

Deep learning speech recognition for residential assistant robot

Robinson Jiménez-Moreno, Ricardo A. Castillo

Department of Mechatronic Engineering, Militar Nueva Granada University, Bogota, Colombia

Article Info

Article history:

Received Mar 16, 2022

Revised Aug 11, 2022

Accepted Sep 10, 2022

Keywords:

Assistant robot

Ceptral coefficients

Convolutional network

Speech recognition

Voice command

ABSTRACT

This work presents the design and validation of a voice assistant to command robotic tasks in a residential environment, as a support for people who require isolation or support due to body motor problems. The preprocessing of a database of 3600 audios of 8 different categories of words like "paper", "glass" or "robot", that allow to conform commands such as "carry paper" or "bring medicine", obtaining a matrix array of Mel frequencies and its derivatives, as inputs to a convolutional neural network that presents an accuracy of 96.9% in the discrimination of the categories. The command recognition tests involve recognizing groups of three words starting with "robot", for example, "robot bring glass", and allow identifying 8 different actions per voice command, with an accuracy of 88.75%.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Robinson Jiménez Moreno

Mechatronics Engineering Program, Faculty of Engineering, Universidad Militar Nueva Granada

Carrera 11 # 101-80, Bogotá D.C., Colombia

Email: robinson.jimenez@unimilitar.edu.co

1. INTRODUCTION

Speech recognition techniques present a very wide field of research with diverse applications such as speech impairment, improving the accuracy of voice command fingerprinting attacks and more, as discussed in [1]–[3]. One of the most representative research projects is in emotion recognition using spectrograms, mel frequency cepstral coefficient (MFCC) and convolutional networks [4], [5]. Cases such as the one presented in [6] employ convolutional networks with 3-dimensional inputs based on the first and second derivative of the spectrogram.

Interaction by voice commands with robots is another field of research interest in speech recognition [7]–[9]. The use of voice assistants such as Amazon's Alexa [10] or Google's [11], allow to obtain a more natural method of human-robot interaction. Thus, highlighting that voice commands are a necessity in interaction with robots [12], where for this research also the use of convolutional networks provides high performance [13]–[15].

Nowadays, the development of intelligent environments is gaining strength, including smart homes [16]. Robotic technology has been included in these environments with different fronts such as people care [17], cleaning [18] and even cooking [19]. In [20] and [21], the development of assistive robots to address patient isolations by COVID is exposed, however, the development is oriented to systems telecommanded by cellular mobile equipment.

Given the relevance of this topic and the need for a more natural and autonomous telecommand system, this work presents an audio command recognition system oriented to an assistive robot in a residential environment, thus integrating what has been found in the state of the art by means of a voice assistant for robotic action, performing audio preprocessing by ceptral coefficients and subsequent recognition by means of convolutional networks. As a contribution to the state of the art, a neuro-

convolutional architecture is designed to be easily embedded in portable electronic systems, performing the separation of the command words by means of a sliding window that calculates the power density of the audio signal.

The document is divided into four sections, the present section exposes the state of the art related to the work developed. Section two presents the methodology used for the separation of the words that make up each command and the neural training. Section 3 presents the analysis of the results achieved and finally section four presents the conclusions reached.

2. METHOD

The proposed objective is to use voice commands consisting of groups of three words, which allow the execution of assistive actions of a mobile robot within a residential environment. The sequence of control words is recorded and each one of them is separated to obtain a two-dimensional map of each audio signal, by means of mel frequency cepstral coefficients (MFCCs). Each map is classified using a convolutional neural network and the coherence of the command to execute the action is validated. The general scheme is presented in Figure 1.

