Pseudo speech command recognition for people with severe Parkinson's disease using machine learning with time series of voice power

Nicolas Ibanez
Departamento de ingeniería eléctrica
Universidad Santiago de Chile
Santiago, Chile
E-mail: nicolas.ibanez.r@usach.cl

Ismael Soto
Departamento de ingeniería eléctrica
Universidad Santiago de Chile
Santiago, Chile
E-mail: ismael.soto@usach.cl

Esteban Toledo-Mercado
dept. departameto de ingeniería eléctrica
Universidad Santiago de Chile
Santiago, Chile
E-mail: esteban.toledo@usach.cl

***Abstract*—A dataset of non-stationary noise-free audios containing specific voice commands in Spanish is created to control IoT devices. Subsequently, a multi-layer neural network is trained for the classification of these voice commands transformed to time series of the voice power, specifically using the LBFGS stochastic gradient descent method. Predictions above 80% were obtained with the model. In addition, a complementary occupancy detection system installed in the home bathroom is proposed for quadriplegics, people with both arms amputated, and especially people suffering from Parkinson's disease stages 4 and 5 of the Hoehn-Yahr scale is proposed with the aim of offering greater independence and privacy.**

***Index Terms*—Voice Recognition, Mel-scaled Spectrogram, MFCC, STFT, Classification Model, Neural Network, GitHub repository, Arduino, Raspberry Pi, VPN, Adobe Audition, Librosa**

I. Introduction

Modern technology has the potential to provide disabled people a future with a much more independent lifestyle than any physically disabled person enjoys today. [1]. People with severe disabilities, such as quadriplegics or people suffering from Parkinson's disease in stage 4 and 5 of the Hoehn-Yahr scale [2] or people with both arms amputated that do not have the possibility of performing activities with their hands , because of this situation, they require special assistance in their homes, even require relative assistance to perform their needs in their bathrooms, although it is necessary, this could be mitigated in the not too distant future with the help of IoT devices to perform simple and basic activities such as turning on a light.

To date, smart home automation systems have been implemented for the control of elderly people assuming that the test individual lives alone in his or her bedroom with repetitive and identifiable habits [3], on the other hand, it has also been tested a multi-modal interface system that allows the use of voice commands and gestures to control household appliances distributed for disabled quadriplegics by means of the use of helmets with motion sensors [1]. Systems have been developed with real time wireless sensors (WSN-Zigbeey) with the special feature of measuring the light intensity in a room, to form a real-time control system in conjunction with BIG-DATA processing which at the higher level is a database server and application [4]. Also, it is important to consider the latest developments in occupancy detection using machine learning techniques shown in the Table I, highlighting public access to time series data from analog and digital sensors in a GitHub repository [5] for accurate detection of the occupancy of an office room from light, temperature, humidity and CO2 measurements using statistical learning models [6], that although the room humidity measured by DTH22 sensor was not occupied for the detection of a study room, this variable could be more relevant in the detection of occupancy in a bathroom due to the use of water.

In regards to speech in people suffering from an early stage of Parkinson's disease, multiple types of speech recordings of three sustained vowels /a/, /o/ and /u/ have been made by extracting Human Factor Cepstral Coefficients (HFCC) and using SVM Classification, KNN along with the LOSO validation scheme [12]. At a speech-to-text recognition level limited sets of words have been used compared to universal speech recognition systems dedicated to work with the whole set of words of one or several of the languages [13].

For this paper, a GitHub repository was created for an audio data-set of Spanish voice commands [14], to train a multi-layer neural network for time-series focused recognition of the power of human vocal register frequencies applied to these basic voice commands, such as "ENCENDER" or "APAGAR", with the idea of applying it to people with limited arm mobility.

