SPEECH RECOGNITION WITH INDUSTRIAL PURPOSES: AN APPROACH USING INTELLIGENT SYSTEMS

Germano Lambert-Torres
Helga G. Martins
germanoltorres@gmail.com
helgagonzaga@gmail.com
Ciro R. Santos
Rômulo A. Carminati
Wagner S. Vieira
ciro_unifei@yahoo.com.br
carminati_unifei@gmail.com
vieira_unifei@gmail.com
Instituto de Sistemas Elétricos e Energia, Federal University at Itajubá (UNIFEI)

Abstract. This paper presents an implementation of Artificial Neural Networks – ANN to recognize voice commands that do not depend on the speaker. From a database having the commands pronounced by different speakers, techniques on voice signal processing were applied in order to have these signals treated within the techniques of Artificial Intelligence, particularly a Neural Network. The definition of which ANN training algorithm to use depends on some factors, such as the complexity of the problem, amount of training data, and precision and accuracy of problems. Multilayer Perceptrons – MLP is a model of Multilayer Neural Network which guarantees good results. Neural Network training uses a powerful voice analysis technique, which is Linear Predictive Coding – LPC. The developed project was considerably empiric, demanding practical tests, graphic and numerical analysis so that the final goal was achieved. The speech recognition, also known as automatic speech recognition or computer speech recognition, converts spoken words into machine-readable input. Speech recognition has great applicability on industry since they can be used to accelerate and optimize trainings and process operation. Beyond that, it can be a useful tool to ensure that people with special needs have more successful job opportunities. Keywords: Speech Recognition, Industrial Processes, Artificial Intelligence, Neural Network, Linear Predictive Coding.

1. INTRODUCTION

Since human civilizations began to use speech as communication, the relationship among people began to grow, and progress has become faster. This is an example that speech improves communication and makes decision making easier. Therefore, with the advent of fast computers, more efficient machines, and new information technologies, humanity began to apply speech in the communication between men and machines. In this context, an Artificial Intelligence technique known as SIASR (Speaker-Independent Automatic Speech Recognition) has been created. It identifies voice commands spoken by any person, based on pattern recognition and ANN (Artificial Neural Networks). It has several applications nowadays, such as telemarketing services, access to functionalities of mobile phone devices, security systems, toys, onboard automobile systems etc. Rabiner (1999). This paper proposes the construction of a computer interface that applies SIASR technology based on signal processing techniques and Neural Networks. This approach is intended to be a helpful toolbox in order to voice-command robots or machines in general, improving the naturalness of these interfaces. A speech recognition system would present a great advance in industrial complexes for the following reasons: shorter training time for employees, faster and more efficient equipment manipulation, optimization of individual and production performance, and the possibility of the operator executing simultaneous actions just by speaking.

2. AUTOMATIC SPEECH RECOGNITION SYSTEMS

SIASR systems are also classified Timoszczuk (2004), Furui (1989) as text dependent, as the speaker must pronounce a pre-defined text for recognition; or text independent, when there is no need for a pre-defined text. The voice signal representation and the characteristics used to form the recognition patterns must be defined. The pattern recognition approach is based on statistic methods Rabiner (1993), consisting of four steps: • Voice signal codification: uses spectral techniques as Linear Predictive Coding (LPC) or Fast Fourier Transform (FFT); • Training phase: reference patterns are created by a method that preserves the statistic characteristics of the class; • Classification phase: comparison between an unknown speech and the reference patterns. Spectral distance is calculated • Selection of the unknown speech class.

3. LINEAR PREDICTIVE CODING

Linear Predictive Coding is one of the most powerful voice analysis techniques and one of the most used voice codification methods Schafer (1975). LPC is used to transmit information of spectra, tolerating filter coefficients transmission errors. When working with a very small error, the spectrum can be fully distorted or, worse, cause prediction filter instability. LPC is a prediction method in which the voice signal sample is based on several previous samples. According to Bezerra (1994) there are the following methods for obtaining linear predictive coefficients: Covariance Method; Autocorrelation Method; Lattice Method; Spectral Estimation Method; and others.

