TecnoLógicas

Simulation Environment for User-Robot Voice Interaction in Product Handling and Packaging Systems

Entorno de simulación para interacción por voz usuario-robot en sistemas de manipulación y empaquetado de productos

Juan C. Guachetá-Alba [1], Robinson Jimenez-Moreno[2], Anny Astrid Espitia-Cubillos[2]

[1]Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil
[2]Universidad Militar Nueva Granada, Bogotá, Colombia

Correspondence: anny.espitia@unimilitar.edu.co

Received: 05 August 2025 **Accepted:** 20 March 2026 **Available:** 6 April 2026

How to cite / Cómo citar

J. C. Guachetá-Alba, R. Jimenez-Moreno, and A. A. Espitia-Cubillos, "Simulation Environment for User-Robot Voice Interaction in Product Handling and Packaging Systems," *TecnoLógicas*, vol. 29, no. 65, e3601, 2026. https://doi.org/10.22430/22565337.3601

Abstract

Advances in artificial intelligence, robotics, and human-machine interaction have permeated industrial processes, making them smarter, more flexible, and more autonomous. Voice interaction is a natural medium that has become a key factor in improving production systems, applicable to processes such as order picking. In this context, this article aimed to develop an intelligent voice-controlled packaging system that integrates natural language processing, computer vision, and robotic product handling, which was then evaluated in a virtual simulation environment. The methodology consisted of designing a simulated environment, developing and integrating voice recognition and synthesis algorithms into an automated workflow that allows the selection and manipulation of finished products to prepare orders. The system was implemented using a robotic platform, a conveyor belt, and a finished goods storage area. An interface was created to facilitate control of the packaging process through a chatbot capable of understanding voice commands and responding to the user utilizing voice recognition and synthesis algorithms. The system's accuracy and robustness were evaluated using 20 commands, analyzed with recordings generated for 13 user profiles with different accents, ages, and vocal characteristics. The results showed an average accuracy of 85.7% in command transcription, with robust performance against voice variations, although the lowest accuracy was observed when using children's voices. Additionally, the packaging and sorting system was validated in the virtual environment, demonstrating its efficient operation in managing box space and available shelf inventory. Thanks to the results obtained, it is possible to conclude that the designed system allows for almost natural and flexible interaction through voice commands, effectively integrating language recognition and robotic manipulation in a simulated environment, which demonstrates its potential application in supply chain automation and Industry 5.0.

Keywords

Artificial intelligence, computer simulation, computer vision, industry 5.0, natural language processing.

Resumen

El avance en inteligencia artificial, robótica e interacción hombre-máquina ha permeado los procesos industriales para que sean más inteligentes, flexibles y autónomos. La interacción por voz es un medio natural que se constituye como un factor clave para mejorar los sistemas productivos, que puede aplicarse a procesos como el alistamiento de pedidos. En ese contexto, el presente artículo tuvo como objetivo desarrollar un sistema inteligente de empaquetado controlado por voz que integra herramientas de procesamiento de lenguaje natural, visión por computador y manipulación robótica de productos, el cual fue evaluado en un entorno virtual de simulación. La metodología consistió en diseñar un entorno simulado, desarrollar e integrar los algoritmos de reconocimiento y síntesis de voz en un flujo de trabajo automatizado, que permite seleccionar y manipular los productos terminados para preparar los pedidos. Para ello, se implementó el sistema con una plataforma robótica, una banda transportadora y el área de almacenamiento de productos terminados y se creó una interfaz que facilita el control del proceso de empaquetado, mediante un chatbot capaz de entender comandos de voz y responder al usuario gracias a algoritmos de reconocimiento y síntesis de voz. Se evaluó la precisión y robustez del sistema usando 20 comandos, analizados con grabaciones generadas para 13 perfiles de usuarios con distintos acentos, edades y características vocales. Los resultados mostraron una precisión promedio del 85.7 % en la transcripción de comandos, con un desempeño robusto frente a variaciones de voz, pese a presentar la menor precisión cuando se usan voces de niños. Adicionalmente, se validó en el entorno virtual que el sistema de empaquetado y clasificación funciona eficazmente, gestionando el espacio de las cajas y el inventario disponible en el estante. Gracias a los resultados obtenidos, es posible concluir que el sistema diseñado permite la interacción casi natural y flexible mediante comandos de voz, integrando de forma efectiva el reconocimiento de lenguaje y la manipulación robótica en un entorno simulado, lo que evidencia su potencial de aplicación en la automatización de cadenas de suministro y la industria 5.0.

Palabras clave

Inteligencia artificial, simulación computacional, visión por computador, industria 5.0, procesamiento de lenguaje natural.

1. INTRODUCTION

Technological advances are impacting the way people interact with technology, so that today a human can converse with a robot receptionist [1], a customer can interact with a robot by voice or touch [2] to manage a service, or replicate interaction with robotic babies so that older adults can engage in interaction and care activities [3]. Visual and auditory perception can be integrated into conversations with robots to make communication more natural [4] and lead to developments in the use of assistive robots in fields as diverse as rehabilitation [5].

Voice interaction with robots facilitates tasks such as directing them spatially [6], controlling wheelchairs [7], or performing risky activities for people using robotic agents [8]. To this end, voice control of robots is supported by other artificial intelligence techniques such as fuzzy logic [9] or reinforcement learning [10]. In turn, testing with robots finds a safe means of validation in virtual environments [11], [12], which even allows for the replication of complex environments with environmental variables such as those required in climbing robots [13] or robots based on multiple robots [14].

These developments have direct applications in industry [15], in areas such as construction [16], [17], decision-making [18], and planning strategies for safe human-robot interaction [19], where even activities such as agriculture can be replicated virtually [20]. This is fundamental for developing Industry 5.0 systems based on human-robot interaction [21], [22] and can be leveraged in the supply chain for product dispatch, an application not directly found in the state of the art with human-robot interaction by voice and which is addressed as a contribution in this research work.

Automated packaging systems in industry have incorporated artificial intelligence algorithms, leading to terms like “smartification” of processes such as counting on packaging lines [23]. Algorithms like neural networks are used for product packaging, as presented in [24], which utilizes information on the number of boxes, their average volume, and box counts by destination. However, automation processes in product selection and storage require video systems for product identification and the use of robotic systems for handling. In [25], [26] robotic platforms are developed for the automatic harvesting and storage of apples, employing

AI algorithms such as the YOLO network. However, in this automated storage environment, there is no evidence of industrial-scale development in integrating advanced voice command recognition for robotic handling on the production line.

Although advances in human-machine interaction, voice control, computer vision, and simulation have driven the development of robotic systems, the integration of these technologies into product packaging and shipping is not widespread at the industrial level; instead, only isolated developments exist. It is observed that current automated systems perform visual object recognition using artificial intelligence algorithms; however, developments are lacking for handling final products in order preparation stages that include interaction with the user through verbal instructions that allow reducing direct user participation.

This preliminary literature review on the subject of study reveals a concentration of documents focused on robotic control for movement, task completion, and audio recognition in non-industrial environments. This study, in turn, seeks to utilize computational simulation as a tool to validate the functionality of the human-robot interface that operates via voice command for order picking of finished products for subsequent shipment.

