A Voice Recognition System for Speech Impaired People

José Leonardo Plaza-Aguilar, David Báez-López, Luis Guerrero-Ojeda and Jorge Rodríguez Asomoza
Departamento de Ingeniería Electrónica
Universidad de las Américas-Puebla
Cholula, Pue. México

Abstract – **On this paper, we will show how MATLAB can be used to realize Voice Recognition; for people with Cerebral Paralysis. The first part of this paper is about what is the Cerebral Paralysis, and its consequences in the way of talk of the people with this illness. After this, we will mention the mathematic background needed to realize this project besides of explaining a first stage of the system, on this stage we could recognize some phonemes. Next, we will describe the functioning of the second stage of the project. On this part, we could recognize words. At the end of the paper, we will give some results and we will discuss them**

I. INTRODUCTION

When we hear someone and we cannot understand what he is saying, we connect the few sounds that we understand with the ones that we know, and taking into account its form, tone, accent and pronunciation, we assign an equivalent word. This presents a problem, how can we know what is the equivalent word to the incomplete phrase? When we spend a great deal of time with the handicapped person, our ear can find some characteristics related with his frequency in the way that the person talks.

Each person has a different voice and different characteristics on his intonation. We can see this on its graphic of frequency. In this way there are a different intonation between the letters that form the word *"hola"* to the ones that form the word ***"o-t-r-a"*** or ***"f-o-c-a"***, that are the cases when the right pronunciation is completely different.

Using our intuition, when we are talking to someone who cannot speak well, we can know if the person is saying ***"hola"*** o ***"foca"***, the reason of this, is that our ear is used to hear the different tones and pronunciations and to give it a right interpretation based on its unique intonation.

Based on these facts, we designed a system that could help us, to differentiate the different words told by people with speech problems. In the beginning, we wanted a system capable of helping three different groups of people. The groups were:

1.	***Understandable***	
	and Hearing 1.	***Recognition of 70%***
2.	***Understandable***	
3.	***and Hearing 2.***	***Recognition of 60%***
4.	***Hearing.***	***Recognition of 50%***

When we speak with people that present Cerebral Paralysis, we can see a group of people that can express themselves so well, that we can understand the things they say. We called this group: *"Understandable and Hearing 1 and 2"*. On the other hand, there is a group that produces only sounds, but they do not pronounce an understandable structure in their phrases. We called this group *"Hearing"*.

Taking this into count, we performed a field job. We took samples of voice. These samples were from people of belonging to groups 2 and 3.

1.1. ¿What is the Cerebral Paralysis?

The term *cerebral* is about the two halves of the brain or hemispheres. Paralysis describes any problem that limit the control of the movement of the human body. These problems are not caused by trouble in the muscles or nerves. In the opposite way, a defective development in the motor areas of the brain interrupts the capability of the brain to control movement and posture.

Opposite to common believe, Cerebral Paralysis is not always the cause of significant disabilities. A child with a strong Cerebral Paralysis cannot walk and needs a lot of care, a child with a low Cerebral Paralysis could pass as someone clumsy. Cerebral Paralysis is not contagious and usually is not hereditary. However, now a days can not be cured.

B. Some Treatments

Therapy is the most important treatment for the Cerebral Paralysis [2]. Physical therapy usually begins in the first years of life of the patient after the diagnosis. The programs of physical therapy use combinations of exercises to achieve important goals, such as: Prevent the deterioration of the muscles, to avoid the contracture that appears when muscles do not present movement.

To children with difficulties to communicate their ideas, speech therapy identifies the specific problems and works to overcome throughout programs of exercises. Such as: if the child has problems to pronounce words that begin with the letter "b" , the doctor can suggest the practice of a list of words that begin with the letter "b", raising the difficulty of the words according to each list [2].

