# Artificial Neural Networks used for Pattern Recognition of Speech Signal based on DCT Parametric Models of Low Order

Priscila Lima Rocha
Federal Institute of Maranhão
São Luís, Maranhão, Brazil
Email: priscilalima_rocha@hotmail.com

Washington Luis Santos Silva
Federal Institute of Maranhão
São Luís, Maranhão, Brazil
Email: washington.silva@ifma.edu.br

*Abstract*—**This paper proposes the development of a Numerical Command Recognition System of Speech Signal based on Neural Networks and DCT models. Thus, two configurations of neural networks, the Multilayer Perceptron and Learning Vector Quantization are evaluated by their performance in speech signal recognition, whose encoding is made by the mel-cepstral coefficients that are used to generate a two-dimensional time matrix by Discrete Cosine Transform (DCT). The selection of the best configuration of neural network for classification of the patterns was carried out by comparative analysis of performance of the MLP and LVQ networks through training, validation and test of the network topology and learning algorithms previously established. For demonstration of the performance of the proposed analysis methodology, the obtained results were compared with other methods of classification given by Gaussian Mixture Models (GMM) and Support Vector Machines (SVM).**

## I. Introduction

The extensive development of research in the speech processing area shows the effort to improve the performance of speech recognition systems for practical applications. The use of such systems allow autonomy in areas such as telephony, wherein service requests are directed through speech commands [1]; in the automotive industry through the activation of devices inside vehicles [2]; in computing systems by utility programs on computers, besides the application in robotics [3], residential and hospital automation for accessibility of people with locomotor and visual disease [4].

Therefore, it is proposed in this paper the development of a Numerical Command Recognition System of Speech Signal with two configurations of neural networks, the Perceptron Multilayer and Learning Vector Quantization. They will be evaluated by its performances in speech signal recognition, whose encoding is done using the mel-cepstral coefficients (MFCC) and discrete cosine transform (DCT). [5]. The mel-cepstral coefficients and the DCT are used for generating of a two-dimensional time matrix which represents, through a reduced set of parameters, the pattern of speech signal to be used as training, validation and testing input of the networks neural.

### A. Proposal Analysis Methodology

The proposed speech recognition system in this paper aims to classify patterns of speech signals, represented by locutions from Brazilian Portuguese of the digit '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', through the recognizer characterized by the neural network. The adopted methodology uses a reduced number of parameters to represent each pattern obtained in the speech signal pre-processing stage for generating two-dimensional time matrices 2, 3 e 4. The time two-dimensional matrix reproduces the global and local variations in time, as well as the spectral envelope of the speech signal. The use of neural network as classifier in speech recognition system using the low-order parameters generated by the two-dimensional matrix of low temporal order and the comparison of the results are the main contributions of this work.

After processing of the speech signal, the topologies and definition of the classifier represented by neural network are carried out through analysis of performance between two configuration in the literature: the MLP and LVQ Neural Networks [6], [7]. This analysis is carried out in two phases: first, the specified topologies the observation of the behavior of the MLP and LVQ networks in training and validation process and the selection of the topologies that accomplish global validation with hit above $80\%$. Then the selected topologies are tested with different parameters of the speech signal that are not used in the training process and the results of classifications of the MLP and LVQ neural networks are shown.

The goal of the tests is to choose the network of best performance. The criteria used to rank the best networks will be the best performance in the correct classification and lower complexity topological. Each of the phases carried out for analysis of performance of the MLP and LVQ networks are executed with the parameters of speech signal encoding through of the two-dimensional time matrix of order 2, 3 and 4. Thus, it is observed the response of the neural networks proposed when is increased the number of parameters that represents the pattern of speech signal to be recognized.

Therefore, according to applied procedures in the prepara-

978-1-5090-2870-2/16/$31.00 ©2016 IEEE

tion of this paper, it can define between studied configurations of neural networks, the network that best adapt to speech recognition system with speech signal represented by a few parameters. The study performance of LVQ Neural Network, presented in this paper, applied in the speech recognition with low-order models provides an alternative approach to the MLP classifier that is the neural network most used in speech recognition problems.

## II. Speech Recognition System based on Networks Neural

It is shown in Figure 1 the block diagram of the proposed speech recognition system.

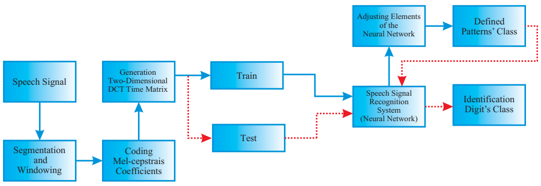



Fig. 1. Block diagram of the Speech Recognition System using Neural Networks

### A. Processing of Speech Signal

*1) Generation of two-dimensional DCT time matrix:* After obtaining the mel-cepstral coefficients from the speech signal, encoding through discrete cosine transform (DCT) was carried out, which enables to synthesize the long-term variations of the spectral envelope of the speech signal. The result of this encoding was the generation of a two-dimensional DCT time matrix, obtained by (1):

$$C_k(n,T) = \frac{1}{N}\sum_{t=1}^{T} \mathrm{mfcc}_k(t)\cos\left[\frac{(2t-1)n\pi}{2T}\right] \tag{1}$$

where $k$, that varies of $1 \leq k \leq K$, is the $k$-th line component of $t$-th segment of the matrix. $K$ is the number of mel-cepstral coefficient; $n$, that varies of $1 \leq n \leq N$, is the $n$-th column. $n$ is the order of the matrix DCT; $T$ is the number of vectors of observation of the mel-cepstral coefficients in time axis; $\mathrm{mfcc}_k(t)$ represent the mel-cepstral coefficients.

