DIGIT RECOGNITION USING WAVELET AND SVM IN BRAZILIAN PORTUGUESE

Adriano de Andrade Bresolin[1], Adrião Duarte Dória Neto[2], Pablo Javier Alsina[2]

[1] UTFPR - Technological Federal University of the Paraná – Brazil
[2] UFRN - Federal University of the Rio Grande do Norte – Brazil
{aabresolin, adriao, pablo}@dca.ufrn.br

ABSTRACT

In this paper we used WPT (*Wavelet Packet Transform*) and neural classifier SVM (*Support Vector Machine*) to recognize spoken digits from 0 to 9 in Brazilian Portuguese. The main objective this work is to find out the Wavelet mother that better represents the speech signal in Brazilian Portuguese. The results obtained were compared with MFCC *(Mel frequency cepstral coefficients)*. We carried out sixteen experiments with different Wavelets in dependent-case and four experiments in independent-case. The database was recorded in three months with 82 eighteen-to-forty years old male speakers. The SVM was used as a classifier in a "one vs. all" strategy. Best results have been obtained using Wavelets Daubechies 5, Meyer and Coiflet 5. Finally, we used a neural network MLP (Multi Layer Perceptron) in order to improve the SVM results.

Index Terms— Speech Recognition, Neural Networks (SVM), Multilayer Perceptrons, Wavelet Transforms.

1. INTRODUCTION

The aim of this paper is to evaluate the performance of the Wavelet Packet with SVM in the isolated word recognition for the Portuguese language. Several papers concerning digit recognition are found in speech recognition literature [1, 2], but few results have been published until now in the particular case of applications using the Portuguese language [3].

The MFCC (Mel Frequency Cepstral Coefficients) feature extraction method was widely adopted in many popular speech recognition systems [4 and 5]. However, it is well-known that the Short Time Fourier Transform (STFT) based on MFCC has achieved uniform resolution over the time-frequency plane. Because of this, it is difficult to detect sudden burst in a slowly varying signal by using STFT. Recently, Wavelets approaches have been proposed for feature extraction [6 and 7] with excellent results.

Although the HMM (Hidden Markov Models) classifier is the most used in speech recognition, the interest in classifiers which can go beyond of its performance models has increased in recent years [8]. So, we used the neural networks SVM in this work, in order to validate this classifier for Portuguese speech recognition.

To carry out this work was necessary to make a new database, because there are few speech databases in Portuguese available to research.

This paper is organized as follows: Section 2 presents the signal pre-processing phase. Section 3 shows the speech features extraction through Wavelet Packet Transform with Mel scale. Section 4 describes the training procedure with SVM. Section 5 presents the obtained results in sixteen experiments with different Wavelets in dependent-case and four experiments with independent-case.

2. SIGNAL PRE-PROCESSING

The audio pre-processing stage is composed of four steps: acquisition, filtering, pre-emphasis and normalization.

In the acquisition step, the voice signal is separated of the video signal. The acquisition rate is 22,050Hz, with a bandwidth of 11,025 Hz. Signal frequencies above 8 kHz and electric power noise are eliminated through a band pass filter with cutoff frequencies of 80 Hz and 8 kHz. After that, the speech signal is pre-emphasized to compensate for spectral tilt. This time domain filter value is 0.97. In the normalization step, the maximum signal amplitude is normalized to one. After, the audio signal is broken into overlapping frames and stored. A 30ms frame size has been used and the step size was 50 percent of the frame size. Each frame is multiplied by Hamming Window, in order to minimize any signal discontinuities in the time domain.

The problem in this kind of work is that each digit (0 until 9) has a different size. Therefore, each digit will have different amounts of windows and each feature vector will have a different size. To use the SVM classifier is necessary that all features have the same size. In order to resolve this problem, we used the dynamic windowing.

2.1. Dynamic Windowing

In order to create only a feature vector for each digit signal, all feature frames will be concatenated. If each digit signal has amounts of different frames, then the size of each feature vector will be different.

