ISOLATED WORD RECOGNITION SYSTEM FOR ROBOT CONTROL USING NEURAL NETWORKS

Rodrigo Jorge Alvarenga, rodalv87@gmail.com
Pedro Paulo Leite do Prado, pplprado@ieee.org
Universidade de Taubaté
Rua Daniel Danelli, s/n, Jardim Morumbi, 12060-440, Taubaté, SP, Brasil

Abstract. *This project was aimed at developing and implementing a system able to recognize and execute voice commands to control the movements of a robot. It is a speaker-independent system for the recognition of isolated words. The six voice commands were: left, right, back, stop and turn. Each voice command was recorded many times by several different speakers. In the pre-processing phase, we used a filter to discard the background noise, in order they would not affect the feature extraction and the neural network training. The speech features were: LPC coefficients, short-term energy and zero cross rating. These features were arranged in batch as the input matrix to the neural network. The output matrix presented the command codewords. The neural network was trained with backpropagation algorithm. The recognition of each command was informed to the PC serial port, which was connected to a microcontrolled electronic circuit. The results of the tests proved to be very good. The worst score for correct recognition was 95%.*

Keywords*: speech processing, neural networks, isolated word recognition, robot control.*

1. INTRODUCTION

This project was aimed at developing a speaker-independent isolated word recognition system, using neural networks, in order to control the movements of a robot. The commands were:

- Direita (right);
- Esquerda (left);
- Em Frente (forward);
- Trás (backward);
- Pare (stop);
- Desligue (turn off).

The application was developed with the Signal Processing Toolbox of MATLAB®. Figure 1 shows the phases of the development (Alvarenga, 2010).

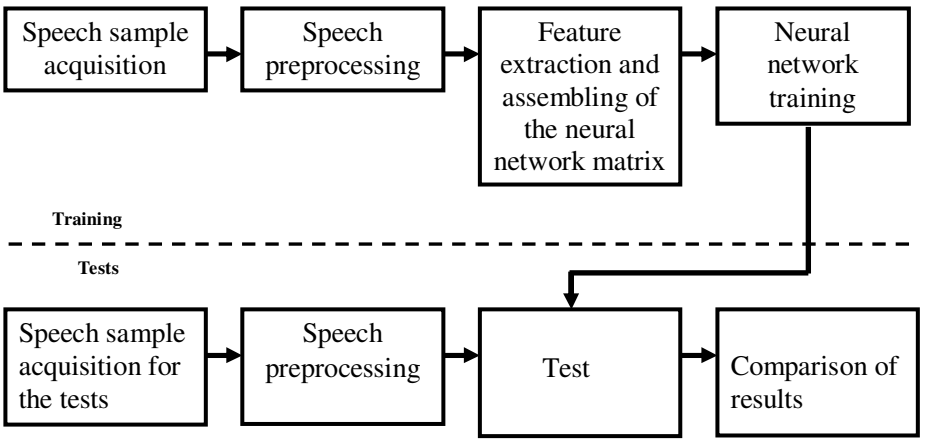

Figure 1. Phases of the Development

2. SPEECH PREPROCESSING

First, the speech sample amplitudes were normalized. Then, an endpoint detector deleted the edges in the signal beginning and end, where there was only noise, as illustrated in Fig. 2 and Fig. 3 (Alvarenga, 2010).