Thus, for each locution of the digit **D** to be recognized there is a two-dimensional DCT time matrix $C_{kn}^{jm}$, where $j = 0, 1, 2, \cdots, 9$ represent the digit to be encoded and $m = 0, 1, 2, \cdots, 9$ represent the example taken for each digit. These matrices were transformed in column vectors $C_N^{jm}$, where $N$ is the number of parameters of the two-dimensional matrix $C_{kn}^{jm}$. The column vector $C_N^{jm}$ preserves the temporal alignment of mel-cepstral coefficients and its general term is given by (2):

$$C_N^{jm} = [c_{11}^{jm},\ c_{12}^{jm},\ \cdots,\ c_{1n}^{jm} c_{21}^{jm},\ c_{22}^{jm},\ \cdots,\ c_{2n}^{jm},\ \cdots,\ c_{kn}^{jm}]' \tag{2}$$

The vectors $C_N^{jm}$ were used as input patterns of the neural network. The dimension of $C_N^{jm}$ defines the number of input of the neural network. In order to compare the performance of the neural network when the number of parameters that compose the input patterns of speech is increased, two-dimensional matrices $C_{kn}^{jm}$ were generated of order $n = 2, 3$ e $4$, thereby obtaining the patterns to be recognized by neural networks represented by $C_N^{jm}$, where $N = 4$, $9$ e $16$, respectively.

*2) Design of Neural Networks:* The design of the neural networks is carried out through simulations of combinations of elements of the network topology and learning algorithms previously established and also based on obtained results in other similar works of patters classification [8], [9], [10]. The configurations of MLP and LVQ network were used to the pattern recognition according to speech signal encoding by two-dimensional time DCT matrix. The choice for analyzing these two configurations in this work is justified because they are neural networks with wide applicability and good results in pattern recognition area, according to the literature [11].

Then, in Table I and Table II present, respectively, the elements related to topology and training algorithms previously established for MLP and LVQ networks that were combined during the training phase.

TABLE I
Training elements of MLP Neural Networks

| Elements | Symbol | Defined Range |
|---|---|---|
| Training Algorithms | - | GD,GDM,RP,LM[a] |
| Nº of Hidden Layers | - | 1 e 2 |
| Nº of Hidden Neuron 1 Hidden Layer | $n_1$ | 20,25,30,35,40,45,50,55,60 |
| Nº of Hidden Neuron 2 Hidden Layers | $n_1$ e $n_2$ | $n_2 = 15$ |
| Learning Rate | $\eta$ | 0.01 |
| Momentum Term | $\alpha$ | 0.8 |
| Nº of Epoch | - | 1000 |
| Activation Function | - | Hyperbolic Tangent |

[a] GD: Gradient Descent; GDM: Gradient Descent with Momentum; RP: Resilient Backpropagation; LM: Levenberg-Marquardt

TABLE II
Training elements of LVQ Neural Networks

| Elements | Symbol | Defined Range |
|---|---|---|
| Training Algorithms | - | LVQ-1 |
| Nº of Hidden Layers | $n_1$ | 20,25,30,35,40,45,50,55,60 |
| Learning Rate | $\eta$ | 0.01 |
| Nº of Epoch | - | 1000 |

The number of inputs of neural network is defined by dimension of vector $C_N^{jm}$, where $N = 4, 9, 16$. The output of the neural network is specified by the number of patterns to be recognized, which is based on the method of one c-classes. How the problem of recognition of this work is to correctly classify 10 digits of the Portuguese Language, the output of the neural network has 10 neurons, ie, a neuron for each class. For MLP configuration networks with two hidden layers were simulated in order to verify the need to increase the number of hidden layers to extract the features contained in the input patterns presented to the network. During the research conducted for the preparation of this paper, it was observed that the initialization of the set of weights for the MLP neural network impacted the final results. Therefore, to realize this fact, four training ($T_1, T_2, T_3, T_4$) with different initializations of the set of weights made over a random uniform distribution between interval $[-0.01, 0.01]$ was carried out for each proposed topology for MLP network. Thus, it was possible to observe the behavior of neural networks in relation to training time and the ability to generalize, because appropriate set of initial weights allows a reduction in training time and a high probability of achieving the overall minimum of error function. In addition, this set can significantly improve performance in generalization.