It is probably that the computer will be the most important example of a new device that can make a difference in the lives of the children with Cerebral Paralysis. Such as a child that can not speak or write, but if the child can move his head, he can learn to control a computer. With a good equipment a child can communicate himself with other people. In other cases, the technology has produced new versions of well proven devices such as the traditional wheelchair.

II. MATHEMATICAL BACKGROUND

Our investigation is based in a strong way, in the Fourier Transform. Next we will explain in a very simple way its principal characteristics.

A. Fourier Analysis

The Discrete Fourier Transform (DFT) is defined by:

$$X(e^{(i\,\theta)}) = \sum_{n=-\infty}^{\infty} x(n)\, e^{(-i\,n\,\theta)} \qquad (2.1)$$

where:

$$\theta = 2\,\pi f T \qquad (2.2a)$$

$$\theta = 2\,\frac{\pi f}{f_s} \qquad (2.2b)$$

Where *T* is the period of sampling and f is the frequency of sampling. The inverse transformation is defined by:

$$x(n) = \frac{1}{2}\,\frac{\int_{-\pi}^{\pi} X(e^{(i\,\theta)})\, e^{(i\,n\,\theta)}\, d\theta}{\pi} \qquad (2.3)$$

B. Fast Fourier Transformation

An easy interpretation requires *N* operations. The FFT suppose *N log N* operations, and if *N* is a factor of two. In general the techniques of the FFT style can be used to any N. First we decompose N in prime factors and after that, we realize a butterfly operation on each factor [1].

We start the Fourier Transformation of highest level described by:

$$X_n = \sum_{n=0}^{N-1} x_n\, e^{\left(2\,\frac{i\,\pi\,n\,k}{N}\right)} \qquad k = 0, \ldots, N-1 \quad (2.4)$$

After writing *X₁(k)* as an addition of even and odd terms:

$$X_1(k) = \left(\sum_{n=0}^{N} x_{2n}\, e^{\left(-8\,\frac{i\,npi\,k}{N}\right)}\right) + N + 1 + \left(\sum_{n=0}^{N} x_{2n}\right) \qquad (2.5)$$

$$X_1(k) = \left(\sum_{n=0}^{N} x_{2n}\, e^{\left(-4\,\frac{i\,npi\,k}{N}\right)}\right) + e^{\left(2\,\frac{i\,\pi\,k}{N}\right)} + N + 1 + \left(\sum_{n=0}^{N} x_{2n}\right) \qquad (2.6)$$

$$X_1(k) = \left(\sum_{n=0}^{N} x_{2n}\, e^{\left(-8\,\frac{i\,npi\,k}{N}\right)}\right) + e^{\left(2\,\frac{i\,\pi\,k}{N}\right)} + N + 1 + \left(\sum_{n=0}^{N} x_{2n}\right) \qquad (2.7)$$

$$X_1(k) = X_{11}(k) + e^{\left(-2\,\frac{i\,\pi\,k}{N}\right)} X_{12}(k) \qquad (2.8)$$

The original Fourier Transformation has been written as two Fourier Transformations operating on the even and odd parts of the data [2].

III FIRST STAGE OF THE INVESTIGATION

MATLAB has the flexibility to implement complex algorithms for digital processing of signals. The basis to our MATLAB programs are:

- Digitalization of a file of sound.
- Implementation of the FFT to represent the signal in the dominion of the frequency.
- Drawing of the results.

It is necessary to convert our samples in a series of data that we can interpret. To do this, we record our samples of sound in an archive with ".wav" extension, and we retrieve it using the instruction wavread, both of them included in MATLAB.

