ISSN: 1692-7257 - Volume 1 – Number 45 - 2025

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

*Digital Object Identifier: 10.24054/rcta.v1i45.3044*

# Deep learning for selection of numerical options by voice as tools for chatbot

## *Aprendizaje profundo para selección de opciones numéricas por voz como herramientas para chatbot*

**PhD. Robinson Jiménez Moreno[1], MSc. Andrés Mauricio Castro Pescador[2], MSc. Anny Astrid Espitia Cubillos[3]**

[1] ***Universidad Militar Nueva Granada,*** *associate professor at the Faculty of Engineering, Mechatronic Engineering program, Bogotá, Colombia.*
[2] ***Universidad Militar Nueva Granada,*** *adjunct professor at the Faculty of Engineering, Mechatronic Engineering program, Bogotá, Colombia.*
[3] ***Universidad Militar Nueva Granada,*** *associate professor at the Faculty of Engineering, Industrial Engineering program, Bogotá, Colombia.*

***Correspondence:*** *anny.espitia@unimilitar.edu.co*

***Received:*** *august 01, 2024.* ***Accepted:*** *december 18, 2024.* ***Published:*** *january 01, 2025.*

***How to cite:*** *R. Jiménez Moreno, A. M. Castro Pescador, and A. A. Espitia Cubillos, "Deep learning for selection of numerical options by voice as tools for chatbot", RCTA, vol. 1, no. 45, pp. 74–81, jan. 2025.*
*Recovered from https://ojs.unipamplona.edu.co/index.php/rcta/article/view/3044*



**Abstract:** This document presents the design of a voice-operated chatbot-type assistant that works following a dialogue model between user and robot, which is trained with deep learning algorithms, using a database of spectrograms constructed from male and female voices, based on the short-time Fourier transform and Mel frequency cepstral coefficients as signal preprocessing techniques. For the recognition and classification of voice patterns, five convolutional network architectures are designed with the same parameters. The performance achieved in the training of the networks is compared, where all degrees of accuracy were greater than 92.8%. It is observed that the number of layers of the networks affects the number of learning parameters, their degree of accuracy and digital weight; in general, a greater number of layers increases both the training time and the classification time. Finally, for validation through a chatbot App, the selected network is applied to the completion of a survey that uses a Likert scale from 1 to 5, where users, in addition to saying the selected option, confirm it with a Yes or No, the App plays the audio of each question, shows its identification, listens and confirms the user's answers. The selected network design is concluded, allowing the development of chatbot applications based on audio interaction.

**Keywords:** deep learning, robotics, artificial intelligence, app, chatbot.

**Resumen:** Este documento presenta el diseño de un asistente tipo chatbot operado por voz que funciona siguiendo un modelo de dialogo entre usuario y robot, el cual es entrenado con algoritmos de aprendizaje profundo usando una base de datos de espectrogramas, construidos a partir de voces tanto masculinas como femeninas, basados en la transformada de Fourier de corto tiempo y los coeficientes cepstrales de frecuencia Mel como técnicas de preprocesamiento de señales. Para el reconocimiento y clasificación de patrones de voz se diseñan cinco arquitecturas de red convolucional con los mismos parámetros. Se compara el desempeño en el entrenamiento de las redes donde todas obtuvieron grados de exactitud superior al 92.8%, se

observa que el número de capas de las redes afecta el número de parámetros de aprendizaje, su grado de exactitud y peso digital, en general mayor cantidad de capas incrementa tanto el tiempo de entrenamiento como el tiempo de clasificación. Finalmente, para su validación mediante un App de chatbot, el diseño de la red seleccionada es aplicado al diligenciamiento de una encuesta que usa una escala de Likert de 1 a 5, en donde los usuarios además de decir la opción seleccionada la confirman con un Sí o un No, la App reproduce el audio de cada pregunta, muestra su identificación, escucha y confirma las respuestas del usuario. Se concluye el diseño de red seleccionado permite desarrollar aplicaciones de chatbot basadas en interacción por audio.

