Support Vector Machines and Mel-Frequency Cepstral Coefficients: an Application for Automatic Voice Recognition

F. G. Barbosa[1] **and W. L. S. Serra**[1]
[1]Department of Eletroelectronics, Federal Institute of Maranhao, Sao Luis, Maranhao, Brazil

Abstract— *The speech recognition problem can be modeled as a classification problem, where one wants to get the best degree of separability between classes representing the voice. In order to apply that concept to build an automated speech recognition system capable of identifying the speaker, many techniques using artificial intelligence and general classification have been developed, which lead to this paper. Here we propose a voice recognition method to recognize keywords in brazilian portuguese for biometric purpose using multiple Support Vector Machines, which builds a hyperplane that separates Mel Frequency Cepstral Coefficients, the MFCC's, for later classification of new data. With a small dataset the system was able to correctly identify the speaker in all cases, having great precision on the task. The machines are based on the Radial Basis Function kernel, the RBF, but were tested with severe different kernels, having also a good precision with the linear one.*

Keywords: Voice Recognition; Support Vector Machines; MFCCs; Brazilian Portuguese.

1. Introduction

The foundation of Support Vector Machines (SVM) was developed by Vapnik[1] and earned a lot of popularity due its promissing characteristics, with better empirical performance. The mathematical formulation uses the *Structural Risk Minimization* (SRM), that has shown itself superior to the *Empirical Risk Minimization* (ERM), used by conventional Neural Nets. SRM minimizes an upper limit over the expected risk, while ERM minimizes the error on the training data. This is the difference that leads SVM to have greater generalization capacity, which is the goal of statistical learning.

SVMs were developed to solve the classification problem, but, recently, have been applied to solve regression ones[2]. The classifiers generated by a *Support Vector Machine* achive good results in general, having that capacity of generalization measured by their eficiency on classifying data that does not belong to the training data set. The main idea of a SVM is building a hyperplane, that is a separating surface, as decision bounds in which the separation of the dichotomous examples is maximum. It is important to highlight from the *Statistical Learning Theory*, that a good classifier accounts all the data set but abstains from the particular cases, and as SVMs constitutes a learning technique, which has been getting attention from the science community because it obtains results comparable to, or even better than *Articial Neural Nets*, a lot of sucessful examples can be mentioned on many fields, like categorizing text[3], image analysis[4], [5] and bioinformatics[6], [7]. Due to its efficiency in working with highdimensional data, it is cited in the literature as a highly robust technique[8].

The Theory of Statistical Learning aims to establish mathematical conditions that allow the selection of a classifier with good performance for the data set available for training and testing. In other words, this theory seeks to find a good classifier with good generalization regarding the entire data set. But, this classifier abstains from particular cases, which defines the capability to correctly predict the class of new data from the same domain in which the learning occured. Machines Learning (ML) employs an inference principle called induction, in which general conclusions are obtained from a particular set of examples. A model of supervised learning based on Theory of Statistical Learning is given in Fig. 1.

The voice recognition, and mostly the automatic one, has been the main goal of many scientists and reseachers for almost five decades, and has inspired many wonders in science fiction. Despite all the glamour around this subjects, and even with many intelligent machines that are able to recognize words and understand its meaning, we're still far from achieving the desirable one, that could understand any speech, independently on the speaker or language spoken, in a noise filled environment[9]. In truth, the actual situation is another one: to recognize a simple word or phrase, one needs an absurd computacional effort, where many areas of knowledge are used on training and final recognition. In order to solve this problem and facilitate the speech recognition, the Support Vector Machines technique is used in large scale on pattern classification , and as for this paper, we propose a biometrical voice recogntion automated system, in which we train ten samples of the digits, from '0' to '9', in brazilian portuguese, for each speaker, and later try to identify which one he is and which word he spoke.