*3) Training and Testing Sets:* The Training and testing of the MLP and LVQ Neural Networks were carried out with patterns of speech signals from selection of locutions from voice bank EPUSP, INATEL and IFMA. These sets are changed depending on the order of the two-dimensional matrix used to generate the patterns. The training and test sets are composed as follows:

1) Training Sets $\Omega_{NL}^{Tr}$: This set represents the number of parameters of patterns of the set, where $N$ = 4, 9, 16, $L$ the total number of locutions and $Tr$ indicates that is training set, is composed of 200 locutions, which has 20 examples of each digit to be recognized ($m = 20$), where half are female speakers and the other half are male speakers. The training set was partitioned in the estimation subset $\Omega_N^E$ and validation subset $\Omega_N^V$. The $\Omega_N^E$ is used to adjust the network weights and it contains 80% of the total of $\Omega_{NL}^{Tr}$; already $\Omega_N^V$ is used to validate the trained topologies and check the generalization of the network and it contains 20% of the total of training set ($\Omega_{N200}^{Tr} = \{\Omega_N^E \cup \Omega_N^V\}$).
2) Testing Set $\Omega_{NL}^{T}$: For this set were selected 20 speakers, where 10 speakers are female ($\Omega_{NL}^{TF}$) and 10 speakers are male ($\Omega_{NL}^{TM}$). All speakers belong to IFMA bank, but are speakers who did not participate with pronunciations for the training set. Each speaker contributed with 10 examples for each digit ($m = 10$), totaling 100 digit pronounced by speaker. Therefore, it has 1000 male locutions and 1000 female locutions for testing ($\Omega_{N2000}^{T} = \{\Omega_{N1000}^{TM} \cup \Omega_{N1000}^{TF}\}$).

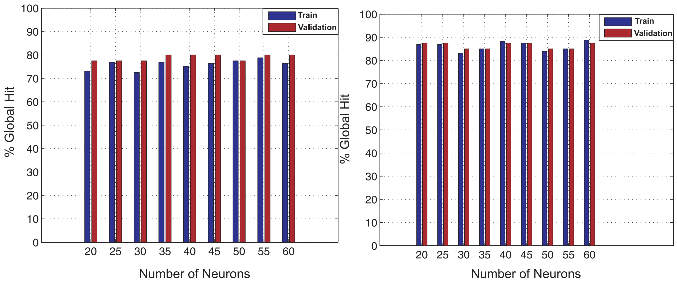
## III. Experimental Results

### A. LVQ Training and Validation

After training the LVQ networks by all combinations of defined topology elements, it is shown in Figure 2a, Figure 2b and Figure 3, respectively, the obtained results with training of networks using the patterns $C_4^{jm}$, $C_9^{jm}$ and $C_{16}^{jm}$.

### B. LVQ Testing

The tests were applied only in trained topologies with correct classification results in the global validation greater than 80%. The obtained results (in percent) with the application of test sets $\Omega_{N1000}^{TM}$ and $\Omega_{N1000}^{TF}$ with $N$ = 4, 9 e 16 are shown, respectively, in Table III, Table IV and Table V. Finishing tests, the topology with best performance was chosen according to criterion that besides presenting the highest mean results in tests must also have a reduced number of neurons.Thus, topologies with mean hit of test detached in blue are those that showed the highest results, however topologies with mean hit detached in red have very close results but with fewer neurons. Therefore, the topology of 40, 20 and 25 neurons shown in Table III, Table IV and Table V are the topology with best result of generalization for LVQ neural network configuration.




(a) (b)

Fig. 2. Result of Training and Validation Global Hit: (a) LVQ $C_4^{jm}$ e (b) LVQ $C_9^{jm}$

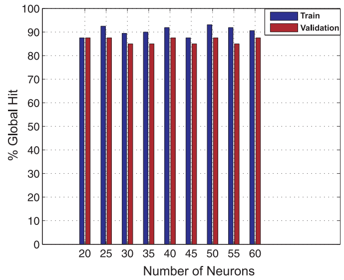



Fig. 3. LVQ $C_{16}^{jm}$: Result of Training and Validation Global Hit

TABLE III
LVQ $C_4^{jm}$: Female and Male Speakers Tests

| | 35 | 40 | 45 | 55 | 60 |
|---|---|---|---|---|---|
| **Loc_F1** | 89 | 91 | 94 | 92 | 94 |
| **Loc_F2** | 95 | 90 | 90 | 87 | 88 |
| **Loc_F3** | 72 | 80 | 67 | 73 | 78 |
| **Loc_F4** | 81 | 84 | 70 | 67 | 76 |
| **Loc_F5** | 90 | 79 | 85 | 83 | 87 |
| **Loc_F6** | 98 | 90 | 95 | 93 | 89 |
| **Loc_F7** | 56 | 57 | 46 | 48 | 55 |
| **Loc_F8** | 72 | 76 | 72 | 64 | 79 |
| **Loc_F9** | 54 | 69 | 53 | 59 | 74 |
| **Loc_F10** | 68 | 78 | 79 | 81 | 77 |
| **Loc_M1** | 71 | 62 | 59 | 60 | 62 |
| **Loc_M2** | 77 | 77 | 86 | 85 | 85 |
| **Loc_M3** | 74 | 77 | 76 | 79 | 78 |
| **Loc_M4** | 72 | 70 | 79 | 83 | 79 |
| **Loc_M5** | 63 | 64 | 63 | 66 | 67 |
| **Loc_M6** | 72 | 65 | 65 | 65 | 69 |
| **Loc_M7** | 62 | 72 | 66 | 71 | 68 |
| **Loc_M8** | 70 | 79 | 74 | 73 | 75 |
| **Loc_M9** | 72 | 63 | 62 | 63 | 65 |
| **Loc_M10** | 76 | 81 | 88 | 78 | 83 |
| **Mean** | 74.2 | 75.2 | 73.45 | 73.5 | 76.4 |