This method is simple: first we find out the biggest signal (digit) and broken by frames with 30ms size and 50% superposition. The maximum quantity obtained was 38 frames. Therefore, we used 40 frames as fixed value.

So, each signal (digit) will be divided in 40 (forty) frames. Now, the frame size is variable according to the signal size measure. But, all signals will always have 40 frames and each frame will never exceed 30ms.

3. SPEECH FEATURE – WAVELET PACKET

The simply Wavelet Transform decomposes the signal through two complementary filters: *High pass* and *Low pass*. This process will generate two new signals (Fig. 1a). The signal "A" (approach) contains the components of low frequency and signal "D" (detail) contains the components high-frequency of the original signal [9].

The Wavelet Packet decomposes the Approximation and Details spaces, originating a binary tree structure [10]. This decomposition facilitates the partitioning of the higher frequency into smaller bands (Fig. 1b), which cannot be achieved by using Wavelet Transform. Figure 1 shows the Wavelet Transform and Wavelet Packet to three levels of the decomposition.

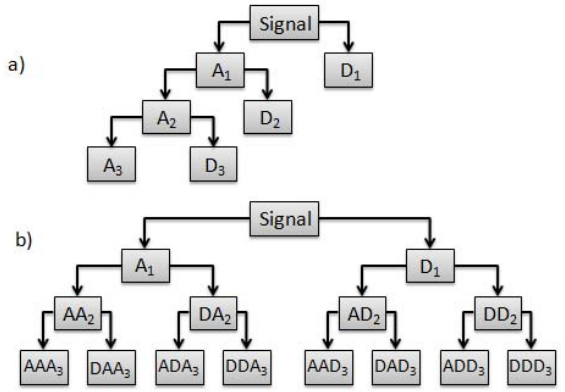

Fig. 1: Three levels of decomposition: (a) simply Wavelet, (b) Wavelet Packet.

The performance of the WPT (Wavelet Packet Transform) using clean speech and noisy speech was compared with MFCC in [6]. The results showed that WPT obtained better recognition rates than MFCC for phoneme recognition task. WPT was superior to MFCC in unvoiced phoneme classification problem [7].

In this work we used the Wavelet Packet with the Mel scale. This method was presented in [11] and results in a vector with 29 elements separated through Mel scale. Figure 2 shows the bands selected through this method.

The WPT decomposition complete tree is formed by 7 levels (Fig. 2). The first decomposition level L1 has 2 bands: Approximation-A_1 (0-5512Hz) and Detail-D_1 (5512-11025Hz), in other words, two bands with bandwidth of the 5512Hz. Level two has 4 bands with 2756Hz. Level three has 8 bands with 1378Hz. Level four has 16 bands with 689Hz. Level five has 32 bands with 344Hz. Level six has 64 bands with 172 Hz and level seven has 128 bands with 82Hz approximately.

We use twelve bands by L7, four bands by L6, five bands by L5, five bands by L4 and three bands by L3. Therefore, the speech signal feature is represented by a vector whose 29 elements represent the energy of each sub-band [11].

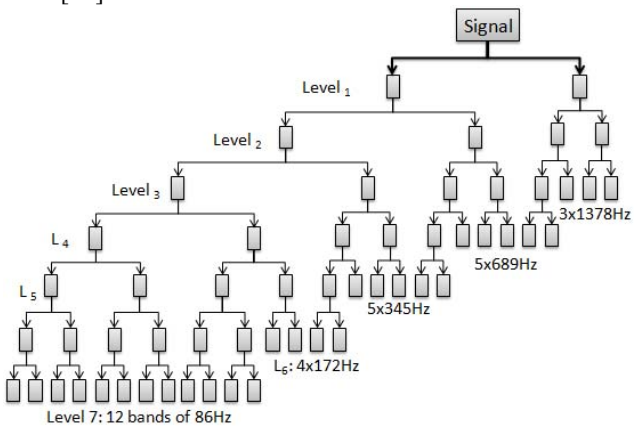

Fig. 2: Twenty nine energy sub-bands selected through Wavelet Packet using Mel-scale.

