Support Vector Machines, Mel-Frequency Cepstral Coefficients and the Discrete Cosine Transform Applied on Voice Based Biometric Authentication

Felipe Gomes Barbosa
Federal Institute of Education, Science and Technology
Department of Electroelectronics
São Luís, Brazil
Email: felipegomes.mj@gmail.com

Washington Luís Santos Silva
Federal Institute of Education, Science and Technology
Department of Electroelectronics
São Luís, Brazil
Email: washington.wlss@ifma.edu.br

***Abstract*—In this paper, the implementation of a Support Vector Machine (SVM) is proposed based on automatic system for voice biometric authentication, recognizing the speaker using Mel-Frequency Cepstral Coefficients and the Discrete Cosine Transform. The voice recognition problem can be modeled as a classification problem, where the objective is to obtain the best degree of separability between the classes which represent the voice. Building an automated speech recognition system capable of identifying the speaker, has many techniques using artificial intelligence and general classification at disposal, Support Vector Machines being the one used in this work. The voice samples used are in the Brazilian Portuguese language and had its features extracted through the Discrete Cosine Transform. Extracted features are applied on the Mel-frequency Cepstral Coefficients to create a two-dimensional matrix used as input to the SVM algorithm. This algorithm generates the pattern to be recognized, leading to a reliable speaker identification using few parameters and a small dataset.**

***Keywords*—Voice Recognition; Support Vector Machines; Mel-Frequency Cepstral Coefficients; Discrete Cosine Transform; Brazilian Portuguese, Biometry.**

I. Introduction

Support Vector Machine (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* achieve good results in general, having that capacity of generalization measured by their efficiency on classifying data that does not belong to the training data set. The foundation of Support Vector Machines (SVM) was developed by Vapnik [1] and earned a lot of popularity due its promising 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.

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 *Artificial Neural Nets*, a lot of successful examples can be mentioned on many fields, like categorizing text [3], image analysis [4]–[6], bioinformatics [7], [8], and even in disease classification and recognition [9], [10]. Due to its efficiency in working with high dimensional data, it is cited in the literature as a highly robust technique [11].

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.

The voice recognition, and mostly the automatic one, has been the main goal of many scientists and researchers 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 [12]. In truth, the actual situation is another one: to recognize a simple word or phrase, one needs an absurd computational 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, with a great variety of works on the field, mainly combining diverse techniques such as Hidden Markov Models (HMMs) [13], and as for this paper, we propose a biometrical voice recognition 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, reducing computational effort, which is one of this work's goals, along with the use of a small dataset for training. While others kernels for the Support

Vector Machines were tested, the machines are based on the Radial Basis Function kernel (RBF) for offering precision and speed on classification in comparison to the others ones. The system was able to correctly identify the speaker in all cases, having great precision on the task.

II. Fundamentals of Support Vector Machines

SVM, as a supervised learning technique, can infer from a set of labelled 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 case, where to illustrate the separation plane generated by it, we can drawn a line that represents the decision boundary that correctly classifies some data set.

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 [14]. 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 labelled $y_i \in \{-1, 1\}$. 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

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

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 multipliers

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

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 Equation 4. It is 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 an infinity dimensional space, where the data can easily be separated, for this reason it is the one used on this paper.

III. Voice Recognition for Biometrical Authentication

From the beginning of its technological and intellectual 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.

On this context, the amazing advances in the last years, mostly in the microelectronics field, made possible to put into practice this line of thought, effectively. At this point, one needs to address 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.

A. 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 [15] and short term analysis by Fourier transform or spectral analysis [16]. 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.

B. 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

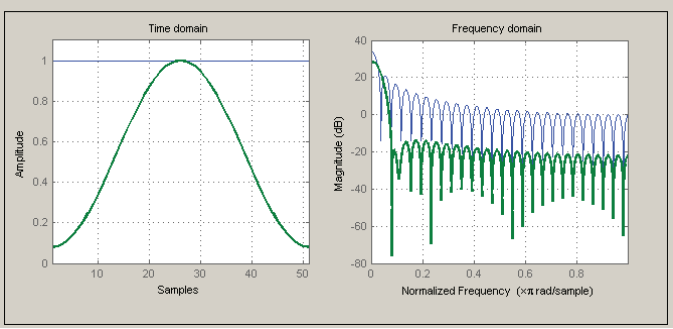


Fig. 1: Hamming Window and equivalent *SNR*

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) [16] 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 [17]. 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. 1 shows the plot of a hamming window in time and frequency domains.

