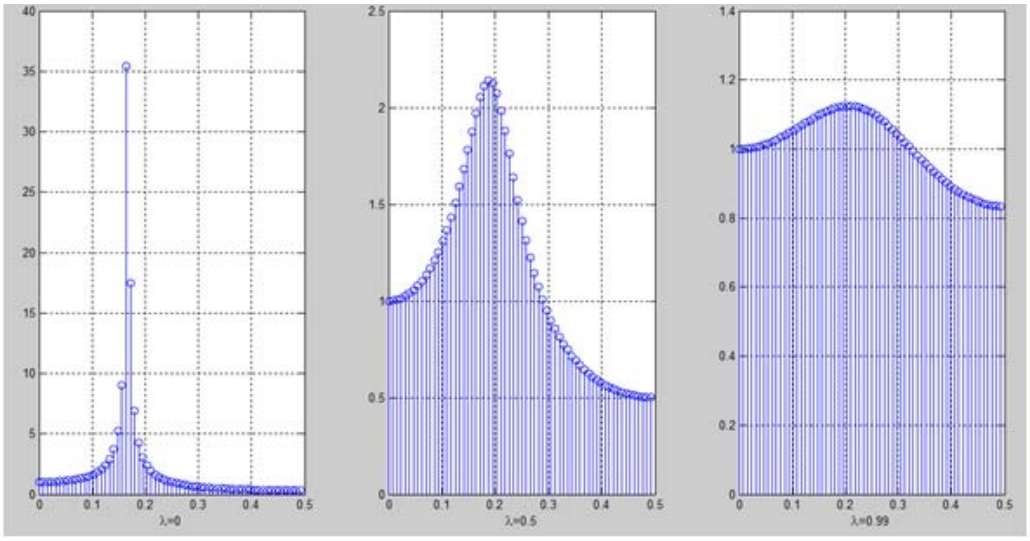
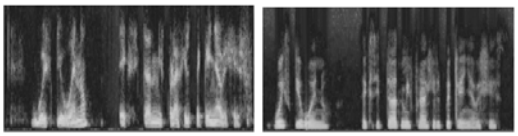


Fig. 5 Filters with different λ

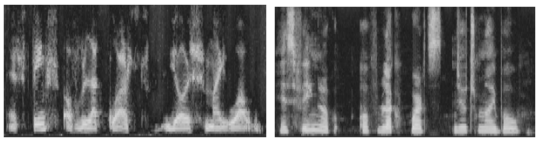
E. Spectrums

For the generation of images of the spectra a routine was developed in MATLAB. The characteristics of the spectrograms were formed using a 64-point comb filter distributed along the bandwidth of the audio recordings.



(a) (b)

Fig. 6 Spectra for a man (a) and a woman (b), phrase in Spanish



(a) (b)

Fig. 7 Spectra for a man (a) and a woman (b) for an English phrase

Finally, the image obtained was stored in a BMP file, grayscale. The image in Fig. 6 is the grayscale spectrogram in BMP format. In addition, the spectrograms for a man and a woman for sentences spoken in Spanish (Fig. 6) and English (Fig. 7) are compared.

The name of the spectrogram includes the identifier of the speaker and the identifier of the declaration. These in turn are related to the information for the speaker in a data file that links this information with its spectrograms.

In Fig. 8 we present one example of spectra of 15 phrases for one speaker.

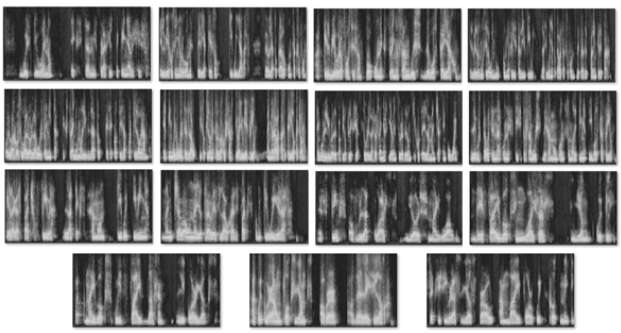


Fig. 8 Spectra of 15 phrases for one speaker

We use the LIRA grayscale neural classifier to recognize a

Open Science Index, Electrical and Computer Engineering Vol:14, No:1, 2020 publications.waset.org/10010982/pdf

speaker for each image. The spectrum has a length of 1000 pixels. For each of these we select in a progressive manner one area of spectrum of 100 pixels of a length equal to the height of the spectrum. We begin from pixel 1 for the first area. The movement can include one percentage of area intersection. Every area from all areas is an image for training of the LIRA grayscale neural classifier.

