SPEAKER RECOGNITION WITH ARTIFICIAL NEURAL NETWORKS

Breno Martins da Costa Corrêa e Souza

Rondinelli Leonardo Jorge

Paulo Eduardo Maciel de Almeida

breno.ec@gmail.com

rondyleonardo@gmail.com

pema@lsi.cefetmg.br

Centro Federal de Educação Tecnológica de Minas Gerais

Av. Amazonas, 7675 - Nova Gameleira, 30510-000, Minas Gerais, Belo Horizonte, Brazil

Abstract. *This paper presents a brief discussion on the speaker recognition task, through short term audio inputs, by means of artificial neural networks. The input signal is treated with the purpose of removing non-pertinent data, as well as reducing the complexity of the neural network that will make the classification. The input of the neural network is the spectrogram of the input signal, acquired through the Fourier transform applied on the treated audio. The studied network maps nonlinear and discontinuous subspaces of the possible input values and the network development, its performance and possible applicability in real scenarios are evaluated. The training time of the analyzed network is of the order of 10 seconds, the classification of all the samples occurs in 1 second. The evaluated neural network correctly identified the speaker of all samples, what make us conclude that the proposed approach is feasible and has a good potential to solve large scale versions of this problem.*

Keywords: *Speaker Recognition, Classification, Artificial Neural Networks*

1 INTRODUCTION

Automatic speaker recognition is a field of technological study that, since the emergence of machines, has always been part of human aspirations. This is due to the benefits inherent in systems capable of succeeding in this task, such as having a machine identifying its owner through speech, without the need for coded inputs or specific languages that are not user friendly.

Artificial neural networks (ANNs) are allied in this sense, since through them it is possible to implement speaker recognition models. There are some records in the literature on applications for speech recognition techniques, such as the assistance to the disabled (Neto et al., 2010) (Thiang, 2007), applications in forensic practice (Cardoso, 2001) and environment automation (Amaral et al., 2004).

On the other hand, there is no great diversity in the area literature, since the systems that allow a more natural communication between man and machine are still not completely dominated (Neto et al., 2010).

This work seeks to demonstrate, through a practical approach, the application of speaker recognition by means of an artificial neural network. Such a network must be capable of recognizing the voice of a certain speaker among a set of different speakers.

The rest of the paper is organized as follows, with methodology, results, experimental analisys and conclusion. Section 3 explains in details the methodology used in this work, with emphasis on the acquisition and manipulation of input data. It exposes how audios are used as input to the projected network. The architecture of the network is also briefly discussed.

2 LITERATURE REVIEW

There is several references in the literature about speech recognition. Many approach are related to processes of recognition of isolated words, as suggested by Gonzalez-Cadenillas and Murrugarra-Llerena (2013), when they used an evaluation of different isolated words recognition techniques on an embedded system using a microcontroller.

Others references using ensemble classifiers that theoretically and experimentally analyze the performance of a weighted plurality voting combination strategy to combine the decisions of multiple classifiers. Where theoretical expressions characterizing the performance of the weighted voting model are derived and the method is applied to the problem of human voice recognition getting results that show the advantage of employing weighted voting based ensemble classifiers in achieving high identification rates (Mu et al. 2009).

Still there are authors who are dedicated to answer questions like which techniques of speech recognition can yield better results, or which ones are the most efficient way of implementing such a system (Murphy, 2014).

In addition to several other approaches to voice recognition, however, the approaches studied did not reveal results as expressive as those achieved in this work. They were obtained by applying the Fourier transform for speech recognition. This technique was chosen for presenting good results for other recognition of patterns, such as the recognition of the human iris, presented by Al-Alaf et al. (2013).

As the results seemed momentarily satisfactory, this work could be relevant to the area of human voice recognition, due to the infinity of applications that exist today.

3 METHODOLOGY

Ten people (Table 1) were selected to make audio recordings to provide data with the purpose of studying and to propose an artificial neural classification network to identify speakers. Once the group of people is characterized, the identity of each speaker is not important. They will, therefore, be numbered from 1 to 10.

Table 1: Characterization of the selected people according to sex and age.