Normally, audio processing applied to speech recognition extracts human factor cepstral coefficients (HFCC) or Mel-scaled cepstral coefficients at Mel-scaled frequencies. What is new, is the treatment with time series data sets of the power of human vocal record frequencies extracted from the Mel-scaled Spectrogram Normalized. In addition, a smart

Table I
OCCUPANCY DETECTION

Source	Classification models used	Sensors/parameters	Occupancy accuracy
[7]	Hidden Markov models, neural networks, Support Vector Machines (SVM)	CO2 inside the room, CO2 outside the room	NONE
[8]	Dirichlet latent assignment	PIR	NONE
[9]	Decision trees (DT)	CO2, luz, PIR, sound	Between 81% and 98.441% (PIR only), Light only: 81.01%, Sound only: 90.78%, CO2 only: 94.68%.
[10]	Neural network with radial basis functions	Lighting, sound, reed sensor, CO2 temperature, RH, PIR	Note: Accuracy for number of occupants 63.23–66.43 %
[11]	Artificial neural networks (MATLAB and WEKA)	CO2, sound, relative humidity, air temperature, computer temperature, PIR	Note: Accuracy for number of occupants: 70,4-72,37 %.

home system is proposed in a private bathroom that would detect from occupancy data "repetitive and identifiable habits" to bypass sensory devices in the head, since they are not useful in people suffering from advanced Parkinson's disease, and so they can turn on the light with their voice.

In the Section II the problem of standard speech recognition in certain handicapped people is described and deepened, a demographic study is carried out to segment a group of people that could be benefited within the theoretical framework of a complementary occupancy detection system installed in a home bathroom. In the Section IV the resolution of the problem and its related terms are described.In the Section V plots of the average signals of the fear series of the power of voice commands and confusion matrices are plotted together with the predictions of the trained neural network. Finally, in the Section VI the conclusion is presented.

II. DESCRIPTION OF THE PROBLEM

Basically, the real problem lies with people with severe Parkinson's disease, specifically stage 4 and 5 of the Hoehn-Yahr scale. It is evident that the sensory helmets used as a complement to voice command recognition are only useful for people who have control of their head movement, for example: quadriplegics or people with both arms amputated, due to the uncontrolled movements of people suffering from severe Parkinson's disease, these people do not have such control over the head, therefore this type of sensory systems on the body do not serve as a complement to a voice command recognition system.

On the other hand, while speech-to-text recognition is a powerful tool, in some cases it can be disastrous for even non-disabled people such as stutterers, although people with severe Parkinson's do not stutter, they do have a distinctive timbre as shown in the Fig. 1.

A. Demographic segmentation

In the Metropolitan Region of Chile, of the distribution of the adult population with a disability, the following are included in the Table II only 23% of adults with severe disabilities have a live-in caregiver to assist them at home [15] [16].

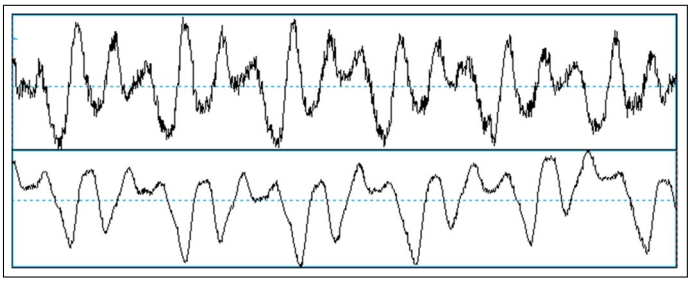

Figure 1. Voice signal of a patient with Parkinson's disease (top) and healthy person (bottom) [12]

Table II
DISTRIBUTION OF THE ADULT POPULATION WITH DISABILITIES

XIII Región Metropolitana	N°	%
Persons without disability status	4,231,646	79.7
People with mild to moderate disabilities moderate	635,355	12
People with severe disabilities	445,722	8.4
People with severe disabilities	1,081,097	20.3
Total population 18 and over	5,312,743	100

While no specific data was found on how many quadriplegic's and severely ill with Parkingson's persons residing in the Metropolitan Region of Chile, the figure of 445,722 Persons with severe disabilities from the Table II is no small number.

III. PROPOSED SOLUTION TO THE PROBLEM

A. Pseudo speech recognition

This paper defines "Pseudo speech recognition" as the recognition of the "intention of the vocal gesture" which is basically the detection of a certain sequence of the power of the voice in time. This is considered convenient to solve the problem described in Section II

B. Proposed installation of a complementary occupancy detection system in a bathroom

The Chilean Electrical Standard `11.0.2.4` [17] makes it easier for paraplegics to install switches 0.8 [m] above floor level so that they are within reach, sockets are not allowed within 0.5 [m] of water sources such as sinks or showers.