4. NEURAL NETWORK

Artificial Neural Networks (ANN) have been successfully implemented in diversified work, having in common the essence of optimized pattern recognition, like: Image Recognition (characters, digital impressions); Voice Recognition (pronounced commands); Financial Tendency Recognition etc. In these applications, typically non-linearly separable problems are treated; therefore, multiple-layered Artificial Neural Networks are used to overcome this difficulty, inspired on the human neural mesh, to ensure good approximations and estimations on pattern recognition. An extremely important step in ANN implementation is training. Neural network training consists of establishing weights from specific training functions. The definition of which training algorithm will be used depends on several factors, like problem complexity, amount of training data, and expected precision and accuracy. A Multilayer ANN model that ensures good results is MLP – Multilayer Perceptrons: • Proposed by Rumelhart (1986); • Networks with two or more layers of Perceptron type neurons; • "Backpropagation error" training algorithm in which inputs and correspondent outputs are used in network training to calibrate an output function correlated to input vectors. This is possible due to error retro-propagation in the network output, in which network weights adjustments are made if necessary Demuth (2007), Perez (2007).

5. METHODOLOGY

5.1. Voice samples preparation

Voice samples were collected by using a microphone that captures perpendicular incident sounds, with minimal capture of surrounding sounds. Signal recording was made with Audacity software (version 1.3), in 16 bits rate, Mono format. Following the untreated voice signal, Fig. 1, editing was carried out by using the same software. Firstly environment noise along the entire length of the signal was detected. The noise perception is performed during the periods of "silence" before or after narration when the relevant signal does not manifest. These periods without the narrator's voice were then removed once they did not represent any interest to the identification of command as seen in Fig 2.

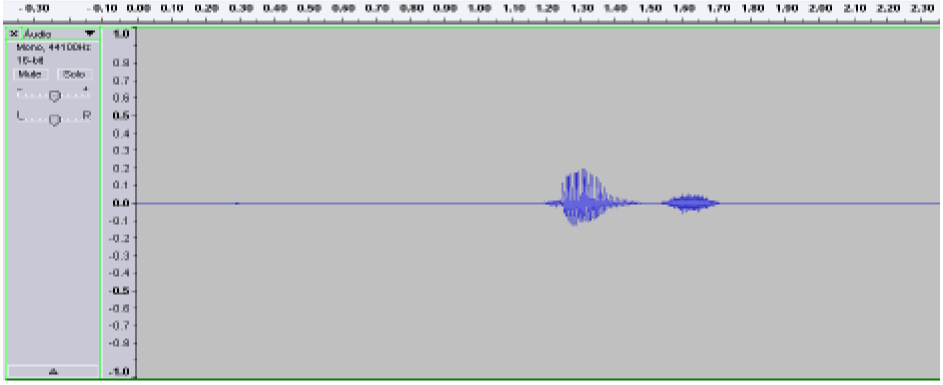


Figure 1 – Unmanufactured voice sign

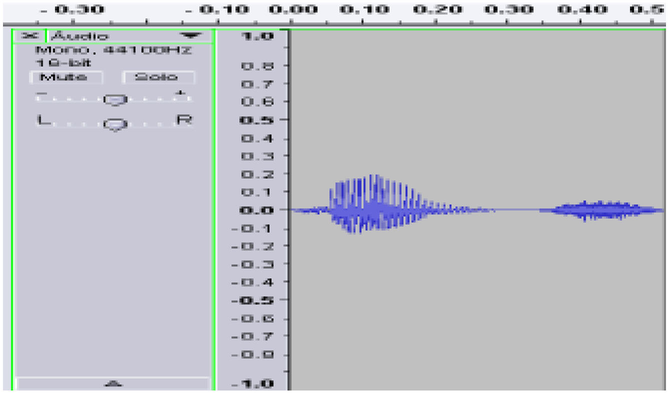


Figure 2 - Filtered and edited voice sign.