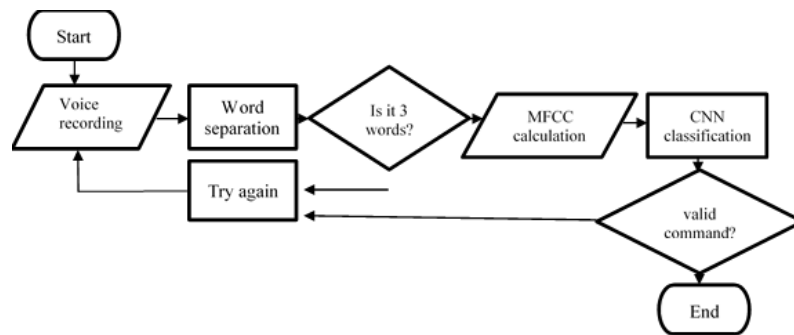


Figure 1. Flowchart of the experimental methods applied

The training of the model is performed by creating a database consisting of 3600 recordings of different users, distributed in 8 classes corresponding to robot, bring, carry, stop, paper, cup, towel, and medicine, where 80% are taken for training and 20% for validation. Each audio is acquired with a sampling frequency of 16000 Hz and each word is separated by the location of the minima found when obtaining the absolute value of the original signal, as shown in Figure 2. By means of a sliding window of ten times the input frequency, the power density of the audio signal is calculated, each one is compared with the previous value and if it presents a decay of 75% it is established as a local minimum, a point used for the separation of each word, where the asterisks on the right side represent the minimum values found.

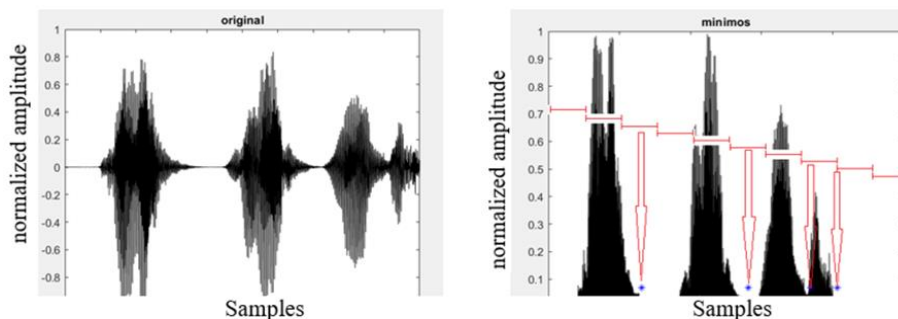


Figure 2. Minimal detection for words separation

The database is made with 10 male and 10 female users, to diversify the learning. Figure 3 shows examples of the database. Figure 3(a) shows an example of a woman's voice and Figure 3(b) shows an

example of a man's voice, both diverge in amplitude and frequency spectrum. The original database is taken in Spanish language.

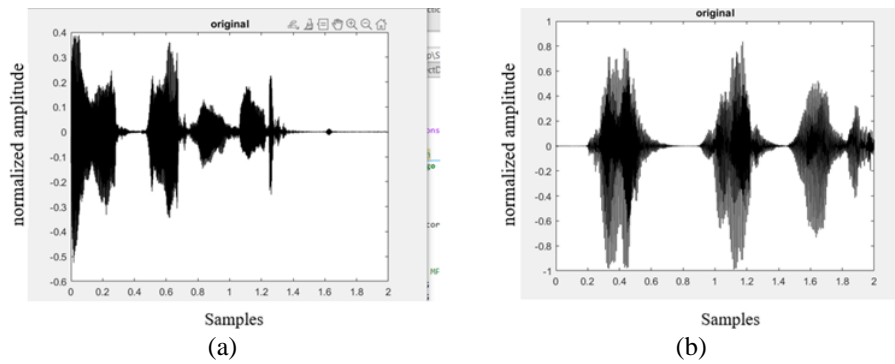


Figure 3. Command robot carrying medicine, (a) Female voice and (b) Male voice

Each of the three words is preprocessed for feature extraction to obtain a two-dimensional, three-channel map, which allows a convolutional neural network [22] to learn the behavior of the voice command over time, to be recognized. The feature map is obtained by calculating the mel frequency cepstral coefficients (MFCCs) using (1) to (3). These are coefficients for speech representation based on human auditory perception [23], widely used in speech analysis systems [24].

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

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

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

By means of 1 it is possible to generate a feature map of 12 coefficients acquired from 199 frames. Being this the first input channel to the network, the first and second derivative (2 and 3 respectively), generate the other two channels. Therefore, the learning input to the network is of dimensions 12×199×3.