In Fig. 2 a demonstration model of a non-isolated bathroom humidity installation for occupancy detection with DHT22

Figure 2. Schematic of the implementation of an occupancy detection system in a bathroom with Arduino nano

humidity sensor, PIR sensor and microphones for pseudo speech recognition using a micro controller, in this case an Arduino nano to make a small and friendly installation not as shown in Fig. 3.

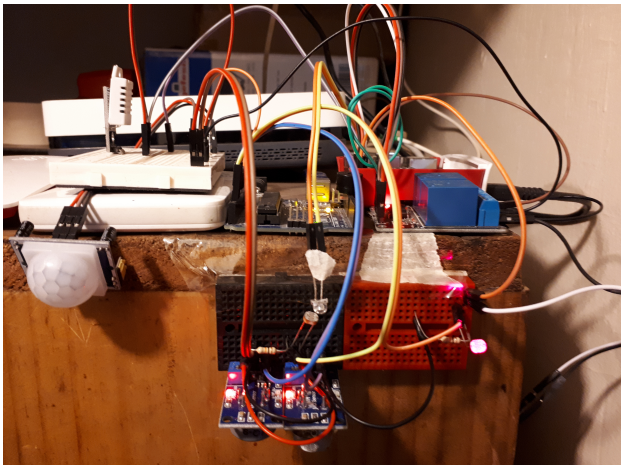

Figure 3. Arduino analog sensor system with serial communication to a RaspberryPi using VPN

In the case of Fig. 2, the program compiled on the Aduino nano uses an Adafruit library [18] that calculates the FFTs of the recorded audios after detection of the occupancy predicted by the system of those installed on an Arduino nano.

In Fig. 3 we present a demonstrative model of an installation for the caregiver and/or family member of disabled to access occupancy data outside the home with the aim of reporting their activity at home. by installing a VPN on a RaspberryPi [19] and serial communication with direct with an Aduino connected to various analog sensors.

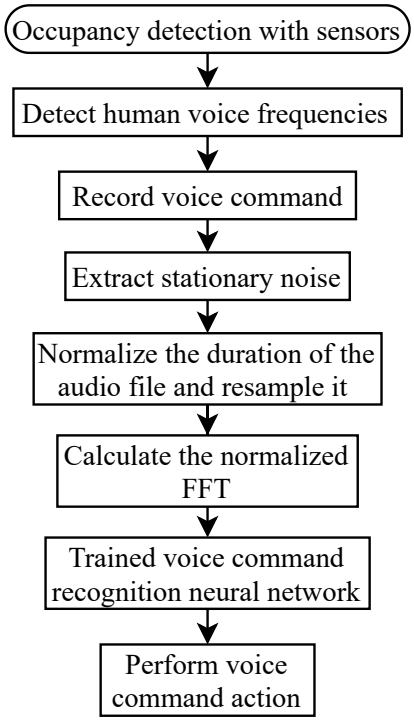

Figure 4. Diagram of Arduino microprocessor process

In the diagram in Fig. 4 the sequence of processes mentioned above that perform the occupancy detection systems in collaboration with pseudo-recognition, using Neural Network API for Arduino [20], all files are exported as DHF5 files [21] in the GitHub repository [14].

IV. Methodology

Presented in Fig. 5 diagram of the processes prior to training a neural network detailed below from sub-Section IV-A. Followed by specifying the multi-layer perdition classifier using LBFGS or stochastic gradient descent described in sub-Section IV-B.

A. Dataset Generation

1) ***Record voice commands to file:*** An audio dataset was created for pseudo speech recognition in Spanish, specifically for 7 voice commands ”ENCENDER”, ”APAGAR”, ”0”, ”25”, ”50”, ”75” and ”100”, these audios were recorded by 9 men and 5 women with a total of 851 .wav files with noise recount processing in the frequency domain and clipped in the time domain. In addition, audios of typical stationary noises in a bathroom are recorded, as well as the noise of an exhaust fan, a heater or the water in the shower.