The edited signal was exported in "wave" format, reducing the resolution rate from 16 bits to 8 bits, Fig. 3. Thus, an even greater simplification of the sampling was achieved as required by the MatLab program. This program deals only with simplified wave format files. A database was constructed consisting of 21 voice samples for supervised neural network training besides 7 samples for further testing phase, among samples of male and female voices. Each sample brings the implicit narrative of four words, recorded separately: "back", "front", "left" and "right".

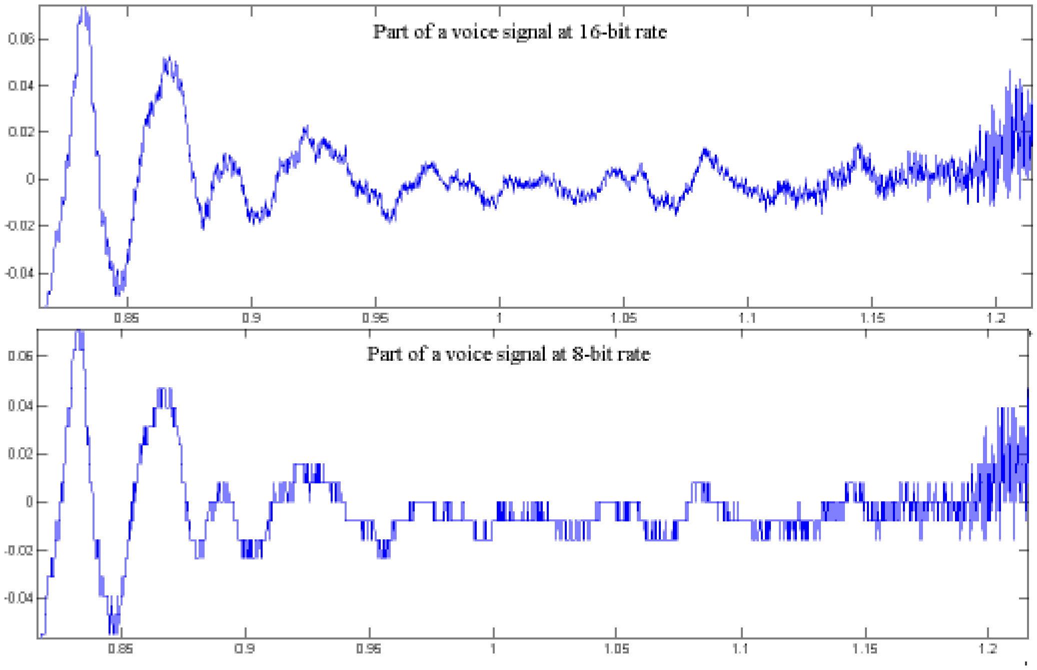


Figure 3. Comparison between 16 bits and 8 bits samples.

5.2. Implementation and determination of the LPC components optimal number

Wave format voice samples were imported into the MatLab work area. The discretized charts of signal LPC were then built. The goal at this point was to determine the number of LPC components that would provide the best outcome in order to optimize the work of the neural network. The 28 discretized charts for each voice command was printed on a single Cartesian plan totaling four chart groups ("back", "front", "left" and "right"). Thus, the degree of coincidence between corresponding sections of the LPC for a single voice command sampled by different speakers can be determined. For the same abscissa, a high concentration of these LPC components indicates the occurrence of a high degree of coincidence. In other words, one can analyze and empirically establish the optimal number of LPC that provides the highest similarity relation among the same voice commands when narrated by different individuals. In addition, the degree of dissimilarity between different commands should also be assessed. This analysis was done empirically from graphs that explain the linear dispersion curve for each voice command, like the one shown in Fig. 4. It was agreed then that the optimal number of components for each LPC is 14.