Now that we are working in the frequency domain, it is necessary that our vector of data presents certain conditions. To represent the signal in the frequency domain we will use the DFT. A way to apply the DFT, is using the FFT. The condition to apply the FFT is that the length of our vector must be a power of two. Next we will get the FFT of our vectors. The real and imaginary components of the FFT of the vectors are stored in the vector *x.* Finally *mag* represents the magnitude of the FFT *(x).*

Next we will show the algorithm to draw the representation of our file of sound in the frequency dominion.

```
f=(0:m/2)*Fs/m;
subplot(221), plot(s), axis tight, grid on, title('Señal
de Voz');
subplot(222),specgram(s),title('Espectograma'),
colorbar;
subplot(223), plot(f,mag(1:m/2+1)), axis ([0 5000 0
10]), grid on, xlabel('Frecuencia
(Hz)'),ylabel('Magnitud'), title('Representación en
Frecuencia');
print -djpeg99 d:\Tesis\a
```

The result of this is an image with the following parts:

a) Voice Signal.- This is the plot of the data of the vector*s.*
b) Spectrum.- The spectrum of the vector*s.*
c) Representation of the Frequencies.- This plot represents the parameters obtained in this stage: frequency (*f*) vs. magnitude (*mag*).

In Fig. 3.1 we present the results obtained for a vowel.

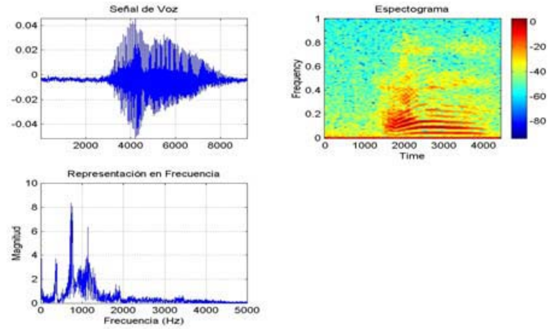


Figure 3.1 Pronunciation of the vowel "A".

We can see now that it is possible to realize a voice recognition using MATLAB.

There are other designs to realize voice recognition. However, they have presented some problems, such as:

- Slow recognition of sounds.
- Slow time of operation.
- Inestability.

A cause of these drawbacks is that the voice is seen as a series of data and is not given a visual interpretation of what it means. Considering that the voice is a signal with a characteristic for in the time and frequency domain, we made a graphic analysis of this signal.

As we could see in the last paragraph, the spectrum of the signal (Frequency Representation) presents maximum and minimum values, with a different distribution in each sound (except those ones how has the same pronunciation). Now if the intensity of the signal changes, the magnitude of the spectrum changes too.

Although the form of the signal is conserved it, the scale of values are not.. This produces confusion in the sound recognition, because these values could fall inside the corresponding range of another sound. To solve this problem, we will do the following:

1) Stabilize of the signal. To realize this, we made a normalization of the signal.
2) After that we use the FFT in our vector *c* , we will get the its magnitude.

Using this, we will keep the maximums and the minimums. With this we could get a better difference between these values and the corresponding values of other sounds. With this we have a low probability of confusion. The results of this procedure are seen in the Fig. 3.2

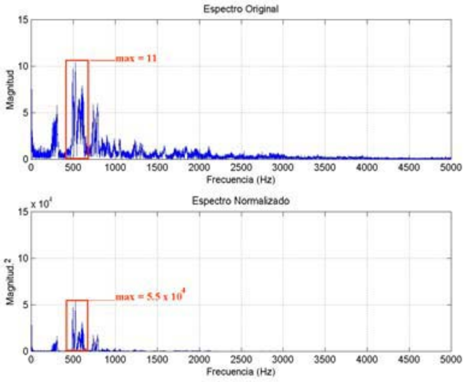


Figure 3.2 Comparison between the original spectrum and the normalized spectrum of the phoneme EME.

We can note, that the noise has disappeared for frequencies over 1000 Hz [3]. Now we will take as dominant maximum the pulse between 400 and 700 Hz. We will find that in the original spectrum there is a little difference (11-10=1). However, in the normalized spectrum the difference between them is larger ($8.5x10^4$ - $5.5x10^4$ = $3x10^4$). Using this procedure we can get a better recognition of the sounds [3]. The maxima and minima of the spectrum change in different ranges of frequency, and we can separate them in bands of 200 Hz (Using passband filters). These bands can be established between even and odd frequencies. That is shown in Fig. 3.3. With the assignation of the bands we split the spectrum in different "zones of action". In this way we can differentiate the sounds by their different magnitudes on these bands. To realize the voice recognition we will use the range of 500 Hz to 4000 Hz. In table 3.1 we define our operation bands.

