Audio Commands Recognition Through Deep Learning for Control Mobile Residential Assistant Robot

Reconocimiento de Comandos de Audio a través de Aprendizaje Profundo para el Robot Asistente Residencial Móvil de Control

Robinson Jimenez Moreno[1], Ricardo Castillo Estepa[2], Javier Martinez Baquero[3]

[1,2]Departamento de Ingeniería Mecatrónica, Universidad Militar Nueva Granada, Colombia,
[3] Msc Tecnología Educativa y Medios Innovadores para la Educación, Docente de Planta, Universidad de los Llanos, Colombia

robinson.jimenez@unimilitar.edu.co, ricardo.castillo@unimilitar.edu.co, jmartinez@unillanos.edu.co

Abstract–This article exposes the training of a convolutional neural network to recognize voice commands that allow an assistant robot to use in a residential environment. Given the needs for isolation due to health factors, development of care robots that can support a human presents a necessary research focus. For this case, the performance of the network is evaluated, which presents a 94.33% accuracy in the recognition of voice commands, through a virtual environment of a co n ventio na l residence.

Keywords: Convolutional Neural Network, Mel-Frequency Cepstral Coefficients, Feature Maps, Speech Recognition.

I. INTRODUCTION

Los sistemas de visión de máquina se han robustecido gracias a los algoritmos de Deep Learning[1]. Las aplicaciones van desde algoritmos de identificación de plantas como se expone[2], de vegetables[3] y frutas[4], muy incidentes en temas de punta como la agricultura de precisión. Así mismo la amplitud del espectro del uso de Deep learning y específicamente, como se evidencia en los trabajos mencionados, de las redes neuronales convolucionales (CNN) abarca reconocimiento de señales electromiografícas[5], detección de materiales peligrosos[6] e incluso en conducción autónoma, para detección de vehículos[7] y líneas guía[8]. De forma tal que incluso sistemas robóticos emplean CNN como parte de su esquema de visión, para aplicaciones como robots cosechadores[9] o de reordenamiento de objetos en bandas transportadoras a nivel industrial[10].

Una aplicación particularmente interesante de las redes convolucionales se encuentra en el reconocimiento de patrones en señales de audio[11]. Para el caso es posible hacer reconocimiento de emociones[12], detección de enfermedades como Parkinson[13], sistemas de atención[14] y clasificación del lenguaje[15][16]. Estas aplicaciones también se pueden integrar al control de sistemas robóticos como se expone en [17], donde se emplea una CNN para identificación del hablante mediante transformada Randon.

El presente trabajo aprovecha las ventajas expuestas de las CNN en reconocimiento de patrones y señales de audio para exponer una aplicación de control robótico asistencial en entornos residenciales, donde los comandos de navegación del robot, para ir a alguna locación especifica dentro de la residencia, son recibidos mediante voz y preprocesados empleando Coeficientes Cepstrales en las Frecuencias de Mel, para ser reconocidos por una red convolucional. Dada las necesidades de aislamiento experimentadas en la pandemia por covid[18], los robots asistenciales entran a jugar un papel fundamental en asistir en sus casas a personas que lo requieran.

II. MATERIALES Y MÉTODOS

El algoritmo de control de un robot asistencial en entorno residencial es desarrollado mediante el uso de comandos de

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2022.1.1.24
ISBN: 978-628-95207-0-5 **ISSN:** 2414-6390

voz, empleando un grupo palabras clave que le permitan saber a qué ubicación de la vivienda debe dirigirse. Dichos comandos son reconocidos mediante una red neuronal convolucional y se evalúa su desempeño por medio de un ambiente virtual, como se expone a continuación.

La base de datos consta de 2640 grabaciones de diferentes usuarios, distribuidas en 8 clases correspondientes a alcoba, baño, cocina, estudio, inicio, jardín, sala y otro. De éstas, se toman el 70% para entrenamiento y el 30% para validación. Cada audio es adquirido con una frecuencia de muestreo de 16000 Hz, realizando un preprocesamiento para extracción de características con el fin de obtener un mapa de dos dimensiones de cada señal de audio, que permite a la red convolucional aprender el comportamiento del comando de voz a través del tiempo. Esta extracción se realiza mediante los Coeficientes Cepstrales en las Frecuencias de Mel (Mel Frequency Cepstral Coefficients) o MFCCs de sus siglas, mediante las ecuaciones 1 a 3. Estos son coeficientes para la representación del habla basados en la percepción auditiva humana [19], ampliamente usados en sistemas de análisis del habla [20].