Number	Sex	Age
1	M	29
2	M	26
3	F	22
4	M	23
5	F	29
6	M	34
7	F	25
8	M	62
9	M	28
10	F	34

A set of 10 short sentences were elaborated for each speaker to say (Table 2), totaling 100 samples: 10 individuals and 10 sentences. Each sentence contains 8 syllables and the set sought to explore different phonemes of the Portuguese Language.

Table 2: Set of 10 phrases recorded by each individual. The recordings were normalized with attenuation of 1 Db. The recordings were edited to present 2 seconds of duration, each.

Number	Phrases in Portuguese	Phrase translated to English
1	Hoje ela mexe muito	Today she moves a lot
2	Feliz aniversário	Happy Birthday
3	Minha mãe gosta de uva	My mom likes grape
4	Suco de limão com pera	Lemon juice with pear
5	Minha axila é limpa	My armpit is clean.
6	Eu tomo chá de gengibre	I have ginger tea
7	O treino da rede neural	The neural network training
8	Envelheço na cidade	I get old in the city
9	Isso é engenharia	This is engineering
10	Base de frases gravadas	Record Phrase Base

In order to improve the generalization capacity of the projected network and to reflect the impossibility of controlling the environment in which the input audio is recorded in real applications, 5 noises are added to the samples (Table 3).

Table 3: Set of 5 noises added to recordings. Duration of 2 seconds, each. Noise signals were normalized with attenuation at 20 Db — about 10 times less intense than the recordings.

Numer	Noise
1	Traffic
2	Office
3	Building site
4	Rain
5	Wind

The sample base contains 600 elements (equation 1); therefore:

$$m = s + r \times s, \tag{1}$$

where $s = 100$ is the number of samples without addition of noise, $r = 5$ is the number of noise base elements and $m = 600$ is the total number of samples. The final base contains 100 samples with no added noise and 500 samples with added noise: a total of 600 samples.

It is desirable that the speaker recognition process does not depend on the text being read, as far as possible, on the speaking time. In order to remove the temporal aspects and extract part of the vocal information, we apply Fourier transformers to the audio signals. The algorithm used was the Fast Fourier Transform (FFT), a discrete and efficient variant of the Fourier transform.

The chosen sampling rate was 44100 Hz. This arbitrary choice is based on the fact that this sampling rate is the most common and accessible, popularized because it is the sample rate used on audio CDs.

The sampling rate of the signal informs us as to the number of points sampled per second. The resulting spectrogram of the Fourier transform application contains information pertaining to half of the signal sampling rate — Nyquist sampling rate (Oran Brigham, 1988). When we apply FFT to a sampled audio sample at 44100 Hz (Figure 1.a), the resulting spectrogram (Figure 2.a) contains data for frequencies up to 22050 Hz, including.

This spectrogram (Figure 2.a) allows us to evaluate that most of the vocal information is concentrated at low frequencies. To simplify the model contemplated by the projected network and to remove noise from the incoming audio signal, we developed a Chebshev Type II low-pass filter with cutoff frequency of 1024 Hz and attenuation of 60 Db — the signal is attenuated by a divisor factor of 1000. The resulting filtered audio signal and spectrogram are shown in the Figures 1.b and 2.b, respectively.

Since the filtered signal has little or no intensity at frequencies above 1024 Hz and preserved a considerable part of the vocal information, it has been re-sampled at 2048 Hz. The resampled filtered audio signal and spectrogram are shown in the Figures 1.c and 2.c, respectively.