**Palabras clave:** aprendizaje profundo, inteligencia artificial, robótica, aplicación, chatbot.

## 1. INTRODUCTION

The objective of this paper is present the design of a basic smart chatbot, based on a convolutional network architecture for the recognition and classification of voice patterns from spectrograms. The use of convolutional neural networks (CNN) for sound recognition derives from their ability to adapt and identify through their tuning parameters [1]. The dialogue facilitates human-computer interaction (HCI). In the case of the proposed application, a voice signal is translated into a series of values and a pair of words (yes and no) to facilitate the participation of people in completing perception surveys.

There are several successful new applications that simultaneously use convolutional neural networks (CNNs) and spectrograms for the study of voice. For example, for the purposes of validating a user's identity [2] with the selection of spectrogram patches. For speech recognition using Google API [3], also for the recognition of regional accents of British English [4]. For the identification of spoken English [1], for word recognition [5], for speech recognition that identifies words even with noisy and low-resolution signals [6]. For music classification [7] and for recognize emotions [8], [9], [10].

In the health sector, [11] presents the high reliability to detect Parkinson's disease from the analysis of voice samples collected by telephone and processed with a CNN using transfer learning, compared to conventional approaches. Voice pathologies are identified early allowing appropriate treatment to be offered [12]. In addition, patients with organic dysphonia have been identified by means to computerized voice analysis [13].

In [14] the gender and age recognition of people is done through voice analysis using CNN, for which they developed and tested three models, they point out that the best results are obtained when multi-attention module (MAM) is located between two feature learning block (FLB) consisting of convolution, pooling and batch normalization layers to extract features.

## 2. METHODOLOGY

For the development of this research, four stages are established. The first aimed at designing convolutional network architectures with different depths and the same parameters. The second consists of training the networks designed for this, starting from the creation of a database that allows the spectrogram to be built and the training to be carried out. The third corresponds to the selection of the network that is supported in the training evaluation. Finally, in the fourth validation stage, a chatbot App with a dialogue model is designed.

To identify the user's responses, a CNN is trained, whose database corresponds to the possible responses based on numbers from one to five and the yes and no options. A total of 100 samples of both male and female voices are used for each of the 7 categories. From the audio files, a spectrogram is obtained based on the Fourier transform of the voice signal sifted through a low-pass filter in the 2.5KHz range, and from this its first and second derivative are obtained to complete a three-channel spectrogram, used as input in the CNN. The database is distributed 70% for training and 30% for validation.

## 3. NETWORK ARCHITECTURE DESIGN

For pattern recognition and classification of each voice spectrogram, five network architectures were designed based on equations 1 to 3, to evaluate the best option for the desired application. The input volume corresponds to each spectrogram where $W_0$=199, $H_0$=12 and $D_0$=3. The depth of the network

($j$) was increased while maintaining the initial hyperparameters.

$$W_j = \frac{(W_{j-1} - F + 2P)}{S} + 1 \quad (1)$$

$$H_j = \frac{(H_{j-1} - F + 2P)}{S} + 1 \quad (2)$$

$$D_j = K_{j-1} \quad (3)$$

The architecture of the networks is summarized in Table 1, where the network design parameters per convolution layer (C) are established, such as the kernel (K), the number of filters (F), the padding (P) and stride (S), each convolution layer is followed by a pooling stage in which the maximum method is used. In the final stage of network classification, three fully connected (FC) phases are used up to the output layer to classify the 7 classes [16].