Sixteen kinds of wavelets were used in the experiments. Figure 3 shows some wavelets mother, like Coiflets, Daubechies and Mayer.

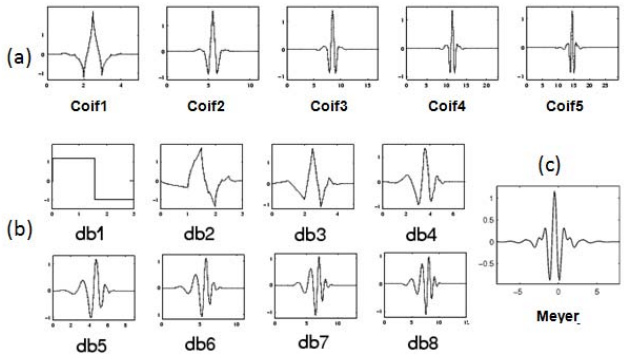

Fig. 3: Wavelet mothers: (a) Coiflet; (b) Daubechies and (c) Meyer.

Two databases have been created. In both, the vocabulary consists of 10 different words (digits 0 until 9), spoken in Portuguese plus the word "silence". First database was recorded in three months with 82 eighteen-to-forty years old male speakers. Each speaker recorded 10 files with all ten digits in three different months. Therefore, this database contains 8,200 word realizations, in other words, 820 for each digit. This database was used in the four independent-case experiments.

Second database, one speaker recorded 700 files with all digits. So, this database has 7,000 words realizations.

This database was used in the sixteen dependent-case experiments.

4. SVM CLASSIFIER

The SVM theory was first introduced by Vapnik in [12]. Support Vector Machines (SVMs) represent a new approach for pattern classification, which has recently attracted a great interest in the machine learning community. The essence of its approach lies in its strong connection with the underlying statistical learning theory, in particular, the theory of Structural Risk Minimization.

The SVM learn the boundary regions between samples belonging to two classes, by mapping the input samples into a high dimensional space, and seeking a separating hyperplane in this space. The separating hyperplane is chosen in such a way that it maximizes its distance to the closest training samples.

In order to validate the use of SVM in the training stage, in this paper, sixteen experiments in independent-case and four ones in dependent-case were carried out. In all experiments the traditional strategy "one" versus "all" was used in association with a decision scheme based on a Committee Machine [13].

The Committee Machine is formed by ten SVMs (Fig. 4). Each SVM was used to separate two classes. For example: the SVM 1 was trained to separate the digit "0" of the other digits. The result is binary "1" or "-1". If the result is "1" this SVM recognized this digit. In all experiments was used the kernel polynomial.

Figure 4 shows the Committee Machine in traditional strategy "one" versus "all" with ten SVMs.

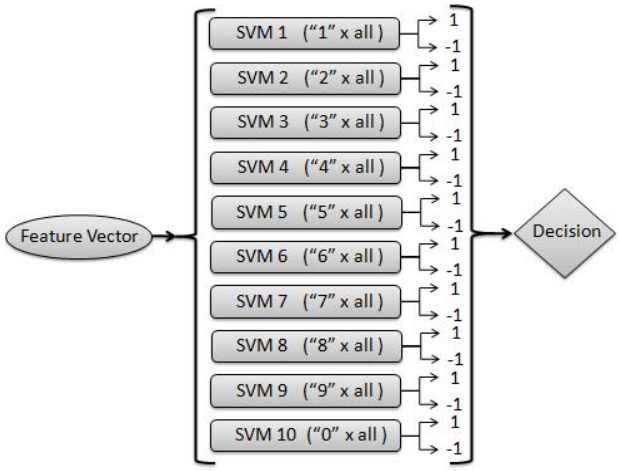

Fig. 4: Committee Machine in traditional strategy "one" versus "all" with ten SVMs.

