BRAZILIAN VOWELS RECOGNITION USING A NEW HIERARCHICAL DECISION STRUCTURE WITH WAVELET PACKET AND SVM

Adriano de Andrade Bresolin[1], Adrião Duarte D. 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 work, a new phoneme recognition system is proposed. The base of decision of the proposed system is the tongue position and roundedness of the lips. The features of the speech are the coefficients of Wavelet Packet Transform with sub-bands selected through the Mel scale. The SVM (Support Vector Machine) is used as classifier in the structure of a Hierarchical Committee Machine. The database used for the recognition was a set of oral vocalic phonemes of the Portuguese language. The experimental results show success rates of 98.07% for the user-dependent case and 91.01% for the user-independent case. This new proposal increased 4.1% and 3.5% the success rate in relation to the "one vs. all" decision strategy, to user-dependent and user-independent case respectively.

Index Terms— *Speech Recognition, Support Vector Machine, Wavelet Packet*

1. INTRODUCTION

A first decision in the development of a speech recognition system is the definition of the unit to be recognized: words, syllables, triphones, diphones or phonemes.

A natural language, such as the Portuguese, possesses about 400.000 words, which demands great amount of processing and storage, a hard problem for continuous recognition. In the last years, research efforts have focused the phonetic unit smaller than the word.

Santos and Alcaim [10] used syllables as units of recognition. However, the authors alert that the syllables are attractive only when the number of patterns is small. Moreover, the syllables can have 2000 patterns and they are not very useful in languages like English, which does not possess a trivial syllabic division. In this case, triphones are more used, but their training is difficult (Young [14]).

This work proposes the use of phonemes as base for the Brazilian Portuguese speech recognition. Since vowels are the nucleus of the syllables that form the words in the Portuguese language, they are the main focus of this paper. The oral vowels (**a, é, i, ó, u, ê, ô**) were used in the recognition process.

The energy coefficients of Wavelet Packet Transform with sub-bands, selected through the Mel scale, were chosen as features of the speech. A new hierarchical Committee Machine decision system is presented. The classification of vowel signals is based on Support Vector Machines (SVM), where the base of decision is the tongue position and the rounding of the lips.

Section 2 presents the signal pre-processing phase. Section 3 shows the speech features extraction through Wavelet Packet Transform using Mel scale. Section 4 describes the training procedure of SVM neural network. Section 5 proposes a new technique for vowel recognition by using seven SVM in a Hierarchical Committee Machine. Section 6 presents some experiments of vowels recognition.

2. PREPROCESSING

The preprocessing stage is composed of four steps: acquisition, filtering, pre-emphasis and normalization. In the acquisition step, the voice signal is sampled at a rate of 22050 Hz, with a bandwidth of 11050 Hz. The signal voice vector is defined as:

$$\mathbf{x_i} = [x_0, x_1, ..., x_N]^{\mathrm{T}} \qquad (1)$$

where **N** is the number of samples.

Signal frequencies above 10 kHz and electric power noise are eliminated through a band pass filter with cutoff frequencies of 80 Hz and 10 kHz.

After that, the speech signal is pre-emphasized. In the normalization step, the maximum signal amplitude is normalized to one.

The signal is broken into K overlapping frames and stored in a K x M matrix Y (2). V is the step size and M the frame size. A 30ms frame size has been used and the step size was 50 percent of the frame size.

$$\mathbf{y_{kj}} = \mathbf{x_{V.k+j}}, \quad \mathrm{k} = 0,1,...,\mathrm{K}-1 \quad \mathrm{j} = 0,1,...,\mathrm{M}-1 \qquad (2)$$

Each frame is multiplied by means of a window function, named Hamming window, in order to minimize any signal discontinuities in the time domain.

3. FEATURE EXTRACTION - WAVELET PACKET

The Wavelet Packet (WP) decomposes the approximation spaces as well as details spaces, originating a binary tree structure. A WP decomposition facilitates the partitioning of the higher frequency side of the frequency axis into smaller bands, which cannot be achieved by using discrete wavelet transform [1]. In this work we used seven levels of decomposition, but only twenty-nine sub- bands are utilized. These sub-bands are selected through the Mel scale.

3.1. Mel Scale

The Mel scale is a signal representation scheme used in the analysis of speech signals. Stevens and Volkmann in [12] defined the Mel scale as a frequency function of the magnitude of an auditory sensation. The Mel scale is linear in the frequency below 1000 Hz and logarithmic above this frequency.

