A NOVEL INTELLIGENT METHODOLOGY FOR SPEECH RECOGNITION

Washington Luis Santos Silva*, Ginalber Luiz de Oliveira Serra†

*Instituto Federal de Educação, Ciência e Tecnologia do Maranhão, Departamento de Eletro-Eletrônica, Laboratório de Inteligência Computacional Aplicada à Tecnologia.
AV. Getúlio Vargas, nº 04 - Monte Castelo
São Luis, Maranhão, Brasil

Emails: washington.wlss@ifma.edu.br, ginalber@ifma.edu.br

Abstract— The concept of fuzzy sets and fuzzy logic is widely used to propose of several methods applied to systems modeling, classification and pattern recognition problem. This paper proposes a genetic-fuzzy recognition system for speech recognition. In addition to pre-processing, with mel-cepstral coefficients, the Discrete Cosine Transform (DCT) is used to generate a two-dimensional time matrix for each pattern to be recognized. A genetic algorithms is used to optimize a Mamdani fuzzy inference system in order to obtain the best model for final recognition. The speech recognition system used in this paper was named Intelligent Methodology for Speech Recognition(**IMSR**). Experimental results for speech recognition applied to brazilian language show the efficiency of the proposed methodology compared to methodologies widely used and cited in the literature.

Keywords— Fuzzy Systems; Automatic Speech Recognition; Genetic Algorithms; Discrete Cosine Transform; Instelligent System.

Resumo— O conceito de conjuntos nebulosos e lógica nebulosa é largamente utilizado no desenvolvimento de diversos métodos aplicados a sistemas de modelageme problemas e classificação e reconhecimento de padrões. Este artigo propõem uma metodologia Genético-Nebulosa para reconhecimento de padrões. Somando-se ao pré-processamentos com coeficientes mel-cepstrais, utiliza-se a Tranformada Cosseno Discreta (TCD) para gerar uma matriz temporal bidimensional para cada padrão a ser reconhecido. O algoritmo genético é utilizado para otimizar o sistema de inferência nebuloso do tipo Mamdani com o objetivo de melhorar o modelo final maximizando a quantidade de acertos de reconhecedor. O sistema de reconhecimento de voz usado neste artigo denomina-se '*Intelligent Methodology for Speech Recognition*'. Resultados experimentais do sistema proposto com os dígito da língua portuguesa mostram a eficiência do metodologia proposta quando comparada a outras metodologias largamente usadas e citadas na literatura.

Keywords— Sistemas Nebulosos, Reconhecimento Automático de Voz; Algoritmos Gnéticos; Transformada Cosseno Discreta; Sistemas Inteligentes.

1 Introduction

The goal of an Automatic Speech Recognition System (ASR) is to accurately and efficiently convert a speech signal into a mathematic coding of the spoken words, independent of the device used to record the speech (i.e., the transducer or microphone), the speakers accent, or the acoustic environment in which the speaker is located (e.g., quiet office, noisy room, outdoors). That is, the ultimate goal, which has not yet been achieved, is to perform as well as a human listener.

Parameterization of an analog speech signal is the first step in speech recognition process. Several popular signal analysis techniques have emerged as standards in the literature. These algorithms are intended to produce a perceptually meaningful parametric representation of the speech signal: parameters that can emulate some behavior observed in human auditory and perceptual systems. Actually, these algorithms are also designed to maximize recognition performance (Picone, 1991).

The selection of the best representation for parametric speech signal is a very important task of developing any speech recognition system. The goal of selecting the best way to encode the signal is to compress the speech data information, eliminating non-phonetic analysis of the signal and improving the aspects of the signal which contribute significantly to detect phonetic differences of speech sounds . We follow here the formalism developed by Andrews (Andrews, 1971). The problem of pattern recognition might be formulated as follows: Let S_k classes, where k = 1,2,3...K, and $S_k \subset \Re^n$. If any pattern space is take with dimension $\Re^x$, where $x \leq n$, it should transform this space into a new pattern space with dimension $\Re^a$, where $a < x \leq n$. Then assuming a statistical measure or second order model for each S_k, through a covariance function represented by $\left[\Phi_x^{(k)}\right]$, the covariance matrix of the general pattern recognition problem becomes:

$$[\Phi_x] = \sum_{k=1}^{K} P(S_k) \left[\Phi_x^{(k)}\right] \quad (1)$$

where $P(S_k)$ is a distribution function of the class S_k, *a priori*, with $0 \leq P(S_k) \leq 1$. A linear transformation operator through the matrix $\mathbf{A}$ maps the pattern space in a transformed space whose columns are orthogonal basis vectors of this matrix $\mathbf{A}$. The patterns of the new space are linear combinations of the original axes as structure of the matrix $\mathbf{A}$. The statistics of second order in

the transformed space are given by:

$$\Phi_{\mathbf{A}} = \mathbf{A}^T[\Phi_x]\mathbf{A} \quad (2)$$

where $\Phi_{\mathbf{A}}$ is the covariance matrix which corresponds to the space generated by the matrix $\mathbf{A}$ and the operator $[\cdot]^T$ corresponds to the transpose of a matrix. Thus, it can extract features that provide greater discriminatory power for classification from the dimension of the space generated (Andrews, 1971).

In this proposal, a speech signal is encoded and parameterized in a two-dimensional time matrix with four parameters of the speech signal. After coding, the mean and variance of each pattern are used to generate the rule base of Mamdani fuzzy inference system. The mean and variance are optimized using genetic algorithm in order to have the best performance of the recognition system. This paper consider as patterns the brazilian locutions (digits): $'0', '1', '2', '3', '4', '5', '6', '7', '8', '9'$.

2 A Hybrid-Intelligent Methodology for Speech Recognition

The proposed Recognition System Block Diagram is depicted in Fig.1.

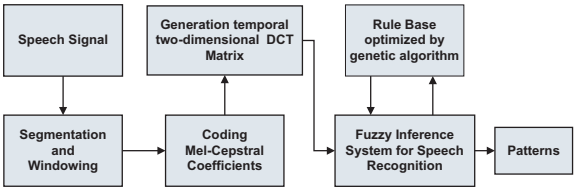

Figura 1: Block diagram of the proposed recongnition system.

2.1 Pre-processing Speech Signal

There is no standard set of features for speech recognition. Instead, various combinations of acoustic, articulatory, and auditory features have been utilized in a range of speech recognition systems. The most popular acoustic features have been the (LPC-derived) mel-frequency cepstrum coefficients and their derivatives. Initially, the speech signal is digitizing, divided in segments, windowed and encoded in a set of parameters defined by the order of mel- frenquency cepstrum coefficients (MFCC). The DCT coefficients are computed and the two-dimensional time DCT matrix is generated, based on each speech signal to be recognized.

2.2 Two-Dimensional Time Matrix DCT Coding

In the encoding stage is used a DCT feature extractor to remove unwanted additional data from the speech samples. Thus, the speech frames are transformed into a DCT space. The DCT-II-E coefficients, as defined in (Chen and Zhou, 2009), and used in this paper, are computed by following relation

$$X(k) = \sum_{n=0}^{N-1} \alpha(n)x(n)cos\frac{(2k+1)n\pi}{2N} \quad (3)$$

$k = 0, 1, 2, ..., N-1$, $n = 0, 1, 2, ..., N-1$, $x(n)$ is a vector of real numbers and

$$\begin{cases} \alpha(n) = \sqrt{1/N}, & \text{if } n = 0 \\ \alpha(n) = \sqrt{2/N}, & \text{else} \end{cases}$$

The two-dimensional time matrix, as the result of DCT in a sequence of T observation vector mel-cepstral coefficients observation vectors on the time axis, is given by:

$$C_k(n,T) = \frac{1}{N}\sum_{t=1}^{T} mfcc_k(t)cos\frac{(2t-1)n\pi}{2T} \quad (4)$$

where $mfcc(.)$ are the mel-frequency cepstral coefficients, $k, 1 \leq k \leq K$, is the k-th (line) component of t-th frame of the matrix and $n, 1 \leq n \leq N$ (column) is the order of DCT. Thus, the two-dimensional time matrix, where the interesting low-order coefficients k and n that encode the long-term variations of the spectral envelope of the speech signal is obtained. Thus, the two-dimensional time matrix $C_k(n,T)$ for each input speech signal. For simplification $C_k(n,T)$ will be represented by C_{kn}. The elements of the matrix are obtained as follows:

1. For a given spoken word P (digit), ten examples of utterances of P are gotten. This way it has itself P_0^0,P_1^0,...,P_9^0, P_0^1, P_1^1...,P_9^1,P_0^2,P_1^2,...,P_9^2,...,P_m^j, where $j \in \{0, 1, 2, ..., 9\}$ is the pattern to be recognized and $m \in \{0, 1, 2, ..., 9\}$ is an exemple of the pattern to be recognized.