The value of the cutoff frequency for the low-pass filter was not at all chosen arbitrarily; neither the duration of the audio: the FFT application presents an efficient computational cost

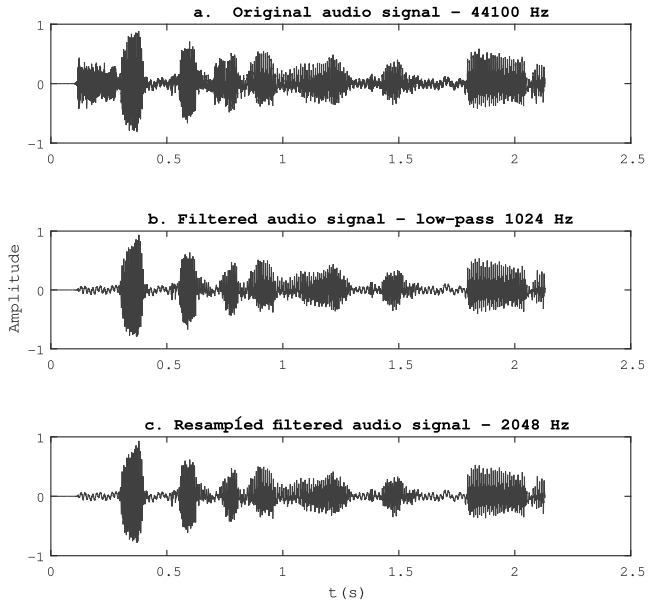

Figure 1: Manipulations performed on an audio signal.

a. Original signal with sampling rate of 44100 Hz.

b. Filtered signal through Chebshev Type II low-pass [1024; 1152] Hz, with attenuation of 1 Db for frequency lower than 1024 Hz and attenuation of 60 Db — attenuation by a factor of divisor of 1000 — for frequencies above 1152 Hz.

c. Resampled filtered audio signal to 2048 Hz.

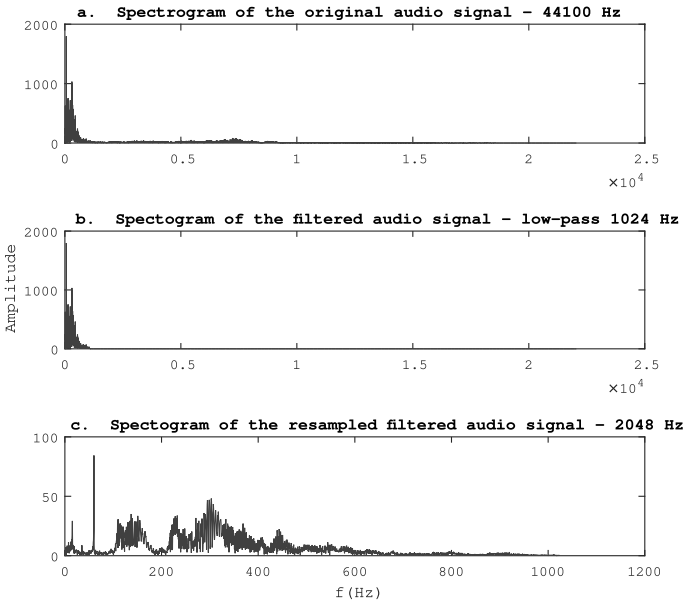

Figure 2: Spectrograms of signals a, b and c from Figure 1, respectively.

for a data sequence whose total size is the result of a power of 2 (Oran Brigham, 1988).

The result of the FFT applied to the filtered and resampled signal is used as input of the presented network. Since it has complex numbers, we chose to use the absolute value of each point, corresponding to its absolute value or magnitude:

$$|a + bi| = \sqrt{a^2 + b^2}. \tag{2}$$

The final size of the input sequence for each sample of the recording base is given by:

$$n = 2048 \times t, \tag{3}$$

where $t = 2$ seconds is the duration of each sample, 2048 is the sampling rate of the audio signal, and $n = 4096$ is the size of the input vector.

The proposed network is a classification artificial neural network, its has an input layer with 4096 points, a hidden layer with 64 neurons, and output layer with 10 neurons — an output neuron for each individual (class). It has a non-linear activation function in the hidden layer and uses the *softmax* activation function. The scaled conjugate gradient as used as the training function.

The input matrix $X_{m \times n}$ has 600 samples, each with 4096 points, of which 360 samples were used for training, 120 for validation and 120 for tests. The sample partitioning occurred randomly. Each sample is associated with its corresponding classification: a binary vector with 10 positions, referring to the 10 individuals — considered classes — where the value 1 in the ith position means that the sample belongs to the ith class; the value 0, in an analogous way, means that the sample does not belong to the ith class. Each sample belongs to a single class, of course. These vectors associated with the m samples form the target matrix $T_{m \times c}$, where $c = 10$ is the number of classes.