***Table 1:*** *Characteristics of the designed network architectures*

| | C | K | F | P | S | FC1 | FC2 | FC3 |
|---|---|---|---|---|---|---|---|---|
| **Network 1** | C1 | 7 | 16 | 2 | 1 | 1024 | 256 | 1024 |
| | C2 | 5 | 16 | 1 | 1 | | | |
| | C3 | 3 | 32 | 1 | 1 | | | |
| | C4 | 2 | 64 | 1 | 1 | | | |
| **Network 2** | C1 | 7 | 16 | 2 | 1 | 1024 | 256 | 1024 |
| | C2 | 5 | 16 | 2 | 1 | | | |
| | C3 | 3 | 32 | 1 | 1 | | | |
| | C4 | 3 | 32 | 1 | 1 | | | |
| | C5 | 2 | 64 | 1 | 1 | | | |
| **Network 3** | C1 | 7 | 16 | 2 | 1 | 1024 | 256 | 1024 |
| | C2 | 5 | 16 | 2 | 1 | | | |
| | C3 | 3 | 32 | 1 | 1 | | | |
| | C4 | 3 | 32 | 1 | 1 | | | |
| | C5 | 3 | 64 | 1 | 1 | | | |
| | C6 | 2 | 64 | 1 | 1 | | | |
| **Network 4** | C1 | 7 | 16 | 2 | 1 | 512 | 256 | 1024 |
| | C2 | 5 | 16 | 2 | 1 | | | |
| | C3 | 3 | 32 | 1 | 1 | | | |
| | C4 | 3 | 32 | 1 | 1 | | | |
| | C5 | 3 | 64 | 1 | 1 | | | |
| | C6 | 2 | 64 | 1 | 1 | | | |
| | C7 | 2 | 64 | 1 | 1 | | | |
| **Network 5** | C1 | 7 | 16 | 2 | 1 | 512 | 256 | 1024 |
| | C2 | 5 | 16 | 2 | 1 | | | |
| | C3 | 3 | 32 | 1 | 1 | | | |
| | C4 | 3 | 32 | 1 | 1 | | | |
| | C5 | 3 | 64 | 1 | 1 | | | |
| | C6 | 2 | 64 | 1 | 1 | | | |
| | C7 | 2 | 64 | 1 | 1 | | | |
| | C8 | 2 | 64 | 1 | 1 | | | |

Table 2 illustrates the training parameters used for all designed network architectures.

***Table 2:*** *CNN training features*

| Hypeparamters | Value |
|---|---|
| Mini Batch Size | 5 |
| Optimizador | SGDM |
| Epochs | 150 |
| Learn Rate | 1e-6 |
| Frecuency | 940 |
| Iteration | 14700 |

To facilitate the analysis of the answers through the chatbot, it uses a Likert scale from 1 to 5. A dialogue model is established between user and robot which is always guided by the robot. The user only must respond verbally with a yes, a no or a number from 1 to 5. The outline of the dialogue is presented in Fig. 1. For its interpretation, the Likert scale [15] with relationships from 1 to 5 is used as the response parameter as follows:

1- Not satisfied at all
2- Slightly satisfied
3- Neutral
4- Very satisfied
5- Extremely satisfied

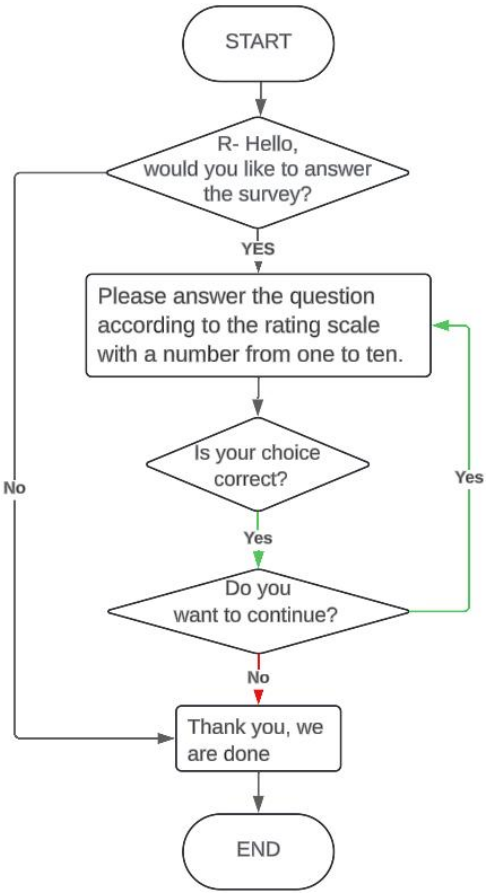



***Fig 1****. Summary dialogue model*

## 4. RESULTS

### 4.1. Network Training

Mel-Frequency Cepstral Coefficients (MFCCs) are used as a technique for the extraction of sound features in automatic speech recognition [17], [18], [19], [20]. To determine these coefficients, a time-frequency representation of the signal called a spectrogram is obtained. The spectrogram shows how the signal frequencies behave over time [17] and is calculated with the Short-Time Fourier Transform (STFT) using (2).