A. Recording of the Phonemes.

We will do the voice recognition using phonemes and vowels. If we group them, we can get phrases and words. The patient will have a extensive vocabulary to communicate his ideas.

Our system is focused to the voice recognition of two kinds of people:

1. Understandable and Hearing. – This is when the patient can articulate some phonemes in a right way.
2. Hearing. – This is when the patient can only say some sounds.

The name of the child that we use to realize this investigation is Arturo Calderón Caro, and his illness is advanced. In addition we choose another child her name is Erika Castelán Caballero and her illness is medium [3].

The record of the phonemes was realized with the program RecAll Versión 2.4a de Sagebrush Systems [4]. Each one of the phonemes was recorded 10 times. Of the recordings we chose the best five to analyze it. These five recordings are those ones that presented less noise.

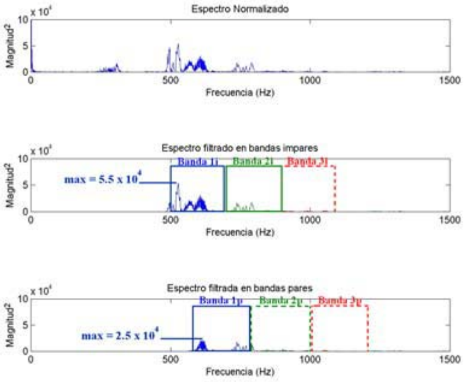


Figure 3.3 Separation in bands of the spectrum of the phoneme EME.

		BANDAS IMPARES				BANDAS PARES	
1 i	🟦	500 Hz -	700 Hz	1 p	🟦	600 Hz -	800 Hz
2 i	🟩	700 Hz -	900 Hz	2 p	🟩	800 Hz -	1000 Hz
3 i	🟥	900 Hz -	1100 Hz	3 p	🟥	1000 Hz -	1200 Hz
4 i	🟦	1100 Hz -	1300 Hz	4 p	🟦	1200 Hz -	1400 Hz
5 i	🟪	1300 Hz -	1500 Hz	5 p	🟪	1400 Hz -	1600 Hz
6 i	🟨	1500 Hz -	1700 Hz	6 p	🟨	1600 Hz -	1800 Hz
7 i	⬛	1700 Hz -	1900 Hz	7 p	⬛	1800 Hz -	2000 Hz
8 i	🟦	1900 Hz -	2100 Hz	8 p	🟦	2000 Hz -	2200 Hz
9 i	🟩	2100 Hz -	2300 Hz	9 p	🟩	2200 Hz -	2400 Hz
10 i	🟥	2300 Hz -	2500 Hz	10 p	🟥	2400 Hz -	2600 Hz
11 i	🟦	2500 Hz -	2700 Hz	11 p	🟦	2600 Hz -	2800 Hz
12 i	🟪	2700 Hz -	2900 Hz	12 p	🟪	2800 Hz -	3000 Hz
13 i	🟨	2900 Hz -	3100 Hz	13 p	🟨	3000 Hz -	3200 Hz
14 i	⬛	3100 Hz -	3300 Hz	14 p	⬛	3200 Hz -	3400 Hz
15 i	🟦	3300 Hz -	3500 Hz	15 p	🟦	3400 Hz -	3600 Hz
16 i	🟩	3500 Hz -	3700 Hz	16 p	🟩	3600 Hz -	3800 Hz
17 i	🟥	3700 Hz -	3900 Hz	17 p	🟥	3800 Hz -	4000 Hz

Table 3.1 Frequency bands.