The process for the LIRA grayscale neural classifier is presented in Fig. 9. This diagram divides the entire system into 4 subprocesses: The first subprocess consists of the creation of initial auxiliary files that contain necessary information for the other subprocesses. The next subprocess is the coding, in which the characteristics of the images are extracted and stored in another auxiliary file. The training is an iterative process and uses the codes obtained from the previous process to modify the weight matrix; in addition, in this process a count of the number of errors per cycle is made which is used to evaluate the performance of the classifier. The last subprocess is the recognition, in which the classifier is evaluated with image codes that did not participate in the training process. In our experiments, 80% of the images were used to train the LIRA grayscale neural classifier, and 20% of the images were used for recognition.

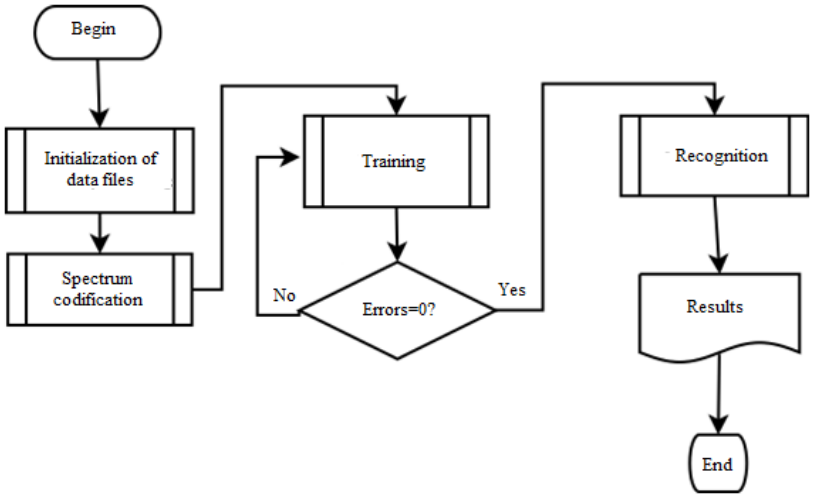


Fig. 9 LIRA work diagram

For the input layer, it is necessary to randomly select N image windows. For this task one only needs to specify the coordinates of a pixel. For each window one must choose 5 pixels, 3 that will serve as input for the neurons ON and 2 pixels for the neurons OFF.

For the intermediate layer I it is necessary to randomly choose the thresholds associated with each ON and OFF neuron. As already specified, N is the number of neurons in the associative layer and in our experiments $N = 64000$.

The coding process is conducted as follows:

1) The size of the overlap between images, the dimensions of the spectrogram image, the dimensions of the image extracted from the spectrogram and the size of the spectrogram are indicated. Window is 20 × 20 pixels.
2) The files with the auxiliary variables of the neural network are loaded into memory. These are the excitation thresholds, the negative and positive points of the I layer, and the positions of the windows.
3) The spectrum image is loaded into memory and a variable (m) is initialized to 0, which is used to scan the spectrogram.
4) The area of the spectrogram determined by the overlap and by m is selected.
5) An index (i) is initialized in 1, which will serve to identify the neuron of the A layer that has been activated.
6) For the first window determined by its position, the values of the positive (pp) and negative (pn) points, also determined by their position, are read. These coordinates have already been read from the archives.
7) An AND operation is applied, with the results of the comparisons of the negative and positive values with their respective thresholds. For negative points the comparison is "greater than" and for positive values the comparison is "less than".
8) If the result of the AND operation is 1, the index (i) is increased and stored in a variable denoted A.
9) If the result is 0, only the value of the index (i) is increased.
10) The process becomes iterative until the total number of neurons (N) is completed.
11) A header is added to variable A, which includes the speaker's identifier and the total number of activated associative neurons. A new row is added to the variable for the next image.
12) The m-index increases by 1 and the process is repeated until the entire spectrogram is scanned.
13) The spectrogram codes denoted by A are stored in a file.

This process is only for a spectrogram of one speaker. It must be repeated for all spectrograms of all speakers. That is, there are 15 spectrograms for 13 speakers.

The training method is an iterative process in which the weights of the connections between the output layer and the

Open Science Index, Electrical and Computer Engineering Vol:14, No:1, 2020 publications.waset.org/10010982/pdf

Open Science Index, Electrical and Computer Engineering Vol:14, No:1, 2020 publications.waset.org/10010982/pdf

associative layer are adapted to ensure that the image shown belongs to its real class. The learning procedure is extreme learning and modifies the weights.