Section 4 presents the results obtained in the experiments by the proposed network.

4 RESULTS

The training of the classification network occurred in time in the order of 10 seconds. The training was interrupted by reaching the lower limit of the gradient (in multiple runs), since it is a stop criteria. The neural network correctly classified all the samples (Figure 3) in a time interval of the order of 1 second. There is no association between target class of Figure 3 and person of Table 1.

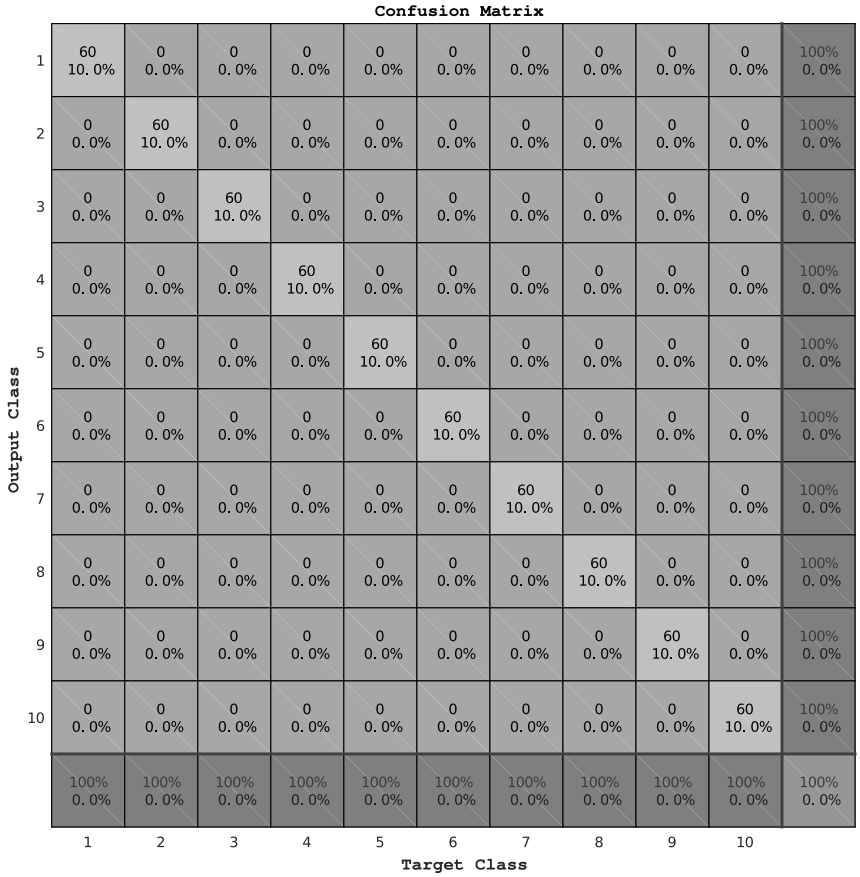

Figure 3: Confusion matrix. Each class has 60 matching samples.

Performance and training gradient decay rapidly (Figures 4 and 5), but, with the lower threshold of the gradient reached, the training is interrupted in order to avoid overfitting. The error histogram (Figure 6) and the correct classification of the samples (Figure 3) indicate that the application of a classification network was adequate to solve the speaker recognition problem.