2. Fundamentals of Support Vector Machines

The *Support Vector Machines* classification technique stands out by its strong theoretical fundamentation, having

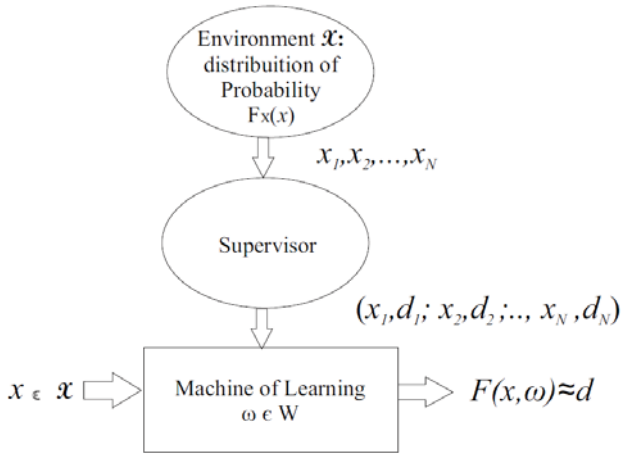

Fig. 1: Flowchart of a model of supervised learning.

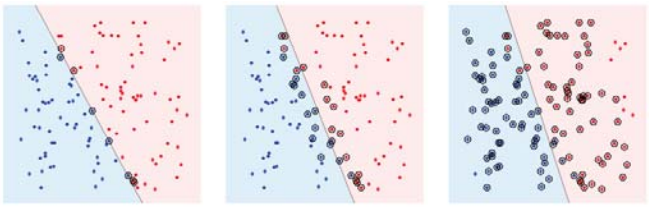

Fig. 2: Comparison of margins to find the optimum separating hyperplane.

on its core the Statitistical Learnig Theory (SLT) being this characteristic a differential above other techniques, as already said, Neural Nets. The ability to work with patterns of high dimensionality is another interesting circunstance of this technique, making it ideal to aplications where a noise data set is the one targetted.

SVM, as a supervised learning technique, can infere from a set of labeled examples, on which the class is known, a function capable of predicting new labels from unknown examples. The simpler derivation of the SVM algorithm is the linear function one, where to ilustrate the separation plane generated by it, we can drawn a line that represents the decision boundary that correctly classifies some data set, such as the one in Fig. 2.

A training dataset composed by two classes, where three decision functions realize the classification correctly between the blue squares and the red circles. However each function determines a different area and different quantity of support vectors for each class represented. Linear decision boundaries consist on a hyperplane (line in two dimensions, plane in 3 dimensions, ...), that separates two regions of the space in question. Such function $g(x)$ can be represented by a mathematical function of vector x e could assume the values $+1$ ou -1.

Support vector machines solve nonlinear problems by transforming the input feature vectors into a dimensionally higher hyperplane, where the linear separation becomes possible. Maximum discrimination is obtained with an optimal placement of the separation plane between the borders of the two classes [10]. If we assume a set H of points $x_i \in R^d$ with $i = 1, 2, 3, \ldots, n$. Each one of the x_i belongs to either of two classes labeled $y_i \in \{-1, 1\}$. Establishing the equation of a hyperplane that divides H is the desired goal, and for this purpose we have some preliminary definitions. By taking the set H, if linearly separable, there exists $w \in R^d$ and $b \in R$ to satisfy

$$y_i \left(w \cdot x_i + b\right) \geq 1 \tag{1}$$

where $i = 1, 2, 3, \ldots, n$.
The pair (w, b)defines a hyperplane

$$\left(w \cdot x_i + b\right) = 0 \tag{2}$$

This defines a separating hyperplane, leading to the problem of finding the optimal separating hyperplane, to which we try to minimize w as the following

$$min \frac{1}{2} \parallel w \parallel^2 \tag{3}$$

where $y_i \left(w \cdot x_i + b\right) \geq 1$.
Then converted to a dual problem by Lagrange multiplies

$$max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \left(\mathbf{X_i} \cdot \mathbf{X_j}\right) \tag{4}$$

where $\sum_{i=1}^{N} \alpha_i y_i = 0, \alpha_i > 0$.
When H cannot be separated linearly, nonnegative slack factor $\xi = (\xi_1, \xi_2, \ldots, \xi_N)$ is introduced. There is

$$y_i \left(w \cdot x_i + b\right) \geq 1 - \xi_i \tag{5}$$

The optimal problem can be described as

$$max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \left(\mathbf{X_i} \cdot \mathbf{X_j}\right) \tag{6}$$

where $\sum_{i=1}^{N} \alpha_i y_i = 0, i = 1, 2, \ldots, N,\ 0 \leq \alpha_i \leq C$.
This is the general form of SVM. If C tends to infinite, (6) becomes a linear separating problem, just like (4). Its a problem that can be solved by quadratic programming using sequential minimal optimization.