C. Voice and Biometry

Biometry its an area that has a wide range of applications and research even in the new and exotic means as ear, gait, odor, etc., but the modalities that are field-proven in large-scale deployments are fingerprint, face, iris, voice, keystroke, and hand geometry [20]. These happen to be the biometric modalities which, today, best meet the tests for uniqueness, permanence, and consistency while also being conducive to capture using sensing devices in an ergonomically and economically practical way. Biometrics are largely statistical in nature, so it follows: the more data we have in a biometric sample, the more likely that it is unique, the same can be applied to the quantity of samples on the dataset. Also there is highly unlikely that an individual will not produce a biometric sample that is very similar to an another sample from a different individual, and there is always the probability of false match from a biometric comparison of samples.

The different modalities of biometric authentication have a variety o characteristics and concerns, some being more permanent over time than others, some more difficult to analyse and some with quality and reliability problems. There is no perfect biometric modality; each has advantages and disadvantages for a given use case. Facial images stand out because they are the biometric modality that humans excel at comparing, and so we can integrate complementary human-and machine-based recognition. Concerns on widespread use of biometric authentication systems are primarily centred around template security, revocability, and privacy [20]. The use of cryptographic primitives to bolster the authentication process can alleviate some of these concerns as shown by biometric cryptosystems [23].

Voice is notable for being behavioural as well as physical, and thus the samples available from a given individual are abundant. Speaker verification or authentication is the use of the voice as a method of determining the identity of a speaker for access control. The speaker claims to be of a certain identity and the voice is used to verify this claim. Speaker verification is a 1:1 match where one speaker's voice is matched to one template (also called a "voice print" or "voice model"). Speaker verification is usually employed as a "gatekeeper" in order to provide access to a secure system (e.g.: telephone banking). These systems operate with the user's knowledge and typically require their cooperation. For example, presenting a person's passport at border control is a verification process - the agent compares the person's face to the picture in the document [22]. Voice biometrics has the most potential for growth, because it requires no new hardware—most PCs already contain a microphone. However, poor quality and ambient noise can affect verification. In addition, the enrolment procedure has often been more complicated than with other biometrics, leading to the perception that voice verification is not user friendly.

Although, even when our biometric samples are unique, permanent, consistent, and physically bonded to us, the sensors and algorithms we have devised to acquire and analyse them are imperfect. Sensors introduce optical and electrical distortion. Information is lost as sample data is converted from analog to digital form, and then again when the digital signal is compressed. Sampling rates significantly impact the quality of biometric samples [21]. The algorithms designed to extract computer-matchable "templates" from a sample vary dramatically in precision and performance, as do algorithms and systems used by computers to rapidly assess their similarity. Machines are good at very fast, reasonably accurate, automated signal processing and template comparison, but they lack a human's ability to visually perceive, analyse and characterize the similarity of two samples. Nevertheless, our physical selves provide many features that are well-suited for biometric comparison and search, and advances in modern sensing and computing technologies continue to improve the ability of a machine to perform biometric identification extremely quickly and accurately [21].

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 method proposed here, 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.

D. Automatic Systems for Voice and Speech Recognition with SVM

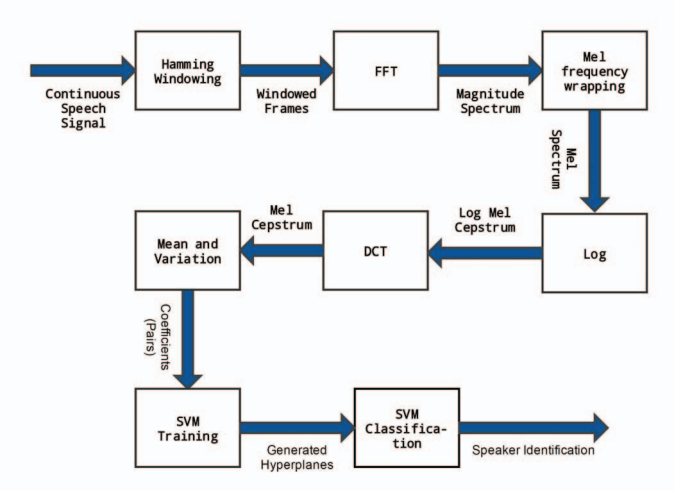
Hidden Markov Models (HMMs) have become the most employed technique for Automatic Speech Recognition (ASR).