2. Each frame of a given example of the word P generates a total of K mel-cepstral coefficients and the significant features are taken for each frame along time. The *N-th* order DCT is computed for each mel-cepstral coefficient of same order within the frames distributed along the time axis, i.e., c_1 of the frame t_1, c_1 of the frame $t_2, ...$, c_1 of the frame t_T, c_2 of the frame t_1, c_2 of the frame $t_2, ...,$, c_2 of the frame t_T, and so on, generating elements $\{c_{11}, c_{12}, c_{13}, ..., c_{1N}\}$, $\{c_{21}, c_{22}, c_{23}, ..., c_{2N}\}$, $\{c_{K1}, c_{K2}, c_{K3}, ..., c_{KN}\}$ of the matrix given in equation (4). Therefore, a two-dimensional time matrix DCT is generated for each example of the word P.

3. Finally, the matrices of mean CM_{kn}^j (5) and variances CV_{kn}^j(6) are generated. The para-

meters of CM_{kn}^{j} and CV_{kn}^{j} are used to produce Gaussians matrices C_{kn}^{j} which will be used as fundamental information for implementation of the fuzzy recognition system. The parameters of this matrix will be optimized by genetic algorithm.

$$CM_{kn}^{j} = \frac{1}{M}\sum_{m=0}^{M-1} C_{kn}^{jm} \tag{5}$$

$$CV_{kn}^{j}(var) = \frac{1}{M-1}\sum_{m=0}^{M-1}\left[C_{kn}^{jm} - \left(\frac{1}{M}\sum_{m=0}^{M-1} C_{kn}^{jm}\right)\right]^{2} \tag{6}$$

2.3 Generation of Fuzzy Patterns

The elements of the matrix C_{kn}^{j} were used to generate gaussians membership functions in the process of fuzzification. For each trained model j the gaussians memberships functions $\mu_{c_{kn}^{j}}$ are generated, corresponding to the elements c_{kn}^{j} of the two-dimensional time matrix $\mathbf{C}_{kn}^{j}$ with $j = 0.1, 2, 3, 4, 5, 6, 7, 8, 9$, where j is the model used in training. The training system for generation of fuzzy patterns is based on the encoding of the speech signal ***s(t)***, generating the parameters of the matrix C_{kn}^{j}. Then, these parameters are fuzzified, and they are related to properly fuzzified output y^{j} by the relational implications, generating a relational surface $\mu_{(Ru)}$, given by:

$$\mu_{Ru} = \mu_{c_{kn}^{j}} \circ \mu_{y^{j}} \tag{7}$$

This relational surface is the fuzzy system rule base for recognition optimized by genetic algorithm to maximize the speech recognition.

2.4 Fuzzy Inference System for Speech Recognition Decision

The decision phase is performed by a fuzzy inference system based on the set of rules obtained from the mean and variance matrices of C_{kn} of all $j-th$ spoken digit. In this paper, a matrix with minimum number of parameters (2×2) is used in order to allow a satisfactory performance compared to pattern recognizers available in the literature. The elements of the matrices C_{kn}^{j} are used by the fuzzy inference system to generate four gaussian membership functions corresponding to each element $c_{kn}^{j}\big|^{k=1,2;n=1,2}$ of the matrix. The set of rules of the fuzzy relation is given by:

Rule Bases

$$\textbf{\textit{IF}}\ \ c_{kn}^{j}\big|^{k=1,2;n=1,2}\ \ \textbf{\textit{THEN}}\ \ y^{j} \tag{8}$$

Modus Ponens

$$\textbf{\textit{IF}}\ \ c_{kn}^{'j}\big|^{k=1,2;n=1,2}\ \ \textbf{\textit{THEN}}\ \ y^{'j} \tag{9}$$