Within the set of recordings, several audio files had different extinctions such as `.wav`, `.mp3`, `.m4a`, and `.ogg` extinction typical of WhatsApp audio files with maximum frequency clipping up to 9 [kHz].

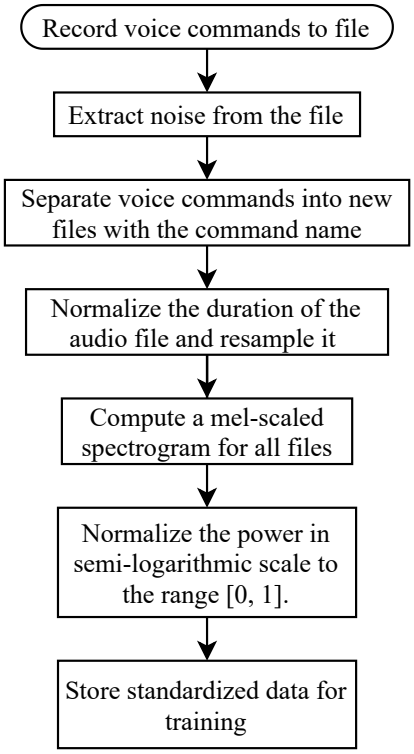

Figure 5. Diagram of data pre-processing before training

*2) **Extract noise from the file**:* Also within this set of audio files, we commonly found recordings with stationary noise noise easy to extract with Python's audio processing library Librosa [22], however high frequency sounds such as birds and nose whistles produced by the same people who recorded the audio voice commands were detected.

To eliminate these uncommon variants of the human voice, which could possibly impair the supervised training of the neural network, a graphical interface visualized in Fig.6 of the audio editing program ADOBE® AUDITION was used to perform non-stationary noise reduction in the frequency domain [23].

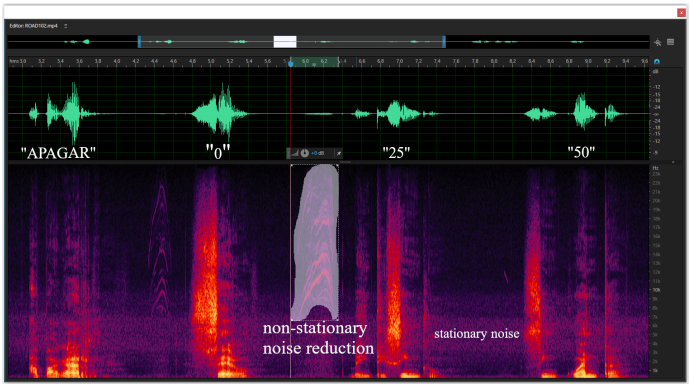

Figure 6. Nonstationary noise reduction of nose whistling in the frequency domain - Adobe Audition.

After the noise reduction process, the voice commands are saved in audio files with extension `.wav` with their respective names and an identifier with the name of the person who recorded the audio and the number of the recording.

*3) **Normalize the duration of theaudio file and resample it**:* This is done to have all the audio files wiht the same number of samples and thus the vector `Xtrain`, this includes all its independent variables or in this case arrays of one dimension and length equal to the number of samples after re-sampling the audios.

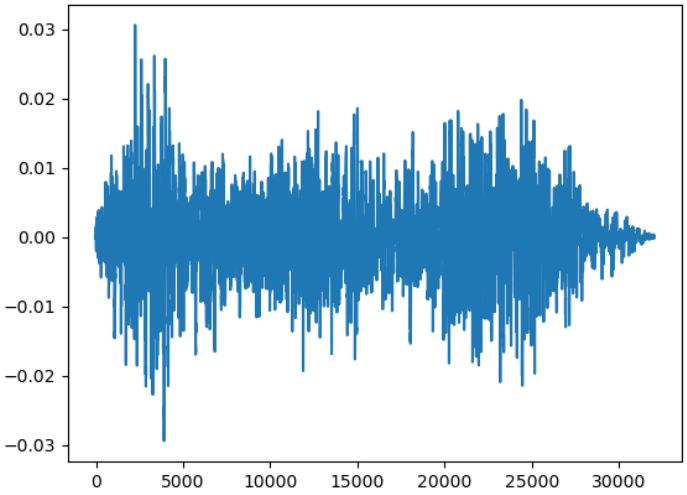

Figure 7. Average signal of all voice commands "ENCENDER" with 32000 mustras for each audio.