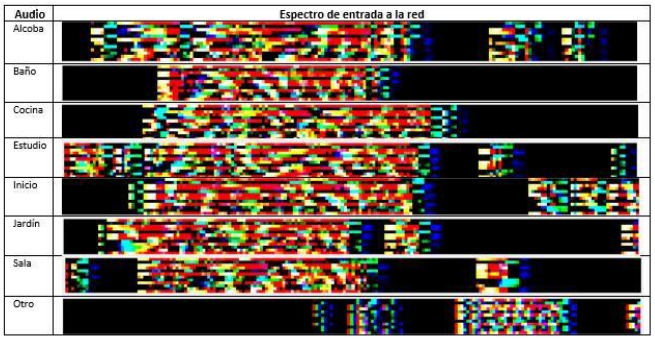
$$Cc'_n = \left(1 + \frac{L}{2}\sin\left(\frac{\pi n}{L}\right)\right) Cc_n, \qquad n = 0, \dots, C \tag{1}$$

$$\Delta Cc' = \frac{\sum_{n=1}^{N} n(Cc'_{t+n} - Cc'_{t-n})}{2\sum_{n=1}^{N} n^2}, \qquad t = 1, \dots, N_f \tag{2}$$

$$\Delta\Delta Cc' = \frac{\sum_{n=1}^{N} n(\Delta Cc'_{t+n} - \Delta Cc'_{t-n})}{2\sum_{n=1}^{N} n^2}, \qquad t = 1, \dots, N_f \tag{3}$$

Cada ecuación anterior permite generar un mapa de características de 12 coeficientes adquiridos de 199 frames, con sus respectivas primera y segunda derivada (ecuaciones 2 y 3 respectivamente). Por lo cual la entrada a la red es de dimensiones 12x199x3. En la tabla 1 se expone una muestra de la base de datos empleada.

Tabla 1. Muestra de la base de datos de entrenamiento



Es importante emplear la categoría otros para aquellos casos en que el robot, en modo de escucha capte audio que no corresponda a una ubicación, no lo asigne a una de las salidas de la red, que dado el caso de no usarse esta categoría, sería la más cercana pertenecientes a la locación entrenada de la vivienda residencial.

La arquitectura de red empleada se muestra en la tabla 2, no es una arquitectura muy profunda a comparación de las empleadas en transferencia de aprendizaje como la VGG o RESNET, dado el limitado número de salidas deseadas y el trabajo puntual que debe desarrollar la red. Los hiper- parámetros de entrenamiento fueron encontrados iterativamente obteniendo rata de aprendizaje de $1e^{-6}$, con 50 épocas.

Tabla 2. Arquitectura CNN empleada

Input: 12 x 199 x 3				
Layer	Kernel	Filters	Padding	Stride
Convolution	5	32	2	0
BatchNorm				
Convolution	5	32	2	0
MaxPooling	2	-	0	2
Convolution	3	64	1	0
Convolution	3	64	1	0
MaxPooling	2	-	0	2
Convolution	2	128	1	0
Convolution	3	128	1	0
MaxPooling	2	-	0	2
Fully-Conn	-	512	-	-
Dropout	-	-	-	-
Fully-Conn	-	2048	-	-
Dropout	-	-	-	-
Fully-Conn	-	8	-	-
Softmax				
Classification				

La figura 1 ilustra el proceso de aprendizaje de la red, con un tiempo de entrenamiento de 65 minutos para 94000 iteraciones, en un computador Intel Core i7 a 2.80GHz con GPU NVIDA Gforce GTX 1050 de 8GB, y finalmente un desempeño del 94.33%.