From the set of rules of the fuzzy relation between antecedent and consequent, a data matrix for the given implication is obtained. After the training process, the relational surfaces is generated based on the rule base and implication method (Zhou and Khotand, 2007). The speech signal is encoded to be recognized and their parameters are evaluated in relation to the functions of each patterns on the surfaces and the degree of membership is obtained. The final decision for the pattern is taken according to the $max - min$ composition between the input parameters and the data contained in the relational surfaces. The process of defuzzification for the pattern recognition is based on the ***mean of maxima (mom)*** method given by:

$$\mu_{y'^{j}} = \mu_{c_{kn}^{'j}} \circ \mu_{(Ru)} \tag{10}$$

$$y^{'} = mom(\mu_{y'^{j}}) = mean\{y|\mu_{y'^{j}} = max_{y\in Y}(\mu_{y'^{j}})\} \tag{11}$$

2.5 Optimization of Relational Surface with Genetic Algorithm

The continuous genetic algorithm is configured with a population size of 100, generations of 300, with mutations probability of 15% and two individuals (chromosomes) with 40 genes each, to optimize a cost function with 80 variables, which are the mean and variances of the patterns (digits) to be recognized by the proposed fuzzy recognition system. The genetic algorithm was used to optimize the variations of mean and variances of each pattern in order to maximize the successful recognition process. For each element of the matrix C_{kn}^{j} coefficients are determined with variations minimum and maximum, and the coefficient $c_{11} \in [c_{11}(min)\ c_{11}(max)]$, $c_{12} \in [c_{12}(min)\ c_{12}(max)]$, $c_{21} \in [c_{21}(min)\ c_{21}(max)]$, $c_{22} \in [c_{22}(min)\ c_{22}(max)]$. Thus, it has eight time varying parameters for each pattern which correspond to eighty parameters to be optimized by genetic algorithm (Weihong et al., 2010).

3 Experimental Results

3.1 System Training

The patterns to be used in the recognition process were obtained from ten speakers who are speaking the digits 0 to 9. After pre-processing of the speech signal and fuzzification of the matrix C_{kn}^{j}, its fuzzifieds components $\mu_{c_{kn}^{j}}$ had been optimized by the GA that maximize the total of successful recognition. The optimization process was performed with 50 realizations of the genetic algorithm.The best result of the recognition processing is shown in Fig.2. The total number of hits using GA was 94 digits correctly identified in the training process from a total of 100 spoken digits. The relational surface generated for this result was used for validation process.

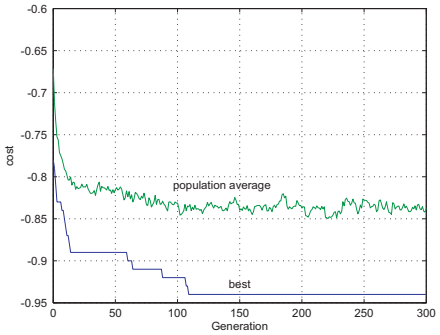

Figura 2: Plot of the best results obtained in the training process.

The best individual in the first generation is shown in Fig.3. In this case the total number of correct answers was 46 digits correctly identified. The relational surface of the best individual in the first generation is shows in Fig.4.

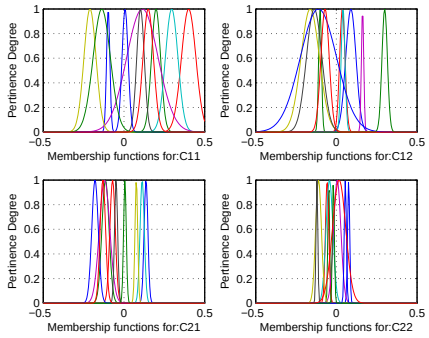

Figura 3: Membership functions for c_{kn}^{j} in the 1st generation.

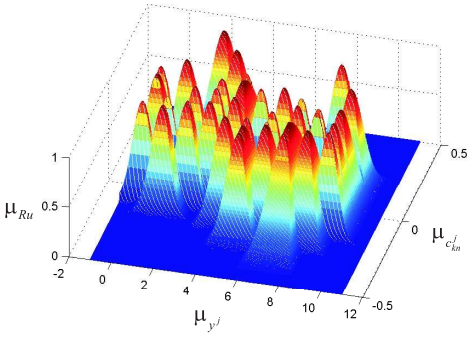

Figura 4: Relational surface (μ_{Ru}) in the 1st generation.