Recognition is the final and fastest process. In it only the codes of a spectrum are presented, and the classifier determines which speaker it belongs to. A cycle can be incorporated to determine the classification of all the spectra from a single execution. The percentage of error is determined by the classified spectra that do not correspond to their original class.

V. Results

Four experiments were performed, 3 of which were performed once and the one that obtained the best results was performed 4 times, varying the training and recognition quality. The overlap of images is 50%, the image height is 500 pixels, the output class number is 13, and for every speaker we have 15 spectrum images.

A. First Experiment

The first experiment demonstrates the best results in the recognition stage. A percentage of errors of 0% was obtained during the training. The weight matrices obtained during the training process could be applied in the recognition stages.

This experiment was divided into 5 tests. In each test, the training and recognition spectra were selected in the manner shown in Table I. The number of spectra per speaker used in the training is shown, and the remaining spectra were used for training.

TABLE I
Tests

Test	Spectrum ID	Results (% of errors)
1	1,2,11	12.82
2	3,4,12	7.69
3	5,6,13	10.25
4	7,8,14	7.69
5	9,10,15	17.95

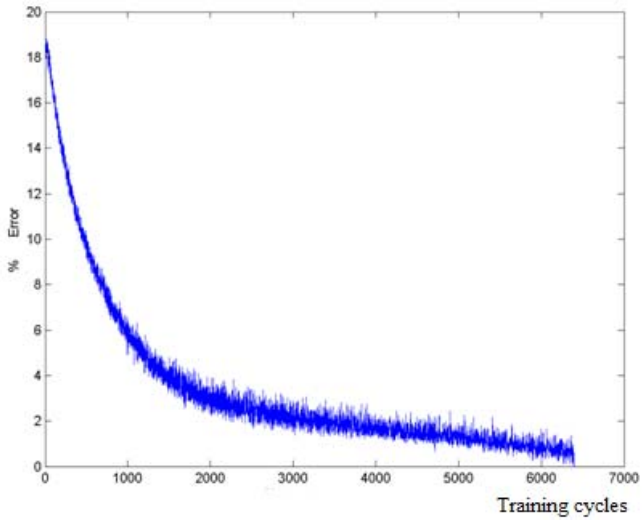


Fig. 10 Results of the first test

In Fig. 10 the error curve for the first test is shown. In the figure it is observed that the maximum error is approximately 18% and that the test converges to 0%. The training ends when no errors are obtained during a cycle, so the error curve vanishes at the last training cycle. This curve represents the average number of errors throughout the training. It is a typical experiment. Four other experiments demonstrate the same behavior.

Fig. 11 shows the percentage of error accumulated in the training cycles as a smoothed curve.

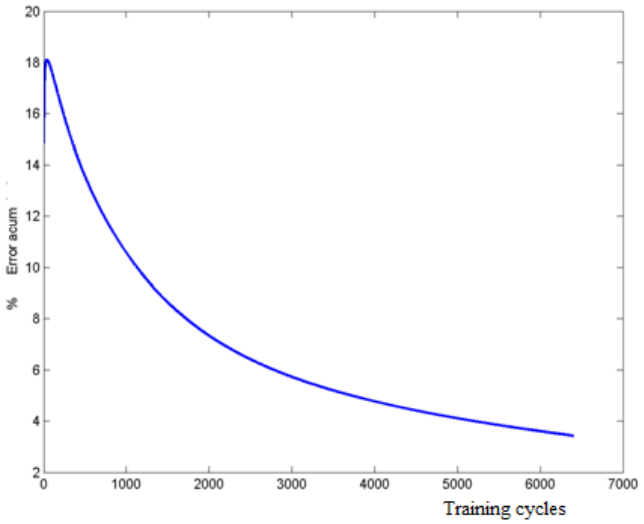


Fig. 11 Errors accumulated in test 1

In each of these tests, the recognition stage was performed. The results are presented in Table I. The selections of the tests 2 and 4 have the best percentage of recognition, having 4 errors out of 39 samples evaluated.

B. The Second Experiment

To improve the efficiency of the classifier, the overlap of images was increased to obtain more information during the training. However, with this modification, the performance of the classifier was decreased because having a greater number of images to process linearly increased the time during coding, training and recognition with the number of processed images.