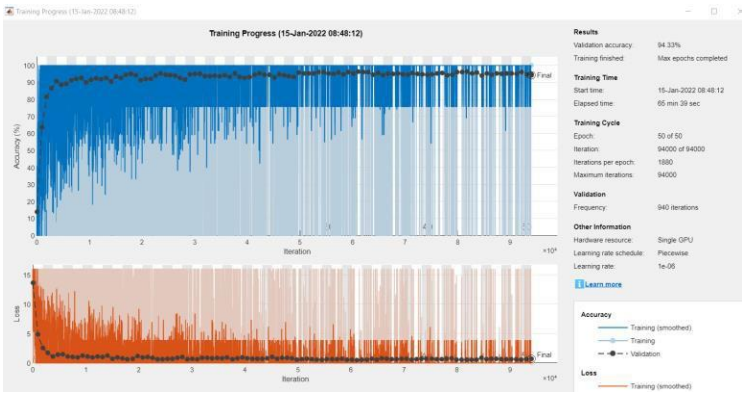


Figura 1. Entrenamiento de la CNN

La figura 2 muestra a la izquierda la matriz de confusión resultante del entrenamiento y a la derecha las activaciones de los filtros de la primera capa de la red. Se logra evidenciar que la locación baño es la que más confusión genera, principalmente con la clase sala. Parte del problema es la dimensión en el espectro dada la similitud de la longitud de ambas palabras. Se logra evidenciar que la etiqueta otro también actúa como filtro en el caso de que no se reconozca una locación valida, para lo cual el robot no se desplazará y deberá ser repetido el comando, lo importante es que no habrá desplazamiento a una locación residencial errónea.

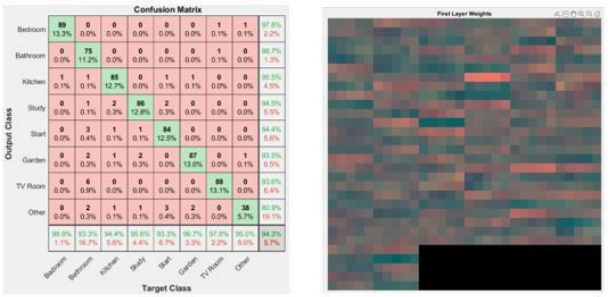


Figura 2. Matriz de confusión y pesos de la primera capa.

La red tiene como objeto permitir el desplazamiento de un robot asistencial dentro de un entorno residencial, de un lugar a otro, al recibir el comando de voz. Para realizar la validación se construyen ambientes virtuales en la herramienta VRML de MATLAB® con las locaciones de interés, como se evidencia en la figura 3. Esto permitirá evidenciar el desplazamiento del robot al recibir el comando de voz.

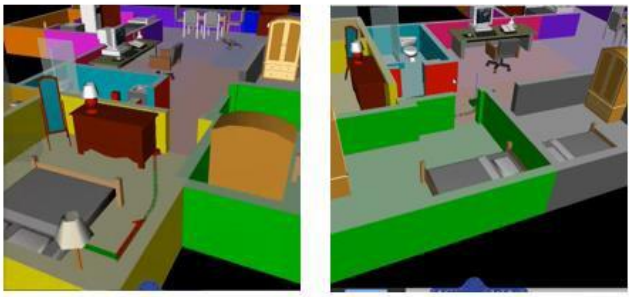


Figura 3. Ambiente virtual en VRML

III. ANÁLISIS DE RESULTADOS

La validación del desplazamiento del robot empleando los comandos de voz, al evaluar la operación conjunta del preprocesamiento mediante MFCC, la clasificación de la red convolucional y el entorno virtual se llevó a cabo mediante la interfaz gráfica de usuario GUIDE que se observa en la figura 4.

Figura 4. GUIDE de pruebas

El GUIDE se inicializa con el botón azul "first recognition", donde se identifica cada una de las locaciones de la residencia, el gráfico superior a este botón es la vista del robot, a la derecha se encuentra en la parte superior las locaciones vacías para identificar fácilmente donde se encuentra el robot mediante un cuadro verde, en la parte inferior el ambiente virtual residencial amoblado. Intermedio a estas dos gráficas se encuentra un indicador de grabación que está en rojo al iniciar y que indica el estado de reconocimiento del comando de voz. Al presionar el botón verde start a la izquierda el botón de grabación pasa a verde, como se aprecia en la figura 4 a la derecha.

La figura 5 ilustra una ampliación de las locaciones residenciales etiquetadas para desplazamiento robótico, hacia donde, según el comando de voz, se dirigirá el robot.