To the 5 recordings of each phoneme we applied a program made in MATLAB, with this program we obtained the spectrum of the filtered signal in different bands. Finally, from these spectra we picked the most common and we call it characteristic standard of the phoneme. From a total of 75 phonemes and 5 vowels, we could differentiate 8 phonemes to each child. The most important reason to only recognize this quantity is that the characteristic standard of the phonemes of the children were almost the same. We can see this in Fig. 3.4. On the other hand, some phonemes we could detect show particularities in their standards.

The phonemes *E* and *NE* can be differentiated by the pulse in the band 3i. In this way we separate the following phonemes:

To Arturo: BA, BE, E, ME, MI, NA, NE, YI
To Erika: A, BI, BO, CHA, DI, FA, JA, PU

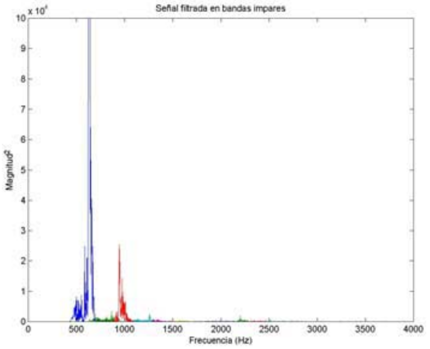


Figure 3.4 Standard of the vowel *I*.

.

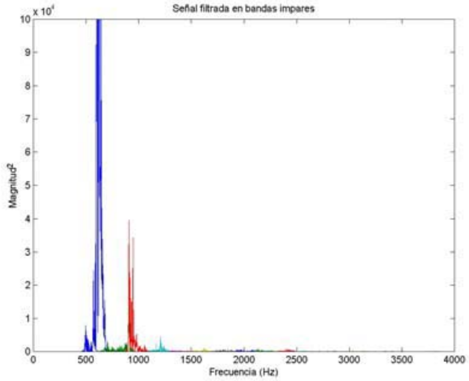


Figure 3.5 Standard of the vowel *E*.

As we can see, the number of vowels and phonemes recognized was very little. The reason was that the patients pronounced exactly the same all the vowels and phonemes. And the spectrum was the same for the even and odd bands of the frequency spectrum.

IV SECOND STAGE

A. Selection of the Patients.

We worked with 3 patients: Graciela, Martín and Perla. The investigation was divided in three parts because each one of our patients presented a different damage in their vocal strings. Graciela falls in the rage of *Understandable and Hearing 2,* Martín and Perla fall in the range that we called *Hearing*.

B. Recognition of the Words

On this second stage of our investigation we decided to create a small groups of words to recognize a larger number of words. Next, we will show the groups:

- ***First Group (Family):*** Mamá, papá, hermano, hermana, tío, tía, primo, prima, abuelo y abuela.

- ***Second Group (Words inside the Foundation):*** Comida, baño, lápiz, plumón, goma, mesa, silla, dormir, libro y cuaderno.
- ***Third Group (Words that we use often):*** Casa, avión, camión, auto, cama, sofá, lámpara, perro, gato, ratón.
- ***Fourth Group (Meals):*** Apple, pear, fish candy, pera, pescado, dulce, pollo, pan, carne, paleta, helado y sopa.
- ***Fifth Group (Position):*** Aquí, cerca, lejos, ven y dame.
- ***Sixth Group (Human body):*** Corazón, cabeza, estómago, pierna y brazo.

We were working with the patients some weeks. Basically, we ask them to repeat 20 times each word of the groups. That was recorded with a tape recorder. From this 20 recordings, we choose the best 15 to analyze them. These 15 recordings were the ones that presented less noise.

On these recordings we use a MATLAB program to obtain the spectrum of the signal in the different filtered bands. Finally, from these spectra, we chose the most common, and we call it the *characteristic standard*. Using these standards, we could watch the words that could be recognized on each group. We can see an example of these standards in Fig. 4.1 and 4.2.