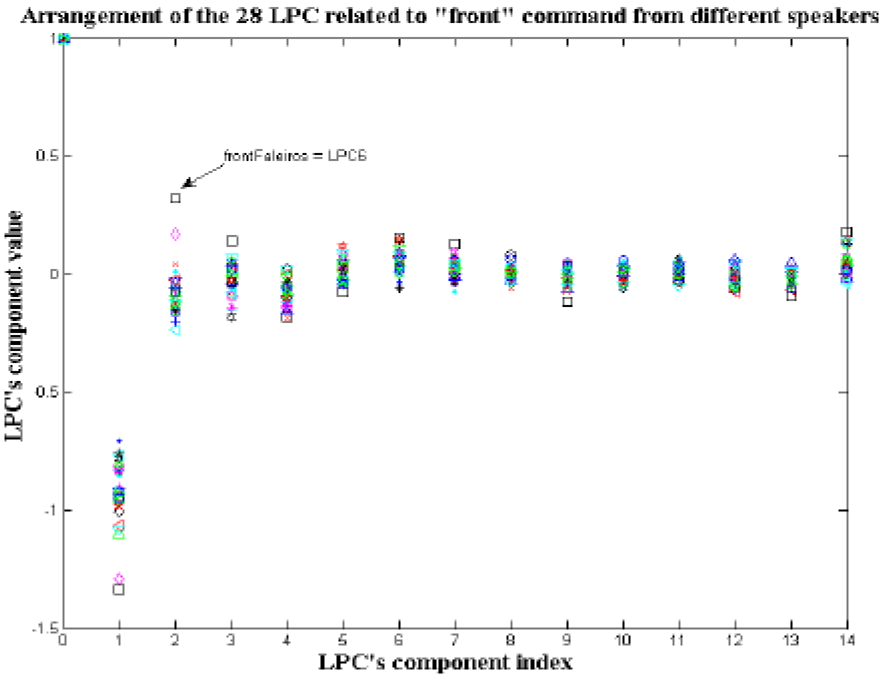


Figure 4. Arrangement of LPC related to "front".

5.3. Neural Network Implementation

After careful analysis in order to determine the optimal values of LPC, MatLab was used with the ANN model selected by the "nntool" provided by the software. Samples of LPC for each command were used as sample vectors in the neural network supervised training. At the end of the training phase, the neural network established comparison patterns for each of the four commands. The last step to be accomplished is finally the series of tests to assess the ability of the structured neural network to properly recognize the proposed voice commands. To this end, the database reserved for experimentation was used. It was composed of 7 samples, which obviously did not participate in training of the neural network. Finally, the signal corresponding to a test sample was introduced into the trained neural network, seeking the recognition pattern that could identify it as one of the four commands in question: "back", "front", "left" and "Right". The procedure was repeated for all the 7 samples, and then the scores and the accuracy rate of the system were determined. It was also observed that the network training could, alternatively, be done with LPC variation rate vectors. These vectors represent the behavior of the ordered variation between each component of a LPC vector, yielding a final sequence with the following values: increasing or decreasing. Value 1 represents increasing variation rate and value 0 decreasing variation rate. As the coefficients never had the same value because of the great number of decimal digits, the condition of zero variation rate was not needed to be considered for this problem. Consider the following LPC vector: LPC = [0.09, 1.2, 0.1, 0.02, 0.03, 2.5]. Its corresponding variation vector is: Δ LPC = [1, 0, 0, 1, 1]. The sequence of the variation vector represents the nuances of the curve which can be outlined by linking the components of a LPC vector. Still, from chart analysis, it was found that the LPC curves followed the same variation tendency with the same words and differentiated with other words. Thus the ΔLPC is also a convenient parameter to work on neural network.

6. NEURAL NETWORK CONFIGURATION

The speech recognition problem is characterized by a problem of nonlinear discriminant analysis. In this case the configuration based on multiple layers is used. The backpropagation configuration in MATLAB has two features, the cascade-forward backpropagation and feed-forward backpropagation.