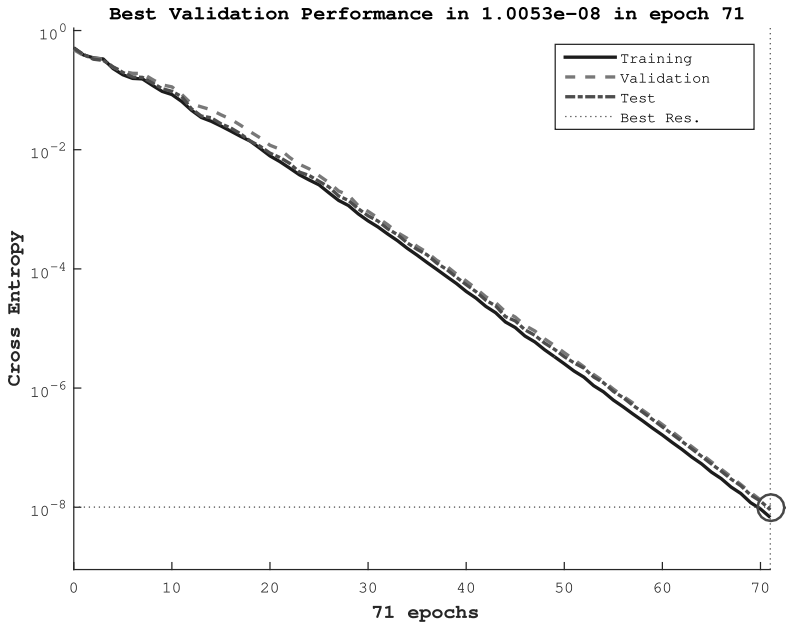

Figure 4: Evolution of cross-entropy by training epochs. Despite the approximately linear decrease, training was stopped in order to avoid overfitting.

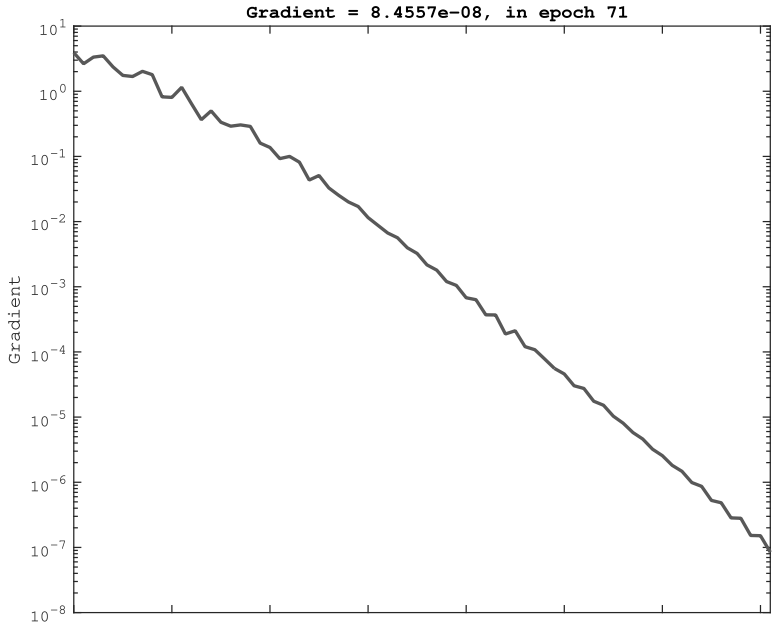

Figure 5: Evolution of the gradient by training epochs. Despite logarithmic decay, training was stopped in order to avoid overfitting.

