ISSN: 1692-7257 - Volumen 1 – Número 45 - 2025

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v1i45.3044

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

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

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

[1] ***Universidad Militar Nueva Granada,*** *profesor asociado de la Facultad de Ingeniería, programa Ingeniería Mecatrónica, Bogotá, Colombia.*
[2] ***Universidad Militar Nueva Granada,*** *profesor ocasional de la Facultad de Ingeniería, programa Ingeniería Mecatrónica, Bogotá, Colombia.*
[3] ***Universidad Militar Nueva Granada,*** *profesora asociada de la Facultad de Ingeniería, programa Ingeniería de Ingeniería Industrial, Bogotá, Colombia.*

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

Recibido: *01 agosto 2024.* ***Aceptado:*** *18 diciembre 2024.* ***Publicado:*** *01 enero 2025.*

Cómo citar: *R. Jiménez Moreno, A. M. Castro Pescador, y A. A. Espitia Cubillos, «Aprendizaje profundo para selección de opciones numéricas por voz como herramientas para chatbot», RCTA, vol. 1, n.º 45, pp. 74–81, ene. 2025. Recuperado de https://ojs.unipamplona.edu.co/index.php/rcta/article/view/3044*

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.

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.

1. INTRODUCCIÓN

El objetivo de este artículo es presentar el diseño de un chatbot inteligente básico, basado en una arquitectura de red convolucional para el reconocimiento y clasificación de patrones de voz a partir de espectrogramas. El uso de redes neuronales convolucionales (CNN) para el reconocimiento de sonido deriva de su capacidad de adaptación e identificación a través de sus parámetros de sintonización [1]. El diálogo facilita la interacción persona-computadora (HCI por sus siglas en inglés). En el caso de la aplicación propuesta, una señal de voz se traduce en una serie de valores y un par de palabras (sí y no) para facilitar la participación de las personas en el diligenciamiento de encuestas de percepción.

Existen varias aplicaciones nuevas exitosas que utilizan simultáneamente redes neuronales convolucionales (CNN) y espectrogramas para el estudio de la voz. Por ejemplo, con el fin de validar la identidad de un usuario [2] con la selección de zonas de espectrograma. Para el reconocimiento de voz utilizando la API de Google [3], también para el reconocimiento de acentos regionales del inglés británico [4]. Para la identificación del inglés hablado [1], para el reconocimiento de palabras [5], para el reconocimiento de voz que logra identificar palabras incluso con señales ruidosas y de baja resolución [6]. Para clasificación de música [7] y para el reconocimiento de emociones [8], [9], [10].

En el sector salud, [11] presenta una alta confiabilidad para detectar la enfermedad de Parkinson a partir del análisis de muestras de voz recolectadas por teléfono y procesadas con una CNN mediante aprendizaje por transferencia, en comparación con los enfoques convencionales. Las patologías de la voz se identifican tempranamente permitiendo ofrecer el tratamiento adecuado [12]. Además, se han identificado pacientes con disfonía orgánica mediante análisis de voz computarizados [13].

En [14] el reconocimiento de género y edad de las personas se realiza mediante análisis de voz utilizando CNN, para lo cual desarrollaron y probaron tres modelos, señalan que los mejores resultados se obtienen cuando el módulo de atención múltiple (MAM) se ubica entre dos bloques de aprendizaje (FLB) que constan de capas de convolución, agrupación y normalización por lotes para extraer características.

2. METODOLOGÍA

Para el desarrollo de esta investigación se establecen cuatro etapas. El primero tenía como objetivo diseñar arquitecturas de redes convolucionales con diferentes profundidades y los mismos parámetros. El segundo consiste en entrenar las redes diseñadas para ello se parte de la creación de una base de datos que permite construir espectrogramas y realizar el entrenamiento. El tercero corresponde a la selección de la red que se apoya en la evaluación de la formación. Finalmente, en la cuarta etapa de validación, se diseña una App de chatbot con un modelo de diálogo.