The optimum individual, presents the features in Fig.5 and Fig.6.

3.2 System Test - Validation

In this step, 100 locutions uttered in a room with controlled noise level and 500 locutions uttered in an environment without any kind of noise control

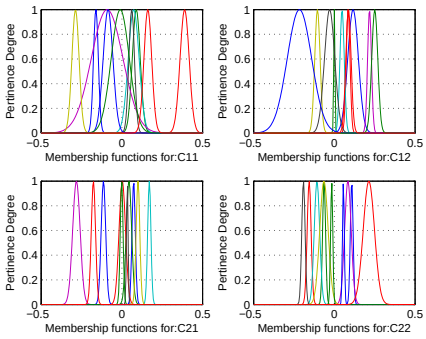

Figura 5: Membership functions for c_{kn}^{j} optimized by GA.

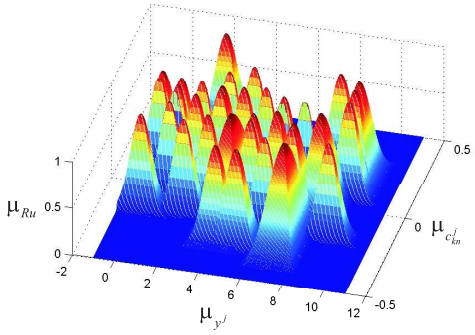

Figura 6: Relational surface (μ_{Ru}) optimized by GA.

were used. For every ten examples of each spoken digit, two-dimensional time matrix cepstral coefficients $\mathbf{C}_{kn}^{j}$ was generated and they were used in the test procedure in which six types of tests were performed: Training: Recognition Optimized by IMSR (5 Female and 5 Male Speakers);
TEST 1: Validation - Strictly speaker dependent recognition, in which the words used for training and testing were spoken by a same group of 10 speakers (5 Female and 5 Male Speakers);
TEST 2: Validation test- Recognition based on the partial dependence of the speaker with 20% of the digits spoken (Female Speaker);
TEST 3: Validation test- Recognition based on the partial dependence of the speaker with 20% of the digits spoken two examples (Male Speaker);
TEST 4: Validation test- Recognition independent of the Speaker, where the speaker does not have influence in the training process (Female Speaker);
TEST 5: Validation test- Recognition independent of the Speaker, where the speaker does not have influence in the training process (Male Speaker);

In the Figures 7 to 12 presents the comparative analysis of the HMM with two, three and four states, two, three and four gaussians mixtures by state and order analysis equal 12, i.e., the number of mel-cepstral parameter equal 12 to HMM and the Intelligent Methodology for Speech Recogni-

tion(**IMSR**) with two, three and four mel-cepstral parameter for speech recognition. With the data points obtained experimentally, a fit curve for all tests was mapped, and the amount of parameters needed to obtain 100% accuracy is estimated by these curves with two tested recognizers. As depicted in Fig.7, with two mel-cepstral parameters, a number of hits equal to 94% is obtained. With three parameters to 98% and 99% with a total of four mel-cepstral parameters. Thus, through the fit curve, a total of 100 % accuracy can be reached, since that tuned properly the parameters of the genetic algorithm. In the Fig.8 and Fig.9 is shown an estimate of 100% accuracy with approximately 5 mel-cepstral parameters. In the Fig.10 is shown an estimate of 100% accuracy with a total of 7 mel-cepstral parameters. It is noteworthy that these results are strictly speaker dependent (Fig.8) and with partial speaker dependence, respectively (Fig.9 and Fig.10). In the Fig.11 and Fig.12, where tests are independent of the speaker gets a higher estimate of the number of mel-cepstral parameters 7 and 16, respectively, to reach the 100% accuracy.

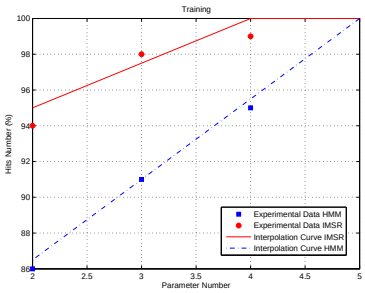

Figura 7: Results in the training.