Both configurations consist of N layers using the dotprod weight function, netsum net input function, the specified transfer function, and the first layer has weights coming from the input. The difference between these configurations is how to set the weights of the layers. In the cascadeforward each subsequent layer has weights coming from the input as well as from all previous layers. However the feed-forward adjusts the weight from the previous layer. The analysis of the obtained results will show the influence of this difference on the precision and convergence of the neural networks.

6.1. Transfer Functions

The sigmoid transfer functions, Logsig and Tansig, are commonly used in backpropagation networks (multiple layers), in part because these functions are differentiable, Fig. 5. The logsig function generates outputs between 0 and 1 as the neuron net input varies from negative to positive infinity. However the tansig function generates outputs between -1 and 1. As the output values were set up between 0 and 1, the logsig function is the transfer function proposed for the problem.

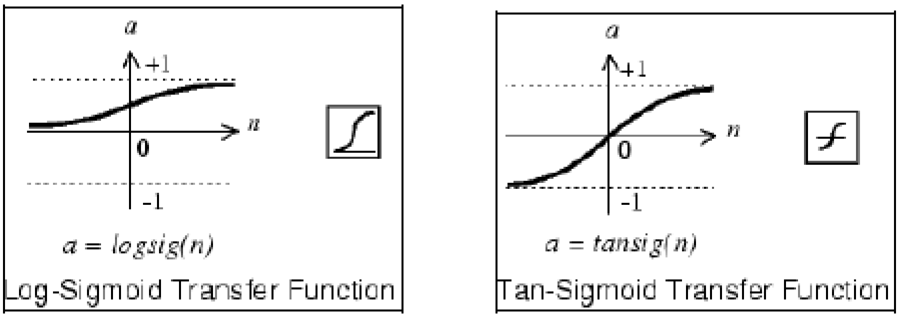


Figure 5. Transfer functions: logsig and transig

6.2. Training Functions

There are several available training functions in MatLab although some of them show better results for specific problems. According to Demuth (2007) the functions trainrp (Resilient Backpropagation) and trainlm (Levenberg-Marquardt) generate better performance neural networks.

7. NEURAL NETWORK TRAINING

The network input table is created by using the LPC coefficients and their variation values, which are the parameters that make up the neural network input vectors. The output table is set up by using the precision questions and the networks convergence as a goal.

7.1. Input Table

Input table is composed of 84 vectors - four samples were used (words back, front, left and right) from 21 people. For the entry of LPC coefficients each vector is composed of 14 elements. The values were defined from studies of LPC coefficient numbers. For the entry of variations rates, each vector is composed of 13 elements, since the number of variation is always minus one unit in relation to the total number of LPC coefficients.

7.2. Output Table

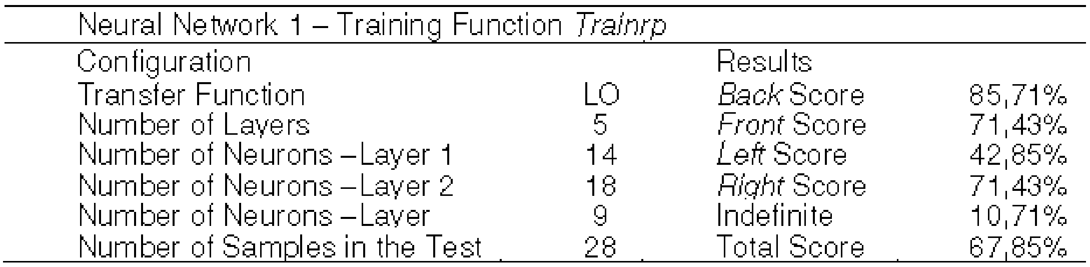
The number of vectors of the output table is the same as that of the input table, but each one with 9 digital elements, that is, it can take values "0" or "1". The vectors are defined with 9 elements each to make the analysis of the networks convergence and precision easier.