Para identificar las respuestas del usuario se entrena una CNN, cuya base de datos corresponde a las posibles respuestas limitadas a números del uno al cinco y las opciones de sí y no. Se utiliza un total de 100 muestras de voces masculinas y femeninas para cada una de las 7 categorías. De los archivos de audio se obtienen espectrogramas a partir de la transformada de Fourier de la señal de voz tamizada a través de un filtro de paso bajo en el rango de 2,5KHz, y de éste se obtienen su primera y segunda derivada para completar un espectrograma de tres canales, utilizado como entrada en la CNN. La base de datos se distribuye en un 70% para capacitación y un 30% para validación.

3. DISEÑO DE ARQUITECTURA DE RED

Para el reconocimiento de patrones y clasificación de cada espectrograma de voz, se diseñaron cinco arquitecturas de red basadas en las ecuaciones 1 a 3, para evaluar la mejor opción para la aplicación deseada. El volumen de entrada corresponde a cada espectrograma donde W_0=199, H_0=12 and D_0=3. La profundidad de la red (j) fue ampliada manteniendo los hiperparámetros iniciales.

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

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

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

La arquitectura de las redes se resume en la Tabla 1, donde se establecen los parámetros de diseño de la red por capa de convolución (C), como son el kernel (K), el número de filtros (F), el padding (P) y stride (S), a cada capa de convolución le sigue una etapa de pooling en la que se utiliza el método de máximos. En la etapa final de clasificación de la red se utilizan tres fases completamente conectadas (FC) hasta la capa de salida para clasificar las 7 clases [16].

Tabla 1: *Características de las arquitecturas de red diseñadas.*

	C	K	F	P	S	FC1	FC2	FC3
Red 1	C1	7	16	2	1	1024	256	1024
	C2	5	16	1	1			
	C3	3	32	1	1			
	C4	2	64	1	1			
Red 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			
Red 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			
Red 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			
Red 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			

La Tabla 2 ilustra los parámetros de entrenamiento utilizados para todas las arquitecturas de red diseñadas.

Tabla 2: *Funciones de entrenamiento de CNN*

Hipeparametros	Valor
Tamaño de lote pequeño	5
Optimizador	SGDM
Épocas	150
Tasa de aprendizaje	1e-6
Frecuencia	940
Iteración	14700

Para facilitar el análisis de las respuestas a través del chatbot se utiliza una escala Likert del 1 al 5. Se establece un modelo de diálogo entre usuario y robot que siempre es guiado por el robot. El usuario sólo debe responder verbalmente con un sí, un no o un número del 1 al 5. El esquema del diálogo se presenta en la Fig. 1. Para su interpretación se utiliza la escala Likert [15] con valores de 1 a 5 como parámetros de respuesta de la siguiente manera:

1- Nada satisfecho
2- Ligeramente satisfecho
3-Neutro
4- Muy satisfecho
5- Extremadamente satisfecho

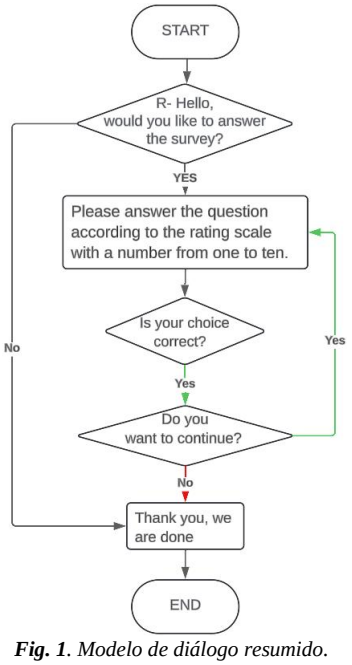

Fig. 1*. Modelo de diálogo resumido.*

4. RESULTS

4.1. Entrenamiento de red

Los coeficientes cepstrales de frecuencia Mel (MFCC) se utilizan como técnica para la extracción de características de sonido en el reconocimiento automático de voz [17], [18], [19], [20]. Para determinar estos coeficientes se obtiene una representación tiempo-frecuencia de la señal denominada espectrograma. El espectrograma muestra cómo se comportan las frecuencias de la señal a lo largo del tiempo [17] y se calcula con la Transformada de Fourier de Tiempo Corto (STFT) usando (2).