Farroq in [4] and Gowdy in [5] evaluated the performance of the Wavelet Packet with Mel scale and compared its performance with MFCC coefficients. The results obtained through Wavelet Packet with Mel scale showed better recognition rates than MFCC for a phoneme recognition task.

In this work, seven levels of decomposition of the WP are utilized and the Mel scale is used to select 29 sub-bands, as depicted in figure 1.

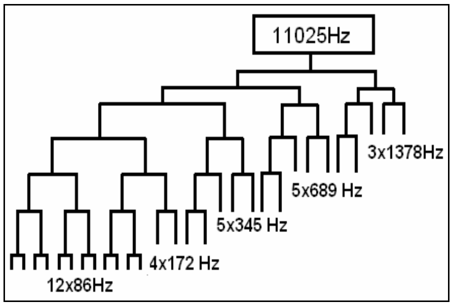
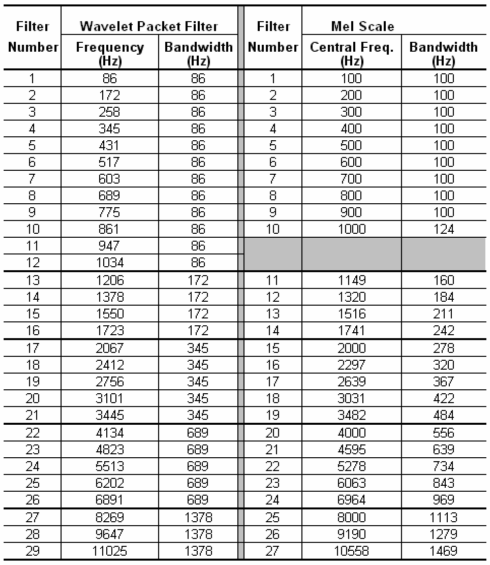


Figure 1. The Wavelet Packet decomposition in 29 sub-bands according to Mel scale

The signal is sampled at 22050 Hz, with a bandwidth of 11025 Hz. First, a full seven level WP decomposition is carried out. Twelve subbands of 86 Hz of the level 7, four subbands of 172 Hz of the level 6, five subbands of 345 Hz of the level 5, five subbands of 689 Hz of the level 4 and three subbands of 1378 Hz of the level 3 are utilized. The bandwidth obtained from each filter using WP decomposition is given in Table 1. Therefore, the speech signal feature is represented by a vector whose 29 elements represent the energy of each sub-band extracted from the WP through the Mel scale. The used Wavelet mother was db5 (Daubechies [2]).

Table 1. Frequency bands achieved by Wavelet Packet Decomposition and Mel scale.



Filter	Wavelet Packet Filter		Filter	Mel Scale	
Number	Frequency (Hz)	Bandwidth (Hz)	Number	Central Freq. (Hz)	Bandwidth (Hz)
1	86	86	1	100	100
2	172	86	2	200	100
3	258	86	3	300	100
4	345	86	4	400	100
5	431	86	5	500	100
6	517	86	6	600	100
7	603	86	7	700	100
8	689	86	8	800	100
9	775	86	9	900	100
10	861	86	10	1000	124
11	947	86			
12	1034	86			
13	1206	172	11	1149	160
14	1378	172	12	1320	184
15	1550	172	13	1516	211
16	1723	172	14	1741	242
17	2067	345	15	2000	278
18	2412	345	16	2297	320
19	2756	345	17	2639	367
20	3101	345	18	3031	422
21	3445	345	19	3482	484
22	4134	689	20	4000	556
23	4823	689	21	4595	639
24	5513	689	22	5278	734
25	6202	689	23	6063	843
26	6891	689	24	6964	969
27	8269	1378	25	8000	1113
28	9647	1378	26	9190	1279
29	11025	1378	27	10558	1469

4. TRAINING

In order to provide a better choice of the frames that represent the speech signal, instead of using all the windows, the signal was segmented using the Kmeans algorithm (Duda and Hart [3]) with two classes. Each signal window possesses a vector characterized by 29-band energies selected by WP. The Kmeans algorithm uses these vectors for signal separation. This procedure results in a significant reduction of the training time and an improvement of the performance of the system.

Figure 2 shows this procedure, where the frames were selected through Kmeans from a vowel "a" signal, using an energy vector.