. From these graphics, we can see that these words can be differentiated in the band 3i. Once we obtained the patterns, we generate the corresponding algorithms to be implemented in MATLAB. We can see this in Fig. 4.3

V RESULTS

Finally we took new voice recordings to probe the system, we used 10 recordings of each word; the results are next As a final adjustment, we can see that the program of recognition has a conditional. If a sound that does not achieve the conditions established by the parameters inside of the algorithm, then the sound is not kept, and the program will return a Symbol. Our program works in a cycle way. With this, the program will be functioning all of the time [3].

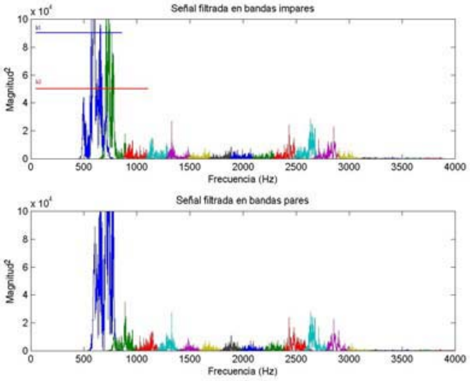


Figure 4.1 Characteristic standard of the word Primo.

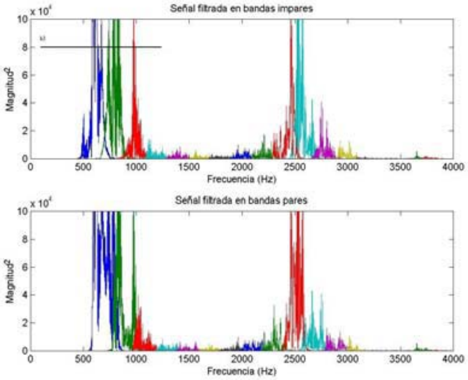


Figure 4.2 Characteristic Standard of the word Tío.

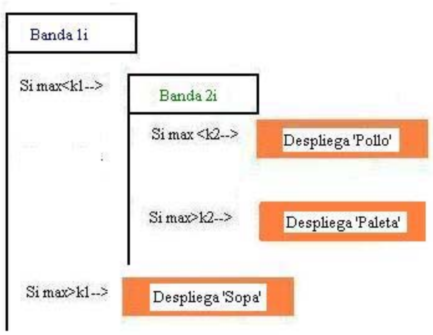
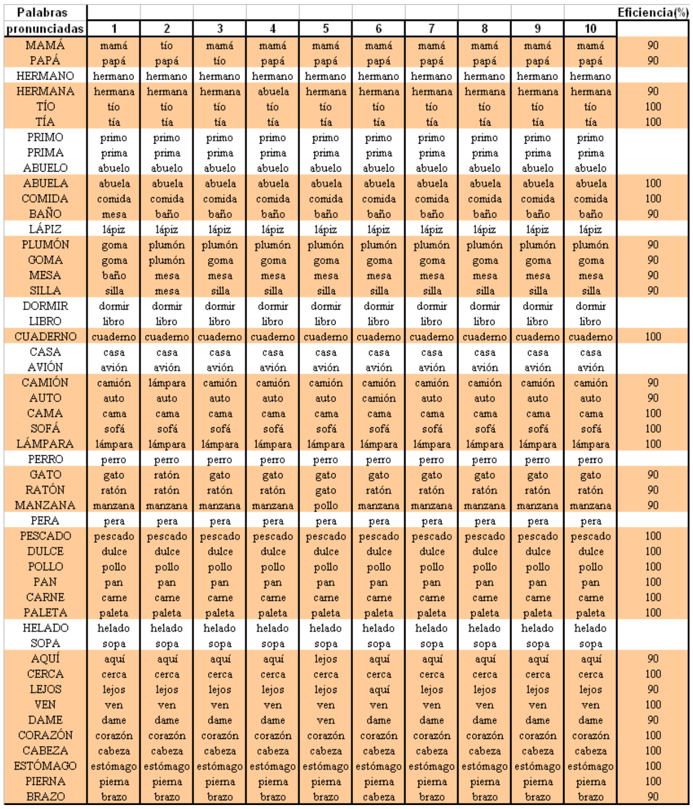


Figure 4.3 Algorithm for the Voice Recognition of the Fourth Group of Perla. (Group A).