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

En esta ecuación, la señal de voz en el dominio del tiempo x[n] se multiplica por una ventana de Hamming h[n] de intervalo de tiempo Tw entre 20 y 40 ms. Suponiendo que la señal es suficientemente estacionaria en intervalos pequeños [21], se calcula la transformada discreta de Fourier X(k_i). Luego, la ventana se desliza a lo largo del eje de tiempo con un tiempo de superposición To. La superposición se realiza para que no se pierda información entre la transición de segmentos [22]. La señal de voz se muestra en la Fig. 2.

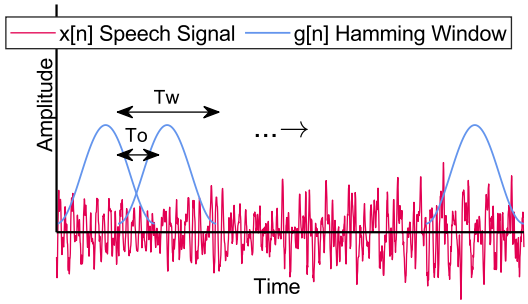

Fig. 2*. Señal de voz con ventanas Hamming.*

Para este trabajo la señal de voz se adquirió con el software Matlab a una frecuencia de muestreo de 16000 [Hz] durante 3 segundos. El tiempo de la ventana de Hamming es Tw=25ms con una superposición entre ventanas de To=10ms. Por lo tanto, el número de ventanas de Hamming fue 199. La transformada corta de Fourier se calculó utilizando la función *stft* de la caja de herramientas de procesamiento de señales de Matlab.

Para obtener un espectrograma, la potencia de la señal de cada intervalo se calcula con (3) donde N es el número de muestras de señal en cada segmento de tiempo.

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

El siguiente paso consiste en aplicar un banco de filtros espaciados en la escala de Mel con el objetivo de extraer características importantes del habla para ser utilizadas en el reconocimiento de voz. Este enfoque busca emular la percepción del oído humano donde la diferencia perceptiva entre tonos o frecuencias no es lineal. La audición humana es menos sensible a las altas frecuencias que a las bajas [23]. La fórmula para convertir una frecuencia f en Hz a Mel se presenta en (4).

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

Se crearon 26 filtros triangulares separados en la escala Mel, como se ilustra en la Fig. 3, y se aplicaron al espectrograma, lo que mejoró la diferencia de percepción de las frecuencias bajas en comparación con las frecuencias altas. El banco de 26 filtros fue diseñado con la función *designAuditoryFilterBank* de la caja de herramientas de audio de Matlab.

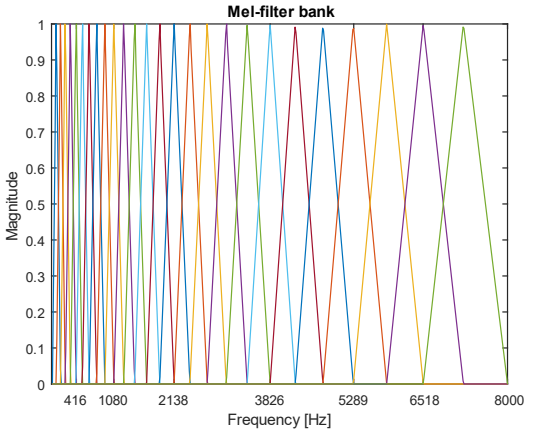

Fig. 3*. Banco de filtros mel*

La figura 4 muestra el espectrograma en la escala Mel de una voz masculina que pronuncia el número uno. Se puede ver en la Fig. 4 (a) la señal de audio y su espectro, en la Fig. 4 (b) el espectrograma de Fourier del canal uno.