This article presented the design, development, and validation of an intelligent product handling and packaging system that operates using voice commands and integrates computer vision, industrial robotics, and natural language processing tools. The system was tested through computer simulation to identify its accuracy, robustness, and responsiveness to variations in the intonation and pronunciation of keywords by different users. The system was developed in MATLAB and validated using CoppeliaSim.

2. METHODOLOGY

With the goal of developing a smart agent that enables human-robot interaction based on voice commands, artificial intelligence algorithms based on deep learning are integrated using MATLAB with an interface in CoppeliaSim. The virtual environment in Coppelia establishes the dynamics and physics required to operate a robotic manipulator in the handling of products for automated packaging.

2.1 System design

The methodology used, which corresponds to applied research development, is based on the construction of the virtual environment in Coppelia, where a full product storage shelf is incorporated, a robotic manipulator with a displacement line is located, and a conveyor belt is used for packaging the product in boxes, as shown in Figure 1. The robotic manipulator corresponds to a UR5 arm controlled by its inverse kinematics and is programmed for product positioning and gripping tasks, which are moved on an XY platform for horizontal travel along the shelf. The storage shelf has a capacity of 50 products organized in rows and columns of 5 x 10, where each cell contains one type of product, which in this application corresponds to containers of cleaning liquids.

A workflow is established for the automated packaging process through a system initialization phase, where the UR5 robot is positioned in front of the shelf and connects to the user interface in MATLAB. This is followed by the interaction phase, where the user selects the products to be packaged using voice commands. Phase three establishes the robot's ability to perform the task, from identifying product availability to positioning the product in the box. Once each box is full, the conveyor belt activates to begin packing the next box, continuing until the selected quantity of boxes, as defined by the user, is completed. The automated packing process concludes with the generation of a PDF report detailing the quantity and products packed (see Figure 2).

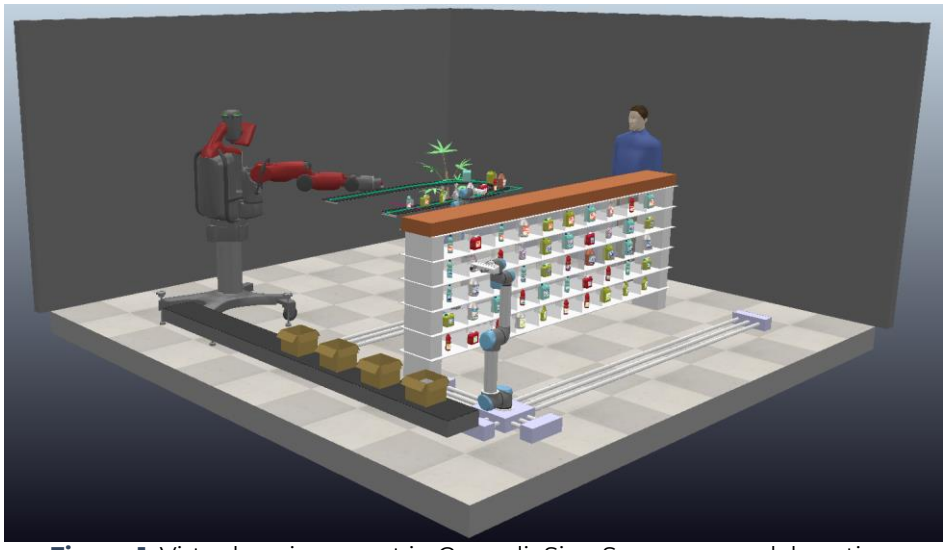

Figure 1. Virtual environment in CoppeliaSim. Source: own elaboration.

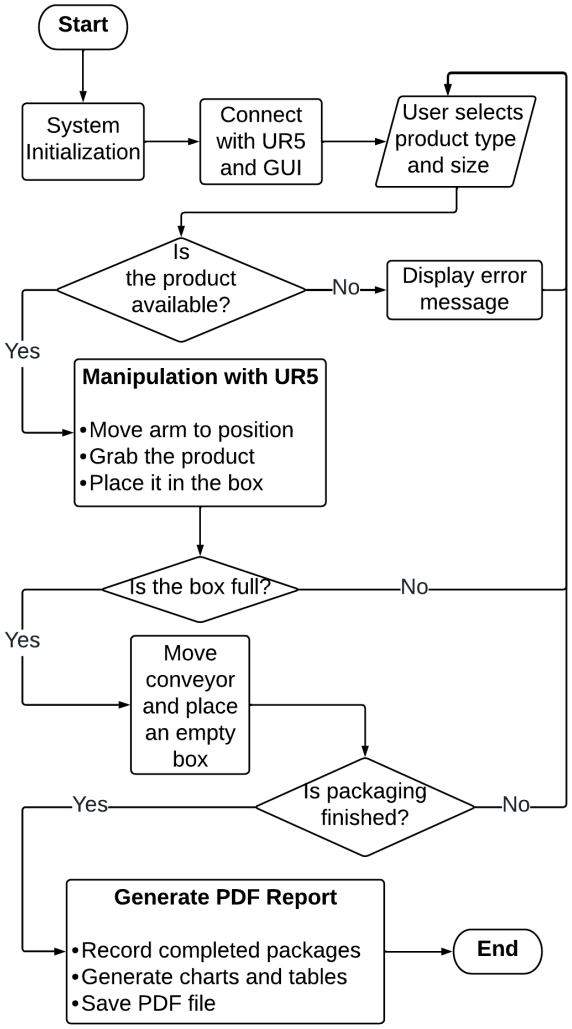

Figure 2. Flow chart of the packaging and inventory management program. Source: own elaboration.

At a technical level, the development involves the use of a video sensor for environmental acquisition, where product location is achieved using pre-trained YOLO version 8 networks, which allow for the classification and spatial location of each product. The version chosen from the available range (version 12 at the time of writing) is based on the inference and recall efficiency provided by Ultralytics, and therefore is not the subject of this study, unlike the voice conditioning described below.

For the speech recognition component facilitating user–robot interaction, a speech-to-text algorithm was employed in MATLAB. This algorithm converts audio signals into text by leveraging state-of-the-art deep learning architectures such as wav2vec 2.0 and Emformer, or alternatively cloud-based services provided by platforms such as Google, IBM, Microsoft, and Amazon.

The speech recognition process begins with the digitization of the acoustic signal at a predefined sampling frequency, followed by the extraction of an internal time–frequency representation, typically based on spectrogram or filterbank features. This representation captures the temporal evolution of spectral characteristics relevant for speech perception and serves as input to a neural inference stage. The latter maps the acoustic features to a sequence of discrete linguistic symbols, which may correspond to characters, subword units, or complete words, depending on the specific model configuration.

Modern automatic speech recognition systems address the inherent temporal misalignment between acoustic frames and linguistic units through end-to-end learning strategies that do not require explicit phoneme-level segmentation. A widely adopted approach in this context is Connectionist Temporal Classification (CTC), which enables robust transcription by marginalizing over multiple valid alignments between the input signal and the output text sequence. Architectures such as wav2vec 2.0 and Emformer have demonstrated strong robustness to speaker variability, accent differences, and background noise, making them suitable for voice-driven human–robot interaction scenarios [27], [28]. In the present work, these models are used without modification to provide reliable transcription of spoken commands that feed the chatbot decision logic described in subsequent sections.