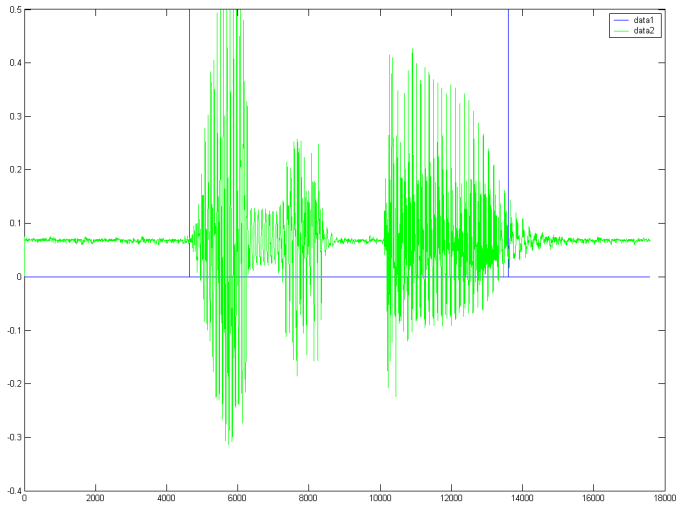

Figure 2. Endpoint detection

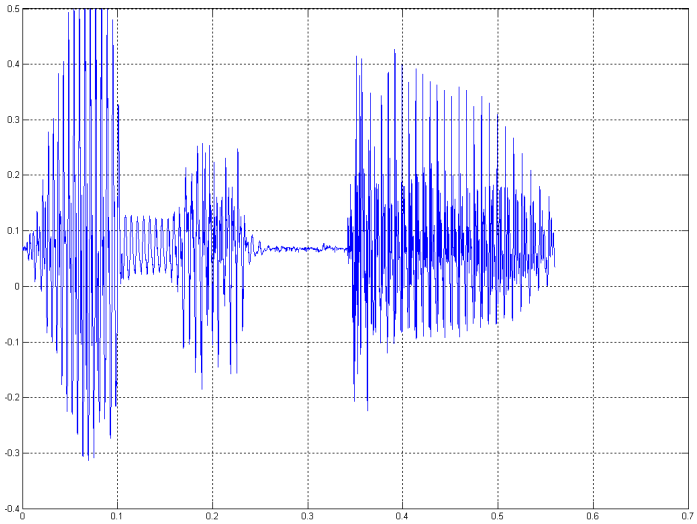

Figure 3. Endpoint extraction

The next stage was a digital low-pass filter, FIR and with cutoff frequency of 3 kHz, intended to reduce or to cancel the environmental noise. Its effects are shown in Fig. 4 and Fig. 5 (Alvarenga, 2010).

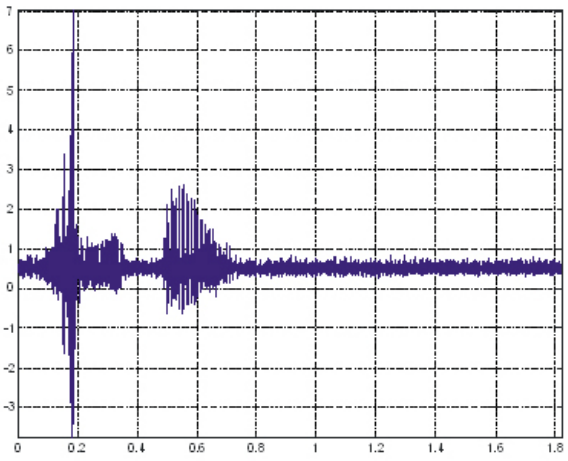

Figure 4. Speech sample before the filter FIR

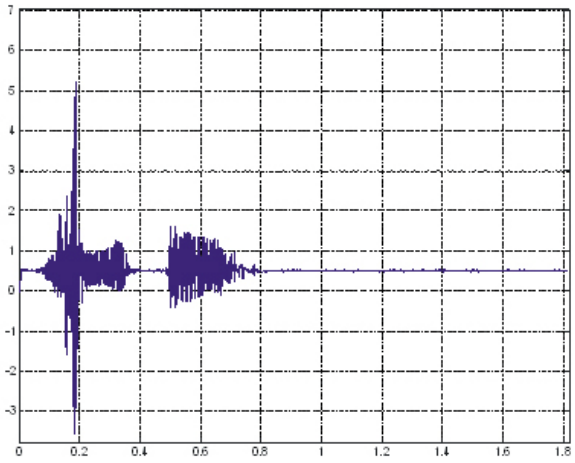

Figure 5. Filtered speech sample

In the last phase, the speech segmentation and windowing enabled the feature extraction task (McLoughlin, 2009). The entire speech was segmented in 50 overlapping windows. The rectangular window was a simple and effective choice.

3. FEATURE EXTRACTION

Three voice features were enough for our proposal: ZCR (Zero Crossing Rate), STE (Short Time Energy) and LPC (Linear Prediction Coding) coefficients. These features were extracted from each window.

3.1. Zero Crossing Rate (ZCR)

The zero crossing rate of a frame is defined by Eq. (1):

$$ZC(m) = \frac{1}{2} \sum_{n=m-N+1}^{m} \left| \text{sgn}(s[n]) - \text{sgn}(s[n-1]) \right| \tag{1}$$

where:
ZC(m): ZCR for a time *m*;
sgn( ): function which returns "1" if the sample is positive or "-1" if the sample is negative;
s(n): speech signal;
N: length of the window in samples.

Equation (1) computes sample pairs and verifies their signs. A zero crossing is computed when the samples have different sign.

ZCR feature usually have higher values for unvoiced (or nasal) phonemes, which have higher frequency content and are more similar to noise.

3.2. Short Time Energy (STE)

Short time energy represents the speech power in a window, as shown in Eq (2) (Rabiner and Schafer, 1978).

$$E(n) = \sum_{n=-\infty}^{\infty} s^2(m)\, h(n-m) \tag{2}$$

where:
E(n): short time energy;
s(n): speech signal;
h(n): rectangular window.

Figure 6 (Now Publishers, 2007) depicts a comparison between STE and ZCR. STE usually have higher values for voiced phonemes.