The network architecture used is shown in Table 1, employing six convolution layers, given the limited number of desired outputs and the punctual work to be performed by the network. The training hyperparameters were found iteratively using a learning rate of 1e-6, with 50 epochs. Figure 4 illustrates the network learning process, with a training time of 31 minutes for 79250 iterations, on a 2.30GHz Intel Core i7 computer with NVIDIA Gforce RTX 3070 8GB GPU, and finally a performance of 96.9%.

Table 1. CNN architecture used

Input: 12×199×3				
Layer	Kernel	Filters	Padding	Stride
Convolution	7	32	2	1
	BatchNorm			
Convolution	5	32	2	1
MaxPooling	2	-	0	2
Convolution	3	64	1	1
Convolution	3	64	1	1
MaxPooling	2	-	0	2
Convolution	2	128	1	1
Convolution	3	128	1	1
MaxPooling	2	-	0	2
Fully-Conn	-	512	-	-
Dropout	-	-	-	-
Fully-Conn	-	2048	-	-
Dropout	-	-	-	-
Fully-Conn	-	8	-	-
	Softmax			
	Classification			

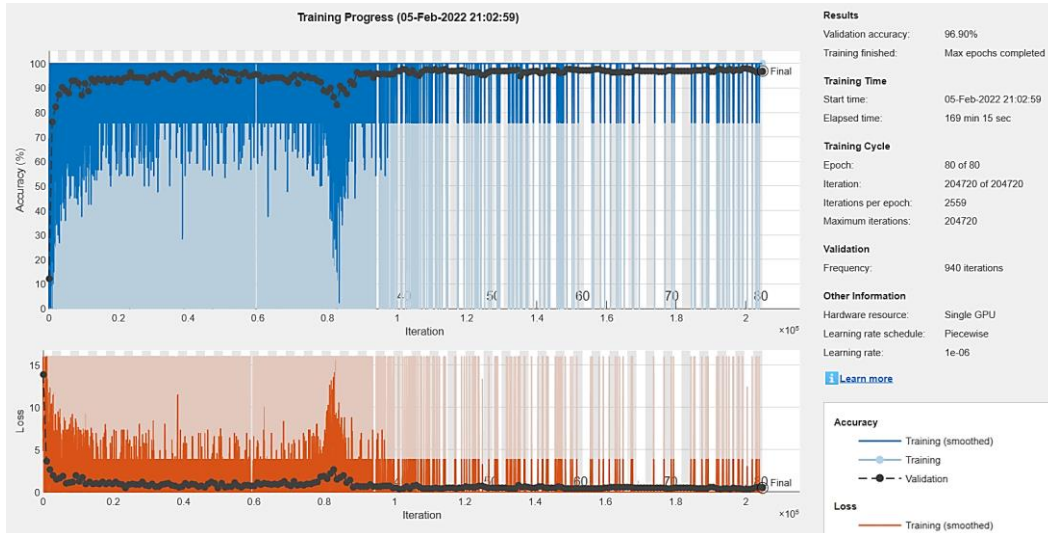


Figure 4. Network accuracy

Figure 5 illustrates the resulting confusion matrix, where it is possible to see the good behavior in the recognition of the words among themselves. Only the category "stop" showing a significant percentage of confusion with the category "towel". The average time for the classification of each word by the network is 0.6 seconds, where each word is submitted to the network for classification, generating an average response time of 1.8 seconds.

		Confusion Matrix								
Output Class	Carry	112 12.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.2%	1 0.1%	0 0.0%	97.4% 2.6%
	Medicine	0 0.0%	113 12.5%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.1%	0 0.0%	99.1% 0.9%
	Robot	0 0.0%	0 0.0%	113 12.5%	0 0.0%	0 0.0%	1 0.1%	5 0.6%	1 0.1%	94.2% 5.8%
	Paper	0 0.0%	0 0.0%	0 0.0%	113 12.5%	0 0.0%	1 0.1%	8 0.9%	0 0.0%	92.6% 7.4%
	Stop	0 0.0%	0 0.0%	0 0.0%	0 0.0%	112 12.4%	5 0.6%	0 0.0%	0 0.0%	95.7% 4.3%
	Towel	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	103 11.4%	0 0.0%	0 0.0%	100% 0.0%
	Bring	1 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	98 10.8%	0 0.0%	99.0% 1.0%
	Glass	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.1%	1 0.1%	0 0.0%	112 12.4%	98.2% 1.8%
			99.1% 0.9%	100% 0.0%	100% 0.0%	100% 0.0%	99.1% 0.9%	91.2% 8.8%	86.7% 13.3%	99.1% 0.9%
	Target Class	Carry	Medicine	Robot	Paper	Stop	Towel	Bring	Glass	