TABLE IV
LVQ $C_9^{jm}$: FEMALE AND MALE SPEAKERS TESTS

| | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 |
|---|---|---|---|---|---|---|---|---|---|
| **Loc_F1** | 88 | 77 | 81 | 81 | 88 | 89 | 78 | 76 | 87 |
| **Loc_F2** | 86 | 88 | 86 | 85 | 78 | 87 | 85 | 85 | 84 |
| **Loc_F3** | 81 | 84 | 81 | 80 | 84 | 83 | 79 | 80 | 82 |
| **Loc_F4** | 87 | 86 | 70 | 71 | 81 | 87 | 69 | 69 | 88 |
| **Loc_F5** | 82 | 82 | 80 | 80 | 73 | 87 | 80 | 84 | 83 |
| **Loc_F6** | 90 | 83 | 80 | 80 | 89 | 90 | 78 | 79 | 84 |
| **Loc_F7** | 72 | 67 | 69 | 72 | 78 | 75 | 70 | 71 | 70 |
| **Loc_F8** | 87 | 83 | 79 | 70 | 90 | 86 | 72 | 78 | 82 |
| **Loc_F9** | 73 | 72 | 70 | 69 | 77 | 85 | 68 | 72 | 78 |
| **Loc_F10** | 64 | 68 | 66 | 62 | 66 | 64 | 59 | 64 | 64 |
| **Loc_M1** | 69 | 64 | 63 | 59 | 69 | 68 | 59 | 66 | 70 |
| **Loc_M2** | 81 | 81 | 76 | 79 | 78 | 82 | 73 | 76 | 82 |
| **Loc_M3** | 80 | 81 | 72 | 75 | 78 | 79 | 65 | 72 | 75 |
| **Loc_M4** | 80 | 69 | 72 | 71 | 71 | 80 | 72 | 75 | 73 |
| **Loc_M5** | 64 | 60 | 69 | 66 | 58 | 61 | 68 | 64 | 69 |
| **Loc_M6** | 90 | 81 | 72 | 75 | 86 | 86 | 74 | 77 | 83 |
| **Loc_M7** | 82 | 74 | 76 | 75 | 83 | 75 | 75 | 74 | 84 |
| **Loc_M8** | 81 | 69 | 73 | 72 | 80 | 71 | 71 | 69 | 81 |
| **Loc_M9** | 69 | 67 | 66 | 69 | 55 | 73 | 64 | 66 | 71 |
| **Loc_M10** | 77 | 73 | 76 | 78 | 73 | 77 | 73 | 76 | 76 |
| **Mean** | 79.15 | 75.45 | 73.85 | 73.45 | 76.75 | 79.25 | 71.6 | 73.65 | 78.3 |

TABLE V
LVQ $C_{16}^{jm}$: FEMALE AND MALE SPEAKERS TESTS

| | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 |
|---|---|---|---|---|---|---|---|---|---|
| **Loc_F1** | 88 | 94 | 89 | 89 | 95 | 81 | 91 | 92 | 90 |
| **Loc_F2** | 86 | 95 | 90 | 88 | 95 | 79 | 94 | 88 | 87 |
| **Loc_F3** | 90 | 90 | 83 | 83 | 90 | 79 | 89 | 84 | 87 |
| **Loc_F4** | 89 | 85 | 87 | 87 | 84 | 84 | 88 | 87 | 88 |
| **Loc_F5** | 85 | 96 | 87 | 89 | 97 | 78 | 87 | 89 | 88 |
| **Loc_F6** | 90 | 99 | 89 | 90 | 99 | 82 | 90 | 89 | 90 |
| **Loc_F7** | 92 | 73 | 80 | 80 | 74 | 77 | 83 | 78 | 82 |
| **Loc_F8** | 98 | 95 | 97 | 99 | 93 | 88 | 97 | 95 | 97 |
| **Loc_F9** | 88 | 80 | 84 | 86 | 80 | 80 | 81 | 87 | 85 |
| **Loc_F10** | 75 | 80 | 70 | 73 | 80 | 76 | 72 | 73 | 71 |
| **Loc_M1** | 69 | 80 | 77 | 80 | 78 | 72 | 75 | 79 | 76 |
| **Loc_M2** | 84 | 87 | 81 | 83 | 88 | 79 | 87 | 89 | 84 |
| **Loc_M3** | 76 | 85 | 77 | 76 | 86 | 76 | 84 | 85 | 76 |
| **Loc_M4** | 87 | 88 | 79 | 83 | 89 | 81 | 89 | 87 | 84 |
| **Loc_M5** | 61 | 74 | 72 | 67 | 79 | 54 | 66 | 75 | 68 |
| **Loc_M6** | 78 | 78 | 80 | 83 | 86 | 73 | 87 | 82 | 78 |
| **Loc_M7** | 79 | 84 | 75 | 82 | 82 | 73 | 85 | 81 | 79 |
| **Loc_M8** | 80 | 84 | 73 | 82 | 82 | 73 | 82 | 77 | 77 |
| **Loc_M9** | 68 | 79 | 68 | 69 | 79 | 60 | 77 | 73 | 68 |
| **Loc_M10** | 76 | 90 | 80 | 78 | 88 | 75 | 77 | 82 | 77 |
| **Mean** | 81.95 | 85.8 | 80.9 | 82.35 | 86.2 | 76 | 84.05 | 83.6 | 81.6 |