For the second task, which involves the robot's verbal communication to the user within the simulation environment, a text-to-speech (TTS) synthesis module was implemented using the Microsoft Speech API (SAPI) via MATLAB. The purpose of this module is to convert system-generated textual messages into audio signals, enabling the delivery of status information, confirmations, and alerts in natural language. As with the speech recognition component, the internal synthesis algorithms are not altered and follow standard implementations described in the literature.

From a functional perspective, the TTS module follows a conventional voice generation pipeline that includes linguistic and phonetic analysis of the input text, assignment of prosodic attributes such as duration, pitch, and amplitude, and subsequent waveform synthesis. In practice, the SAPI engine produces a monaural pulse-code modulated (PCM) audio signal with 16-bit resolution and a sampling frequency of 16 kHz, which is used for real-time playback within the simulation environment. This approach ensures that the robot's verbal communication remains clear and intelligible to the user while allowing control over speech rate and intonation [29].

Considering all these functionalities, a state flow diagram was designed to enable user navigation and interaction within the simulation environment. This diagram is directly linked to the graphical user interface flow chart but additionally incorporates specific questions and responses presented by the chatbot to guide the task. The user makes decisions through yes/no options, product selections, or packaging confirmations, and the system executes the corresponding actions based on the interpreted instructions. To clearly represent this interaction, Figure 3 depicts chatbot prompts as rectangles, while circles indicate the expected user responses, which determine the subsequent decision according to the interpreted intent.

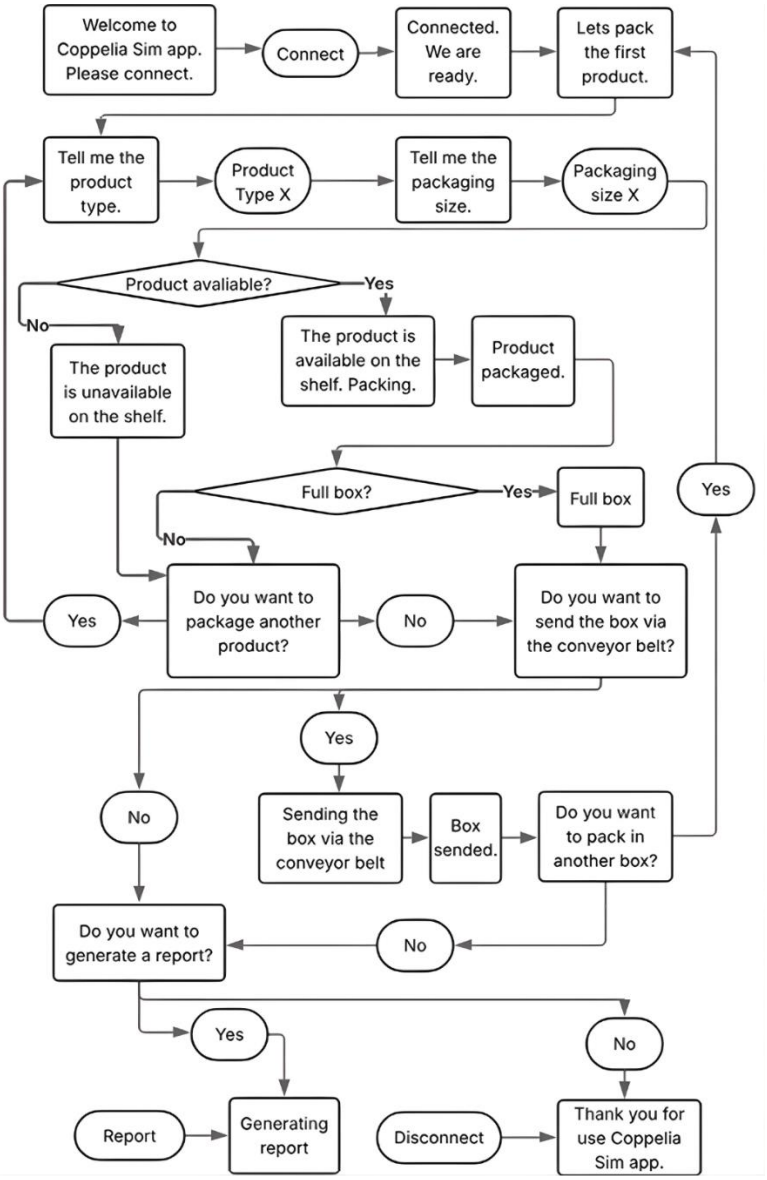

Figure 3. User–chatbot interaction flow chart for the packaging operation in the simulation environment. Source: own elaboration.

A predefined set of commands recognized by the chatbot was selected to guide the user through the interaction flow. These commands include basic terms such as connect, disconnect, yes, no, among others. In total, 20 possible responses were defined, enabling the user to navigate all stages of the dialogue and ensuring clear and effective communication. It is important to note that some commands, such as disconnect or report, have a global scope within the conversation flow, meaning they are correctly interpreted by the system regardless of the current state and trigger the corresponding action without altering the ongoing interaction state.

2.2 Chatbot architecture

A graphical user interface (GUI) was developed in MATLAB and connected to the CoppeliaSim virtual environment (Figure 4), giving the user direct control over the packaging process. At the top edges are the functional “Connect” and “Disconnect” buttons, which initiate

and terminate the connection to the simulation system, respectively. A key feature of the interface is the integration of a "Talk" voice interaction button located at the top center, which activates the user's audio capture. This module allows control of the decision-making flow through natural language commands, offering a more intuitive and flexible user experience while directly triggering actions executed through the traditional GUI. In the central left panel of this interface is the module that allows users to select the product type and size using drop-down menus. These user actions generate a visual response from the system, displaying an image of the product in the selected packaging on the right. Also centrally located are two functional buttons: "Pack" to start the packaging process, and "Next Box" to advance to the next box, ensuring process continuity. The "Message" area displays informational text about the system status, such as stock shortages or full boxes. Finally, at the bottom, there is a visual module with a real-time representation of the available inventory on the shelf. Also located here is the "Generate Report" button, which initiates the creation of process and stock reports. The interface supports process control and facilitates effective user interaction with the product handling and packaging system.

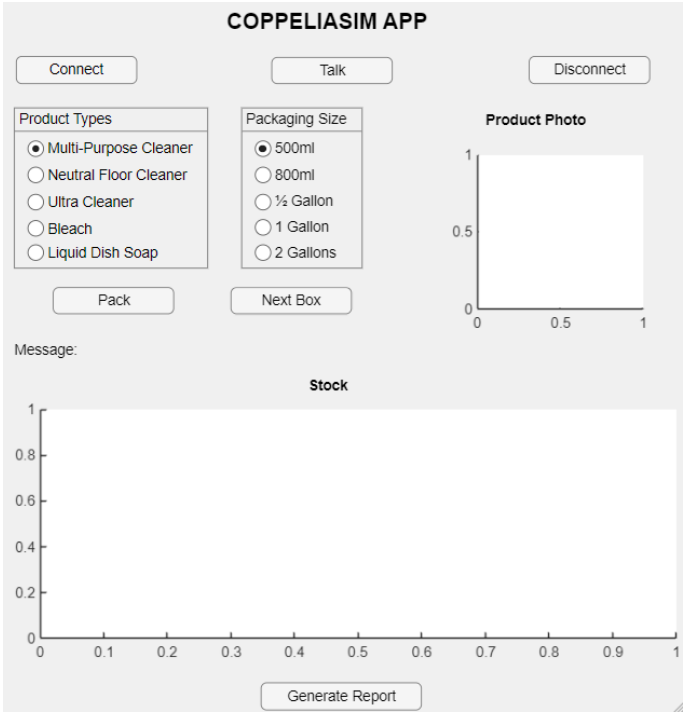

Figure 4. Graphical user interface designed in MATLAB for controlling the packaging system. Source: own elaboration.