Figure 5. final confusion matrix

3. RESULTS AND DISCUSSION

The algorithm is validated by evaluating the action commands to be developed by the mobile robot. For this purpose, the variants of the commands are established according to the words to be recognized as shown in Table 2. The number of true positives versus false positives that the algorithm exhibits is determined. A true positive corresponds to a valid command of the desired action, a false positive corresponds to a valid command, but not according to the desired action.

Table 2. Valid commands

Command	True positives	False positives
Robot bring glass	16	1
Robot bring paper	18	2
Robot bring towel	16	1
Robot bring medicine	20	5
Robot carrying glass	17	0
Robot carrying paper	18	3
Robot carrying towel	17	0
Robot carrying medicine	20	6

The algorithm filters by software the validity of a command initially evaluating the existence of the three words, in this case by means of the minima of the signal spectrum. Figure 6 illustrates the case two examples of the commands "robot bring glass" shows in Figure 6(a) and "robot carry paper" shows in Figure 6(b) with the location of the minima that result in the division of the words, where the difference between each command is appreciated. The first word must always be robot, otherwise it will not be validated. From Table 2, it is possible to derive an efficiency of 88.75% in the discrimination of the commands to the robot, where the characteristics of confusion of classes between carry and bring stand out: 16.6% of false positives and 55.5% of false positives correspond to confusing the object (glass, paper and towel) with the class medicine.