### C. MLP Training and Validation

Because the MLP network have a greater number of topology elements and training algorithms to be combined, many simulations were carried for this configuration. Through these simulations it was possible to observe the behavior of proposals topologies and so to define the best result. It was verified during simulations that $GD$ and $GDM$ training algorithms did not achieve good results for the pattern recognition problem with the proposal encoding, showed results of training and validation below $50\%$. Thus, it is concluded that these algorithms have not been able to extract the features of the patterns presented to the neural network and to generate a satisfactory classification. These results are not presented in this paper. In Figure 4, Figure 5 and Figure 6 are shown, respectively, the global hit results of training and validation obtained to MLP networks with one hidden layer, trained by $LM$ and $RP$ training algorithms using the patterns $C_4^{jm}$, $C_9^{jm}$ and $C_{16}^{jm}$. From the results shown in Figure 4, Figure 5 and Figure 6, it is possible to verify the influence of the set of initial weights for the MLP network. It is observed that for the same topology, one set of weights initialized randomly in a training led to satisfactory response, but in another training, the set of initial weights resulted in an undesirable response.

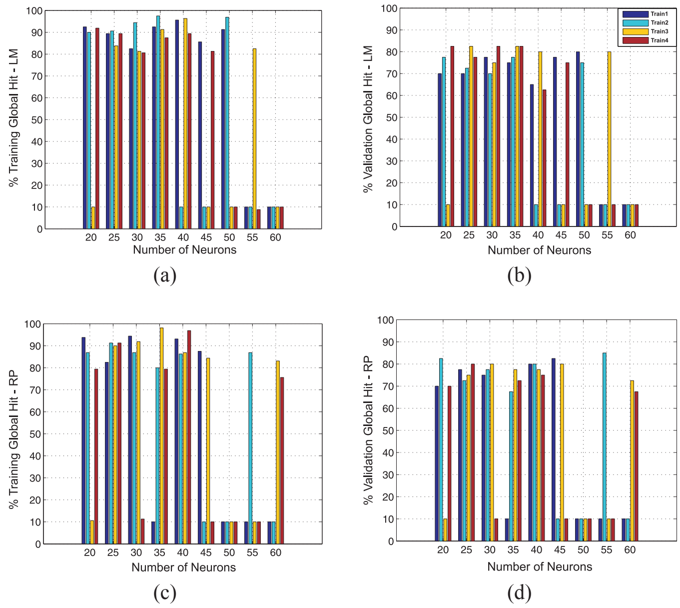



Fig. 4. MLP $C_4^{jm}$: Result of Training and Validation Global Hit of the neural network with 1 hidden layer for the LM and RP algorithm

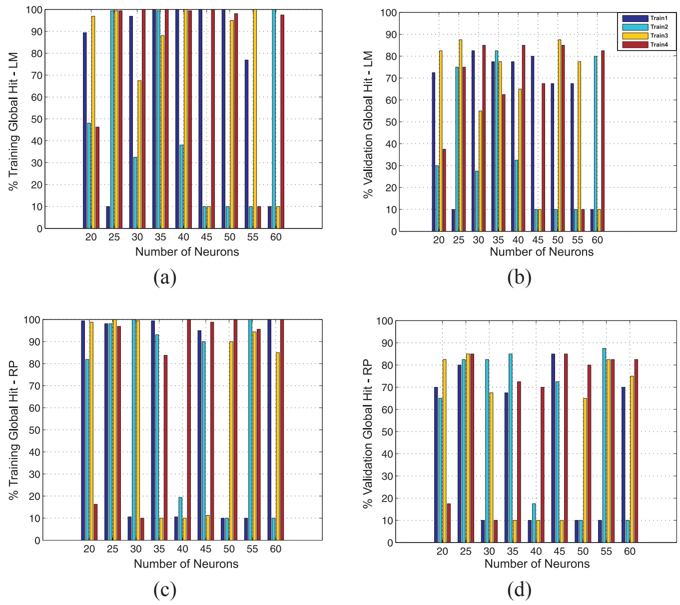



Fig. 5. MLP $C_9^{jm}$: Result of Training and Validation Global Hit of the neural network with 1 hidden layer for the LM and RP algorithm

### D. MLP Testing

As well as carried out for the LVQ neural network, after completion of training and validation stage, between the obtained results in the simulations, the topologies of MLP Networks that showed global hit results of validation greater than $80\%$ were selected for application testing. The best results (in percent) found in executed tests, considering the networks trained with one and two layer by RP and LM algorithms, using the test sets $\Omega_{N1000}^{TM}$ and $\Omega_{N1000}^{TF}$ with $N = 4$, 9 e