As shown in the interface, a "Talk" button functions as a chatbot, allowing the user to navigate and interact with the system via voice commands instead of relying solely on traditional graphical buttons and options. This enables the execution of all available interface actions through natural language commands spoken by the user. To achieve this, two key functionalities were implemented: first, a speech recognition module to interpret the user's spoken commands and guide system decisions accordingly, and second, a speech synthesis module to communicate system status information, assist the user during operation, and prompt for input to continue task execution.

2.3 Experimental protocol

Given the variability in accents, intonations, noise levels, and other factors affecting voice signal perception and comprehension, it is essential to evaluate the robustness of the algorithm to ensure correct interpretation of commands from diverse users. Additionally, a strategy is sought whereby the system can adapt and continue the dialogue flow effectively even when minor deviations occur in the transcription of expected words.

At the phonetic level, certain words are particularly similar and may cause confusion during transcription, necessitating comparison criteria that quantify such similarity. To this end, the Levenshtein distance, denoted as Lev (s_1, s_2), is adopted as a standard metric to evaluate similarity between two text strings. This distance measures the minimum number of edit operations (insertions, deletions, or substitutions of characters) required to transform string s_1 into string s_2. Its formal definition, expressed in (1), ensures that for each position i in s_1 and position j in s_2, the cumulative distance accounts for all possible alignments to achieve the optimal matching between the two strings.

$$Lev(i,j) = \begin{cases} i, & j = 0 \\ j, & i = 0 \\ Lev(i-1, j-1), & if\ s_1[i] = s_2[j] \\ 1 + \ min \begin{cases} Lev(i-1,j) \\ Lev(i,j-1) \\ Lev(i-1,j-1) \end{cases} & if\ s_1[i] \neq s_2[j] \end{cases} \quad (1)$$

To achieve a clear and quantitative assessment of the accuracy of the system's command recognition, (2) normalizes the Levenshtein distance to calculate the percentage of similarity between strings s_1 and s_2. Based on this percentage, the system determines whether to continue execution when the similarity level is high or requests the user to confirm or repeat the instruction, thus ensuring a robust and reliable interaction flow.

$$Similarity(s_1, s_2) = \left(1 - \frac{Lev(s_1, s_2)}{\max(|s_1|, |s_2|)}\right) \times 100\% \quad (2)$$

To verify the chatbot's sensitivity to accurately interpret the 20 predefined commands in response to speech variations, audio recordings of 13 users with various US, British, and Australian accents were studied, with a wide age range including children and adults, resulting in variability in intonation, accent, and vocal features.

The voice spectrograms shown in Figure 5 correspond to users with different accents for the commands “Product” and “Disconnect.” These spectrograms allow visualization of differences in amplitude, frequency spectrum, and phase, as well as pauses, breaths, and other factors that affect recognition and transcription. Despite the variability, the characteristics of each command remain consistent (right and left sides of Figure 5), which allows for differentiation between the commands.

Table 1 succinctly presents the main characteristics of experimental design related to the number and accents of users, number and description of commands evaluated, command input mode, metrics, objective and evaluation environment.

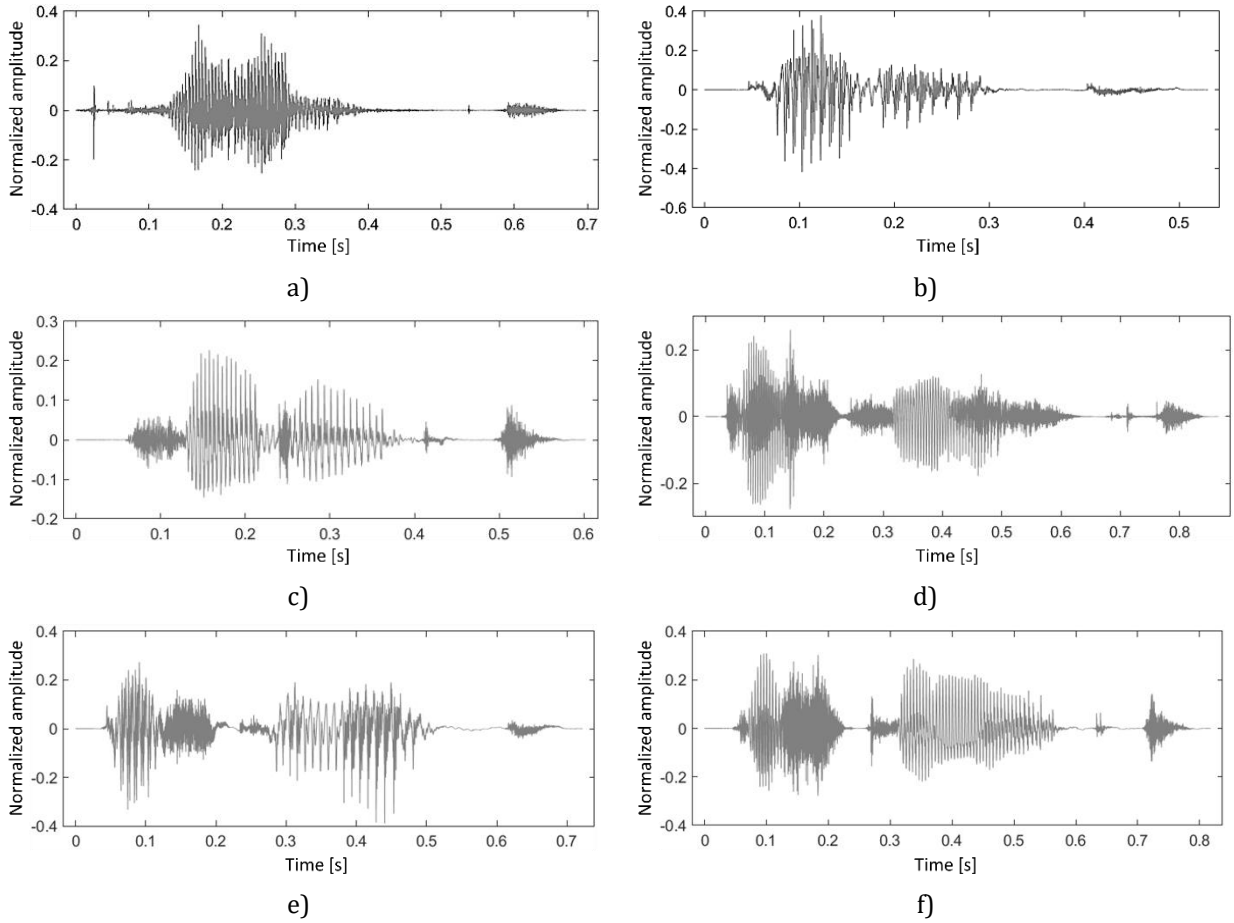

Figure 5. Voice spectrograms of users Justin -a) d)-, Russell -b), e)-; and Amy – c), f)- of the commands "Product" -a), b), c)- and "Disconnect" -d), e), f)-. Source: own elaboration.