These variables will not be used to train the model, time series of the speech power will be computed based on these audio arrays with equal number of samples. (See time series of the average speech power for audios containing the "ENCENDER" speech commands Fig. 10)

*4) **Compute a mel-scaledspectrogram for all files**:* Mel Frequency Cepstral Coefficients [24] are generally calculated by means of a filter bank applied to spectrograms calculated with STFT sequences, which represent a signal in the time frequency domain by calculating discrete Fourier transforms (DFT) in short, overlapping time windows. [25]. The spectrogram used in this paper uses a semi-logarithmic power scale, centered in the main frequencies of the human voice between (300 [Hz] - 3400 [Hz]).

*5) **Store standardized data fortraining**:* The first step is to normalize the power in semi-logarithmic scale to the range [0, 1]. Then, the time series of the voice power is extracted as the average value of the rows of the matrix representing the transformation of the spectrogram mentioned above displayed in the Fig. 8.

Finally, the time series of the voice power are converted into one-dimensional arrays that are stored in a file `.hdf5`[1].

[1]The file paths of the audio files are also stored. (see [21])

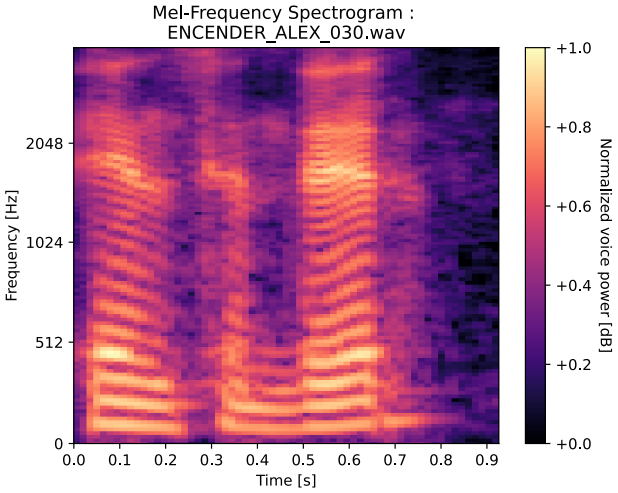

Figure 8. Mel-scaled Spectrogram Normalized in Power and Time with 512 samples between successive frames of STFT

B. Supervised neural network training

The Multi-layer Perceptron (MLP) supervised learning algorithm will be used as it has the ability to learn real-time nonlinear model [26]. Next, we specify the multilayer perdition classifier using LBFGS or stochastic gradient descent already described above with a the generic structure displayed in the Fig. 9.

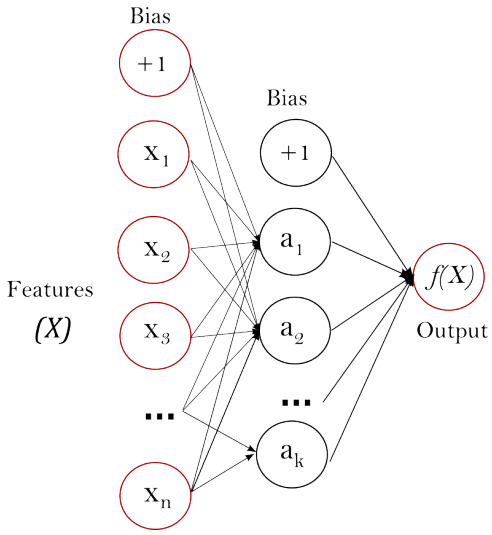

Figure 9. (MLP) model of a hidden layer

The time series of the power mentioned in the sub-Section IV-A are extracted from the `.hdf5` files, and then concatenated to create the vector `X_train` to train the supervised neural network with the comparison of the vector `y_train` that contains the names of the voice command classes “ENCENDER”, “APAGAR”, “0”, “25”, “50”, “75”, “100” which were extracted from the filename taken of the audio files.