When the data is easily linearly separable, the previous equations are able to classify with minimum error, but when the data is highly nonlinear one needs to use the kernel method, in which the data is put in a higher dimensional plane, where it can be linearly separated. This is possible when we take the dot product of $X_i \cdot X_j$ and apply another function, validated by the Mercer's Conditions, that in some cases, like the *Radial Basis Function* (RBF), can place the data in a infinity dimensional space, where the data can easily be separated, for this reason it is the one used on this paper.

3. Voice Recognition for Biometrical Authentication

The voice exists for the human desire of verbalizing its thoughts, emotions and opinions, being part of our identity. It is one of the strongest extensions of our personality, and many times it's possible to recognize someone just by their voice.

From the beginning of its technological and intelectual development, the human beings intended to create machines that were able to produce and understand the human speech. Using voice to interact with automatic systems has a vast field of application. The combination with phone network allows remote access to databases and new services, like, for example, an e-mail check from anywhere on the globe and consultations of flight schedule without needing an operator.

Recently, several methods of Speech Recognition have been proposed using mel-frequency cepstral coefficients and Neural Networks Classifiers [11], [12], [13], Sparse Systems for Speech Recognition [14], Hybrid Robust Voice Activity Detection System [15], Wolof Speech Recognition with Limited vocabulary Based HMM and Toolkit [16], Real-Time Robust Speech Recognition using Compact Support Vector Machines.

On this context, the amazing advances in the last years, mostly in the microeletronics field, made possible to put into pratice this line of thought, effectively. At this point, one needs to adress the field of *Digital Signal Processing,* that is the core of many areas in science. From the engineering point of view, signals are functions or series used to carry information from a source to the recipient. The signals specific characteristics depend on the communication used for the transmission. They are processed on the transmission side to be produced and configured, and on the receptor they're decodified to extract the information contained, with maximum efficiency, if possible.

3.1 Digital Signal Processing

Method that consists in analysing real world signals (represented by a numerical sequence), extract its features through mathematical tools, in order to extract the essential information. There are many purposes on the matter such as biometric authentication, image processing and recognition and even preventing diseases. As for speech, subject of this paper, we follow basically three steps: sampling, followed by segmentation of words or phonemes[17] and short term analysis by Fourier transform or spectral analysis[18]. After this step, responsible for digital processing of the speech signal, we need to recognize and correctly classify a word, and for that there are some existing techniques, capable of extracting parameters based on a certain model and then applying a transformation to represent the signal in a more convenient form for recognition.

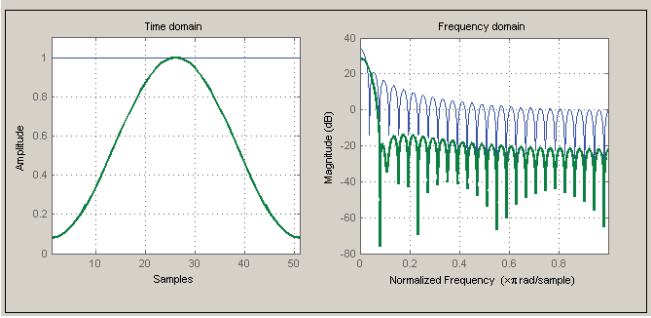

Fig. 3: Hamming Window and equivalent *SNR*

3.2 Pre-processing of the Speech Signal

The moment the segmentation of the speech is passed through the process of windowing, responsible for 'dividing' the signal with minimum power loss and noise, the speech signal is sampled and segmented into frames and is encoded in a set of mel-cepstral parameters. The number of parameters obtained is determined by the order of mel-cepstral coefficients. The obtained coefficients are then encoded by Discrete Cosine Transform (DCT) [18] in a two dimensional matrix that will represent the speech signal to be recognized. The process of windowing, hamming windowing for this case, in a given signal, aims to select a small portion of this signal, which will be analysed and named frame. A short-term Fourier analysis performed on these frames is called signal analysis frame by frame. The length of the frame T_f is defined as the length of time upon which a parameter set is valid. The term frame is used to determine the length of time between successive calculations of parameters. Normally, for speech processing, the time frame is between 10ms and 30ms[19]. There's also the superposition of the windowing, which determines where the window will start in order to reduce the power loss, initiating before the previous window reaches its end. Fig. 3 shows the plot of a hamming window in time and frequency domains.