Therefore, in each experiment Committee Machine is formed by ten SVMs and there will be recognition, if and only if, only one SVM results= "1". If two or more SVMs the results is "1", then it is inconsistent or error.

5. EXPERIMENTS AND RESULTS

In dependent case, sixteen experiments were carried out with different Wavelet mothers.

In order to resolve SVM inconsistency, the seventeenth experiment (using the MLP - Multi Layer Perceptron) was carried out with the best Wavelet (Daubechies 5). MLP neural network with three layers (1160-20-10) was used, in other words, input layer has 1160, the second layer (hidden layer) has 20 and output layer has 10 neurons. The training algorithm used was back-propagation algorithm [13]. We used the MLP only in the files in which there was inconsistent recognition by SVM.

The same MLP was tested in the eighteenth experiment. In this case we did not use the SVM, only MLP was used as a classifier.

Finally, the nineteenth experiment with MFCC (Mel Frequency Cepstral Coefficients) was carried out in order to compare with Wavelets experiments.

All files were divided in four random groups. 25% of the files are used to training and 75% were used to Validation. This process is called Cross-Validation.

Table 1 show the results obtained in all nineteen experiments in dependent-case. Column 2 shows the features signal used. Column 3, 4, 5 and 6 shows the results obtained through cross-validation groups. Column 7 shows the mean results obtained. All results are expressed in success rate percentage (%SR).

Table 1: Dependent speaker case results: (1 to 16) sixteen different Wavelets; (17) Daubechies 5 with MLP; (18) MLP with classifier experiment and (19) MFCC experiment. (% SR) Success rate percentage.

Test	Feature signal	%SR G1	%SR G2	%SR G3	%SR G4	% SR Mean
1	Coiflet 1	93.11	92.39	93.67	92.83	93.00
2	Coiflet 2	95.56	93.72	94.06	94.11	94.36
3	Coiflet 3	95.28	94.28	94.11	94.00	94.42
4	Coiflet 4	95.94	94.56	95.11	94.06	94.92
5	Coiflet 5	97.06	96.50	97.11	96.56	**96.81**
6	Daubechies 1	92.06	92.11	91.50	91.17	91.71
7	Daubechies 2	93.89	93.72	93.44	93.83	93.72
8	Daubechies 3	94.17	92.89	93.89	93.06	93.50
9	Daubechies 4	94.94	93.56	93.67	93.56	93.93
10	Daubechies 5	97.44	98.22	97.67	98.28	**97.90**
11	Daubechies 6	95.72	94.17	94.22	93.61	94.43
12	Daubechies 7	94.72	93.78	94.33	93.83	94.17
13	Daubechies 8	95.33	94.83	94.83	93.56	94.64
14	Meyer	97.83	96.61	97.44	96.78	**97.16**
15	Biort 1.1	92.06	92.11	91.50	91.17	91.71
16	Biort 2.2	90.94	89.33	91.11	89.67	90.26
17	Daubechies 5 (SVM + MLP)	98.94	99.19	98.67	99.23	**99.00**
18	Daubechies 5 (Only MPL Classifier)	82.89	84.28	83.56	80.78	82.87
19	MFCC features with SVM classifier	97.53	97.85	98.30	97.33	**97.75**

The best results were obtained with Daubechies 5, MFCC, Meyer and Coifelet 5. Only Daubechies 5 was better than MFCC. MLP with classifier, using Daubechies 5, obtained the worse result. But, when we used the MLP to resolve the SVM inconsistency (test 17), the success rate improved more than 1%.

In independent-case, four experiments were carried out. Three with the best Wavelet mothers found in dependent-case experiments (Daubechies 5, Meyer and Coifelet 5) and one with MFCC features.

Again, the files were divided in four groups. 25% of the files are used to training and 75% were used to Validation (Cross-validation).

Table 2 shows the results obtained for four experiments in independent-case. All results are expressed in success rate percentage (%SR).

Table 2: Independent speaker case results for Coiflet 5, Daubechies 5, Mayer and MFCC features. (% SR) Success rate percentage.