$$X(k_i) = \sum_{n=0}^{N-1} x_i[n]h[n]e^{-j\frac{2\pi}{N}kn}$$
$$k = 0, \dots, n-1 \qquad (2)$$

In this equation the time domain speech signal x[n] is multiplied with a Hamming window h[n] of time interval Tw between 20 to 40 ms. Assuming that the signal is sufficiently stationary in small intervals [21], the discrete Fourier transform $X(k_i)$ is calculated. The window then slides along the time axis with an overlap time To. The overlap is done so that no information is lost between the transition of segments [22]. The voice signal is shown in Fig. 2.

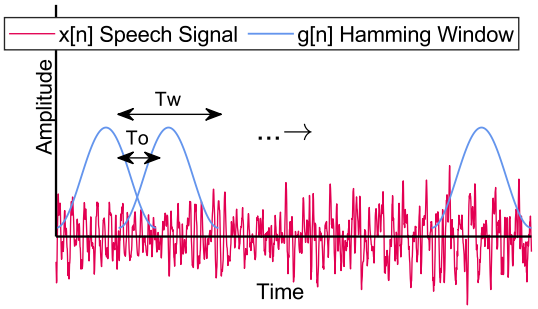



***Fig. 2**. Speech signal with Hamming windows.*

For this work, the voice signal was acquired with the Matlab software at a sampling frequency of 16000 [Hz] for 3 seconds. The Hamming window time is Tw=25ms with an overlap between windows of To=10ms. Therefore, the number of Hamming windows were 199. The short Fourier transform is calculated using the *stft* function of Matlab's Signal Processing Toolbox.

To obtain a spectrogram, the signal power of each interval is calculated with (3) where N is the number of signal samples in each time segment.

$$P[k_i] = 20\log_{10}\frac{|X[k_i]|^2}{N} \qquad (3)$$

The next step consists of applying a bank of filters spaced on the Mel scale with the objective of extracting important speech characteristics to be used in speech recognition. This approach seeks to emulate the perception of the human ear where the perceptual difference between tones or frequencies is non-linear. Human hearing is less sensitive at high frequencies than at low frequencies [23]. The formula to convert a frequency f in Hz to Mel is presented in (4).

$$Mel(f) = 2595\log_{10}\left(1+\frac{f}{700}\right) \qquad (4)$$

26 separate triangular filters were created on the Mel scale, as illustrated in Fig. 3, and applied to the spectrogram, which improved the perceptual difference of low frequencies compared to high frequencies. The bank of 26 filters were designed with the designAuditoryFilterBank function of Matlab's Audio Toolbox.

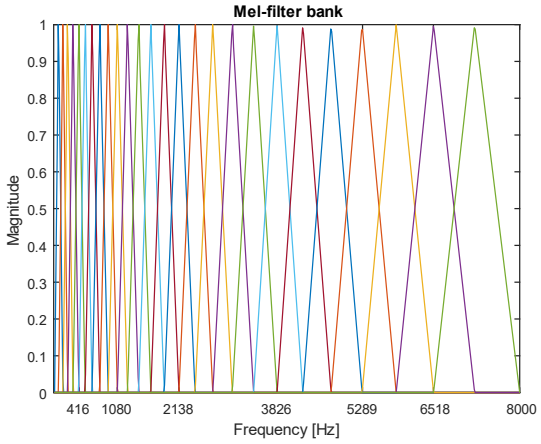
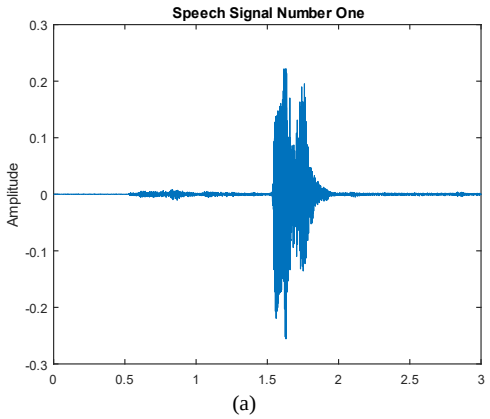
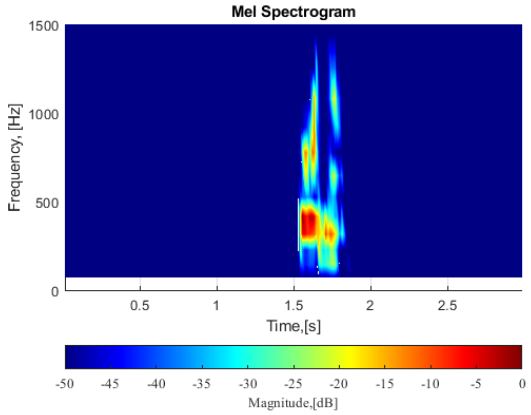