V. Results

The python files in the `code` folder of the GitHub repository were executed, and a clear differentiation of the average signals of the speech time series displayed in Fig. 11, and also obtained predictions higher than 83% for the binary classification between ‘ENCENDER’ and “APAGAR” voice commands and 80% for the classification between numeric voice commands (see the confusion matrices in Fig. 12 and Fig. 13).

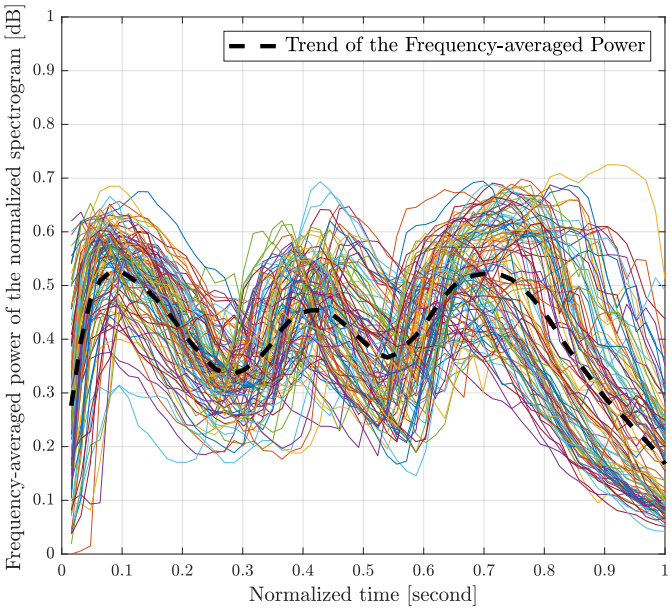

Figure 10. Frequency-averaged Power of the normalized spectrogram - Voice commands “ENCENDER”

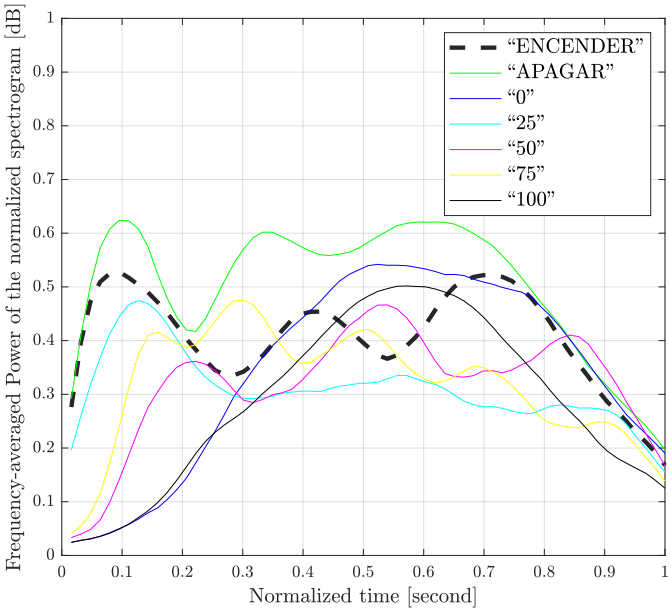

Figure 11. Trends of all voice commands Frequency-averaged Power of the normalized spectrogram