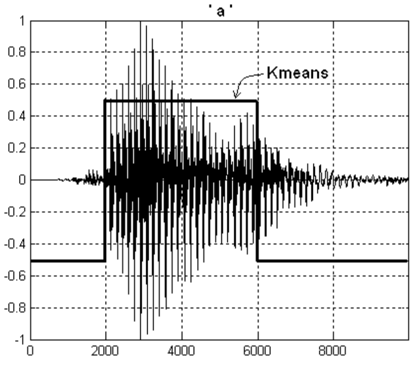


Figure 2. Segmentation vowel '**a**' through Kmeans, using an energy

It is perfectly clear that the frames were selected in the nearness of the center of the signal. This selection process avoids the use of frames that can represent variations of pronounce or noise, which generally occurs in the beginning and in the end of the locution.

4.1. Support Vector Machines

Support Vector Machines (SVMs) represent a new approach for pattern classification, what has recently attracted a great interest in the machine learning community. Their appeal lies in their strong connection with the underlying statistical learning theory, in particular, the theory of Structural Risk Minimization.

The SVM theory was first introduced by Vapnik in [13]. The SVM learns 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.

Juneja [8] demonstrated the utility of the SVM in the classification of phonemes. Russell and Bilmes [9] affirm that in the last years it was verified a growing interest on classifiers that can go beyond the performance of the HMM.

To validate the use of SVM in the training stage, two experiments were carried out. First, the traditional strategy "one vs. all" was used in association with a decision scheme based on a Committee Machine formed by a mixture of specialists (Haykin [6], pp. 402). In second test, a new strategy, called Hierarchic Committee Machine – HCM, was used. This new strategy based on the articulatory phonetic is presented in Section 5.

5. HIERARCHIC COMMITTEE MACHINE - HCM

5.1. Articulatory Phonetic: Vowels Classification

The proposed HCM is based on the characteristic vowel articulation of the Portuguese language. In phonetics, a vowel is a sound in spoken language, characterized by an open configuration of the vocal tract, without obstruction of air pressure above the glottis [11]. The articulatory features that distinguish different vowels in a language are said to determine the vowel's quality. The vowels are described in terms of the common features: height (vertical tongue position), backness (horizontal tongue position) and roundedness (lip position), as shown in Table 2.

Height refers to the vertical position of the tongue relative to either the roof of the mouth or the aperture of the jaw. In high vowels, such as [i] and [u], the tongue is positioned high in the mouth, whereas in low vowels, such as [a], the tongue is positioned low in the mouth.

Backness refers to the horizontal tongue position during the articulation of a vowel relative to the back of the mouth. In front vowels, such as [i], the tongue is positioned forward in the mouth, whereas in back vowels, such as [u], the tongue is positioned towards the back of the mouth.

Roundedness refers to whether the lips are rounded or not. In most languages, roundedness is a reinforcing feature of mid to high back vowels, and not distinctive.

Table 2. Frequency bands achieved by Wavelet Packet Decomposition and Mel scale.

Horiz. Position Vertical Position		Front		Central		Back	
	Lips	rounded	not	rounded	not	rounded	not
High		i				u	
Mid-high		ê				ô	
Mid-low		é				ó	
Low				a			

5.2. Hierarchical Decision

Hosom in [7] used three neural networks specialists to detect the manner of articulation, place of articulation and height of the tongue in the production of phonemes. The outputs of the three neural networks were evaluated by a classifier using the Bayes rule.

Figure 3 shows the proposed new classification structure, in which characteristics like the tongue height and roundedness of the lips are the base for decision process.

The system is composed by seven SVM specialists, in which machine 01 selects phoneme /a/ through the strategy "one vs. all". The phoneme /a/ is classified as Central and Low. Since vowel /a/ differs from the other vowels, its classification is made in first place.

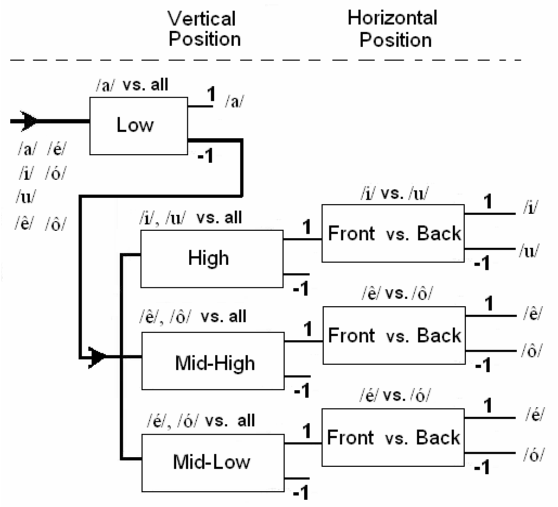


Figure 3. New Classification system based on tongue position and roundedness of the lips.