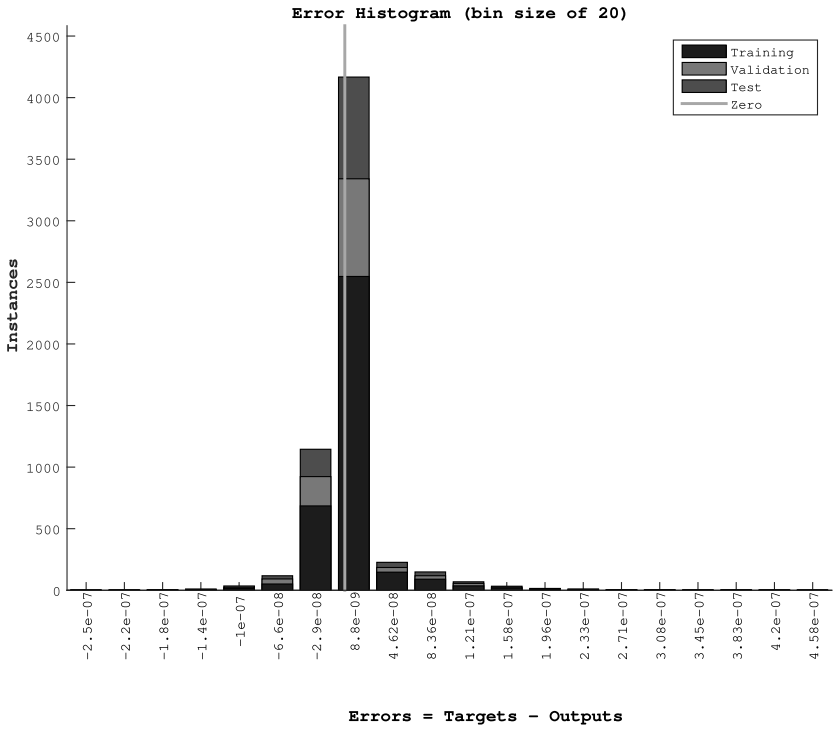

Figure 6: Error histogram (bins of 20).

5 EXPERIMENTAL ANALISYS

The present work was carried out with a reduced time frame. As a result, some factors threaten the universality of the results obtained. Speakers do not represent extensive coverage of vocal possibilities and the number of classes is relatively small; for certain applications.

The attenuation of 20 Db in the noise signals was arbitrary and not supported by studies of any nature. The reduction of the sampling rate from 44100 Hz to 2048 Hz has similar weakness. A sample rate of 8000 Hz is sufficient, however, since we have interests in the frequencies emitted by the human voice. In order to make it possible to classify an input, it is necessary to resample it to 2048 Hz. This work did not evaluate the computational impact that this entails.

6 CONCLUSIONS

Artificial neural networks are often used in cases where there is little information about a given problem and there is no way to establish an analytical solution for it.

The The proposed approach, based on an MLP artificial neural network using one hidden layer, presented training time feasible for real applications (in the order of 10 seconds) and presented equally feasible performance (600 samples classified in time of the order of 1 second), which could be implemented in embedded systems and real time. All samples were correctly classified by the proposed network, what make us conclude that the proposed approach is feasible and has a good potential to solve large scale versions of this problem.

7 AKNOWLEDGMENTS

We would like to thank the CEFET-MG for its support by grant.

REFERENCES

Al-Alaf, O. N. A., Abdalkader, S. A., & Tamimi, A. A., 2013. Pattern recognition neural network for improving the performance of Iris recognition system. In *Journal of Scientific & Engineering Research*, vol. 4, n. 6, pp 661-667.

Amaral, M. A., Barriviera, R., & Teixeira, E. C., 2004. Reconhecimento de voz para automação residencial baseado em agentes inteligentes. *Revista Eletrônica de Sistemas de Informação*, vol. 3, n. 1.

Oran Brigham, E., 1988. The Fast Fourier Transform and Its Applications. *UK: Prentice Hall*,pp. 83-87, 131-136, 148-156.

Gonzalez-Cadenillas, C., & Murrugarra-Llerena, N., 2013. Isolated words recognition using a low cost microcontroller. In: *Computing Systems Engineering (SBESC)*, III Brazilian Symposium on. IEEE, pp. 77-82.

Cardoso, S. E., 2001. *A inteligência artificial no judiciário*. Dissertação, Universidade de Florianópolis/ Santa Catarina.

Mu, X., Lu, J., Watta, P., & Hassoun, M. H., 2009. Weighted voting-based ensemble classifiers with application to human face recognition and voice recognition. In: *Neural Networks, International Joint Conference on IEEE*, pp. 2168-2171.

Murphy, A., 2014. *Implementing Speech Recognition with Artificial Neural Networks*. Tese de Doutorado. Algoma University.

Neto, J. A., Castro, M. A., & Felix, L. B., 2010. Reconhecimento de comandos de voz para o acionamento de cadeira de rodas. *Anais do XVIII Congresso Brasileiro de Automática*, pp. 3819-3824.

Thiang, D. W., 2007. Implementation of speech recognition on MCS51 microcontroller for controlling wheelchair. *Intelligent and Advanced Systems*, pp. 1193-1198.