***Fig. 3**. Mel filter bank.*

Fig. 4 shows the spectrogram on the Mel scale for a male voice pronouncing the number one. It can be seen in Fig. 4(a) the audio signal and its spectrum, in Fig. 4(b) the Fourier spectrogram of channel one.




(a)




(b)

***Fig. 4.*** *Spectrogram on the Mel scale*

Finally, the Discrete Cosine Transform (DCT) is applied to de-correlate the Mel filter bank and obtain the 26 Mel-frequency Cepstral Coefficients (MFCCs), of which only coefficients 2 to 12 are used for speech recognition, as presented in Fig. 5.

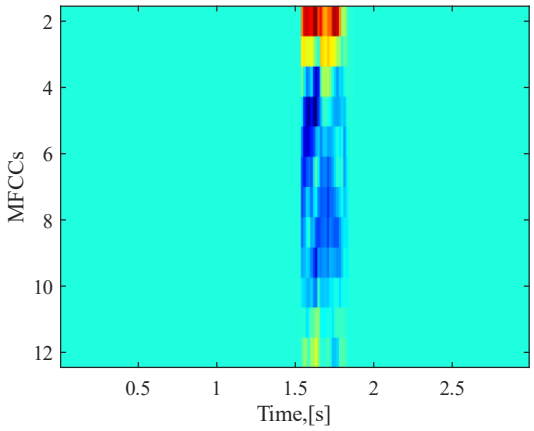



***Fig. 5****. Mel-Frequency Cetral Coefficients*

To improve and capture information about changes in the voice signal, the Delta and Delta-Delta coefficients are calculated. Taking the first and second temporal derivative of Mel's Ceptral coefficients, shown in Fig. 6 for channel two and Fig. 7 for channel three, respectively. The three resulting matrices of the MFCCs and their deltas are of size 12x199.