Table 1. Experimental design Source: own elaboration.

Element	Description
Number of users	13 (from children to adults)
User characteristics	8 accents from different US states 3 British accents 2 Australian accents
Number of commands evaluated	20 predefined commands
Commands	"Half gallon", "two gallons", "500 milliliters", "800 milliliters", "bleach", "box", "connect", "conveyor belt", "disconnect", "one-gallon", "liquid dish soap", "multi-purpose cleaner", "neutral floor cleaner", "next", "no", "pack", "product", "send", "ultra cleaner", "yes".
Input mode	Voice commands
Main metric	Levenshtein distance between command and transcription
Derived metric	Percentage similarity
Evaluation objective	Evaluate the robustness of the chatbot against changes in accent, intonation, and pronunciation of verbal commands
Evaluation environment	Virtual environment developed in CoppeliaSim integrated with a system designed in MATLAB

3. RESULTS AND DISCUSSION

The voice-to-text transcription function was run to evaluate the 20 predefined words, using 13 different voice types. Some of the results obtained for these tests are presented in Table 2. As can be seen, in certain cases the transcribed word matches the expected word exactly, as is the case with Kendra's voice. However, for other people, the transcription is not always accurate, although there is still a clear similarity between the detected word and the original. This highlights the importance of applying the metric defined in (2), which allows for a quantitative evaluation of the degree of similarity between the expected word and the transcription obtained. In this way, it is possible to assess the quality of the transcription and make decisions in the interaction flow, such as continuing with the operation or requesting confirmation from the user to ensure the correct interpretation of the command.

Table 2. Comparison between expected word and transcribed word for different users. Source: own elaboration.

Words	Joey	Justin	Kendra	Kimberly
Multi purpose cleaner	“moulti purpose cleaner”	“multy purpose cleaner”	“multi purpose cleaner”	“moulty purpose gleaner”
Neutral floor cleaner	“neudral floor cleaner”	“neutral flor cleaner”	“neutral floor cleaner”	“neutril floor cleaner”
Connect	“connect”	“came out”	“connect”	“connect”
Bleach	“bleac”	“bweet”	“bleach”	“bleech”

The similarity metric was evaluated for the 20 keywords using the 13 different user voices available, considering that each expected word was known as a priori to evaluate with greater certainty the degree of match between the original word and the transcription obtained. Based on this information, it was possible to calculate the average similarity of each keyword spoken by each user with various accents and speech variations, values shown in Figure 6. These results allow us to identify the most difficult and easiest keywords for the chatbot to recognize. The word “Yes” was recognized in 100% of the tests, the word “two gallons” had an average similarity of 98%, and the word “product” had a similarity of 95%. The most difficult keywords were “ultra clean” with a similarity of 71%, and the least similar (67%) was “bleach.” The overall average keyword recognition rate was 85.7%, indicating satisfactory overall performance.

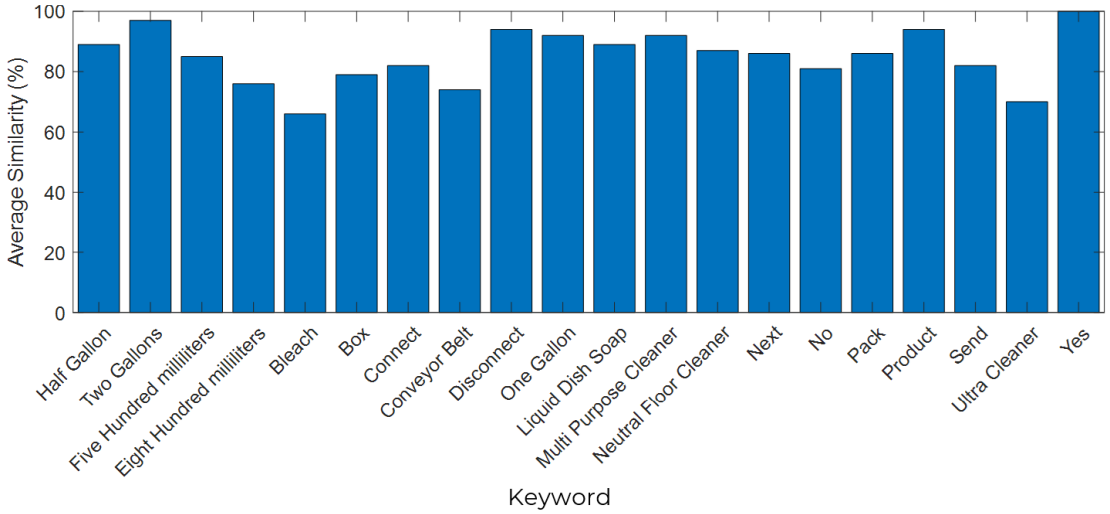

Figure 6. Average similarity achieved for each keyword analyzed in 13 users. Source: own elaboration.

It is also important to evaluate how the chatbot's performance varies for different people. To do this, we calculated the average similarity considering all the words predicted by each user, instead of averaging by word as in the previous analysis. In this way, it is possible to observe how well the system worked for everyone, whose results are reflected in Figure 7. The users with the highest similarity rates were Kendra, Joey, and Brian, reaching 95%, 94%, and 94% respectively, while the users with the lowest performance were Ivy and Justin, with 53% and 79%. It is important to note that the most unfavorable results correspond to the audio recordings of two children (one girl and one boy), which shows that age and vocal characteristics affect the quality of recognition. Likewise, although the audio recordings of men showed a slight advantage in terms of comprehension by the system, this difference was not significant. On the other hand, when considering the different accents analyzed (American, British, and Australian), no relevant differences were found in the chatbot's performance.

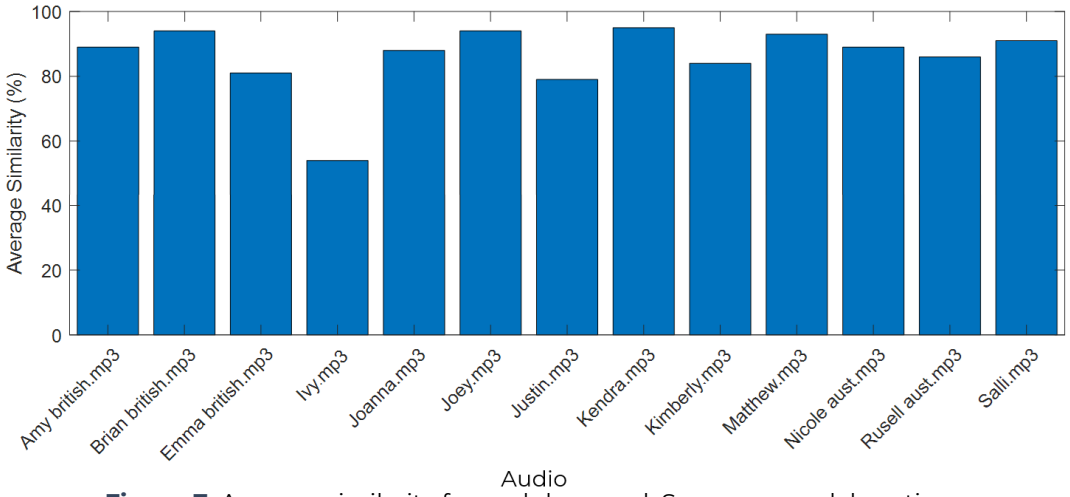

Figure 7. Average similarity for each keyword. Source: own elaboration.