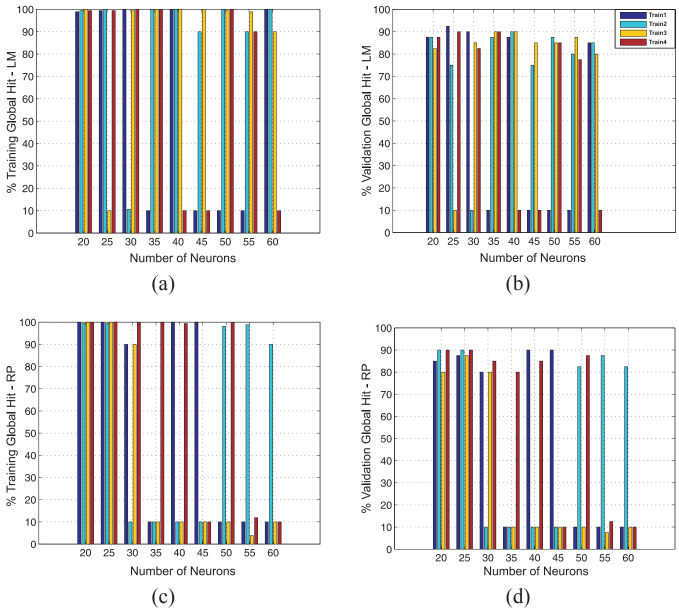



Fig. 6. MLP $C_{16}^{jm}$: Result of Training and Validation Global Hit of the neural network with 1 hidden layer for the LM and RP algorithm

16 are shown, respectively, in Table VI, Table VII and Table VIII. Finishing tests, as carried out for the LVQ networks, the topology of MLP network with best performance was chosen according to criterion that besides presenting the highest mean results in tests must also to have less complexity in the topological structure. Thus, topologies with mean hit of test detached in blue are those that showed the highest results, however topologies with mean hit detached in red have very close results but with fewer neurons and only layer. Therefore, the topology of 30, 25 and 20 neurons with only one hidden layer trained by LM algorithm shown in Table VI, Table VII and Table VIII are the topology with best result of generalization for LVQ neural network configuration.

TABLE VI
SELECTION OF BEST TEST RESULTS MLP $C_4^{jm}$

| Column 1 | Column 2 | Column 3 | Column 4 | Column 5 |
|---|---|---|---|---|
| | 1 layer | | 2 layers | |
| | LM | RP | LM | RP |
| | 30-T4 | 45-T3 | 45-T1 | 45-T3 |
| **Loc_F1** | 94 | 82 | 88 | 89 |
| **Loc_F2** | 95 | 97 | 90 | 94 |
| **Loc_F3** | 83 | 84 | 82 | 84 |
| **Loc_F4** | 87 | 79 | 82 | 83 |
| **Loc_F5** | 91 | 88 | 88 | 88 |
| **Loc_F6** | 95 | 97 | 94 | 99 |
| **Loc_F7** | 67 | 57 | 65 | 63 |
| **Loc_F8** | 91 | 91 | 85 | 90 |
| **Loc_F9** | 75 | 59 | 73 | 70 |
| **Loc_F10** | 84 | 67 | 77 | 81 |
| **Loc_M1** | 72 | 77 | 77 | 74 |
| **Loc_M2** | 89 | 81 | 81 | 85 |
| **Loc_M3** | 86 | 82 | 86 | 85 |
| **Loc_M4** | 82 | 80 | 78 | 80 |
| **Loc_M5** | 72 | 72 | 71 | 68 |
| **Loc_M6** | 67 | 64 | 66 | 64 |
| **Loc_M7** | 73 | 68 | 73 | 75 |
| **Loc_M8** | 80 | 74 | 76 | 80 |
| **Loc_M9** | 73 | 73 | 70 | 70 |
| **Loc_M10** | 87 | 91 | 93 | 93 |
| **Mean** | 82.15 | 78.15 | 79.75 | 80.75 |

TABLE VII
SELECTION OF BEST TEST RESULTS MLP $C_9^{jm}$

| Column 1 | Column 2 | Column 3 | Column 4 | Column 5 |
|---|---|---|---|---|
| | 1 layer | | 2 layers | |
| | LM | RP | LM | RP |
| | 25-T3 | 45-T1 | 50-T4 | 60-T1 |
| **Loc_F1** | 90 | 93 | 92 | 95 |
| **Loc_F2** | 94 | 90 | 93 | 93 |
| **Loc_F3** | 93 | 93 | 91 | 92 |
| **Loc_F4** | 95 | 94 | 95 | 96 |
| **Loc_F5** | 91 | 97 | 96 | 96 |
| **Loc_F6** | 89 | 92 | 93 | 91 |
| **Loc_F7** | 76 | 87 | 85 | 85 |
| **Loc_F8** | 91 | 91 | 92 | 95 |
| **Loc_F9** | 82 | 76 | 77 | 83 |
| **Loc_F10** | 78 | 79 | 66 | 77 |
| **Loc_M1** | 74 | 76 | 83 | 79 |
| **Loc_M2** | 82 | 84 | 80 | 82 |
| **Loc_M3** | 89 | 87 | 85 | 91 |
| **Loc_M4** | 88 | 90 | 89 | 88 |
| **Loc_M5** | 65 | 65 | 64 | 65 |
| **Loc_M6** | 83 | 86 | 78 | 84 |
| **Loc_M7** | 69 | 73 | 77 | 68 |
| **Loc_M8** | 69 | 71 | 72 | 69 |
| **Loc_M9** | 88 | 82 | 84 | 85 |
| **Loc_M10** | 82 | 80 | 87 | 90 |
| **Mean** | 83.4 | 84.3 | 83.95 | 85.2 |