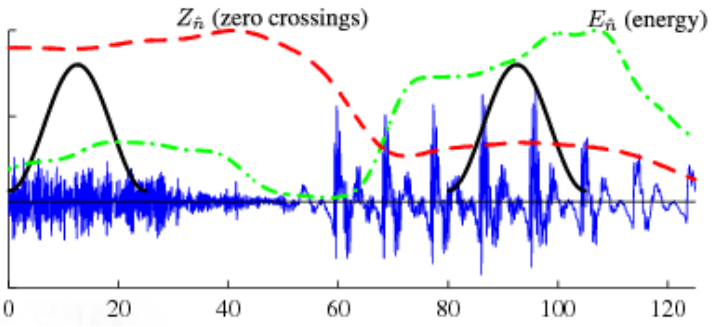

Figure 6. Comparison of ZCR and STE behavior on speech signals

3.3. LPC Coefficients

LPC is one of the most powerful tools in speech processing and it is used to obtain other important parameters, such as, speech pitch and formant values.

It is based on the prediction of a sample, using a linear combination of the previous ones, as represented in Eq. (3) (Rabiner and Juang, 1993).

$$s(n) \approx a_1\, s(n-1) + a_2\, s(n-2) + a_3\, s(n-3) + \ldots + a_p\, s(n-p) \tag{3}$$

where:
a_1, a_2, … a_p: LPC coefficients.

Equation (4) represents the model of a speech generator based on LPC (Rabiner and Juang, 1993).

$$s(n) = \sum_{i=1}^{p} a_i\, s(n-1) + G\, u(n) \tag{4}$$

where:
$u(n)$: excitation;
G: gain.

4. NEURAL NETWORK

The features extracted from the windows composed a matrix which was the input for a neural network (Fig 7). The output matrix columns C_j (j: 1, 2, 3, 4, 5, 6) are the target codes of the six commands (Fig. 8) (Alvarenga, 2010).

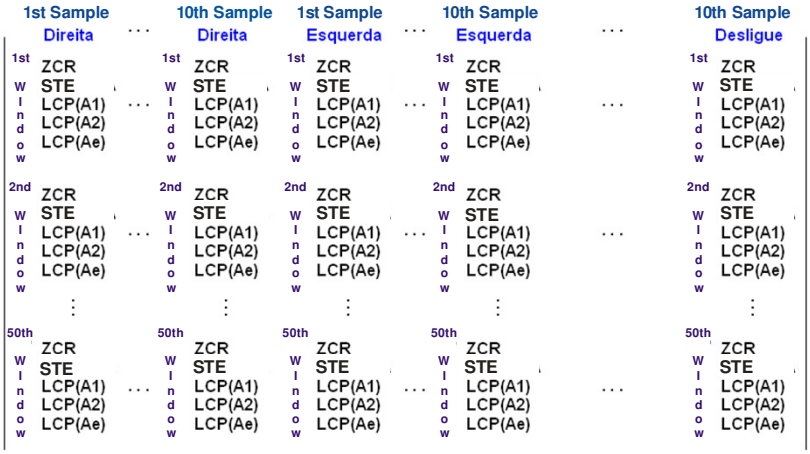

Figure 7. Example of the input matrix for the neural network

Direita	Esquerda	Frente	Tras	Pare	Desligue
1	1	1	1	1	-1
1	1	1	1	-1	1
1	1	-1	-1	1	-1
1	1	-1	-1	-1	1
1	-1	1	-1	1	-1
1	-1	1	-1	-1	1
1	-1	-1	1	1	-1
1	-1	-1	1	-1	1

Figure 8. Output matrix of the neural network

The neural network used a backpropagation algorithm (Nunes, 2000). The system was trained with 60 voice commands of 6 different speakers.

5. TESTS AND RESULTS

During the test, 3 speakers who do not belong to training set, pronounced 10 times each one of the six commands. This voice command was then processed by the speaker recognition system and resulted in an 8-bit vector T, similar to one of the output matrix columns shown in Fig. 8.

The final step consisted in verifying which command C_j was the nearest to the test sample T, by calculating the cross-correlation between T and each C_j. The largest cross-correlation value represented the target code (one of the columns of the output matrix). There is a correct recognition when the voice command matches the right target code.

The results proved to be very good because the worst score for correct recognition was 95%.

6. REFERENCES

Alvarenga,R.J., 2010, “Reconhecimento de Palavras Isoladas para Comando de Robô” Universidade de Taubaté.

McLoughlin I., 2009, “Applied Speech Audio Processing”, Cambridge University Press

Now Publishers, 2007 “Foundations and Trends in Signal Processing”
<http://www.nowpublishers.com/product.aspx?product=SIG&doi=2000000001&section=x1-56r1>

Nunes, L.E.N.P. , 2000, “Redes Neurais Aplicadas ao Reconhecimento de Padrões” ,Tese de Mestrado, Universidade de Taubaté.

Rabiner, L.R. and Schafer, R.W, 1978, “Digital Processing of Speech Signals”,Prentice-Hall.

Rabiner, L. R. and Juang, B. H., 1993, “Fundamentals of Speech Recognition” Prentice Hall, Englewood Cliffs, New Jersey.

7. RESPONSIBILITY NOTICE

The authors are the only responsible for the printed material included in this paper.