Previous studies on speech recognition and transcription in human-computer interactions indicate that results are significantly affected by users' pronunciation and accent, as well as the vocabulary used. Accuracy rates between 80% and 95% are reported when verbal commands are limited, decreasing with younger users and greater variation in pronunciation. In particular, several works have reported reduced recognition accuracy for children due to higher pitch and articulation variability, while restricted command sets are known to improve robustness and decision reliability even when transcription errors occur [30], [31].

Next, the similarity metric was evaluated for the 20 words considering the 13 different types of voices, this time analyzing all possible combinations to determine whether this metric is useful for the chatbot. The purpose of this evaluation is to verify whether there is confusion between the different words used in the system and to assess whether taking the greatest similarity as a decision criterion is an appropriate strategy for continuing with the task. To do this, the average similarity between each predicted or transcribed word and the expected word was calculated for all users, the results of which are shown in Figure 8 using a heat map.

As expected, the main diagonal shows the highest similarity indices, reaching an average of 86%, which confirms the validity of the metric for our case study. On the other hand, the elements outside the diagonal reached an average similarity of only 13.3%, which is desirable to ensure that there are no ambiguous decisions for unexpected words. However, it is important to note that for some specific cases, such as half gallon, two gallons, and one gallon, the similarity is high, which could cause confusion by sharing a common word (gallon) and make it difficult to distinguish them correctly. For this reason, it is recommended to alert the user to use clear diction, especially for words that differ in critical elements, such as numbers indicating the size of the container or key commands such as connect and disconnect.

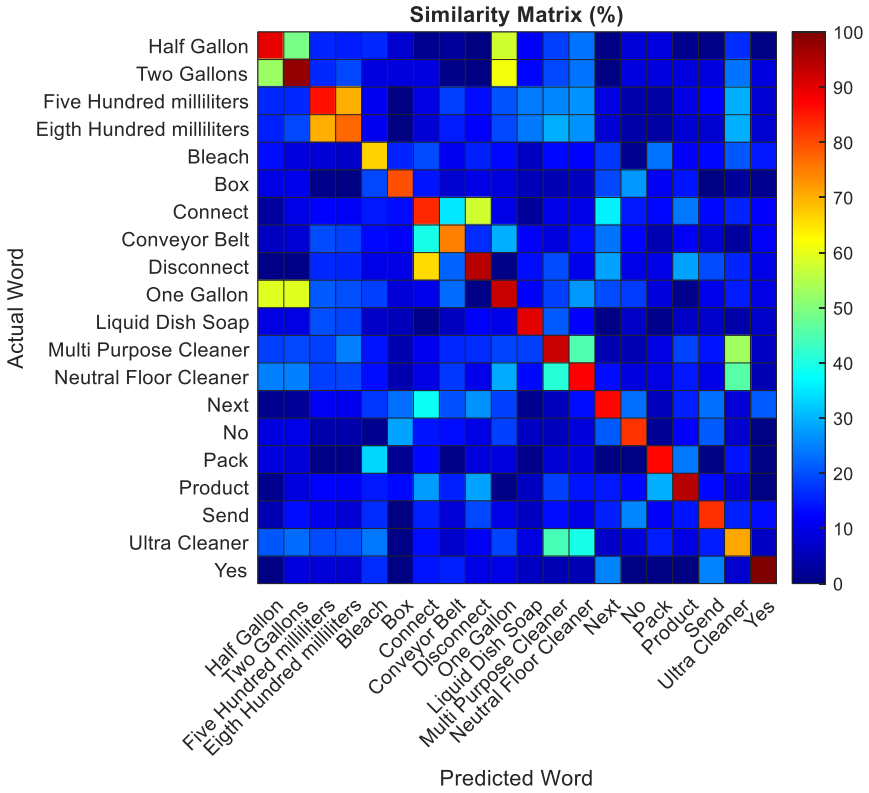

Figure 8. Heat map showing the average similarity between desired and transcribed words for all users. Source: own elaboration.

It should be noted that the reported performance is inherently linked to the use of a predefined and restricted set of voice commands. This design increases the system's robustness by reducing uncertainty in the robot's actions; however, it sacrifices flexibility in language. This rigidity works well in industrial settings with standardized processes, where there is strict control over the verbal instructions given to the robot.

In contrast to [32], where only eight words are used for robotic control, it can be observed that increasing the number of words reduces the network's accuracy level; in this case, it decreases by 10% with more than double the number of words. This was expected due to the similarity between spectrograms; however, using the similarity metric allows for obtaining a correlation factor in command identification that complements the network's response.

After implementing these elements, the chatbot is integrated into the graphical user interface, and tests are run to evaluate its functionality in a virtual simulation environment created in CoppeliaSim. Figure 9 shows the interface connected to the virtual environment. It displays a product selected for packaging and the available inventory on the shelf, confirming that the identification and packaging system functions correctly in real time, resulting in effective interaction between the user and the system, and proper product handling.

Since boxes have a limited capacity for storing products of varying sizes, it is necessary to distribute the products in a way that optimizes the use of available space. If a box does not have enough space for another product, the system generates a notification for the user to approve the use of a new box. Figure 10 shows four boxes with different orders automatically saved according to the user's instructions in the graphical interface. These results demonstrate that the system is efficient for organizing and packaging products in various boxes.

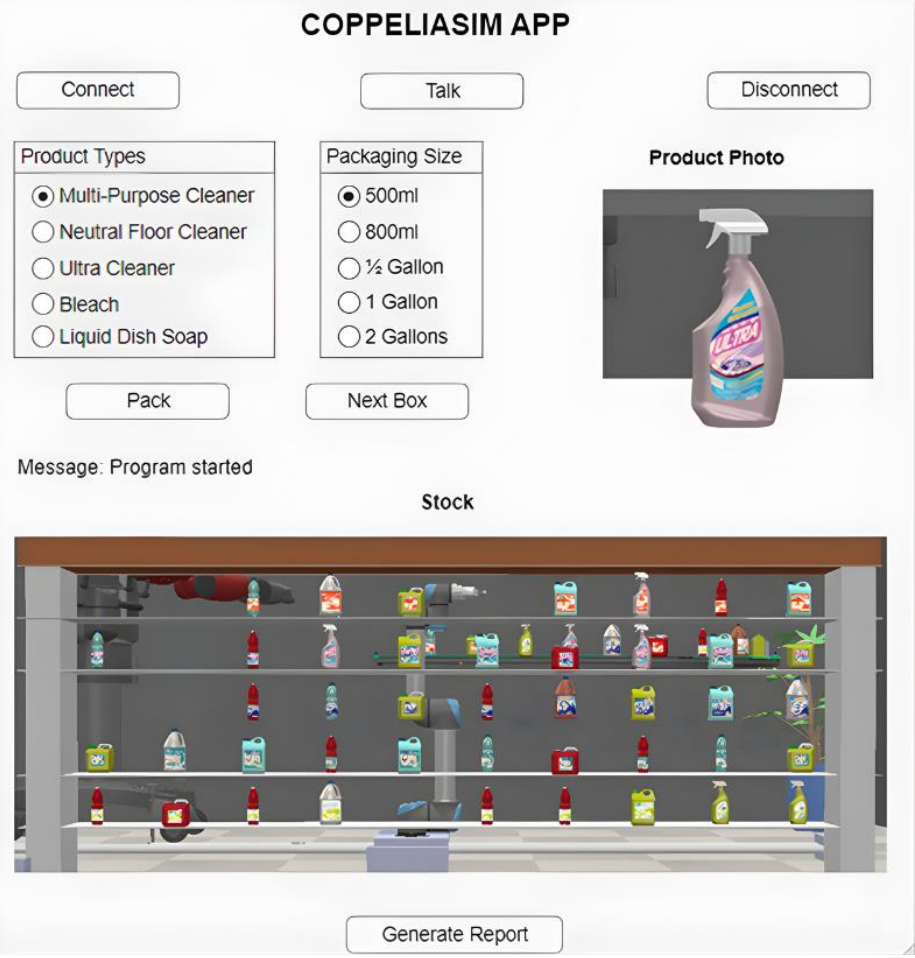

Figure 9. Simulation of the packaging and inventory distribution process in CoppeliaSim. Source: own elaboration.