3.3 SVMs and Biometry

Biometric authentication is any form of human biological measurement or metric that can be used to identify and authenticate an authorized user of a secure system. Biometric authentication- can include fingerprint, voice,S iris, facial, keystroke, and hand geometry[22]. Concerns on widespread use of biometric authentication systems are primarily centered around template security, revocability, and privacy[22]. The use of cryptographic primitives to bolster the authentication process can alleviate some of these concerns as shown by biometric cryptosystems[23]. Support Vector Machines or SVM is one of the most successful and powerful statistical learning classification techniques and it has been also implemented in the biometric field[24]. As for voice recognition, the technique has shown excellent results, hence not only it can generalize, but it can also restrict the

parameters if correctly made, leading to a great biometrical authentication voice based system.

4. Methodology

As a recognition default we proposed the classification and identification of the voice of a speaker by a keyword, in a text-dependent system. The speech signal is sampled and encoded in mel-cepstral coefficients and coefficients of Discrete Cosine Transform (DCT)[18] in order to parameterize the signal with a reduced number of parameters. Then, it generates two dimensional matrices referring to the Discrete Cossine Transform coefficients. The elements of these matrices representing two-dimensional temporal patterns will be classified by Support Vector Machines (SVMs)[20]. The innovation of this work is in the reduced number of parameters which lies in the SVM classifier and in the reduction of computational load caused by this reduction of parameters. The classification is made based o the *Radial Basis Function.*

4.1 The DCT matrix

After being properly parameterized in mel-cepstral coefficients, the signal is encoded by DCT performed in a sequence of T observation vectors of mel-cepstral coefficients on the time axis. The coding by DCT is given by the equation following:

$$C_k\left(n,T\right)=\frac{1}{N}\sum_{t=1}^{T}MFCC_k\left(t\right)cos\frac{\left(2t+1\right)n\pi}{2T} \qquad (7)$$

where k, $1\leq k\leq K$, refers to the k-th line (number of Mel frequency cepstral coefficients) of t-th segment of the matrix n, $1\leq n\leq N$ component refers to the n-th column (order of DCT), $MFCC_k(t)$ represents the mel-cepstral coefficients. Thus, one obtains the two-dimensional matrix that encode the long term variations of the spectral envelope of the speech signal [21]. This procedure is performed for each spoken word. Thus, there is a two-dimensional matrix $C_k(n,T)\equiv C_{k^n}$ for each input signal. The matrix elements are obtained as the following:

1) For a given model of spoken word W (digit), ten examples of this model are pronounced. Each example is properly divided into T frames distributed along the time axis. Thus, we have: P_i^j, where $i=0,1,2,...,9$ is the number of patterns to be recognized and $j=1,2,3,...,10$, is the number of samples to generate each pattern.

2) Each frame of a given example of model W generates a total of K mel-frequency cepstral coefficients, and then, significant characteristics are obtained within each frame over this time. The DCT of order N is then calculated for each mel-cepstral coefficient of the same order within the frame, that is, C_1 in the frame t_1, C_1 in the frame $t_2,...,C_1$ in the frame t_T, and so on, generating elements $\{C_{11},...,C_{1N}\}$, $\{C_{21},...,C_{2N}\}$, $\{C_{K1},...,C_{KN}\}$ in the matrix given in (7).

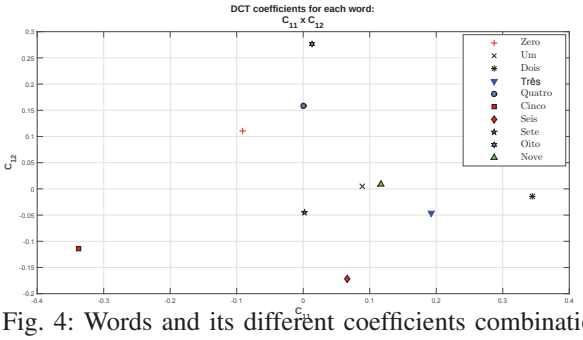

Fig. 4: Words and its different coefficients combinations

Thus, a two-dimensional temporal array DCT is generated for each j example of model W, represented by C_{KN}^{ij}.