This value could not be very low or high because it would complicate the already mentioned analyses, but it should be sufficient to distinguish the output vector.

8. TESTS AND RESULTS

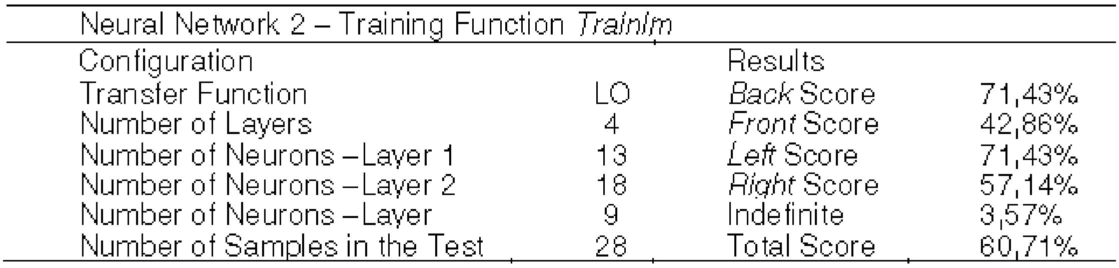
The tests were conducted by using voice samples of people who did not participate in the network training. This work analyzed two types of responses, one relating to individual neural networks and another to the coupling of two distinct networks. Each neural network can provide the words back, front, left, right and indefinite as output. The purpose of two neural networks coupling is to examine the network responses and determine the most convenient output response. This determination uses only the network precision as a parameter. That is, if the network response was the word 'back' and the other network was the word 'right', the coupling of networks determines that the answer will be the word 'back' if the first network has a higher percentage score for the word 'back' than the percentage score of the second network for the word 'right', otherwise the answer is the word 'right'. If one of the networks presents 'indefinite' as an answer and the other network presents any of the other words, the coupling determines that the answer will be any of those words and not 'indefinite'. The answer will be 'indefinite' only when both networks provide the 'indefinite' answer. And when the percentage of success of a specific word of a network is equal to another word of the other network, the coupling determines that the answer will be the word of the network which provides a higher percentage in total (arithmetic average of the percentages of the four words). The training of neural networks through the trainlm and trainrp training functions with backpropagation configuration is made by using two types of inputs, one is the LPC coefficients and the other is the variation rates of these coefficients. Several tests were made, from the coupling of networks of same function and same training type to simple neural networks. The higher performance results came from the coupling of a network with trainrp training function and a network with trainlm training function, both with the LPC coefficients as input data. The results are in tables 1, 2, and 3.

Table 1. Neural Network 1 results



Neural Network 1 – Training Function *Trainrp*			
Configuration		Results	
Transfer Function	LO	*Back* Score	85,71%
Number of Layers	5	*Front* Score	71,43%
Number of Neurons –Layer 1	14	*Left* Score	42,85%
Number of Neurons –Layer 2	18	*Right* Score	71,43%
Number of Neurons –Layer	9	Indefinite	10,71%
Number of Samples in the Test	28	Total Score	67,85%

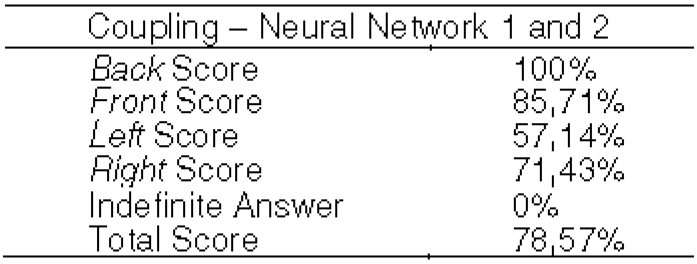
Table 2. Neural Network 2 results



Neural Network 2 – Training Function *Trainlm*			
Configuration		Results	
Transfer Function	LO	*Back* Score	71,43%
Number of Layers	4	*Front* Score	42,86%
Number of Neurons –Layer 1	13	*Left* Score	71,43%
Number of Neurons –Layer 2	18	*Right* Score	57,14%
Number of Neurons –Layer	9	Indefinite	3,57%
Number of Samples in the Test	28	Total Score	60,71%