However, the HMM-based ASR systems may reach their limit of performance. Hybrid systems based on a combination of artificial intelligence techniques provide significant improvements of performance. However, the progress in this paradigm has been hindered by their training computational requirements, which were excessive when these systems were proposed. Recently, several methods of Speech Recognition have been proposed using mel-frequency cepstral coefficients and Neural Networks Classifiers [25]–[27], Sparse Systems for Speech Recognition [28], Hybrid Robust Voice Activity Detection System [29], Wolof Speech Recognition with Limited vocabulary Based HMM and Toolkit [30], Real-Time Robust Speech Recognition using Compact Support Vector Machines [31]. Thus, the SVM has many functions; it is a binary algorithm, based in the Theory of Statistical Learning and in the Functional of Risk. And, finally, it has many functions for classification, such as in the case of multiple classes.

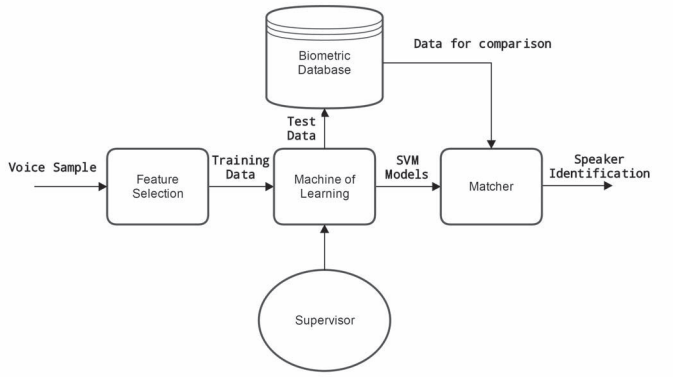
On the other hand we have voice authentication, which is not based on voice recognition but on voice-to-print authentication. The aforementioned processes also are largely employed on this matter, along with DTW (dynamic time warping), SVM-GMM (Support Vector Machines and Gaussian Mixture Models), SVM-HMM (Support Vector Machines and Hidden Markov Models), VQ (Vector Quantisation), and many others.

IV. 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) [16] in order to parametrize the signal with a reduced number of parameters. Then, it generates two dimensional matrices referring to the Discrete Cosine Transform coefficients. The elements of these matrices representing two-dimensional temporal patterns will be classified by Support Vector Machines (SVMs) [18]. 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 on the *Radial Basis Function.* Fig. 2 shows a condensed flowchart explaining the process used and system model, and Alg. 1 contains the pseudo code.



(a) Generic flowchart of whole method



(b) Block diagram for system model

Fig. 2: Detailed steps on the method used for recognition

A. The Discrete Cosine Transform matrix

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

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

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 Discrete Cosine Transform), $MFCC_k(t)$ 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 [19]. 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). Thus, a two-dimensional temporal array DCT is generated for each j example of model W, represented by C_{KN}^{ij}.

Algorithm 1 Pseudo code of the process

```
load Voice Bank
for Coefficients Combinations do

    for each speaker on database do

        CurrentCombination =
        ParametersExtraction(VoiceBank);

        TheClass(1 : 10) = 1
        TheClass(11 : 20) = −1

        for each word do

            SVMModels =
            trainSVM(CurrentCombination,
            TheClass, KernelFunction)

        end for

    end for

end for

Now, for prediction of new input:

for Coefficients Combinations do

    NewInputParameters =
    ParametersExtraction(NewInput);

    for each SVMModel do

        for each word do

            Output = predictSVM(SVMModel,
            NewInputParameters)

        end for

    end for

end for
```

B. Generating the Support Vector Machines

As SVM calls for a bidimensional space, two parameters will place the speech signal's characteristics on a $2D$ representation of space, where the hyperplane will try and separate them in the best possible way. For these characteristics we have the Discrete Cosine Transform 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 characteristics (the coefficients) in six different combinations for a 2 by 2 matrix, as shown in the Fig. 3, 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 characteristics 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 [32], the use of SVM with wavelet digital voice recognition in Brazilian Portuguese, obtained an average of 97.76% using 26 mel-frequency cepstral coefficients 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 generalize instead of restricting. In comparison to this work, the results of this remain more effective, because 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 objective intended of each one.