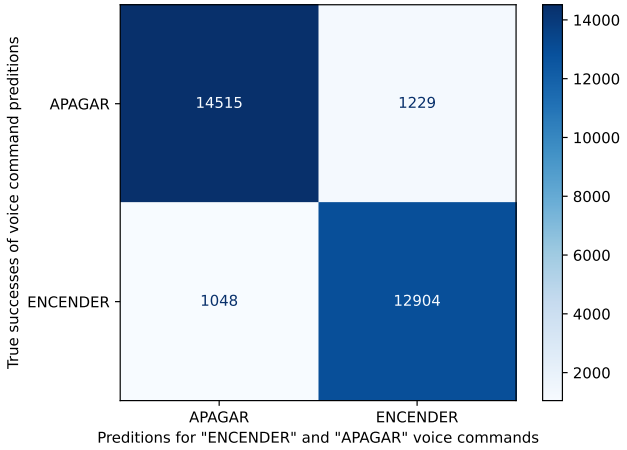

Figure 12. Confusion matrix for classification between “ENCENDER” and ‘

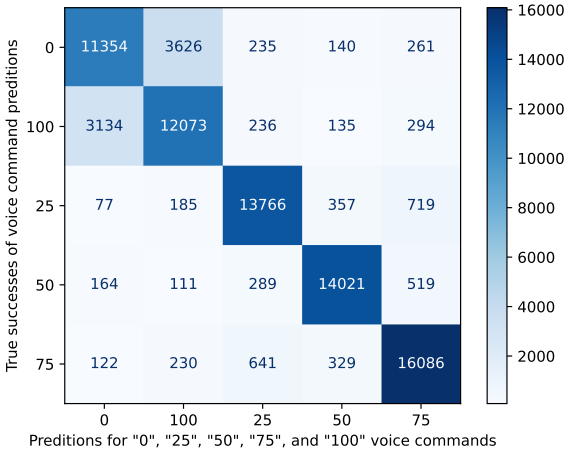

Figure 13. Confusion matrix for classification between “0”, “25”, “50”, “75” and “100” voice commands

VI. Conclusion

This research undertook the challenge of developing a pseudo speech recognition system in Spanish, targeting specific voice commands. The methodology followed a comprehensive step-by-step process, starting with the meticulous generation and pre-processing of a dataset. This dataset, comprised of voice commands, was carefully refined by eliminating stationary and non-stationary noises, normalizing duration, resampling, and then converting into mel-scaled spectrograms. The end-result provided a standardized set of data ideal for supervised training.

The choice of the Multi-layer Perceptron (MLP) as a supervised learning algorithm was pivotal. Not only is the MLP known for its capability to grasp non-linear models in real-time, but when combined with a meticulous pre-processed dataset, its efficiency in voice command recognition was commendable.

The results obtained from the research are encouraging. There was a significant accuracy in the binary classifications of voice commands, particularly between the commands ’ENCENDER’ and ’APAGAR’. These high predictive scores affirm the effectiveness of the chosen methodology and neural network.

In conclusion, the combination of thorough data preprocessing and the selection of a suitable neural network can lead to efficient and accurate speech recognition systems, even for specific commands in a language as rich and complex as Spanish. Such systems hold immense potential in various applications, be it in smart home systems or voice-assisted technologies. It is vital for further research to focus on enhancing the accuracy, expanding the command list, and testing in diverse environments to ensure robustness.

References

[1] C. Thornett and A. Brown, “Interfacing modern technology to disabled patients,” in *IEE Colloquium on Engineering Design for the Disabled*, Nov. 1988, pp. 9/1–9/2.

[2] M. M. Hoehn and M. D. Yahr, “Parkinsonism,” *Neurology*, vol. 17, no. 5, pp. 427–427, 1967. [Online]. Available: https://n.neurology.org/content/17/5/427

[3] M. Chan, C. Hariton, P. Ringeard, and E. Campo, “Smart house automation system for the elderly and the disabled,” in *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, vol. 2, 1995, pp. 1586–1589 vol.2.

[4] F. Leccese, G. Salvadori, M. Rocca, C. Buratti, and E. Belloni, “A method to assess lighting quality in educational rooms using analytic hierarchy process,” *Building and Environment*, vol. 168, p. 106501, Jan. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132319307139

[5] L. Candanedo, “LuisM78/Occupancy-detection-data,” Jun. 2021, original-date: 2015-08-19T13:15:32Z. [Online]. Available: https://github.com/LuisM78/Occupancy-detection-data

[6] L. M. Candanedo and V. Feldheim, “Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models,” *Energy and Buildings*, vol. 112, pp. 28–39, Jan. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778815304357

[7] K. Lam, M. Höynck, B. Dong, B. Andrews, Y.-S. Chiou, R. Zhang, D. Benitez, and J. Choi, “Occupancy detection through an extensive environmental sensor network in an open-plan office building,” 2009, pp. 1452–1459.