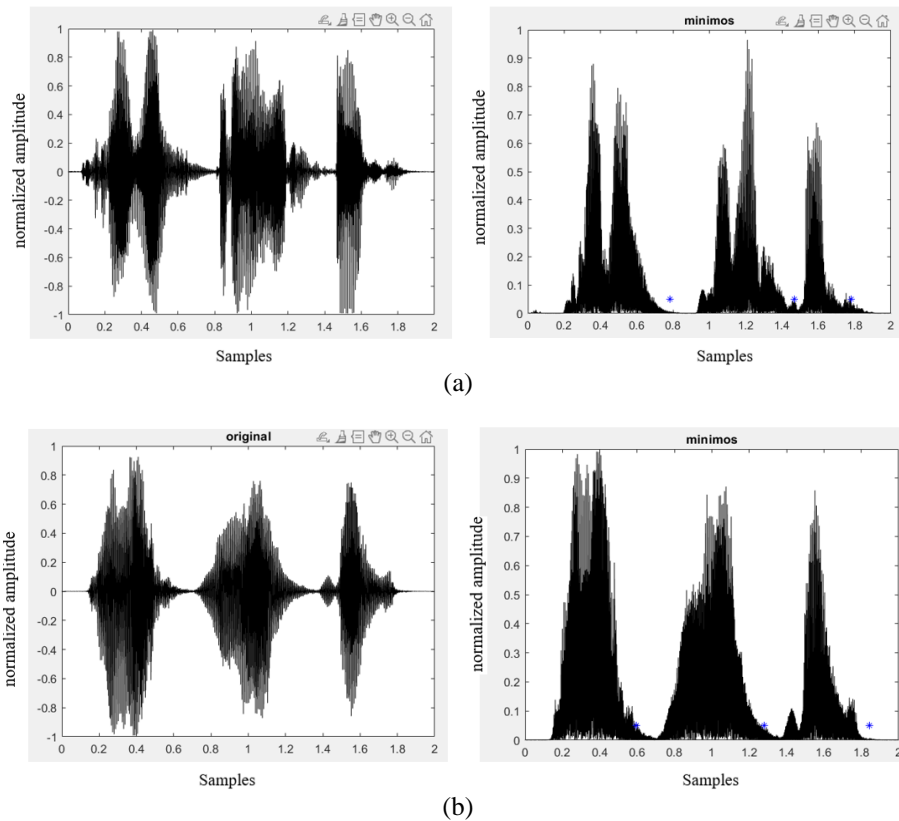


Figure 6. Example commands, (a) robot bring glass and (b) robot carry paper

Figure 7 illustrates the results of the prediction of the network by discriminating each word and evaluating it, to the right of each separated word there is evidence of noise generated by complementing the size of the information vector. This is because the duration of the input to the network is 2 seconds, which at a sampling frequency of 16000 Hz implies 32000 samples. When trimming each word, the vector is shortened and, since it cannot be filled in a concerted manner, due to the MFCC derivatives, a random filling of ± 0.01 is generated. Figure 7(a) and Figure 7(b) illustrate two examples of different commands from the original signal to their separation into words and recognition of each (robot bring paper and robot carry paper respectively).

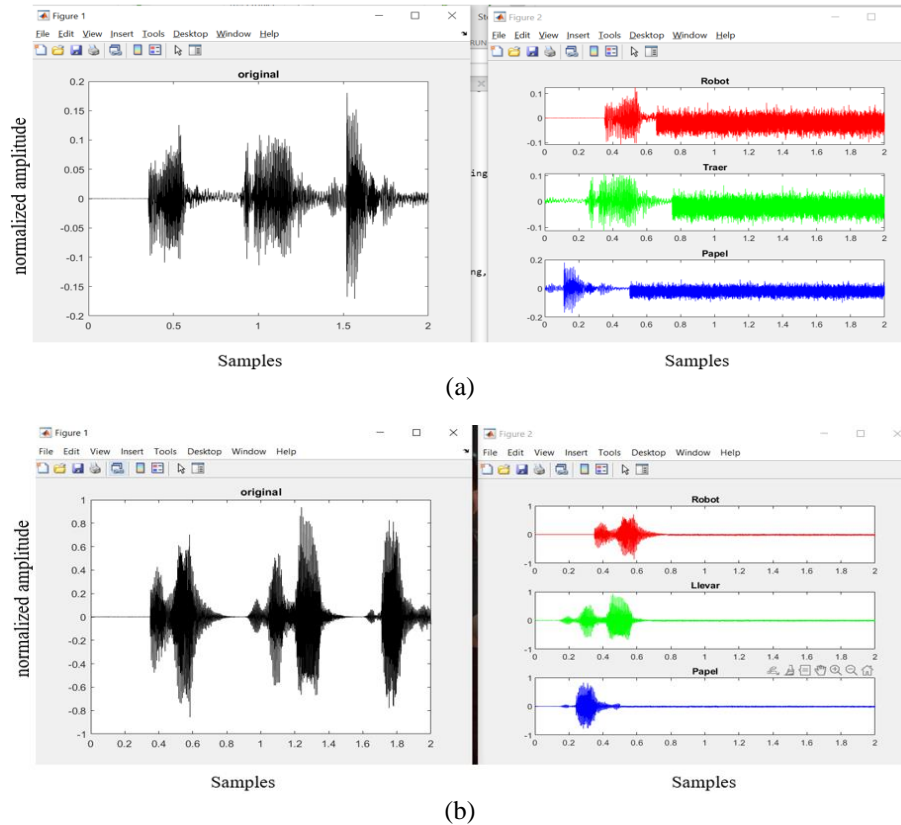


Figure 7. Commands correctly recognized, (a) robot bring paper and (b) robot carry paper

In contrast to Figure 7, Figure 8 illustrates a case of erroneous detection in the action command "robot carry paper". The similarity of the first and second word is evidenced, varying mainly in amplitude. So it is recognized as the same word generating in the network the output "robot robot paper", which is classified as an invalid command. In this case, the error was associated to environmental noise at the time of recording the command.