4.2 Generating the Support Vector Machines

As SVM calls for a bidimensional space, two parameters will place the speech signal's characteristcs on a $2D$ representation of space, where the hyperplane will try and separate them in the best possible way. For these characteristcs we have the DCT n-order square matrix with its elements, each set composing a word, where the $n's$ used were $2, 3$ and 4.

One of the differentials of this paper is the use of Brazilian Portuguese language, an area that lacks works of this kind and has limited database. The keywords used on the experiment are the digits from '0' to '9'.

The coefficients generated for each person and each word are compared one by one with each other, in a methodology that's called one versus all technique. For example, one speaker has a dataset composed by ten samples of each of the ten digits, and the coefficients of the ten samples are extracted and disposed on the plane for later separation. They are put on the plane in pairs of characteristcs (the coefficients) in six different combinations for a 2 by 2 matrix, as shown in the Fig. 4, the plot of the mean for each coefficient and each word. First we extract the coefficients of each word for two different speakers, then we compare one of the words spoken by one of the speakers with all the words spoken by the other speaker, and so on for all the other speakers. The pairs of characteristcs are every possible nonrepeated combination of the DCT-matrix elements, first the $C_{11}xC_{12}$ then $C_{11}xC_{21}$, and so on: $C_{11}xC_{22}$, $C_{12}xC_{21}$, $C_{12}xC_{22}$ and $C_{21}xC_{22}$. Each combination expresses a part of the biometrical authentication, and the algorithm classifies a voice based on the majority of the matches. In Brasolin[25], the use of SVM with wavelet digital voice recognition in Brazilian Portuguese, obtained an average of 97.76% using 26 MFCC's in the pre-processing of voice and SVM machine's with the following characteristics: MLP as Kernel functions, ten machines (one for each class) and "one vs. all" as method of multiple classes. Also, the author tried to generalise instead of restricting. In comparison to this work, the results of this remain more effective, because

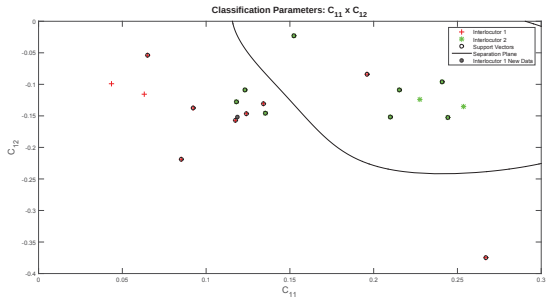

Fig. 5: One of the generated hyperplanes correctly classifying the new data

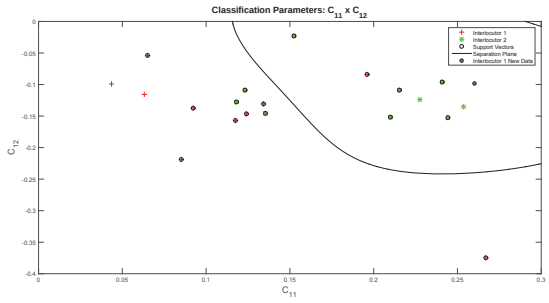

Fig. 6: One of the few misclassifications

the amount of MFCC's is smaller (only a 2 by 2 matrix) and, also, the input of parameters in the machines are lower. Consequently, the computational load is lower. But one cannot really compare because of the objetive intended of each one.

5. Training and testing

After performing the extraction of the parameters and putting them in the pairs, the Support Vector Machine algorithm is applied in order to generate the hyperplane and classify the new data. As shown in Fig. 5 and Fig. 6.

The black dots represent the new data input entering the system, and, for most of the cases, was correctly classified, showing in overall a precise voice recognition system, able to identify the speaker with higher than 90% probability, misclassifying few of the parameters, later compensated by the other combinations of coefficients as show in Table 1 and 2. The mentioned tables contain the percentage of the words correctly classified, for example, for the first speaker the system correctly classified all the keywords spoken, regarding $C_{11}xC_{12}$, as for the fourth speaker, the system matched correctly nine of the ten words, hence 90 percent.