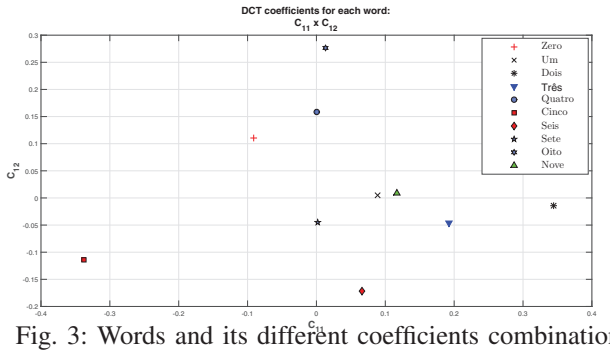


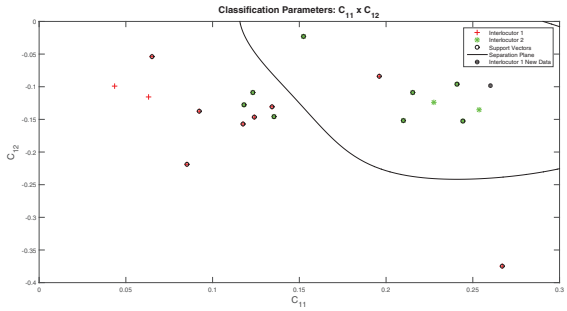
Fig. 3: Words and its different coefficients combinations

V. Training and Testing

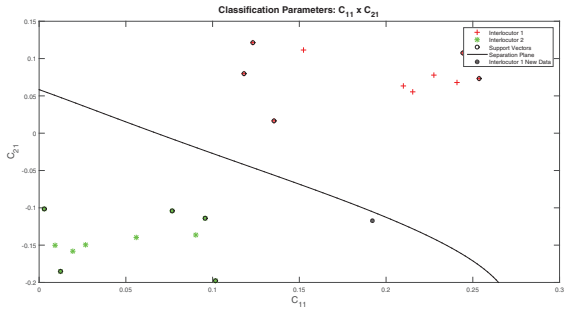
After performing the extraction of the parameters and putting them in the pairs, the Support Vector Machine algorithm is applied in order to generate the hyperplane and classify the new data. As shown in Fig. 4 and Fig. 5.

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 I and II. 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 III shows the result after the 6 combinations of pairs

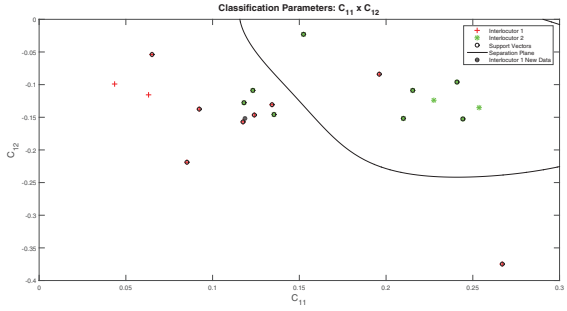


(a) $C_{11}xC_{22}$

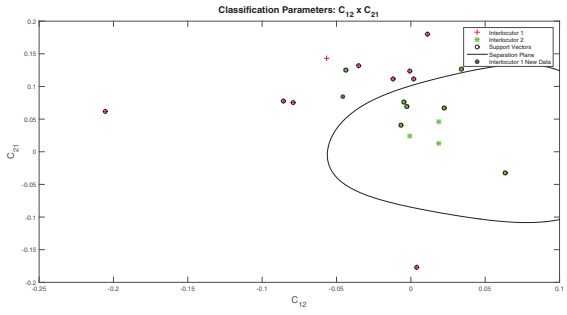


(b) $C_{11}xC_{21}$

Fig. 4: Hyperplanes correctly classifying the new data



(a) $C_{11}xC_{22}$



(b) $C_{12}xC_{21}$

Fig. 5: Some misclassifications

TABLE I: 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 II: 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

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, três, quatro, cinco, seis, sete, oito, nove, in brazilian portuguese), each one spoken ten times by the same speaker in order to generate good parameters for each digit.