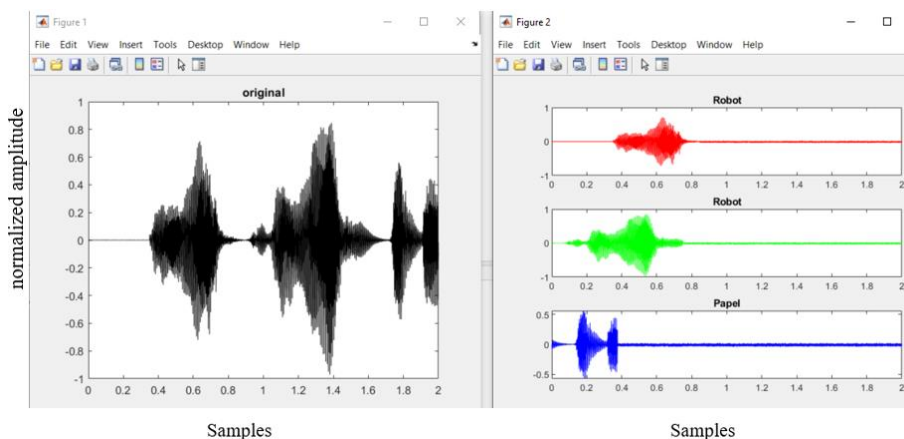


Figure 8. Error in detecting robot carrying paper

Similar work is presented in [25], where the robotic action commands also employ word separation and generate feature extraction by MFCC, using a single channel, but combining the CNN with an LSTM network, they report up to 90.37% accuracy for word recognition. The 6% improvement achieved by the

CNN network developed in this work is due to a higher number of training audios and the use of the MFCC derivatives of each word. It is validated that the use of 7×7 filters in the first convolutional layer, instead of 5×5 as in [25], also helped to improve the accuracy by 3%.

A virtual environment was designed for evaluation of the robotic navigation task and voice command discrimination, as shown in Figure 9. The response time of the robot in discriminating the actions is about 8 seconds. This time include the robot responses of valid command and identification of the place where the desired action will be carried out in the residential environment.

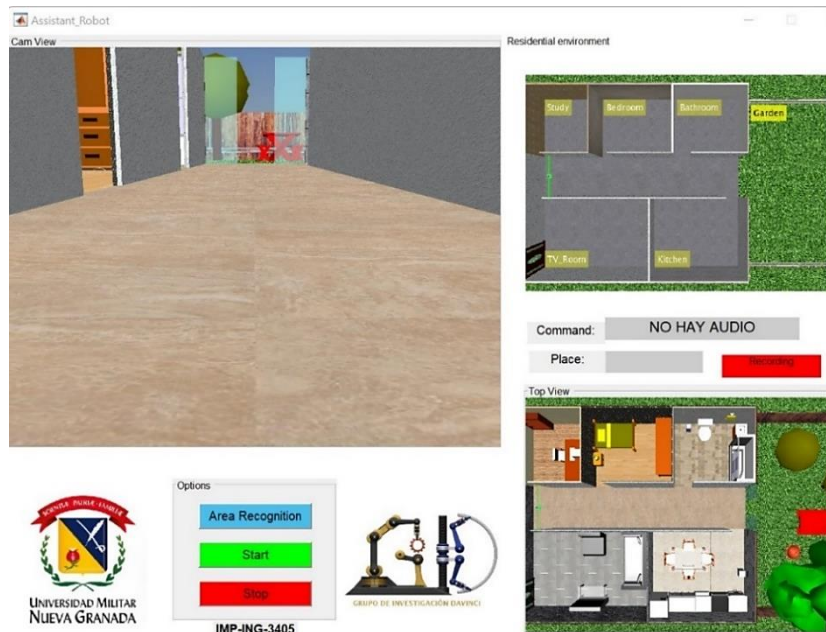


Figure 9. Virtual test environment

4. CONCLUSION

The use of convolutional networks for voice command generation in mobile robotics offers a natural field of human-machine interaction. It is concluded that MFCC discrimination allows to generate a map of recognizable features by the network that results in a functional voice assistant for robotic command. It was found that speech recognition accuracy depends on a low ambient noise factor and on generating an adequate vocalization of each word. This factor decreases its incidence when enlarging the database with background noise and varying the speed and volume of pronunciation. It was concluded from the training of the network that the use of few classes facilitates the discrimination of the spoken word, suggesting that future training can use identification trees, for example, a network of identification of actions (verbs) and one of objects (nouns), to expand the number of commands received by the robot.