Palabras											Eficiencia(%)
pronunciadas	1	2	3	4	5	6	7	8	9	10	
MAMÁ	mamá	tío	mamá	mamá	mamá	mamá	mamá	mamá	mamá	mamá	90
PAPÁ	papá	papá	tío	papá	papá	papá	papá	papá	papá	papá	90
HERMANO	hermano	hermano	hermano	hermano	hermano	hermano	hermano	hermano	hermano	hermano	
HERMANA	hermana	hermana	hermana	abuela	hermana	hermana	hermana	hermana	hermana	hermana	90
TÍO	tío	tío	tío	tío	tío	tío	tío	tío	tío	tío	100
TÍA	tía	tía	tía	tía	tía	tía	tía	tía	tía	tía	100
PRIMO	primo	primo	primo	primo	primo	primo	primo	primo	primo	primo	
PRIMA	prima	prima	prima	prima	prima	prima	prima	prima	prima	prima	
ABUELO	abuelo	abuelo	abuelo	abuelo	abuelo	abuelo	abuelo	abuelo	abuelo	abuelo	
ABUELA	abuela	abuela	abuela	abuela	abuela	abuela	abuela	abuela	abuela	abuela	100
COMIDA	comida	comida	comida	comida	comida	comida	comida	comida	comida	comida	100
BAÑO	mesa	baño	baño	baño	baño	baño	baño	baño	baño	baño	90
LÁPIZ	lápiz	lápiz	lápiz	lápiz	lápiz	lápiz	lápiz	lápiz	lápiz	lápiz	
PLUMÓN	goma	plumón	plumón	plumón	plumón	plumón	plumón	plumón	plumón	plumón	90
GOMA	goma	plumón	goma	goma	goma	goma	goma	goma	goma	goma	90
MESA	baño	mesa	mesa	mesa	mesa	mesa	mesa	mesa	mesa	mesa	90
SILLA	silla	mesa	silla	silla	silla	silla	silla	silla	silla	silla	90
DORMIR	dormir	dormir	dormir	dormir	dormir	dormir	dormir	dormir	dormir	dormir	
LIBRO	libro	libro	libro	libro	libro	libro	libro	libro	libro	libro	
CUADERNO	cuaderno	cuaderno	cuaderno	cuaderno	cuaderno	cuaderno	cuaderno	cuaderno	cuaderno	cuaderno	100
CASA	casa	casa	casa	casa	casa	casa	casa	casa	casa	casa	
AVIÓN	avión	avión	avión	avión	avión	avión	avión	avión	avión	avión	
CAMIÓN	camión	lámpara	camión	camión	camión	camión	camión	camión	camión	camión	90
AUTO	auto	auto	auto	auto	auto	camión	auto	auto	auto	auto	90
CAMA	cama	cama	cama	cama	cama	cama	cama	cama	cama	cama	100
SOFÁ	sofá	sofá	sofá	sofá	sofá	sofá	sofá	sofá	sofá	sofá	100
LÁMPARA	lámpara	lámpara	lámpara	lámpara	lámpara	lámpara	lámpara	lámpara	lámpara	lámpara	100
PERRO	perro	perro	perro	perro	perro	perro	perro	perro	perro	perro	
GATO	gato	ratón	gato	gato	gato	gato	gato	gato	gato	gato	90
RATÓN	ratón	ratón	ratón	ratón	gato	ratón	ratón	ratón	ratón	ratón	90
MANZANA	manzana	manzana	manzana	manzana	pollo	manzana	manzana	manzana	manzana	manzana	90
PERA	pera	pera	pera	pera	pera	pera	pera	pera	pera	pera	
PESCADO	pescado	pescado	pescado	pescado	pescado	pescado	pescado	pescado	pescado	pescado	100
DULCE	dulce	dulce	dulce	dulce	dulce	dulce	dulce	dulce	dulce	dulce	100
POLLO	pollo	pollo	pollo	pollo	pollo	pollo	pollo	pollo	pollo	pollo	100
PAN	pan	pan	pan	pan	pan	pan	pan	pan	pan	pan	100
CARNE	carne	carne	carne	carne	carne	carne	carne	carne	carne	carne	100
PALETA	paleta	paleta	paleta	paleta	paleta	paleta	paleta	paleta	paleta	paleta	100
HELADO	helado	helado	helado	helado	helado	helado	helado	helado	helado	helado	
SOPA	sopa	sopa	sopa	sopa	sopa	sopa	sopa	sopa	sopa	sopa	
AQUÍ	aquí	aquí	aquí	aquí	lejos	aquí	aquí	aquí	aquí	aquí	90
CERCA	cerca	cerca	cerca	cerca	cerca	cerca	cerca	cerca	cerca	cerca	100
LEJOS	lejos	lejos	lejos	lejos	lejos	aquí	lejos	lejos	lejos	lejos	90
VEN	ven	ven	ven	ven	ven	ven	ven	ven	ven	ven	100
DAME	dame	dame	dame	dame	ven	dame	dame	dame	dame	dame	90
CORAZÓN	corazón	corazón	corazón	corazón	corazón	corazón	corazón	corazón	corazón	corazón	100
CABEZA	cabeza	cabeza	cabeza	cabeza	cabeza	cabeza	cabeza	cabeza	cabeza	cabeza	100
ESTÓMAGO	estómago	estómago	estómago	estómago	estómago	estómago	estómago	estómago	estómago	estómago	100
PIERNA	pierna	pierna	pierna	pierna	pierna	pierna	pierna	pierna	pierna	pierna	100
BRAZO	brazo	brazo	brazo	brazo	brazo	cabeza	brazo	brazo	brazo	brazo	90