TABLE VIII
SELECTION OF BEST TEST RESULTS MLP $C_{16}^{jm}$

| Column 1 | Column 2 | Column 3 | Column 4 | Column 5 |
|---|---|---|---|---|
| | 1 layer | | 2 layers | |
| | LM | RP | LM | RP |
| | 20-T1 | 50-T2 | 40-T1 | 30-T2 |
| **Loc_F1** | 95 | 88 | 88 | 92 |
| **Loc_F2** | 91 | 87 | 89 | 89 |
| **Loc_F3** | 97 | 98 | 98 | 97 |
| **Loc_F4** | 93 | 88 | 86 | 94 |
| **Loc_F5** | 90 | 93 | 93 | 95 |
| **Loc_F6** | 97 | 89 | 91 | 93 |
| **Loc_F7** | 92 | 89 | 94 | 93 |
| **Loc_F8** | 92 | 93 | 97 | 93 |
| **Loc_F9** | 82 | 85 | 82 | 79 |
| **Loc_F10** | 73 | 80 | 82 | 84 |
| **Loc_M1** | 84 | 74 | 74 | 86 |
| **Loc_M2** | 91 | 89 | 89 | 90 |
| **Loc_M3** | 90 | 86 | 86 | 91 |
| **Loc_M4** | 85 | 92 | 94 | 93 |
| **Loc_M5** | 59 | 64 | 67 | 69 |
| **Loc_M6** | 89 | 88 | 88 | 89 |
| **Loc_M7** | 74 | 79 | 83 | 88 |
| **Loc_M8** | 71 | 80 | 85 | 83 |
| **Loc_M9** | 82 | 82 | 83 | 82 |
| **Loc_M10** | 93 | 91 | 92 | 92 |
| **Mean** | 86 | 85.75 | 87.05 | 88.6 |

At the end the tests carried out with the MLP and LVQ networks and selected the topology of best performance for each configuration, for each of the three sets of patterns $C_N^{jm}$, $N = 4$, 9 and 16, the obtained results are summarized in the Table IX

TABLE IX
FINAL PERFORMANCE SUMMARY OF THE MLP AND LVQ NETWORKS

| | $C_4^{jm}$ | | $C_9^{jm}$ | | $C_{16}^{jm}$ | |
|---|---|---|---|---|---|---|
| | Nº of Neurons | % Test Hit | Nº of Neurons | % Test Hit | Nº of Neurons | % Test Hit |
| LVQ | 40 | 75.2 | 20 | 79.15 | 25 | 85.8 |
| MLP | 30 | 82.15 | 25 | 83.4 | 20 | 86 |

## IV. Comparison to other methods used in speech recognition

To check the performance of the methodology presented in this work, it was carried out the comparison of obtained results with MLP and LVQ neural networks with recognizers based on Gaussian Mixture Models (GMM) and Support Vector Machine (SVM). Thus, for comparison, the input parameters for the three recognizers were the same, that is, the elements $C_{kn}$ for each pattern $j$. In Tables X and XI test results are presented for female and male speakers, respectively [12], [13].

TABLE X
COMPARISON OF SPEECH RECOGNITION RESULTS USING NEURAL NETWORKS, SVM E GMM FOR FEMALE SPEAKERS

| | | Order of Matrix = 2 | Order of Matrix = 3 | Order of Matrix = 4 |
|---|---|---|---|---|
| **Loc_F1** | LVQ | 91 | 88 | 94 |
| | MLP | 94 | 90 | 95 |
| | SVM-Poli | 68 | 62 | 65 |
| | SVM-RBF | 74 | 76 | 78 |
| | GMM | 92 | 84 | 88 |
| **Loc_F2** | LVQ | 90 | 86 | 95 |
| | MLP | 95 | 94 | 91 |
| | SVM-Poli | 65 | 65 | 66 |
| | SVM-RBF | 80 | 80 | 80 |
| | GMM | 94 | 89 | 82 |
| **Loc_F3** | LVQ | 80 | 81 | 90 |
| | MLP | 83 | 93 | 97 |
| | SVM-Poli | 60 | 60 | 77 |
| | SVM-RBF | 78 | 78 | 80 |
| | GMM | 88 | 88 | 95 |
| **Loc_F4** | LVQ | 84 | 82 | 96 |
| | MLP | 91 | 91 | 90 |
| | SVM-Poli | 66 | 72 | 72 |
| | SVM-RBF | 78 | 72 | 82 |
| | GMM | 70 | 72 | 83 |
| **Loc_F5** | LVQ | 79 | 90 | 99 |
| | MLP | 95 | 89 | 97 |
| | SVM-Poli | 66 | 72 | 72 |
| | SVM-RBF | 78 | 72 | 82 |
| | GMM | 70 | 72 | 83 |