ACKNOWLEDGEMENTS

The authors are grateful to Universidad Militar Nueva Granada and the vicerrectoría de investigaciones de la Universidad Militar, for the funding of this project with code IMP-ING-3405 (validity 2021-2022) and titled "Prototipo robótico móvil para tareas asistenciales en entornos residenciales".




REFERENCES

- [1] R. Kishore Kumar and K. Sreenivasa Rao, "A novel approach to unsupervised pattern discovery in speech using Convolutional Neural Network," *Computer Speech and Language*, vol. 71, 2022, doi: 10.1016/j.csl.2021.101259.
- [2] G. Guven, U. Guz, and H. Gürkan, "A novel biometric identification system based on fingertip electrocardiogram and speech signals," *Digital Signal Processing: A Review Journal*, vol. 121, 2022, doi: 10.1016/j.dsp.2021.103306.
- [3] L. A. Kumar, D. K. Renuka, S. L. Rose, M. C. Shummuga priya, and I. M. Wartana, "Deep learning based assistive technology on audio visual speech recognition for hearing impaired," *International Journal of Cognitive Computing in Engineering*, vol. 3, pp. 24–30, 2022, doi: 10.1016/j.ijcce.2022.01.003.
- [4] O. Atila and A. Şengür, "Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition," *Applied Acoustics*, vol. 182, 2021, doi: 10.1016/j.apacoust.2021.108260.