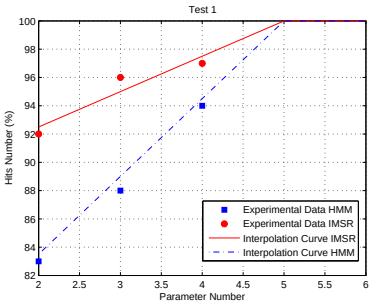

Figura 8: Validation Test 1.

Experimental results and mel-cepstral parameter were presented in Table 1 for tests performed in this paper.From data obtained, it is observed that even with a lower number of parameters, resulting from the encoding of the speech signal, similar results are obtained and are compared with methodologies more complex with a larger number of parameter.

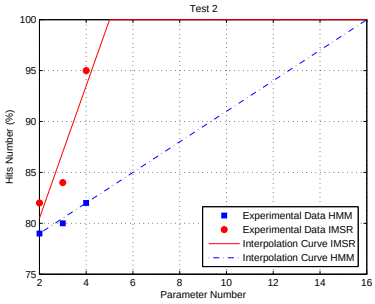

Figura 9: Validation Test 2.

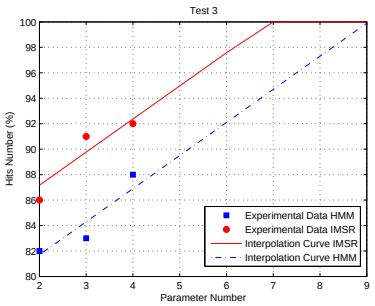

Figura 10: Validation Test 3.

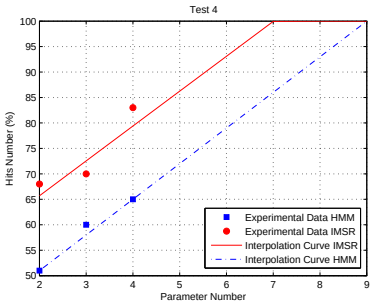

Figura 11: Validation Test 4.

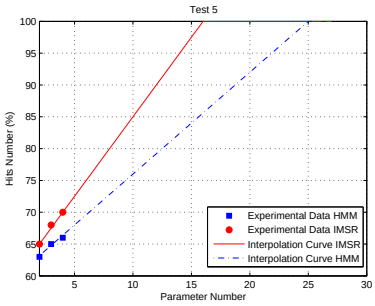

Figura 12: Validation Test 5.

Tabela 1: Results for Proposed Methodology

IMSR Parameters	IMSR Traning	IMSR Test-01	IMSR Test-02	IMSR Test-03	IMSR Test-04	IMSR Test-05
MFCC=2, DCT=2	94	92	82	86	68	65
MFCC=3, DCT=3	98	96	84	91	70	68
MFCC=4, DCT=4	99	97	95	92	84	70
(Estimate)MFCC=4, DCT=4	100	-	-	-	-	-
(Estimate)MFCC=5, DCT=5	-	100	-	-	-	-
(Estimate)MFCC=5, DCT=5	-	-	100	-	-	-
(Estimate)MFCC=7, DCT=7	-	-	-	100	-	-
(Estimate)MFCC=7, DCT=7	-	-	-	-	100	-
(Estimate)MFCC=16, DCT=16	-	-	-	-	-	100

3.3 Comparison with other intelligent methods

The goal with this comparison is to show that the proposed method, even with a minimal num-

ber of input parameters, produces similar results comparable with other intelligent techniques with a substantial amount of input data. For this comparison the results presented in the article (Silva et al., 2010) were used. The work cited (Silva et al., 2010) use the same patterns used in proposed work, i.e, brazilian locutions (digits). The database used in (Silva et al., 2010), consists of spoken digits in portuguese collected during a period of three months, from eighty-two men aged between 18 and 42 years-old. The sampling rate of the recording is 22050Hz. Altogether, the database has 216 sequences of 10 digits (0 - 9) each, totalling 10 classes and 2160 examples. Thus, it is a balanced dataset considering the class distribution. That it uses a MFCC with 13 coefficients and line spectral frequencies with orders 24 and 48. In the proposed methodology, it is use a IMSR com order 2, 3 and 4, with a amount of 4, 9 and 16 input data for the recognition. In the Table 2 there are describ scenarios used by (Silva et al., 2010) and in the Table 3 there are described the obtained results in the work perfomed by (Silva et al., 2010) for comparasion.