Fig. 12 shows the error curve during training. It is observed that it was not possible to obtain 0 errors even by increasing the number of cycles. The highest percentage was approximately 11%, and the same parameters were taken from the previous experiment, changing only the overlap between images. Fig. 13 shows the cumulative error curve for all cycles performed. The parameters used for the implementation are identical to those in the previous experiment with the following modifications: a) Overlap: 75% b) Height of the image: 64 pixels.

Unfortunately, many of the errors from the recognition stage were not equal to zero and the results were not better than previous ones.

Open Science Index, Electrical and Computer Engineering Vol:14, No:1, 2020 publications.waset.org/10010982/pdf

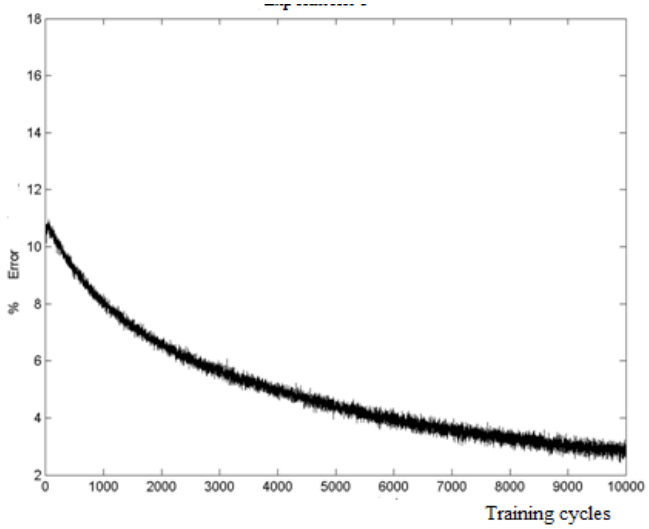


Fig. 12 Results of experiment 2

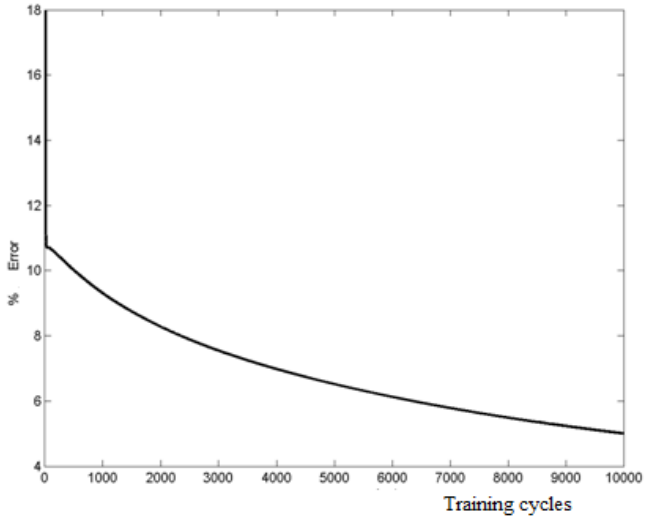


Fig. 13 Errors accumulated in experiment 2

C. The Third Experiment

For this experiment, a spectrum of 500 pixels in height and an overlap of 75% was used.

The results were not better than previous results, but with this finding the idea that the efficiency of the classifier could be improved by increasing the overlap was ruled out. The error curve during training and the cumulative error curve are shown in Figs. 14 and 15, respectively.

The recognition percentage was better, but it did not exceed that obtained in the tests of experiment 2. The recognition was 25.64%. In Table II we present the classifier results. In addition, an extra experiment was performed, with the same phrases used during the training but spoken 3 months after those used for the training. In this case, only two speakers and 12 sentences per speaker were used, and 100% recognition was obtained using the weight matrix of experiment 1.