Tables IV and V 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 are visually

TABLE III: 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 IV: 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 V: 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

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 IV are the conjoint time of extraction and hyperplane generation, and the values of Table V are times of prediction for one keyword.

VI. Conclusion

Biometrical classification utilizing voice as an input parameter a SVMs to classification, has shown success 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 approximately two hours, one must remember the absurd quantity of Support Vector Machines, exactly 912600, hence the delay. However for one keyword and one speaker at time, the training can last an insignificant time when compared with other techniques, so as the classification, revealing the real time application possible, fast, precise and very reliable. The computer used for training and prediction has 6 GB of ram and an Intel Core i5^{TM}. The data was sampled at a 22050Hz frequency, with 16 bits of resolution.

With an overall greater than 90% of success rate, the system was accomplished 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.

For future works and references, the dataset used for this paper shall be shared on the institute's website, in order to facilitate others in the matter, mainly because speech datasets for brazilian portuguese are scarce and difficult to access. Its also intended to continue this work, finding new parameters along with the mel coefficients to do voice and speech recognition, identifying the speaker and the word spoken, and changing the dataset from words to phonemes, removing the limitation brought by the need of each word to be placed on the dataset.

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] Lining Zhang, L.P. Wang and W. Lin, "Geometric Optimum Experimental Design for Collaborative Image Retrieval", IEEE Transactions on Circuits and System for Video Technology, vol.24, pp.346-359, 2014.

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

[8] B. Scholkopf, 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.

[9] Feng Chu and L.P. Wang, "Applications of support vector machines to cancer classification with microarray data," International Journal of Neural Systems, vol.15, no.6, pp.475-484, 2005.

[10] A.H. Khandoker, M. Palaniswami, and C.K. Karmakar, "Support Vector Machines for Automated Recognition of Obstructive Sleep Apnea Syndrome From ECG Recordings," IEEE Trans. Information Technology in Biomedicine, Vol.13, pp.37-48, 2009.

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

[12] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. Available at: http://www.csie.ntu.edu.tw/˜cjlin/libsvm/. Accessed in: 09/2014.

[13] A. Sloin and D. Burshtein, "Support Vector Machine Training for Improved Hidden Markov Modeling", IEEE Trans. Signal Processing, vol.56, pp.172-188, 2008.

[14] 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.

[15] P. Fantinato, Segmentacao de Voz baseada na Analise Fractal e na Transformada Wavelet. Prentice Hall, October 2008.

[16] L. Rabiner and R. Schafer, Digital Processing of Speech Signals. Prentice Hall, 1978.

[17] J. Picone, "Signal modelling techniques in speech recognition." IEEE Transactions on Computer, April 1991, pp. 1215–1247.

[18] S. Haykin, Redes Neurais: Princípio e prática. Bookman, 2002.

[19] P. Fissore and E. Rivera, "Using word temporal structure in hmm speech recognition." ICASSP 97, April 1997, pp. 975–978.

[20] 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.

[21] White Paper – What Are Biometrics?. Published by Aware, Inc., January 2014. Available at: http://www.aware.com/biometrics/whitepapers/wab_biometric-modalities.html. Accessed in: 01/2015.

[22] Biometrics Institute Privacy Code, September 2006. Available at: http://www.biometricsinstitute.org/pages/types-of-biometrics.html. Accessed in: 09/2014.

[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] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] R. Urena, A. Moral, C. Moreno, M. Ramon, and F. Maria, "Realtime robust automatic speech recognition using compact support vector machines." IEEE Transactions on Audio, Speech, and Language Processing, May 2012, pp. 1347–1362.

[32] A. Brasolin, A. Neto, and P. Alsin, "Digit recognition using wavelet and svm in brazilian portuguese". ICASSP 2008, April 2008, pp. 1–4.