In the next decision step, the system verifies if the pattern is High, Mid-high or Mid-low (Vertical Tongue Position). Having the biggest number of positive classifications, the specialist machine is declared winner. According to the winner, the system will classify the phoneme based on to the horizontal tongue position (Front vs. Back) in a strategy "one vs. one". The classification based on the roundedness of the lips is equivalent to that one based on the horizontal tongue position.

6. EXPERIMENTAL RESULTS

In order to validate the proposed classification scheme, two experiments were performed: the first one, with the traditional strategy (one vs. all); the second, with the proposed hierarchical strategy.

For the user-dependent case, the training set was composed of 900 patterns. A set of 5600 patterns was utilized for testing. The traditional strategy results showed success rates of 93.93%. The hierarchical strategy results showed success rates of 98.07%. Table 3 shows the confusion matrix for the hierarchical strategy.

Table 3. Confusion Matrix - Hierarchical strategy for user-dependent case.

-	a	é	i	ó	U	ê	ô	% Success rate
a	800	0	0	0	0	0	0	100
é	0	800	0	0	0	0	0	100
i	0	0	800	0	0	0	0	100
ó	0	20	0	780	0	0	0	97.50
u	0	0	10	0	790	0	0	98.75
ê	0	0	10	0	25	765	0	95.62
ô	0	13	0	0	0	30	757	94.62

In the user-independent case, the training set was composed of 1080 patterns. A set of 8120 patterns was utilized for testing. The traditional strategy results show success rates of 87.44%. The hierarchical strategy results show success rates of 91.01%.

7. CONCLUSIONS

In this work, a new vowel recognition system is proposed, where the tongue position and roundedness of the lips are adopted as base of decision. The coefficients of Wavelet Packet Transform with subbands selected through the Mel scale were selected as speech features.

The Support Vector Machine was used as classifier in the structure of a Hierarchical Committee Machine. The database used for the recognition was a set of oral vowel phonemes of the Portuguese language.

Our main objective was to validate the new proposed classification scheme.

The experimental results showed success rates of 98.07% for the user-dependent case and 91.01% for the user-independent case. This new proposal increased 4.1% and 3.5% the success rate in relation to the "one vs. all" decision strategy, to user-dependent and user-independent case respectively.

Therefore, we conclude that the new proposal presented better recognition taxes than the traditional strategy (one vs. all). Moreover, for the phonemes /a/, /é/ and /i/, the recognition rate was 100% for the user-dependent case.

The new hierarchical strategy decision scheme proved to be more efficient, faster and robust, achieving a significant reduction in the complexity of the decision process.

8. REFERENCES

[01] Burrus, Sidney C. Gopinath, R. A. and Guo, Haitao. *Introduction to Wavelets and Wavelets Transforms.* Prentice Hall, New Jersey. (1998).

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

[03] Duda R.O. and Hart, P.E. *Pattern classification and scene analysis*. John Wiley & Sons, New York, (1973).

[04] Farooq, O. and Datta, S. "Mel filter-like admissible wavelet packet structure for speech recognition". *IEEE Signal Processing Letters*. Vol. 08, Issue 07, pp. 196-198. July (2001).

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

[06] Haykin, Simon. *Redes Neurais: Princípios e Prática*. Editora Bookman 2ª Edição, Porto Alegre, (2001).

[07] Hosom, John P. "Automatic Phoneme Alignment Based on Acoustic-Phonetic Modeling". *International Conference on Spoken Language Processing-ICSLP'02*, Boulder, Co., vol. I, pp 357-360, Sep. (2002).

[08] Juneja, A. and Espy-Wilson, C. "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines". *Proceedings of International Joint Conference on Neural Networks*, Portland, Oregan, (2003).

[09] Russell, Martin J. and Bilmes, Jeff A. "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).

[10] Santos, S. C. and Alcaim, Abraham. "Sílabas como unidades fonéticas para o reconhecimento de voz em Português", *SBA Controle & Automação.* vol. 12, nº 01. (2001).

[11] Silva, T. C.. *Fonética e Fonologia do Português.* Ed. Contexto, 7º Edição, São Paulo-Brazil, (2003).

[12] Stevens, S. S. Volkman, J. e Newman, E. B. "A Scale for Measurement of the Psychological Magnitude Picth". *Journal of the Acoustical Society of America,* vol. 08, pp. 185-190, January (1937).

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

[14] Young, S. "A Review of Large-Vocabulary Continuous-Speech Recognition". *IEEE Signal Processing Magazine,* pp. 45-57. Set. (1996).