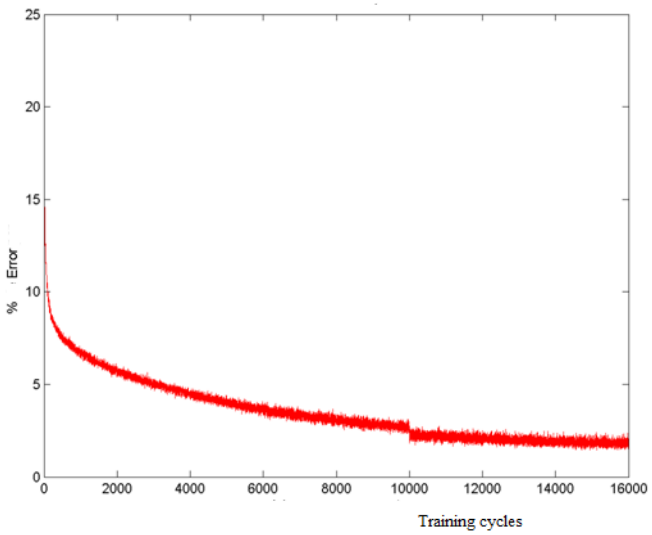


Fig. 14 Results of experiment 3

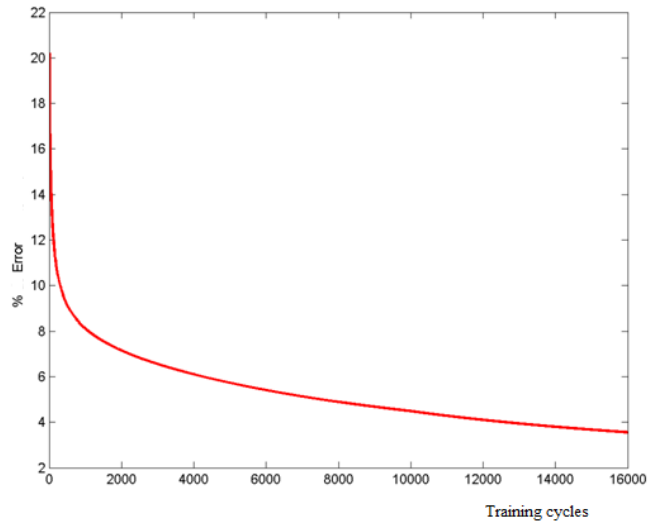


Fig. 15 Errors accumulated in experiment 3

TABLE II
CLASSIFIER RESULTS

Experiment	Height	Overlap	Errors
1	500	50	7.69%
2	64	25	28.2%
3	500	25	25.64%

A user interface was created to have a usable application. Fig. 16 contains the box of the user interface. It consists of the following parts: a listbox to select the spectrum to be analyzed; a button to start the recognition process: an image in which the spectrum that is being evaluated is shown; a spectrum window that will be recognized individually; a label that shows the speaker of the individual window; a label that shows the speaker obtained from the fashion of the recognized speakers in each window. This is the final result and it contains the identified speaker.

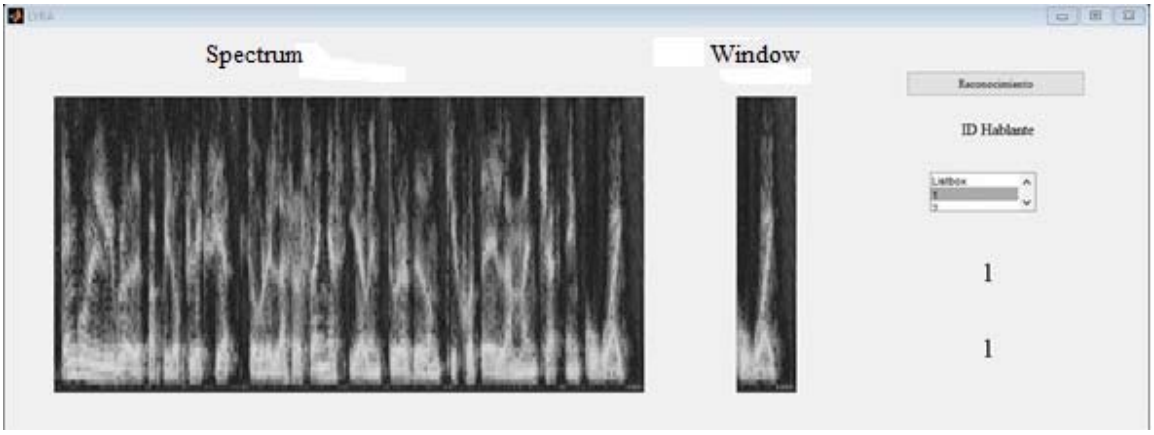


Fig. 16 Interface

Open Science Index, Electrical and Computer Engineering Vol:14, No:1, 2020 publications.waset.org/10010982/pdf

VI. Conclusions

A base of voices was prepared to solve the speaker recognition task. From the process of selection of the phrases said by each speaker, it was essential to use those phrases that contained phonetic information to distinguish the different speakers. The voices of a total of 13 people were recorded, and each person spoke 15 sentences, of which 10 were in Spanish and 5 were in English. The participants were men and women between 20 and 40 years of age. The phrases obtained were saved in the standard WAV format, which is the format used in voice signal processing systems. All the voice samples were filtered to obtain their corresponding spectrograms, and these spectrograms were used as input images for the neuronal classifier. Subsequently, the spectrograms were transformed into images in the BMP format. This format ensures that there is no loss of information due to compression.