Tabela 2: Description of the scenarios of the experiments performed by (Silva et al., 2010)

Scenario	Inducer/Settings
1-NN	Nearest Neibor
5-NN	5- Nearest Neighbor weighted by inverse distance (one)
7-NN	7- Nearest Neighbor weighted by inverse distance
9-NN	9- Nearest Neighbor weighted by inverse distance
SVM-Poly1	Suport Vector Machine with Polynomial Kernel with Degree 1
SVM-Poly2	Suport Vector Machine with Polynomial Kernel with Degree 2
SVM-Poly3	Suport Vector Machine with Polynomial Kernel with Degree 3
SVM-RBF0.01	Suport Vector Machine with RBF Kernel with Gamma =0.01
SVM-RBF0.05	Suport Vector Machine with RBF Kernel with Gamma =0.05
SVM-RBF0.1	Suport Vector Machine with RBF Kernel with Gamma =0.1
NB	Naive Bayes
RF	Random Forest

Tabela 3: Mean accuracy for three analyzed methods on 12 scenario performed by (Silva et al., 2010)

Scenario	MFCC	24LSF	48 LSF
1-NN	86.33	92.92	93.03
5-NN	89.52	95.57	95.66
7-NN	89.61	95.82	95.98
9-NN	90.20	96.13	95.67
SVM-Poly1	97.96	98.85	99.30
SVM-Poly2	97.88	98.77	99.31
SVM-Poly3	97.91	98.75	99.17
SVM-RBF0.01	93.62	97.93	98.64
SVM-RBF0.05	96.88	98.54	98.70
SVM-RBF0.1	97.19	98.32	98.02
NB	90.63	94.86	94.72
RF	91.83	96.36	95.89

4 Conclusion

Evaluating the results, it is observed that the proposed speech recognizer, even with a minimal number of parameters in the generated patterns, was reliably able to extract the temporal characteristics of the speech signal and produce good recognition results compared with the traditional HMM. To obtain equivalent results with HMM is necessary to increase the state number and/or mixture number. In tests performed by the authors it was verified that an increase in the order of the analysis above 12 does not improve significantly the performance of HMM. Other intelligent techniques, such as neural networks, svm, or hybrid intelligent techniques, can achieve good results in speech recognition, however, they usually suffer from the curse of dimensionality, with a high number of parameters and a heavy computational load. The proposed methodology can work with a small number of parameters, maintaining a reasonable number of hits, which indicates its ability to discard redundant information not necessary to the process of recognition. It is believed that with proper treatment of the signal to noise ratio in the process of training and testing, the proposed speech recognizer may improve its performance: Increase the speech bank with different accents, the use Nonlinear Predictive Coding for feature extraction in speech recognition, the use digital filter in the speech signal to be recognized and increase the parameters number used.

Acknowledgment

The authors would like to thank FAPEMA, research group of computational intelligence applied to technology at the Federal Institute of Education, Science and Technology of the Maranhão and Master and PhD program in Eletrical Engineering at the Federal University of Maranhão (UFMA).

Referências

Andrews, H. (1971). *Multidimensional Rotations in Feature Selection*, IEEE Transaction on Computers.

Chen, P. and Zhou, J. (2009). *Generalized Discrete Cosine Transform*, Pacific-Asia Conference on Circuits,Communications and System.

Picone, J. (1991). *Signal modeling techiniques in speech recognition*, vol.79, 2ed. edn, IEEE Transactions on Computer.

Silva, D., de Souza, V., Batista, G. and Giusti, R. (2010). *Spoken Digit Recognition in Portuguese Using Line Spectral Frequencies*, vol.45, nº01 edn, Lectures Notes in Artificial Intelligence-LNAI 7637. Springer-Verlag Berlin Heidelberg.

Weihong, Z., Shunqing, X. and Ting, M. (2010). *A Fuzzy Classifier Based on Mamdani Fuzzy Logic System and Genetic Algorithm*, IEEE Conference on Information Computing and Telecommunications.

Zhou, E. and Khotand, A. (2007). *Fuzzy Classifier Design Genetic Algorithms*, vol.18, nº03 edn, Pattern Recognition Journal-Elsevier.