Table 3 shows the result after the 6 combinations of pairs are made, to achieve more confiability on the identification. Were used on the training, as mentioned before, thirteen different voices from thirteen different speakers and ten different keywords, the digits from '0' to '9' (zero, um, dois, trÃłs, quatro, cinco, seis, sete, oito, nove, in brazilian

Table 1: Overall results for $C_{11}xC_{12}, C_{11}xC_{21}, C_{11}xC_{22}$

Speaker	C11x12	C11xC21	C11xC22
1	100	90	100
2	100	100	100
3	100	90	100
4	90	100	100
5	100	90	100
6	90	90	90
7	100	100	100
8	90	80	90
9	90	100	100
10	100	100	100
11	100	80	90
12	80	80	90
13	100	80	80

Table 2: Overall results for $C_{12}xC_{21}, C_{12}xC_{22}, C_{21}xC_{22}$

Speaker	C12x21	C12xC22	C21xC22
1	100	90	100
2	100	100	100
3	90	100	90
4	100	100	100
5	100	100	100
6	80	100	100
7	100	90	100
8	100	100	100
9	100	100	100
10	100	100	100
11	100	90	80
12	90	90	100
13	90	100	90

portuguese), each one spoken ten times by the same speaker in order to generate good parameters for each digit. Tables 4 and 5 show the computational time needed for one of the runs of the algorithm, and may vary depending on other tasks executed at the same time. The tests were made based on the same procedure, where no other tasks or softwares, but the operational system fundamentals, were initialized, hence reducing delays due processing sharing.

When using the combinations of DCT coefficients, most of the times it is easy for the program to generate the hyperplane, thus little time of training. That happens because the points are well placed on the plane, creating clusters that

Table 3: Percentage results for all pairs combined

Speaker 1	95
Speaker 2	100
Speaker 3	95
Speaker 4	98
Speaker 5	98
Speaker 6	91
Speaker 7	98
Speaker 8	93
Speaker 9	98
Speaker 10	100
Speaker 11	90
Speaker 12	88
Speaker 13	90

Table 4: Computation time for training

Speaker	Time in seconds
1	5,60
2	5,28
3	5,62
4	5,56
5	6,18
6	5,59
7	6,06
8	5,57
9	5,63
10	5,62
11	5,58
12	6,12
13	5,61

Table 5: Overall prediction time

Speaker	Time in milliseconds
1	9,30
2	8,14
3	8,65
4	8,20
5	8,25
6	9,53
7	10,64
8	12,65
9	11,68
10	13,66
11	9,25
12	8,86
13	9,22

are visually easy to separate, where sometimes even a linear kernel can obtain excellent results. Most of the computing time it's on the extraction of the parameters from the voice signal, i. e., calculating the MFCCs and the DCT transform. The values from Table 4 are the conjoint time of extraction and hyperplane generation, and the values of Table 5 are times of prediction for one keyword.

6. Conclusion

Biometrical classification utilizing voice as a input parameter a SVMs to classification, has shown sucess in general for identifying the speaker. Also the restrictions set by the classifiers, restricts in such a way that prevents false positive to rule over the actual positive results. The dichotomical nature of the technique leads to a excellent response time of computational execution, although the time the algorithm took for training all the datasets and comparing then with new data was aproximatelly two hours, one must remeber the absurd quantity of Suppor Vector Machines, exactly 912600, hence the delay. However for one keyword and one speaker at time, the training can last a insignificant time when compared with other techniques, as presented on the tables afore mentioned, so as the classification, revealing the real time application possible, fast, precise and very reliable. The computer used for training and prediciton has 6 GB of ram and a Intel Core i5^{TM}. The data was sampled at a 22050Hz frequency, with 16 bits of resolution.

With an overall greater than 90% of sucess rate, the system was acomplished and the premise validated, and in order to improve the work, more training data can be given to the system. As for the use of *Kernel Functions*, the used one for the final results was the *Radial Basis Function*, but in order to reduce training and predicting time, one can use the *Linear Kernel* with little loss of precision and reliability.

Acknowledgment

The authors thank the Scientific Initiation Program of the Federal Institute of Maranhao for financial support through grant aid and also the eletrical engineering student Libanio Vieira who helped substantially on the paper.

References

[1] B. E. Boser, I. L. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the 5th Annual Workshop on Computational Learning Theory, pages 144–152, Pittsburg, Pennsylvania, US, 1992.

[2] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. Knowledge Discovery and Data Mining, 2(2):1–43, 1998.

[3] T. Joachims. Learning to classify texts using support vector machines: methods, theory and algorithms. Kluwer Academic Publishers, 2002.