[8] F. Castanedo, D. López-de Ipina, H. Aghajan, and R. Kleihorst, “Building an occupancy model from sensor networks in office environments,” *ICDSC*, vol. 3, pp. 1–6, 2011.

[9] E. Hailemariam, R. Goldstein, R. Attar, and A. Khan, “Real-time occupancy detection using decision trees with multiple sensor types,” *Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design*, pp. 141–148, 2011.

[10] Z. Yang, N. Li, B. Becerik-Gerber, and M. Orosz, “A multi-sensor based occupancy estimation model for supporting demand driven HVAC operations,” *Proceedings of the 2012 Symposium on Simulation for Architecture and Urban Design*, no. 2, pp. 49–56, 2012.

[11] T. Ekwevugbe, N. Brown, and V. Pakka, “Realt-Time Building Occupancy Sensing for Supporting Demand Driven HVAC Operations,” 2013, accepted: 2014-01-10T20:21:08Z Publisher: Energy Systems Laboratory (http://esl.tamu.edu). [Online]. Available: https://oaktrust.library.tamu.edu/handle/1969.1/151431

[12] A. Benba, A. Jilbab, and A. Hammouch, “Using Human Factor Cepstral Coefficient on Multiple Types of Voice Recordings for Detecting Patients with Parkinson’s Disease,” *IRBM*, vol. 38, no. 6, pp. 346–351, Nov. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1959031817301136

[13] S. PLESHKOVA, A. BEKYARSKI, and Z. ZAHARIEV, “Reduced Database for Voice Commands Recognition Using Cloud Technologies, Artificial Intelligence and Deep Learning,” in *2019 16th Conference on Electrical Machines, Drives and Power Systems (ELMA)*, Jun. 2019, pp. 1–4.

[14] N. I. NicoGitSoft, “voice command recognition in spanish for handicapped restrooms,” 2021. [Online]. Available: https://github.com/NicoGitSoft/recognition_of_voice_commands_in_Spanish.git

[15] SENADIS, "Estudio nacional de la discapacidad," *Un nuevo enfoque para la inclusión*, pp. 20–22, 2015. [Online]. Available: https://www.senadis.gob.cl/descarga/i/3315

[16] ——, "Segundo estudio nacional de la discapacidad," *Resultados Regionales para la Población Adulta*, pp. 9–10, 2015. [Online]. Available: https://www.senadis.gob.cl/pag/355/1197/ii_estudio_nacional_de_discapacidad

[17] *Norma Eléctrica Chilena NCh. Elec. 4/2003,* ***11.0.2.4*** *"Electricidad. Instalaciones de consumo en baja tensión", Superintendencia de Electricidad y Combustibles S.E.C.*, Std., 2003. [Online]. Available: https://www.sec.cl/sitioweb/electricidad_norma4/objetivo.pdf

[18] "Adafruit ZeroFFT Library," Aug. 2021, original-date: 2018-01-05T22:15:43Z. [Online]. Available: https://github.com/adafruit/Adafruit_ZeroFFT

[19] "VPN on Raspberry Pi - It's That Simple!" [Online]. Available: https://www.experte.com/vpn/raspberry-pi

[20] "Neural Network API for Arduino." [Online]. Available: https://annhubhelp.anscenter.com/NeuralNetworkAPIforArduino.html

[21] "Hdf5 for python - documentation h5py v3.3.0." [Online]. Available: https://docs.h5py.org/en/stable/index.html

[22] "librosa — librosa v0.8.1 documentation," 2021. [Online]. Available: https://librosa.org/doc/latest/index.html

[23] "Reduce audio noise in recordings | Adobe." [Online]. Available: https://www.adobe.com/products/audition/noise-reduction.html

[24] "MFCC," Aug. 2020, page Version ID: 128384562. [Online]. Available: https://es.wikipedia.org/w/index.php?title=MFCC&oldid=128384562

[25] "Spectrograms." [Online]. Available: https://ccrma.stanford.edu/~jos/st/Spectrograms.html

[26] "1.17. Neural network models (supervised) — scikit-learn 0.24.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/neural_networks_supervised.html