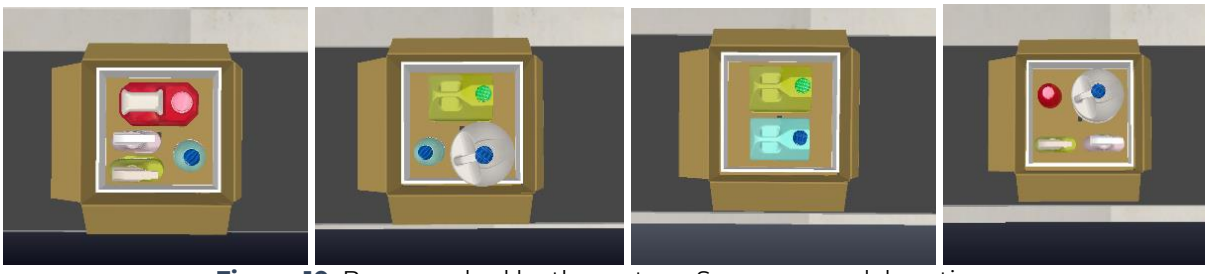

Figure 10. Boxes packed by the system. Source: own elaboration.

The results above demonstrate that the developed system achieves its objective by correctly and completely executing the order picking and packing of boxes through appropriate interaction via voice commands that were recognized with an average similarity of 85.7%, making it a valid alternative for order control in an automated industrial environment. It is acknowledged that some commands are more difficult to identify, particularly when pronounced by children; however, the system exhibits not only robust but also stable performance in the face of variations in accents and pronunciation.

Since the tests were conducted using computer simulation, which does not account for all real-world elements such as ambient noise, microphone placement, and user fatigue, and only 13 voices were used, further research is proposed to validate the system in a real-world environment with a larger number of users.

Globally, the results of the validation in a virtual environment indicate that the designed system correctly interprets the verbal commands given by the user, correctly recognizes and manipulates the products, maintains adequate inventory control, and adjusts to the storage capacity of the boxes, making it a functional and efficient system for the automatic preparation of orders

4. CONCLUSIONS

The product handling and packaging system, which operates through a graphical interface (chatbot) designed in MATLAB, allows voice interaction between the user and the robot. It effectively integrates natural language processing tools, computer vision, and robotic product handling. The system was tested in a virtual simulation environment created in CoppeliaSim, where the results confirm that the user can interact effectively with the system by voice to select, classify, and organize products.

The main contributions of this work are the integration of voice-based interaction into a complete simulated packaging and inventory workflow, the implementation of a chatbot that translates predefined spoken commands into operational decisions, and the experimental validation of the system using multiple voices, accents, and age groups. System validation included testing with multiple voices, accents, and ages, evaluating the robustness of speech recognition using objective metrics such as phonetic similarity and error rate. The results demonstrated high accuracy in language comprehension, as well as effective integration between the virtual components and the packaging control logic.

Future work involves using long language models that allow for general process interaction in the product supply line with variable commands that allow the operator to perform emergency stops, system pauses, automatic report generation, among others.

5. ACKNOWLEDGEMENTS AND FUNDING

Product derived from the research project “Fortalecimiento de procesos de recepción de pedidos y control de inventario de materias primas soportado en industria 4.0” INV-ING-4150 funded by the Vice-Rectorate for Research at the Universidad Militar Nueva Granada, year 2025. The authors thank Universidad Militar Nueva Granada where they are full-time associate professors.

6. REFERENCES

[1] X.-X. Liu, C.-Y. Yin, and M.-R. Li, "The power of voice! The impact of robot receptionists’ voice pitch and communication style on customer value cocreation intention," Int. J. Hosp. Manag., vol. 122, no. 12, p. 103819, Sep. 2024. https://doi.org/10.1016/j.ijhm.2024.103819

[2] C. Liu, L. Zhang, S. L. Xin Liu, and T. Zhu, "Speaking versus touching: How consumers respond to robot communication modality in hospitality services," Int. J. Hosp. Manag. vol. 126, p. 104017, Apr. 2025. https://doi.org/10.1016/j.ijhm.2024.104017

[3] S. Feng, N. Yamato, H. Ishiguro, M. Shiomi, and H. Sumioka, "Baby schema in human-robot physical interaction: Influence of baby likeness in a communication robot on caregiving behavior," Comput. Human Behav., vol. 4, p. 100150, May. 2025. https://doi.org/10.1016/j.chbah.2025.100150

[4] M. Li, F. Guo, X. Wang, J. Chen, and J. Ham, "Effects of robot gaze and voice human-likeness on users’ subjective perception, visual attention, and cerebral activity in voice conversations," Comput. Human Behav., vol. 141, p. 107645, Apr. 2023. https://doi.org/10.1016/j.chb.2022.107645

[5] H. Ka, "Voice-Controlled Vision-based Semi-Autonomous Assistive Robotic Manipulation Assistance," Arch. Phys. Med. Rehab., vol. 96, no. 10, pp. e10-e11, Oct. 2015. https://doi.org/10.1016/j.apmr.2015.08.028