[4] K. I. Kim, K. Jung, S. H. Park, and H. J. Kim. Support vector machines for texture classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(11):1542–1550, 2002.

[5] M. Pontil and A. Verri. Support vector machines for 3-D object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(6):637–646, 1998.

[6] W. S. Noble. Support vector machine applications in computational biology. In B. SchÃűlkopf, K. Tsuda, and J.-P. Vert, editors, Kernel Methods in computational biology, pages 71–92. MIT Press, 2004.

[7] B. , I. Guyon, and J.Weston. Statistical learning and kernel methods in bioinformatics. In P. Frasconi and R. Shamir, editors, Artificial Intelligence and Heuristic Methods in Bioinformatics, pages 1–21. IOS Press, 2003.

[8] C. Ding and I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 2001.

[9] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Acessed in: 09/2003, 2004.

[10] Jian Zhou; Wang, Guoyin; Yong Yang; Peijun Chen, Speech Emotion Recognition Based on Rough Set and SVM, Cognitive Informatics, 2006. ICCI 2006. 5th IEEE International Conference on , vol.1, no., pp.53,61, 17-19 July 2006.

[11] D. Hanchate, M. Nalawade, M. Pawar, V. Pohale, and P. Maurya, "vocal digit recognition using artificial neural network"," 2nd International Conference on Coumputer Engineering and Technology, April 2010, pp. 88–91.

[12] R. Aggarwal and M. Dave, "application of genetically optimized neural networks for hindi speech recognition system"." World Congress on Information and Communication Technologies (WICT), December 2011, pp. 512–517.

[13] S. Azam, Z. Mansor, M. Mughal, and S. Moshin, "urdu spoken digits recognition using classifield mfcc and backpropagation neural network"." 4th International Conference on Computer Graphics, Imaging and Visualization (CGIV), August 2007, pp. 414–418.

[14] M. Mohammed, E. Bijov, C. Xavier, A. Yasif, and V. Supriya, "robust automatic speech recognition systems:hmm vesus sparse"." Third International Conference on Intelligent Systems modelling and Simulation, February 2012, pp. 339–342.

[15] C. Ganesh, H. Kumar, and P. Vanathi, "performance analysis of hybrid robust automatic speech recognition system"." IEEE International Conference on Signal Processing, Computing and Control (ISPCC), March 2012, pp. 1–4.
[16] J. Tamgo, E. Barnard, C. Lishou, and M. Richome, "wolof speech recognition model of digits and limited-vocabulary based on hmm and toolkit"." 14th International Conference on Computer Modelling and Simulation (UKSim), March 2012, pp. 389–395.
[17] P. Fantinato, Segmentacao de Voz baseada na Analise Fractal e na Transformada Wavelet. Prentice Hall, Outubro 2008.
[18] L. Rabiner and R. Schafer, Digital Processing of Speech Signals. Prentice Hall, 1978.
[19] J. Picone, "Signal modeling techniques in speech recognition." IEEE Transactions on Computer, April 1991, pp. 1215–1247.
[20] S. Haykin, Redes Neurais:Principio e pratica. Bookman, 2002.
[21] P. Fissore and E. Rivera, "Using word temporal structure in hmm speech recongnition." ICASSP 97, April 1997, pp. 975–978.
[22] M.A.,Kowtko, Biometric authentication for older adults, Systems, Applications and Technology Conference (LISAT), 2014 IEEE Long Island , vol., no., pp.1,6, 2-2 May 2014.
[23] Upmanyu, M.; Namboodiri, A.M.; Srinathan, K.; Jawahar, C.V., Blind Authentication: A Secure Crypto-Biometric Verification Protocol, Information Forensics and Security, IEEE Transactions on , vol.5, no.2, pp.255,268, June 2010
[24] Fahmy, M.S.; Atyia, A.F.; Elfouly, R.S., Biometric Fusion Using Enhanced SVM Classification, Intelligent Information Hiding and Multimedia Signal Processing, 2008. IIHMSP '08 International Conference on , vol., no., pp.1043,1048, 15-17 Aug. 2008
[25] A. Brasolin, A. Neto, and P. Alsin, ""digit recognition using wavelet and svm in brazilian portuguese"." ICASSP 2008, April 2008, pp. 1–4.