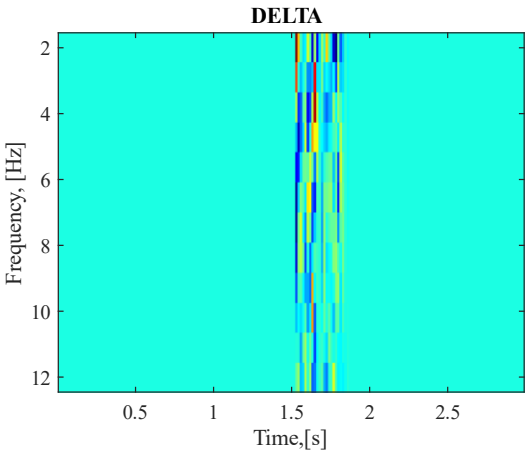



***Fig. 6.*** *Delta coefficients*

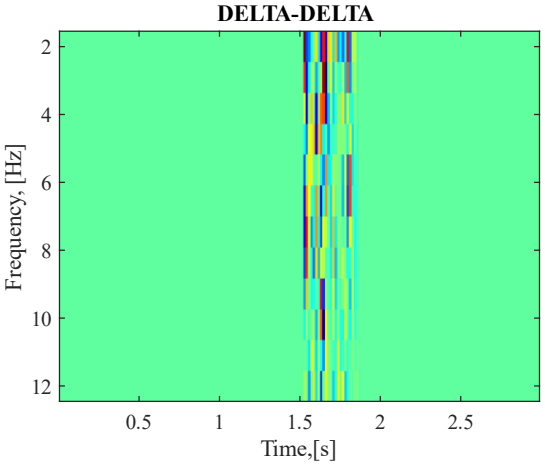



***Fig. 7.*** *Delta-Delta coefficients*

### 4.2. Network Selection

Table 3 shows the training results in relation to the difference in architectures, where it is evident that the number of layers of the networks affects the number of learning parameters, their degree of accuracy, in general a greater number of layers increases both the training time and the classification time. At the same time, it is evident that the resulting network with a greater number of parameters is digitally heavier, which is not favorable either. In this case, network number 4 is determined as the best, because it has a high precision in identifying the response with a low digital network weight.

***Table 3**: CNN training results*

| Network | Number of Layers | Learning parameters | Network weight | Accuracy | Training time |
|---|---|---|---|---|---|
| **1** | 16 | 554.4k | 2.024 KB | 94.76 % | 5 min 47 sec |
| **2** | 17 | 563.7k | 2.058 KB | 95.24 % | 6 min 6 sec |
| **3** | 18 | 590.3k | 2.154 KB | 93.81 % | 7 min 11 sec |
| **4** | 20 | 475.7k | 1.737 KB | 97.1% | 7 min 58 sec |
| **5** | 22 | 426.6k | 1.558 KB | 92.4 % | 7 min 52 sec |

Fig. 8 illustrates the training performance of network 4, which was the one that had the best behavior. However, for all cases the rapid learning achieved around 1200 iterations and with high degrees of accuracy is evident, in all cases, greater than 92.4%. The graph shows that a much smaller number of epochs could have been handled, close to 80.

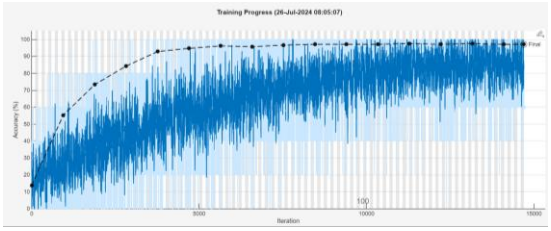


***Fig. 8.** CNN training of numbers using network 4*

Table 4 shows the classification time from each architecture. Where due to the low classification time of each case, for a real-time application, the byte size of the network and its precision are more significant. Network 4 was therefore chosen as it was the most balanced in these two factors.

***Table 4:** Classification times*

| Network | Classification time |
|---|---|
| **1** | 0.0428 seconds |
| **2** | 0.0462 seconds |
| **3** | 0.0502 seconds |
| **4** | 0.0529 seconds |
| **5** | 0.0730 seconds |

Fig. 9 illustrates the confusion matrix of the chosen network. The best classifications are obtained by the classes: three, five and yes, the least efficient class was four, it is observed that the reason is due to the confusion between classes four and one, which can be given by the similarity of the patterns when strong and clear vocalization is not generated.

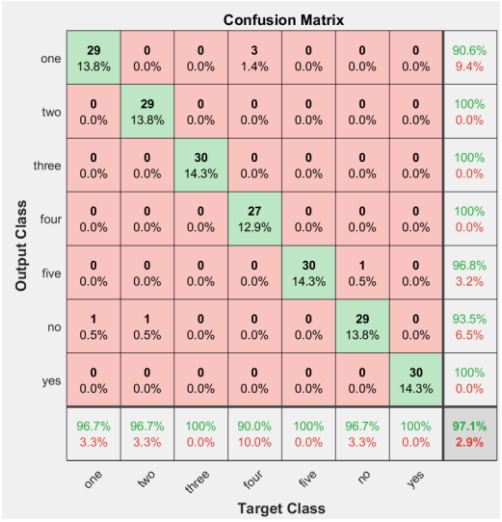



***Fig. 9.** Matrix confusion of network 4.*

### 4.3. Validation

The functionality of the selected network is validated through a chatbot App (see Fig. 10), which displays the scale to be used and the numerical value associated with each one, the App shows the number of the question in which the survey is, has a text box that can be read and another box that shows the user which option they chose and recognized the network. Likewise, the bot plays an audio with each question. The use of a voice-operated chatbot-type assistant is designed according to the flow chart presented in Fig. 1. For the example, the questionnaire is based on the following questions:

1- How do you rate the activity in general?
2- How do you value the activity carried out in terms of time efficiency?
3- How do you value the activity carried out in terms of quality?
4- How do you value the guidance received in the development of the activity?
5- How do you value the activity carried out in terms of efficiency of use?