Table 3. Result of Network 1 and Network 2 Coupling

Table 3. Result of Network 1 and Network 2 Coupling



Coupling – Neural Network 1 and 2	
Back Score	100%
Front Score	85,71%
Left Score	57,14%
Right Score	71,43%
Indefinite Answer	0%
Total Score	78,57%

9. CONCLUSION

This article presents the potential use of Artificial Neural Networks (ANN) to solve the problem of Speaker- Independent Automatic Speech Recognition - SIASR once ANN can establish comparable patterns for voice commands. Good results lie on the ANN as well as the use of reliable input vectors. Therefore the thorough work of signal processing, which was intended to reduce the resolution rate from 16 bit to 8 bit, eliminate the noise and implement the LPC, was important for the analysis of the problem. However, as the aim of the SIASR proposal is to recognize the speech and not the speaker, the information lost in this treatment did not influence the results. With the observation that the variation rate of LPC coefficients could appear as pattern, in addition to the LPC coefficients, efforts were also made to target the analysis of this parameter. It was observed that the variation rate is really a pattern, but the ANNs with LPC coefficients input vectors showed a better performance. It was concluded that the LPC coefficients are presented more as a pattern than their variation rates; even so they are convenient to work with ANN. Is important to express that the cascade backpropagation-forward and feed-forward backpropagation did not show significant differences for this problem, as results showed very close, even when using different ways to adjust the weight of neurons layers. The coupling of neural networks was a good way to improve the performance. The best coupling presented 0% of 'indefinite' answers and obtained greater percentage scores than the individual use of networks. It is important to mention that the best couplings did not emerge from ANNs with good results for all commands, but from ANNs that were specialists in other commands which other networks were not, and vice versa. They were trained with different training functions, one with trainrp and the other with trainlm, but using the same LPC coefficients as input vectors. Computers controlled by human speech can also present a social background. The physically impaired people can use this technology to replace usual data input/output systems in regular industrial equipment. In this way they will be able to work on machines that are still inaccessible to them as efficiently as any other person. In fact, the practical implementation of the technology treated in this research will make the relations between employees and machines of a factory different from what they are seen nowadays: closer, more natural and more productive. Acknowledgements The authors would like to thank CNPq, CAPES, and FAPEMIG - Brazilian research funding agencies, for the research scholarships, which supported this work.

REFERENCES

Rabiner, L. R., Applications of Voice Processing to Telecommunications, Proceedings of the IEEE, v. 82, n. 2, p 197-228, Feb. 1999.

Timoszczuk, A. P., Reconhecimento Automático do Locutor com Redes Neurais Pulsadas, Ph. D. Thesis, Escola Politécnica - USP, SP, 2004.

Furui, S., Digital Speech Processing, Synthesis, and Recognition, New York, Marcel Dekker, 1989.

Rabiner, L. and Juan, B. H., Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, NJ, 1993.

Schafer, R. W. and Rabiner, L. R., Digital Representations of Speech Signals, Proceedings of the IEEE, v.63, n. 4, p662-677, April, 1975.

Bezerra, M. R., Reconhecimento Automático de Locutor para fins Forenses, Utilizando Técnicas de Redes Neurais, (in Portuguese) Master Degree Dissertation, Instituto Militar de Engenharia, RJ, 1994.

Demuth, H., Beale, M., and Hagan, M., Neural Network Toolbox 5 – User's Guide. Matlab, http://www.mathworks.com/access/helpdesk/help/pdf_doc/nnet/nnet.pdf, Accessed on Aug 5th, 2007.

Perez,C. Ricardo, Ferramenta Para Reconhecimento de Padrões de Defeitos em Linhas de Transmissão, GQE, April, 2007.