- [5] L. Wijayasingha and J. A. Stankovic, "Robustness to noise for speech emotion classification using CNNs and attention mechanisms," *Smart Health*, vol. 19, 2021, doi: 10.1016/j.smhl.2020.100165.
- [6] Z. Zhao *et al.*, "Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition," *Neural Networks*, vol. 141, pp. 52–60, 2021, doi: 10.1016/j.neunet.2021.03.013.
- [7] A. Pipitone and A. Chella, "What robots want? Hearing the inner voice of a robot," *iScience*, vol. 24, no. 4, 2021, doi: 10.1016/j.isci.2021.102371.
- [8] P. Fiati, "SMILE: A verbal and graphical user interface tool for speech-control of soccer robots in Ghana," *Cognitive Robotics*, vol. 1, pp. 25–28, 2021, doi: 10.1016/j.cogr.2021.03.001.
- [9] N. Iwahashi, "Language acquisition through a human-robot interface by combining speech, visual, and behavioral information," *Information Sciences*, vol. 156, no. 1–2, pp. 109–121, 2003, doi: 10.1016/S0020-0255(03)00167-1.
- [10] C. Deuerlein, M. Langer, J. Seßner, P. Heß, and J. Franke, "Human-robot-interaction using cloud-based speech recognition systems," *Procedia CIRP*, vol. 97, pp. 130–135, 2020, doi: 10.1016/j.procir.2020.05.214.
- [11] R. M. Bommi, J. Vijay, V. Murali Manohar, J. P. Dinesh Kumar, and D. Sriram, "Speech and gesture recognition interactive robot," *Materials Today: Proceedings*, vol. 47, pp. 37–40, 2021, doi: 10.1016/j.matpr.2021.03.503.
- [12] P. Gustavsson, A. Syberfeldt, R. Brewster, and L. Wang, "Human-robot Collaboration Demonstrator Combining Speech Recognition and Haptic Control," *Procedia CIRP*, vol. 63, pp. 396–401, 2017, doi: 10.1016/j.procir.2017.03.126.
- [13] A. Shafik *et al.*, "Speaker identification based on Radon transform and CNNs in the presence of different types of interference for Robotic Applications," *Applied Acoustics*, vol. 177, 2021, doi: 10.1016/j.apacoust.2020.107665.
- [14] D. Yongda, L. Fang, and X. Huang, "Research on multimodal human-robot interaction based on speech and gesture," *Computers and Electrical Engineering*, vol. 72, pp. 443–454, 2018, doi: 10.1016/j.compeleceng.2018.09.014.
- [15] H. Chaurasiya and G. Chandra, "Ambience Inhaling: Speech Noise Inhaler in Mobile Robots using Deep Learning," *Procedia Computer Science*, vol. 133, pp. 864–871, 2018, doi: 10.1016/j.procs.2018.07.108.
- [16] S. Reig, T. Fong, J. Forlizzi, and A. Steinfeld, "Theory and Design Considerations for the User Experience of Smart Environments," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 3, pp. 522–535, 2022, doi: 10.1109/THMS.2022.3142112.
- [17] J. Berrezueta-Guzman, V. E. Robles-Bykbaev, I. Pau, F. Pesantez-Aviles, and M. L. Martin-Ruiz, "Robotic Technologies in ADHD Care: Literature Review," *IEEE Access*, vol. 10, pp. 608–625, 2022, doi: 10.1109/ACCESS.2021.3137082.
- [18] P. Veerajagadheswar, S. Yuyao, P. Kandasamy, M. R. Elara, and A. A. Hayat, "S-Sacrr: A Staircase and Slope Accessing Reconfigurable Cleaning Robot and its Validation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4558–4565, 2022, doi: 10.1109/LRA.2022.3151572.
- [19] G. Sochacki, J. Hughes, S. Hauser, and F. Iida, "Closed-Loop Robotic Cooking of Scrambled Eggs with a Salinity-based 'Taste' Sensor," *IEEE International Conference on Intelligent Robots and Systems*, pp. 594–600, 2021, doi: 10.1109/IROS51168.2021.9636750.
- [20] S. Sowrabh, A. Rameshkumar, P. V. Athira, T. B. Jishnu, and M. N. Anish, "An Intelligent Robot Assisting Medical Practitioners to Aid Potential Covid-19 Patients," *Proceedings of the IEEE International Conference Image Information Processing*, vol. 2021-Novem, pp. 413–417, 2021, doi: 10.1109/ICIIP53038.2021.9702538.
- [21] Y. Krishnan, V. Udayan, and S. Akhil, "Hospital Assistant Robotic Vehicle (HARVi)," *Proceedings of the 2021 IEEE 18th India Council International Conference, INDICON 2021*, 2021, doi: 10.1109/INDICON52576.2021.9691738.
- [22] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
- [23] S. Young, "The HTK book," *Camb. Univ. Eng. Dep.*, vol. 2, no. 175, p. null, 1997.
- [24] Y. Qian and P. C. Woodland, "Very deep convolutional neural networks for robust speech recognition," *2016 IEEE Workshop on Spoken Language Technology, SLT 2016 - Proceedings*, pp. 481–488, 2017, doi: 10.1109/SLT.2016.7846307.
- [25] M. C. Bingol and O. Aydogmus, "Performing predefined tasks using the human-robot interaction on speech recognition for an industrial robot," *Engineering Applications of Artificial Intelligence*, vol. 95, 2020, doi: 10.1016/j.engappai.2020.103903.

BIOGRAPHIES OF AUTHORS



Robinson Jiménez-Moreno    is an Electronic Engineer graduated from the Francisco José de Caldas District University in 2002. He received a M.Sc. in Engineering from the National University of Colombia in 2012 and D. Eng (Doctor in Engineering) from the Francisco José de Caldas District University in 2018. His current working as assistant professor of Universidad Militar Nueva Granada. In addition, he served as Director of Research for the faculty of engineering (2019-2021). His research interests focus on the use of convolutional neural networks for object recognition and image processing, for robotic applications such as assistive robotics and human-machine interaction. He can be contacted at email: robinson.jimenez@unimilitar.edu.co.



Ricardo A. Castillo    was born in Colombia in 1980. He received his B.Eng. in Mechatronics Engineering in Universidad Militar Nueva Granada in 2004, and his M.Sc. and Ph.D. in Mechanical Engineering in the University of Campinas-UNICAMP, Brazil in 2010 and 2015 respectively for his work on coordination and indirect communication strategies for mechatronic systems. Since 2005 he is a full-time Professor and Researcher in the Department of Mechatronics Engineering at Universidad Militar Nueva Granada (Bogotá-Colombia). His current research projects deal with collaborative modular robotics, applied artificial intelligence, and mobile autonomous robotics. He can be contacted at email: ricardo.castillo@unimilitar.edu.co.