A New Hierarchical Decision Structure Using Wavelet Packet and SVM for Brazilian Phonemes Recognition

Adriano de A. Bresolin[1], Adrião Duarte D. Neto[2], and Pablo Javier Alsina[2]

[1] UTFPR - Technological Federal University of the Paraná – Brazil
Postal Box: 271, Av. Brasil, 4232
CEP 85.884-000, Medianeira, PR, Brazil
[2] UFRN - Federal University of the Rio Grande do Norte – Brazil
Postal Box: 1524, Campus Universitário Lagoa Nova
CEP 59072-970, Natal, RN, Brazil
{aabresolin, adriao, pablo}@dca.ufrn.br

Abstract. In this work, a new phonemes 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 97.50% for the user-dependent case and 91.01% for the user-independent case. This new proposal increased 3.5% the success rate in relation to the "one vs. all" decision strategy.

Keywords: 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, what demands great amount of processing and storage, a hard problem for continuous recognition. In the last years, research efforts have focused the unit smaller than the word. Santos and Alcaim [10], used syllables as units of recognition. However, 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. The oral vowels **(a, é, i, ó, u, ê, ô**), were used in the recognition.

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.

I. King et al. (Eds.): ICONIP 2006, Part II, LNCS 4233, pp. 159 – 166, 2006.

Section 2, presents the signal pre-processing phase. Section 3, shows the speech features extraction. Section 4, describes the training procedure of SVM neural network. Section 5, proposes a new technique for vowel recognition. 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.

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

3 Features Extraction Using Wavelet Packets and Mel Scale

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 what cannot be achieved by using discrete wavelet transform [1].

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 and Datta in [4] had used Wavelet Packet with the Mel scale, which was found to be superior to Mel Frequency Cepstral Coefficients (MFCC) in unvoiced phoneme classification problem.

Gowdy and Tufekci 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.

First, a full seven level WP decomposition is carried out. Twelve subbands 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]).