Figura 5. Ubicaciones para desplazamiento robótico

Una vez es reconocido el comando de voz, se visualiza el nombre del lugar a desplazarse bajo el indicador de grabación, como se ve en la figura 6. La figura a su vez ilustra cuatro ejemplos de desplazamiento exitoso del robot. La figura superior izquierda en que se reconoce la palabra baño y se da el desplazamiento adecuado, la figura superior derecha en que se reconoce la palabra sala y se da el desplazamiento adecuado, la figura inferior izquierda en que se reconoce la palabra alcoba y se da el desplazamiento adecuado, la figura

inferior derecha en que se reconoce la palabra estudio y se da el desplazamiento adecuado.

Figura 6. Pruebas desplazamiento robótico

Se realizaron 25 pruebas por cada una de las seis locaciones para evaluar el desempeño del robot en el desplazamiento en el entorno residencial, pronunciando cada palabra con variaciones de tono y velocidades al pronunciar.

Para el caso se obtuvo un 100% de respuesta acertada al reconocer cada comando de voz, se evidencia que no hay confusión entre ellos.

Se realizaron 50 pruebas para evaluar el desempeño del robot en el desplazamiento en el entorno residencial, pronunciando palabras que no se encuentran en el entrenamiento, obteniendo el resultado ilustrado en la tabla 3.

Tabla 3. Pruebas comandos desconocidos

Locación	Identificación
Alcoba	1
Estudio	4
Sala	2
Baño	9
Cocina	6
Jardín	2
Otro	26

Esto permite determinar que en la prueba se presentó un 48% de identificaciones erróneas que hacen que el robot se desplace sin desearlo. Se logra identificar que los casos de error presentado se dan por palabras con fonética similar y/o duración del sonido similar.

IV. CONCLUSIONES

Se logra validar que el método de preprocesamiento empleado permite discriminar los comandos de voz utilizados para desplazamiento robótico en cada locación dentro de la residencia. La red genera muy pocas confusiones entre las clases entrenadas, de forma que las pruebas realizadas no generaron desplazamientos erróneos.

Al validar comandos de voz fuera del entrenamiento, se evidencia desplazamientos no deseados, lo que permite concluir que se requiere ampliar el número de muestras de entrenamiento de la clase otros significativamente. Esto también muestra la importancia de incluir una clase adicional donde toda aquella entrada no deseada, no se asigne a una clase valida por similitud.

AGRADECIMIENTOS

Los autores agradecen a la Universidad Militar Nueva Granada y la Universidad de los Llanos. Agradecimiento especial al Vicerrectorado de Investigación de la Universidad Militar, por el financiamiento de este proyecto con código

IMP-ING-3405 (vigencia 2021-2022) y titulado "Prototipo robótico móvil para tareas de asistencia en ambientes residenciales".

REFERENCIAS

[1] Xiao Bai, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, Been Kim, Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments, Pattern Recognition, Volume 120, 2021, 108102, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2021.108102.

[2] G. Valarmathi, S.U. Suganthi, V. Subashini, R. Janaki, R. Sivasankari, S. Dhanasekar, CNN algorithm for plant classification in deep learning, Materials Today: Proceedings, Volume 46, Part 9, 2021, Pages 3684-3689, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.01.847.

[3] Christyan Cruz Ulloa, Anne Krus, Antonio Barrientos, Jaime del Cerro, Constantino Valero, Robotic Fertilization in Strip Cropping using a CNN Vegetables Detection-Characterization Method, Computers and Electronics in Agriculture, Volume 193, 2022, 106684, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2022.106684.

[4] Weikuan Jia, Yuyu Tian, Rong Luo, Zhonghua Zhang, Jian Lian, Yuanjie Zheng, Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot, Computers and Electronics in Agriculture, Volume 172, 2020, 105380, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2020.105380.

[5] Liukai Xu, Keqin Zhang, Genke Yang, Jian Chu, Gesture recognition using dual-stream CNN based on fusion of sEMG energy kernel phase portrait and IMU amplitude image, Biomedical Signal Processing and Control, Volume 73, 2022, 103364, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2021.103364.