Table 4.5 Results of Martín ‘s Programs

V CONCLUSIONS

This investigation gave us interesting results. The most important is that if we have a great number of groups we can recognize more words. Besides if the groups are small our opportunities of success are better.

On the other hand, we can say to future investigations, that if you have a big number of samples of a word, you can get a better characteristic standard. Because with 20 recordings was easier to find it than with 10 recordings.

We could not recognize all the words that we propose to all the groups, because some words are almost the same in their frequency spectrum.

Each patient presented different complications, such as Perla, because of her condition, she said all the words in the same way. In her case, we made two programs to each group of words (except groups 5 and 6) to recognize more words.

We could watch a common problem with our patients, and it is that they have problems to pronounce the letter “r”. This is caused by their problem in their vocal strings.

The recognition of these groups of words must not be limited to the connection of complex phrases and could be use to control machines that are of common use by them.

This system must be completely efficient, because we are talking of people that sometimes can not move around by themselves. That makes impossible that they can use something that will not be activated by their voice.

On the other hand we can not realize a universal system, because every person has a different voice.

REFERENCES

[1] J. G. Proakis and G. D. Manolakis. “Tratamiento Digital de Señales: Principios, Algoritmos y Aplicaciones”. Tercera Edición, Prentice-Hall, México, D.F., 2002.

[2] M. Sanclemente Puyuelo, “Parálisis Cerebral Infantil”. Ediciones Aljí, 1999.

[3] J. L. Plaza, and O. Caballero,. “Reconocimiento de Voz para personas con Discapacidad en el Habla.” Tesis Profesional, Universidad de las Américas-Puebla, 2002.

[4] J. A.Contreras,. “Sistema de ayuda a discapacitados mediante el reconocimiento de voz.” Tesis Profesional, Universidad de las Américas-Puebla, 2001.