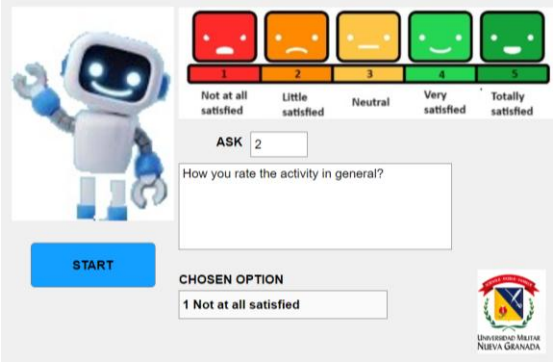



***Fig. 10.** Example app*

## 4. CONCLUSIONS

In this paper, the Short-Time Fourier Transform (STFT) and Mel-Frequency Cepstral Coefficients (MFCCs) were employed as preprocessing signal techniques. The MFCCs, Delta, and Delta-Delta coefficients, aimed to capture the relevant information from the human speech signal. Additionally, the noise in the signal was attenuated before being applied to the convolutional neural network. These preprocessing steps enhanced the quality of the input information to the convolutional network.

It was possible to extract the relevant audio features that allowed evaluating convolutional network architectures to develop chatbot applications based on audio interaction. Network architectures designed for this purpose present high degrees of precision in all cases greater than 92.4%, evidencing the good performance of convolutional networks in speech pattern learning.

To select the most convenient network, the following were considered as decision criteria: the highest degree of accuracy, the smallest byte size of the network and the shortest classification time. Since no option simultaneously meets all the criteria, the most balanced one was chosen.

The developed application allows to demonstrate the scope of using deep learning techniques in natural interactions with bots, being a more user-friendly tool.

As future work, the number of learning words can be increased to expand the conversational scope between user and robot.

## ACKNOWLEDGMENTS

Product derived from the research project titled “Design of a human-robot interaction model using deep learning algorithms” INV-ING-3971 financed by the vice-rector for research of the Universidad Militar Nueva Granada, year 2024.

## REFERENCES

[1] P. Rashmi and M. P. Singh, "Convolution neural networks with hybrid feature extraction methods for classification of voice sound signals," World Journal of Advanced Engineering Technology and Sciences, vol. 8, no. 2, pp. 110-125, doi: 10.30574/wjaets.2023.8.2.0083, 2023.

[2] S. A. El-Moneim, M. A. Nassar and M. Dessouky, "Cancellable template generation for speaker recognition based on spectrogram patch selection and deep convolutional neural networks," International Journal of Speech Technology, vol. 25, no. 3, pp. 689-696, doi: 10.1007/s10772-020-09791-y, 2022.

[3] P. H. Chandankhede, A. S. Titarmare and S. Chauhvan, "Voice recognition based security system using convolutional neural network," in 2021 International Conference on Computing, Communication and Intelligent Systems (ICCCIS), 2021.

[4] O. Cetin, "Accent Recognition Using a Spectrogram Image Feature-Based Convolutional Neural Network," Arabian Journal for Science and Engineering, vol. 48, no. 2, pp. 1973-1990, doi: 10.1109/SLT.2018.8639622, 2023.

[5] A. Soliman, S. Mohamed and I. A. Abdelrahman, "Isolated word speech recognition using convolutional neural network," in 2020 international conference on computer, control, electrical and electronics engineering (ICCCEEE), 2021.

[6] A. Alsobhani, A. H. M. and H. Mahdi, "Speech recognition using convolution deep neural networks," in Journal of Physics: Conference Series, 2021.