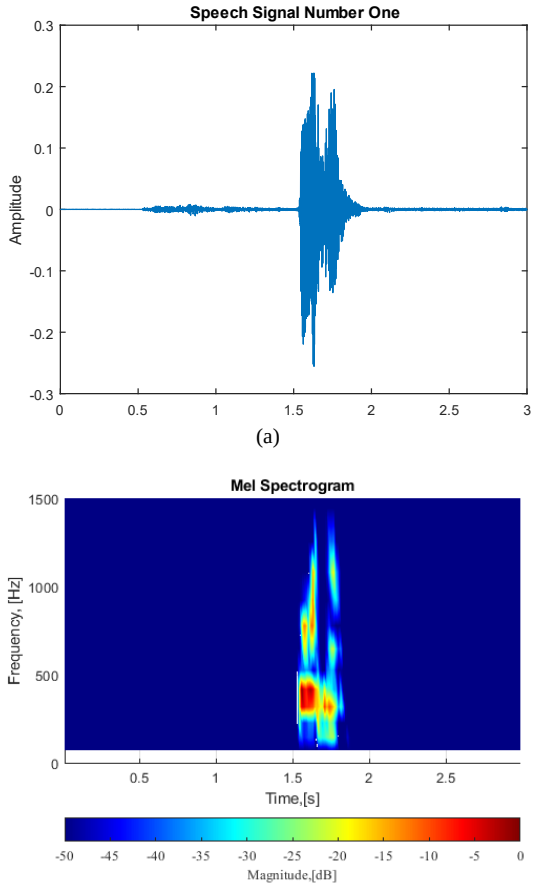

Fig. 4. *Espectrograma en la escala Mel.*

Finalmente, se aplica la Transformada de Coseno Discreto (DCT) para descorrelacionar el banco de filtros Mel y obtener los 26 Coeficientes Cepstrales de frecuencia Mel (MFCC), de los cuales solo se utilizan los coeficientes 2 a 12 para el reconocimiento de voz, como se presenta en la Fig. 5.

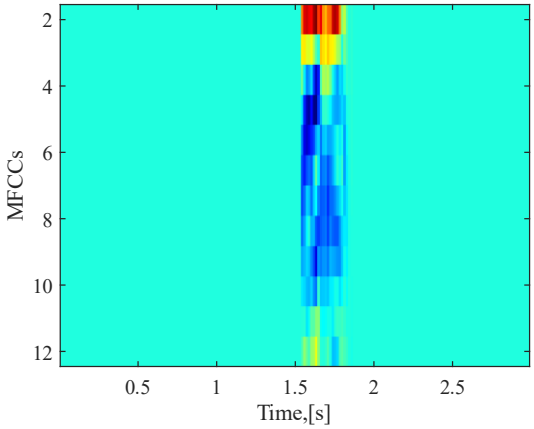

Fig. 5*. Coeficientes centrales de frecuencia Mel*

Para mejorar y capturar información sobre cambios en la señal de voz, se calculan los coeficientes Delta y Delta-Delta. Tomando la primera y segunda derivada temporal de los coeficientes Ceptral de Mel, que se muestran en la Fig. 6 para el canal dos y en la Fig. 7 para el canal tres, respectivamente. Las tres matrices resultantes de los MFCC y sus deltas son de tamaño 12x199.