[6] Amir Sharifi, Ahmadreza Zibaei, Mahdi Rezaei, A deep learning based hazardous materials (HAZMAT) sign detection robot with restricted computational resources, Machine Learning with Applications, Volume 6, 2021, 100104, ISSN 2666-8270, https://doi.org/10.1016/j.mlwa.2021.100104.

[7] Wankou Yang, Ziyu Li, Chao Wang, Jun Li, A multi-task Faster R-CNN method for 3D vehicle detection based on a single image, Applied Soft Computing, Volume 95, 2020, 106533, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2020.106533.

[8] Yassin Kortli, Souhir Gabsi, Lew F.C. Lew Yan Voon, Maher Jridi, Mehrez Merzougui, Mohamed Atri, Deep embedded hybrid CNN-LSTM network for lane detection on NVIDIA Jetson Xavier NX, Knowledge-Based Systems, 2022, 107941, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2021.107941.

[9] Longsheng Fu, Yaqoob Majeed, Xin Zhang, Manoj Karkee, Qin Zhang, Faster R–CNN–based apple detection in dense-foliage fruiting-wall trees using RGB and depth features for robotic harvesting, Biosystems Engineering, Volume 197, 2020, Pages 245-256, ISSN 1537-5110, https://doi.org/10.1016/j.biosystemseng.2020.07.007.

[10] Julián E. Herrera B., Robinson Jimenez-Moreno and Jorge A. Aponte-Rodríguez. Extrinsic camera calibration and inverse kinematics calculation through root finding problem. ARPN Journal of Engineering and Applied Sciences, October 2021, Vol. 16 No. 20.

[11] Kishore Kumar R., K. Sreenivasa Rao, A novel approach to unsupervised pattern discovery in speech using Convolutional Neural Network, Computer Speech & Language, Volume 71, 2022, 101259, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2021.101259.

[12] Mustaqeem, Soonil Kwon, MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach, Expert Systems with Applications, Volume 167, 2021, 114177, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.114177.

[13] Mehmet Bilal Er, Esme Isik, Ibrahim Isik, Parkinson's detection based on combined CNN and LSTM using enhanced speech signals with Variational mode decomposition, Biomedical Signal Processing and Control, Volume 70, 2021, 103006, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2021.103006.

[14] Lahiru Wijayasingha, John A. Stankovic, Robustness to noise for speech emotion classification using CNNs and attention mechanisms, Smart Health, Volume 19, 2021, 100165, ISSN 2352-6483, https://doi.org/10.1016/j.smhl.2020.100165.

[15] Yogesh Sharma, Bikesh Kumar Singh, One-dimensional convolutional neural network and hybrid deep-learning paradigm for classification of specific language impaired children using their speech, Computer Methods and Programs in Biomedicine, Volume 213, 2022, 106487, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2021.106487.

[16] Himanish Shekhar Das, Pinki Roy, A CNN-BiLSTM based hybrid model for Indian language identification, Applied Acoustics, Volume 182, 2021, 108274, ISSN 0003-682X, https://doi.org/10.1016/j.apacoust.2021.108274.

[17] Amira Shafik, Ahmed Sedik, Basma Abd El-Rahiem, El-Sayed M. El-Rabaie, Ghada M. El Banby, Fathi E. Abd El-Samie, Ashraf A.M. Khalaf, Oh-Young Song, Abdullah M. Iliyasu, Speaker identification based on Radon transform and CNNs in the presence of different types of interference for Robotic Applications, Applied Acoustics, Volume 177, 2021, 107665, ISSN 0003-682X, https://doi.org/10.1016/j.apacoust.2020.107665.

[18] Lisa M. Mann, Benjamin R. Walker, The role of equanimity in mediating the relationship between psychological distress and social isolation during COVID-19, Journal of Affective Disorders, Volume 296, 2022, Pages 370-379, ISSN 0165-0327, https://doi.org/10.1016/j.jad.2021.09.087.

[19]YOUNG, Steve, et al. The HTK book. Cambridge university engineering department, 2002, vol. 3, p. 175.
[20]QIAN, Yanmin; WOODLAND, Philip C. Very deep convolutional neural networks for robust speech recognition. En Spoken Language Technology Workshop (SLT), 2016 IEEE. IEEE, 2016. p. 481-488.