[6] M. T. Tariq, Y. Hussain, and C. Wang, "Robust mobile robot path planning via LLM-based dynamic waypoint generation," Expert Syst. Appl. vol. 282, p. 127600, Jul. 2025. https://doi.org/10.1016/j.eswa.2025.127600
[7] M. Bakouri, "Development of Voice Control Algorithm for Robotic Wheelchair Using MIN and LSTM Models," Comput. Mat. Contin., vol. 73, no. 2, pp. 2441-2456, Jun. 2022. https://doi.org/10.32604/cmc.2022.025106
[8] M. Meghana et al., "Hand gesture recognition and voice controlled robot," Mater. Today Proc., vol. 33, no. 7, pp. 4121-4123, Jun. 2020. https://doi.org/10.1016/j.matpr.2020.06.553
[9] M. H. Haider et al., "Robust mobile robot navigation in cluttered environments based on hybrid adaptive neuro-fuzzy inference and sensor fusion," J. King Saud Univ. Comput. Inf. Sci., vol. 34, no. 10, Part B, pp. 9060-9070, Nov. 2022. https://doi.org/10.1016/j.jksuci.2022.08.031
[10] S. Jamshidi, A. Nikanjam, K. W. Nafi, F. Khomh, and R. Rasta, "Application of deep reinforcement learning for intrusion detection in Internet of Things: A systematic review," Internet of Things, vol. 31, no. 8, p. 101531, May. 2025. https://doi.org/10.1016/j.iot.2025.101531
[11] B. Bogaerts, S. Sels, S. Vanlanduit, and R. Penne, "Connecting the CoppeliaSim robotics simulator to virtual reality," SoftwareX, vol. 11, p. 100426, Jan. 2020. https://doi.org/10.1016/j.softx.2020.100426
[12] A. Farley, J. Wang, and J. A. Marshall, "How to pick a mobile robot simulator: A quantitative comparison of CoppeliaSim, Gazebo, MORSE and Webots with a focus on accuracy of motion," Simul. Model. Pract. Theor., vol. 120, p. 102629, Nov. 2022. https://doi.org/10.1016/j.simpat.2022.102629
[13] R. Rajendran, and J. Arockia Dhanraj, "Measurement of payload under variable adhesive force through coppeliasim and validating the design of a wall climbing robot," Measur. Sensors, vol. 29, p. 100872, Oct. 2023. https://doi.org/10.1016/j.measen.2023.100872
[14] R. Duarte Lamperti, and L. V. Ramos de Arruda, "A strategy based on Wave Swarm for the formation task inspired by the Traveling Salesman Problem," Eng. Appl. Artif. Intell., vol. 126, no. Part B, p. 106884, Nov. 2023. https://doi.org/10.1016/j.engappai.2023.106884
[15] M. Rinaldi, V. Di Pasquale, P. Farina, R. Iannone, R. Macchiaroli, and E. H. Grosse, "Human–robot interaction in industry: a tertiary study," Procedia Comput. Sci., vol. 253, no. 22, pp. 1691-1701, 2025. https://doi.org/10.1016/j.procs.2025.01.231
[16] J. Liu, H. Luo, and D. Wu, "Human–Robot collaboration in construction: Robot design, perception and Interaction, and task allocation and execution," Adv. Engin. Inform., vol. 65, no. Part A, p. 103109, May. 2025. https://doi.org/10.1016/j.aei.2025.103109
[17] S. Guan, J. Wang, X. Wang, C. Ding, H. Liang, and Q. Wei, "Dynamic gesture recognition during human–robot interaction in autonomous earthmoving machinery used for construction," Adv. Engin. Inform., vol. 65, no. Part C, p. 103315, May. 2025. https://doi.org/10.1016/j.aei.2025.103315
[18] W. T. Lima Junior, R. A. Welter, W. Pacheco Ferreira, R. Ferreira Souza, and T. Eduardo, "Human robot interaction (HRI): An artificial cognitive autonomy approach to enhance Decision-Making," Cogn. Syst. Res., vol. 91, p. 101336, Jun. 2025. https://doi.org/10.1016/j.cogsys.2025.101336
[19] B. Feng, Z. Wang, L. Yuan, Q. Zhou, Y. Chen, and Y. Bi, "Towards safe motion planning for industrial human-robot interaction: A co-evolution approach based on human digital twin and mixed reality," Robot. Comput-Int. Manuf., vol. 95, p. 103012, Oct. 2025. https://doi.org/10.1016/j.rcim.2025.103012
[20] D. Udekwe, and H. Seyyedhasani, "Virtual Reality-Enabled remote Human-Robot interaction for strawberry cultivation in greenhouses," Comput. Electron. Agric., vol. 237, no. Part A, p. 110567, Oct. 2025. https://doi.org/10.1016/j.compag.2025.110567
[21] A. Apraiz et al., "The user experience in industrial human-robot interaction: A comparative analysis of Unimodal and Multimodal interfaces for disassembly tasks," Robot. Comput.-Int. Manuf., vol. 95, p. 103045, Oct. 2025. https://doi.org/10.1016/j.rcim.2025.103045
[22] A. Bonello, E. Francalanza, C. A. Brown, and P. Refalo, "Cognitive Design of Collaborative Human-Robot workstations in Industry 5.0: A Kansei Engineering approach to quantifying emotions in problem decomposition," Procedia Comput. Sci., vol. 253, pp. 2811-2820, Jan. 2025. https://doi.org/10.1016/j.procs.2025.02.005
[23] P. Novák, J. Vyskočil, J. Kubalík, P. Kadera, M. Jílek, and V. Jirkovský, "Smart Counting Machines for Modular Industry 4.0 Packing Lines," IFAC-PapersOnLine, vol. 56, no. 2, pp. 2976-2981, 2023. https://doi.org/10.1016/j.ifacol.2023.10.1422
[24] C. Liu, J. Lyu, and K. Fang, "Integrated packing and routing: A model and its solutions," Comput. Oper. Res., vol. 172, p. 106790, Dec. 2024. https://doi.org/10.1016/j.cor.2024.106790
[25] X. Liu et al., "Design, integration, and evaluation of a low-cost system for automatic apple picking and infield sorting," Comput. Electron. Agric., vol. 239, no. Part A, pp. 110933, Dec. 2025. https://doi.org/10.1016/j.compag.2025.110933

[26] Y. Zhang, L. Chen, X. Li, Q. Li, and J. Li, "Multi-arm robotic system and strategy for the automatic packaging of apples," Artif. Intell. Agric., vol. 16, no. 1, pp. 578-591, Mar. 2026. https://doi.org/10.1016/j.aiia.2025.11.006

[27] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Adv. Neural Inf. Proces. Syst., vol. 33, pp. 12449-12460, Oct. 2020. https://doi.org/10.48550/arXiv.2006.11477

[28] Y. Shi et al., "Emformer: Efficient Memory Transformer for Streaming Speech Recognition," in IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2021, Toronto, ON, Canada, 2021, p. 6783-6787. https://doi.org/10.1109/ICASSP39728.2021.9414560

[29] P. Taylor, Text-to-Speech Synthesis, Cambridge, U.K.: Cambridge University Press, 2012. https://books.google.com.co/books/about/Text_to_Speech_Synthesis.html?id=T0O-NHZx7kIC&redir_esc=y#:~:text=Text%2Dto%2DSpeech%20Synthesis%20provides,assumes%20no%20specialised%20prior%20knowledge

[30] P. G. Shivakumar, and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," Computer speech & language, vol. 63, p. 101077, Sep. 2020. https://doi.org/10.1016/j.csl.2020.101077

[31] M. Malik, M. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," Multimedia Tools and Applications, vol. 80, no. 6, p. 9411–9457, Mar. 2021. https://doi.org/10.1007/s11042-020-10073-7

[32] R. Jiménez-Moreno, and R. A. Castillo. "Deep learning speech recognition for residential assistant robot," IAES Int. J. Artif. Intell., vol. 12, no. 2, pp. 585-592, Jun. 2023. https://doi.org/10.11591/ijai.v12.i2.pp585-592

CONFLICT OF INTEREST

All authors declare that there are no conflicts of interest.

AUTHORSHIP CONTRIBUTION

Juan Camilo Guachetá-Alba: Software, Validation, Investigation, and Data Curation,

Robinson Jiménez-Moreno: Idea, Conceptualization, Methodology, Formal analysis, and Writing (Original Draft).

Anny Astrid Espitia-Cubillos: Idea, Conceptualization, Methodology, Writing (Review & Editing), Resources, Supervision, and Project administration.