TABLE XI
COMPARISON OF SPEECH RECOGNITION RESULTS USING NEURAL NETWORKS, SVM E GMM FOR MALE SPEAKERS

| | | Order of Matrix = 2 | Order of Matrix = 3 | Order of Matrix = 4 |
|---|---|---|---|---|
| **Loc_M1** | LVQ | 77 | 81 | 87 |
| | MLP | 89 | 82 | 91 |
| | SVM-Poli | 70 | 66 | 73 |
| | SVM-RBF | 76 | 80 | 80 |
| | GMM | 57 | 64 | 72 |
| **Loc_M2** | LVQ | 77 | 80 | 85 |
| | MLP | 86 | 89 | 90 |
| | SVM-Poli | 67 | 63 | 71 |
| | SVM-RBF | 76 | 63 | 81 |
| | GMM | 80 | 87 | 91 |
| **Loc_M3** | LVQ | 77 | 80 | 88 |
| | MLP | 82 | 88 | 85 |
| | SVM-Poli | 62 | 63 | 70 |
| | SVM-RBF | 78 | 80 | 78 |
| | GMM | 52 | 67 | 77 |
| **Loc_M4** | LVQ | 77 | 64 | 74 |
| | MLP | 72 | 65 | 59 |
| | SVM-Poli | 68 | 63 | 69 |
| | SVM-RBF | 76 | 80 | 80 |
| | GMM | 66 | 70 | 71 |
| **Loc_M5** | LVQ | 79 | 77 | 90 |
| | MLP | 87 | 82 | 93 |
| | SVM-Poli | 66 | 66 | 74 |
| | SVM-RBF | 76 | 80 | 82 |
| | GMM | 72 | 74 | 86 |

## V. Conclusion

According presented results, it is concluded that the MLP and LVQ configurations were able to extract the features of two-dimensional time DCT matrices of low order provided as patterns of digits to be classified. Both configurations were able to present considerable performance with neural topologies reduced in the testing phase. They showed high generalization capacity when patterns provided are distinct from those used in the training phase. The parametrization of the speech signal generated by two-dimensional time matrix through the mel-cepstral coefficients and DCT, proposed in methodology, was efficient in forming the set of input patterns presented to neural network during the training and validation phase. An important point verified in this paper is the influence of the values of initial weights in the result achieved by MLP Neural Network. The random initialization of the set of weights can direct the MLP Network to local minimum values in the surface of the cost function that is not the most appropriate solution. In addition, the convergence time and generalization of the neural network are compromised. Comparison of results for the speech recognition system using MLP and LVQ Neural Networks with other methodologies, such as SVM-polynomial, the SVM-RBF and the GMM, showed that the proposed methodology in this paper has good performance in solving of the problem in question, becoming feasible to use in speech recognition systems.This approach can be extended to other languages. For this to be done, a database of speech in the language of interest is necessary for the methodology presented is applied.

## Acknowledgment

The authors would like to thank the Federal Institute of Maranhão (IFMA) for providing the infrastructure of the Digital Systems Laboratory to carry out this research and obtaining the experimental results and Electronic Instrumentation and Applied Technology Research Group from IFMA.

## References

[1] S. A. Cardoso, J. E. C. Castanho, M. N. Franchin, and I. R. Fontes, "Sesame: sistema de reconhecimento de comandos de voz utilizando pds e rna," in *Anais do XVIII Congresso Brasileiro de Automática*, Mato Grosso, Brazil, 2010, pp. 1316–1323.

[2] L. Weifeng, Z. Yicong, N. Poh, Z. Fei, and L.Qingmin, "Feature denoising using joint sparse representation for in-car speech recognition," *Signal Processing Letters, IEEE*, vol. 20, no. 7, pp. 681–684, 2013.

[3] Y.-M. Koo, J.-S. Yang, M.-Y. Park, E.-U. Kang, W.-J. Hwang, W.-S. Lee, and S.-H. Han, "An intelligent motion control of two wheel driving robot based voice recognition," in *Control, Automation and Systems (ICCAS), 2014 14th International Conference on*, Seoul,South Korea, Oct 2014, pp. 313–315.

[4] A.Cubukcu, M. Kuncan, K. Kaplan, and H. M. Ertunc, "Development of a voice-controlled home automation using zigbee module," in *Signal Processing and Communications Applications Conference (SIU), 2015 23th*, May 2015, pp. 1801–1804.

[5] W.L.S.Silva, "Sistema de inferência genético-nebuloso para reconhecimento de voz: uma abordagem em modelos preditivos de baixa ordem utilizando a transformada cosseno discreta," Ph.D. dissertation, Univ. Federal do Maranhão, São Luís, march 2015.

[6] C.M.Bishop, *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.

[7] A. P. Braga, *Redes neurais artificiais: teoria e aplicações*. LTC Editora, 2007.

[8] I. SILVA, D.H.SPATTI, and R.A.FLAUZINO, *Redes Neurais Artificiais para Engenharia e Ciências Aplicadas- Curso Prático*. ARTLIBER, 2010.

[9] S.Katagiri, *Handbook of Neural Networks for Speech Processing*. Artech House, 2000.

[10] L.V.Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*, ser. Prentice-Hall international editions. Prentice-Hall, 1994.

[11] S.S.HAYKIN, *Redes Neurais*. BOOKMAN COMPANHIA ED, 2001.

[12] G.C.Batista and W.L.S.Silva, "Application of support vector machines to recognize speech patterns of numeric digits," in *Natural Computation (ICNC), 2015 11th International Conference on*, Aug 2015, pp. 831–836.

[13] A.A.V.Beserra, W.L.S.Silva, and G.L.O.Serra, "A gmm/cpso speech recognition system," in *Industrial Electronics (ISIE), 2015 IEEE 24th International Symposium on*, June 2015, pp. 26–31.