Test	Feature signal	%SR G1	%SR G2	%SR G3	%SR G4	% SR Mean
1	Coiflet 5	85.06	86.70	87.23	86.45	86.36
2	Daubechies 5	93.24	92.12	93.16	93.38	**92.97**
3	Meyer	88.38	89.78	91.14	91.17	90.12
4	MFCC	92.11	93.15	93.70	92.55	92.87

In this case, the best result was obtained with Daubechies 5. Again, only Daubechies 5 was more successful than MFCC.

6. CONCLUSION

In this paper we used WPT (Wavelet Packet Transform) and neural classifier SVM (Support Vector Machine) to recognize spoken digits from 0 to 9 in Brazilian Portuguese.

The main objective of this work was to find out the Wavelet mother that better represents the speech signal in Brazilian Portuguese. The results obtained were compared with MFCC (Mel frequency cepstral coefficients).

In dependent-case, sixteen experiments were carried out with different Wavelets. The best results were obtained with Daubechies 5, MFCC, Mayer and Coifelet 5, respectively. Only Daubechies 5 was better than MFCC. The experiment using only MLP like classifier obtained weak performance, but when the MLP was used to resolve SVM inconsistent was obtained the better result. The success recognition rate was increased in more than 1% in this case.

In independent-case, four experiments were carried out. Three with the best Wavelet mothers found in dependent-case experiments (Daubechies 5, Meyer and Coifelet 5) and one with MFCC features. Again, only Daubechies 5 was better than MFCC. In this case the performance was weak. We believe that it has occurred due to the great diversity in Brazilian accent. Therefore, the best Wavelet was Daubechies 5 in both tests.

Acknowledgments: This work was supported by CAPES (Federal government of Brazil).

11. REFERENCES

[1] V. Tabarabaee, B. Azimisadjadi, S. B. Zahirazami, C. Lucas, "Isolated word recognition using a hybrid neural network," *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-94,* vol. II, p. 649-652, April, 1994.

[2] K. Kim, D. H. Youn, and L. Chulhee, "Evaluation of wavelet filters for speech recognition," *IEEE International Conference on Systems, Man, and Cybernetics,* vol. 4, pp: 2891-2894. Oct. 2000.

[3] V. Pera, F. Sa, P. Afonso and R. Ferreira, "Audio-visual speech recognition in a Portuguese language based application," *IEEE Intern. Conf. on Ind. Technology*, vol. 2, pp. 688-692, Dec. 2003.

[4] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, "Phoneme recognition using time-delay neural networks," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 37, nº 03, pp. 328-339, March 1989.

[5] M. N. Stuttle, and M. J. F. Gales, "A Mixture of Gaussians Front End for Speech Recognition". *Eurospeech,* Scandinavia, 2001.

[6] J. N. Gowdy, and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition," *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing,* p. 1351-1354, 2000.

[7] O. Farooq and S. Datta, "Phoneme recognition using wavelet based features," *Elsevier: Information Sciences*, vol. 150, Issues 1-2, pp. 5-15, March 2003.

[8] M. J. Russell, and J. A. Bilmes, "Introduction to the special issue on new computational paradigms for acoustic modeling in speech recognition," *Editorial, Computer Speech and Language*, nº 17, pp. 107-112, March 2003.

[9] I. Daubechies, "The Wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inf. Theory,* pp. 961-1005, 1990.

[10] S. C. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms*. Prentice Hall, New Jersey. 1998.

[11] A. A. Bresolin, A. D. D. Neto, and P. J. Alsina, "Brazilian Vowels Recognition using a New Hierarchical Decision Structure with Wavelet Packet and SVM," *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP'07*, vol. 2, pp: 493-496, April 2007.

[12] V. N. Vapnik, "Principles of risk minimization for learning theory," *Advances in Neural Information Processing Systems,* vol. 04, pp.831-838, San Mateo, CA 1992.

[13] S. Haykin, *Neural Networks: A Comprehensive Foundation,* 2nd edition. Macmillan, New York. 1998.