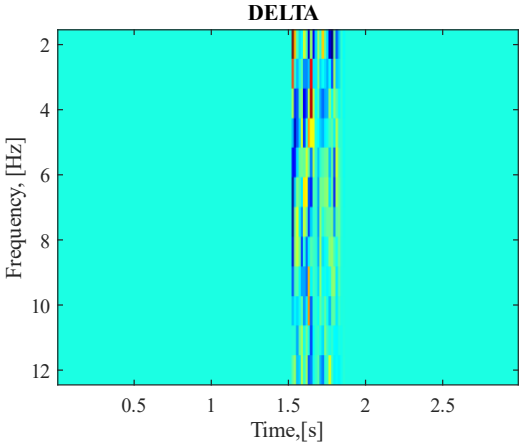

Fig. 6. *Delta coefficients*

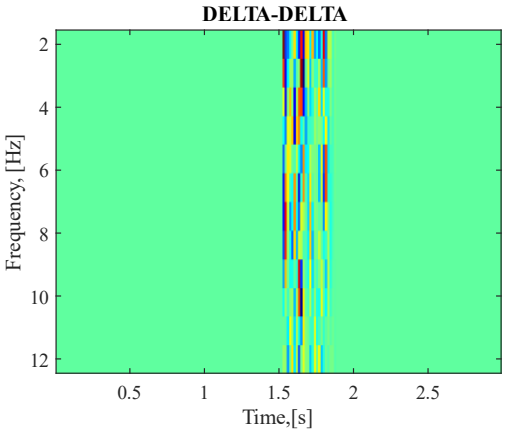

Fig. 7. *Delta-Delta coefficients*

4.2. Selección de la red

En la Tabla 3 se muestran los resultados del entrenamiento de las diferentes arquitecturas, donde se evidencia que el número de capas de las redes afecta la cantidad de parámetros de aprendizaje, su grado de precisión, en general a mayor número de capas aumenta tanto el tiempo de entrenamiento como el tiempo de clasificación. Al mismo tiempo, es evidente que la red resultante con un mayor número de parámetros es digitalmente más pesada, lo que tampoco es favorable. En este caso, la red número 4 se determina como la mejor, porque tiene una alta precisión en identificar la respuesta con un peso de red digital bajo.

***Tabla 3**: Resultados del entrenamiento de CNN*

Red	Número de capas	Parámetros de aprendizaje	Peso de la red	Precisión	Tiempo de entrenamiento
1	16	554.4k	2.024 KB	94.76 %	5 min 47 seg
2	17	563.7k	2.058 KB	95.24 %	6 min 6 seg
3	18	590.3k	2.154 KB	93.81 %	7 min 11 seg
4	20	475.7k	1.737 KB	97.1%	7 min 58 seg
5	22	426.6k	1.558 KB	92.4 %	7 min 52 seg

La Fig. 8 ilustra el desempeño del entrenamiento de la red 4, que fue la que tuvo mejor comportamiento. Sin embargo, para todos los casos es evidente el rápido aprendizaje logrado, en alrededor de 1200 iteraciones, y con altos grados de precisión, en todos los casos, superior al 92,4%. El gráfico muestra que se podría haber manejado un número mucho menor de épocas, cercano a 80.

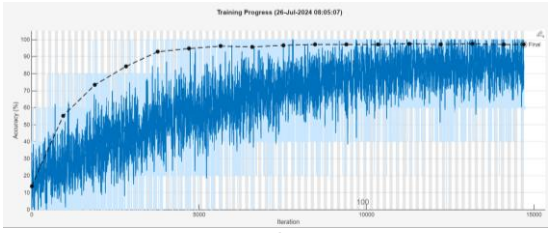

***Fig. 8.** Entrenamiento de números usando la red CNN 4*

La Tabla 4 muestra el tiempo de clasificación de cada arquitectura. Donde debido al bajo tiempo de clasificación de cada caso, para una aplicación en tiempo real, el tamaño de bytes de la red y su precisión son más significativos. Por lo tanto, se eligió la red 4 por ser la más equilibrada en estos dos factores.

***Tabla 4:** Classification times*

Red	Tiempo de clasificación
1	0.0428 segundos
2	0.0462 segundos
3	0.0502 segundos
4	0.0529 segundos
5	0.0730 segundos

La Fig. 9 ilustra la matriz de confusión de la red elegida. Las mejores clasificaciones se obtienen por las clases: tres, cinco y sí, la clase menos eficiente fue la cuatro, se observa se debe a la confusión entre las clases cuatro y uno, lo que puede ser resultado de una similitud de los patrones al no contar con una vocalización fuerte y clara.