[7] J. Li, L. Han, X. Li, J. Zhu, B. Yuan and Z. Gou, "An evaluation of deep neural network models for music classification using spectrograms," Multimedia Tools and Applications, vol. 81, pp. 4621- 4627, doi: 10.1007/s11042-020-10465-9, 2022.

[8] V. Gupta, S. Juyal and Y. C. Hu, "Understanding human emotions through speech spectrograms using deep neural network," The Journal of Supercomputing, vol. 78, no. 5, pp. 6944-6973, doi: 10.1007/s11227-021-04124-5, 2022.

[9] D. Issa, M. F. Demirci and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," Biomedical Signal Processing and Control, vol. 59, pp. 101894, doi: 10.1016/j.bspc.2020.101894, 2020.

[10] K. Bhangale and K. Mohanaprasad, "Speech emotion recognition using mel frequency log spectrogram and deep convolutional neural network," in International Conference on Futuristic Communication and Network Technologies, Singapore.

[11] A. Iyer, A. Kemp, Y. Rahmatallah, L. Pillai, A. Glover, F. Prior, L. Larson-Prior and T.

Virmani, "A machine learning method to process voice samples for identification of Parkinson's disease," Scientific Reports, vol. 13, pp. 20615, doi: 10.1038/s41598-023-47568-w, 2023.

[12] M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. Khanapi Abd Ghani, M. S. Maashi, B. Garcia-Zapirain and F. T. Al-Dhief, "Voice pathology detection and classification using convolutional neural network model," Applied Sciences , vol. 10, no. 11, pp. 3723, doi: 10.3390/app10113723, 2020.

[13] L. Vavrek, M. Hires, D. Kumar and P. Drotár, "Deep convolutional neural network for detection of pathological speech," in IEEE 19th world symposium on applied machine intelligence and informatics (SAMI), 2021.

[14] A. Tursunov, Mustaqeem, J. Y. Choeh and S. Kwon, "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms," Sensors, vol. 21, no. 17, p. 5892, 2021.

[15] C. Cheng, K.-L. Lay, H. Yung-Fong and T. Yi-Miau , "Can Likert scales predict choices? Testing the congruence between using Likert scale and comparative judgment on measuring attribution," Methods in Psychology, vol. 5, pp. 100081,doi: 10.3390/ai5030048., 2021.

[16] R. Liu, G. Yibei , J. Runxiang and Z. Xiaoli , "A Review of Natural-Language-Instructed Robot Execution Systems," AI 5, no. 3, pp. 948-989, doi: 10.1016/j.metip.2021.100081., 2024.

[17] A. Koshy and S. Tavakoli, "Exploring British Accents: Modelling the Trap–Bath Split with Functional Data Analysis," Journal of the Royal Statistical Society Series C: Applied Statistics, vol. 71, pp. 773–805, doi: 10.1111/rssc.12555, 2022.

[18] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan and A. Q. Ohi, "A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities," IEEE Access, vol. 9, pp. 79236-79263, doi: 10.1109/ACCESS.2021.3084299, 2021.

[19] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880-2894, doi: 10.1109/TASLP.2020.3030497, 2020,.

[20] J. Martinsson and M. Sandsten, "DMEL: The Differentiable Log-Mel Spectrogram as a Trainable Layer in Neural Networks," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, 2024.

[21] J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," Applied Acoustics, vol. 179, p. doi.org/10.1016/j.apacoust.2021.108046, 2021.

[22] M. Samaneh, C. Talen, A. Olayinka, T. John Michael, P. Christian, P. Dave and S. Sandra L, "Speech emotion recognition using machine learning — A systematic review," Intelligent Systems with Applications, vol. 20, p. doi.org/10.1016/j.iswa.2023.200266, 2023.

[23] A. Yenni, H. Risanuri and B. Agus, "A Mel-weighted Spectrogram Feature Extraction for Improved Speaker," International Journal of Intelligent Engineering and Systems, vol. 15, no. 6, p. 74–82 DOI: 10.22266/ijies2022.1231.08, 2022.