The neural classifier LIRA was used to distinguish speakers. Each spectrogram image was scanned with different window sizes. The different stages of the classifier work were presented as three parts: coding, training and recognition. It was also possible to measure the processor time consumed in each stage: coding, training and recognition require 20%, 78% and 2% of the processor time, respectively. Good results were obtained in the training processes, for example after 700 cycles 0 errors were obtained. In the recognition process, 7.69% of identification errors were obtained as the best result.

The choice of the window was the most complicated task. Very short segments increase the coding and training time considerably and worsen the results during the recognition. Finally, for the recognition, the identification of the speaker had no errors, showing that the algorithm is adaptable for recognition tasks.

Although few test subjects were used as speakers, it is possible to add new elements. Only the training process with these elements must be executed again, the recognition process is not altered, and time increases linearly with the number of speakers. The recognition can be tested for sentences other than those used in the experiments, which only requires coding the new spectrum. The voice recognition is one of the important tasks of biometric investigation. We propose to use for this purpose neural classifiers with extreme learning.

References

[1] E. Kussul, T. Baidyk, D. Wunsch, *Neural Networks and Micromechanics*, New York: Springer-Verlag, 2010.

[2] O. Makeyev, E. Sazonov, T. Baidyk, A. Martin, "Limited receptibe area neural classifier for texture recognition of mechanically treated metal surfaces," *Neurocomputing*, vol. 71, no 7-9, pp. 1413-1421, March 2008.

[3] O. Makeyev, E. Sazonov, S. Schuckers, P. Lopez, T. Baidyk, E. Melanson, M. Neuman, "Recognition of swallowing sounds using time frequency decomposition and limited receptive area neural classifier, " *in Proc. Of AI-2008, The twenty-eight SGAI Intern Conf on Innovetive Techniques and Applications of Artificial Intelligence*, Eds. Tont Allen, Richard Ellis and Miltos Petridis, Springer, Cambridge, UK, December, pp. 33-46, 2008.

[4] J. P. Campbell, "Speaker recognition: a tutorial," *in Proc. IEEE*, vol. 85, no 9, pp. 1437-1462, 1997.

[5] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun*., vol. 17, pp. 91–108, 1995.

[6] D. A. Reynolds, "A Gaussian mixture modeling approach to text independent speaker identification," Ph.D. Thesis, Georgia Institute of Technology, 1992.

[7] A. L. Higgins and W R. E. Ohlford, "A new method of text independent speaker recognition," *in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Tokyo, Japan, pp. 869–872, 1986.

[8] N. Z. Tishby, "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, no 3, pp. 563–570, 1991.

[9] D. Reynolds and B. Carlson, "Text-dependent speaker verification using decoupled and integrated speaker and speech recognizers," in *Proc. EUROSPEECH*, pp. 647–650, 1995.

[10] C. Che and Q. Lin, "Speaker recognition using HMM with experiments on the YOHO database," *in Proc. EUROSPEECH*, pp. 625–628, 1995.

[11] J. Colombi, D. Ruck, S. Rogers, M. Oxley, and T. Anderson, "Cohort selection and word grammer effects for speaker recognition," *in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 85–88, 1996.

[12] D. A. Reynolds, "M.I.T. Lincoln Laboratory site presentation," *Speaker Recognition Workshop*, A. Martin, Ed., 1996.

[13] Instituto Politécnico de Madrid, Escuela Universitaria Técnica de Telecomunicación, *Manual Técnico de sonido***.** Recuperado el 08 Julio del 2013 http://www.diac.upm.es/escuela, 2000.

[14] F. Miyara, *Acústica y Sistemas de sonido*, Argentina: UNR, 1999.

[15] O. Makeyev, "Automatic method of acoustical swalling detection for monitoring of ingestive behavior," Ph.D Thesis, Clarkson University, Potsdam, NY, USA, April 2010.

[16] E. Sazonov, O. Makeyev, S. Schuckers, P. Meyer, E. Melanson, M. R Neuman, "Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior," *IEEE Trans on Biomed Eng,* vol 57, no 3, pp. 626-633, 2010.