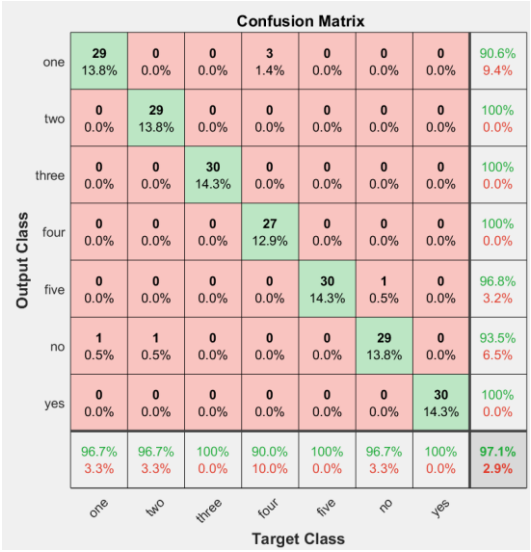

***Fig. 9.** Matriz de confusión de la red 4.*

4.3. Validación

La funcionalidad de la red seleccionada se valida a través de una App chatbot (ver Fig. 10), la cual muestra la escala a utilizar y el valor numérico asociado a cada una, la App muestra el número de la pregunta en la que se encuentra la encuesta, tiene un cuadro de texto que se puede leer y otro cuadro que muestra al usuario qué opción eligió y reconoció la red. Asimismo, el bot reproduce el audio de cada pregunta. El uso de un asistente tipo chatbot operado por voz se diseña de acuerdo con el diagrama de flujo presentado en la Fig. 1. Para el ejemplo, el cuestionario se basa en las siguientes preguntas:

1- ¿Cómo evalúa la actividad en general?
2- ¿Cómo valora la actividad realizada en términos de eficiencia de tiempo?
3- ¿Cómo califica la actividad realizada en términos de calidad?
4- ¿Cómo califica la orientación recibida en el desarrollo de la actividad?
5- ¿Cómo valora la actividad realizada en términos de eficiencia de uso?

***Fig. 10.** Aplicación de ejemplo*

4. CONCLUSIONES

En este artículo, se emplearon la transformada de Fourier de corto tiempo (STFT) y los coeficientes cepstrales de frecuencia Mel (MFCC) como técnicas de preprocesamiento de señales. Los coeficientes MFCC, Delta y Delta-Delta tenían como objetivo capturar la información relevante de la señal del habla humana. Además, el ruido de la señal se atenuó antes de aplicarse a la red neuronal convolucional. Estos pasos de preprocesamiento mejoraron la calidad de la información de entrada a la red convolucional.

Se logró extraer las características de audio relevantes que permitieron evaluar arquitecturas de redes convolucionales para desarrollar aplicaciones de chatbot basadas en interacción de audio. Las arquitecturas de red diseñadas para este fin presentan altos grados de precisión en todos los casos superiores al 92,4%, evidenciando el buen desempeño de las redes convolucionales en el aprendizaje de patrones de habla.

Para seleccionar la red más conveniente se consideraron como criterios de decisión: el mayor grado de precisión, el menor tamaño de bytes de la red y el menor tiempo de clasificación. Dado que ninguna opción cumple simultáneamente todos los criterios, se eligió la más equilibrada.

La aplicación desarrollada permite demostrar los alcances del uso de técnicas de aprendizaje profundo en interacciones naturales con bots, siendo una herramienta más amigable para el usuario.

Como trabajo futuro, se puede aumentar la cantidad de palabras de aprendizaje para ampliar el alcance de la conversación entre el usuario y el robot.

AGRADECIMIENTOS

Producto derivado del proyecto de investigación titulado “Diseño de un modelo de interacción humano-robot utilizando algoritmos de aprendizaje profundo” INV-ING-3971 financiado por la vicerrectoría de investigaciones de la Universidad Militar Nueva Granada, vigencia 2024.

REFERENCIAS

[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.