POLITECNICO DI TORINO
Repository ISTITUZIONALE

Marvin: An Innovative Omni-Directional Robotic Assistant for Domestic Environments

Original
Marvin: An Innovative Omni-Directional Robotic Assistant for Domestic Environments / Eirale, A., Martini, M., Tagliavini, L., Gandini, D., Chiaberge, M., Quaglia, G.. - In: SENSORS. - ISSN 1424-8220. - ELETTRONICO. - 22:14(2022), pp. 5261-5282. [10.3390/s22145261]

Availability:
This version is available at: 11583/2970143 since: 2022-07-15T17:27:25Z

Publisher:
MDPI

Published
DOI:10.3390/s22145261

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

27 November 2026

sensors

MDPI

Article

Marvin: An Innovative Omni-Directional Robotic Assistant for Domestic Environments

Andrea Eirale [1], Mauro Martini [1], Luigi Tagliavini [2], Dario Gandini [1], Marcello Chiaberge [1,*] and Giuseppe Quaglia [2]

[1] Department of Electronics and Telecommunications (DET), Politecnico di Torino, 10129 Torino, Italy; andrea.eirale@polito.it (A.E.); mauro.martini@polito.it (M.M.); dario.gandini@polito.it (D.G.)
[2] Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino, 10129 Torino, Italy; luigi.tagliavini@polito.it (L.T.); giuseppe.quaglia@polito.it (G.Q.)
* Correspondence: marcello.chiaberge@polito.it; Tel.: +39-331-671-4686

Abstract: Population aging and pandemics have been shown to cause the isolation of elderly people in their houses, generating the need for a reliable assistive figure. Robotic assistants are the new frontier of innovation for domestic welfare, and elderly monitoring is one of the services a robot can handle for collective well-being. Despite these emerging needs, in the actual landscape of robotic assistants, there are no platforms that successfully combine reliable mobility in cluttered domestic spaces with lightweight and offline Artificial Intelligence (AI) solutions for perception and interaction. In this work, we present Marvin, a novel assistive robotic platform we developed with a modular layer-based architecture, merging a flexible mechanical design with cutting-edge AI for perception and vocal control. We focus the design of Marvin on three target service functions: monitoring of elderly and reduced-mobility subjects, remote presence and connectivity, and night assistance. Compared to previous works, we propose a tiny omnidirectional platform, which enables agile mobility and effective obstacle avoidance. Moreover, we design a controllable positioning device, which easily allows the user to access the interface for connectivity and extends the visual range of the camera sensor. Nonetheless, we delicately consider the privacy issues arising from private data collection on cloud services, a critical aspect of commercial AI-based assistants. To this end, we demonstrate how lightweight deep learning solutions for visual perception and vocal command can be adopted, completely running offline on the embedded hardware of the robot.

Keywords: mobile robotics; assistive indoor robotics; modularity; Artificial Intelligence; vocal assistant; system design

check for updates

Citation: Eirale, A.; Martini, M.; Tagliavini, L.; Gandini, D.; Chiaberge, M.; Quaglia, G. Marvin: An Innovative Omni-Directional Robotic Assistant for Domestic Environments. *Sensors* **2022**, *22*, 5261. https://doi.org/10.3390/s22145261

Academic Editor: Ahmad Rad

Received: 20 June 2022
Accepted: 12 July 2022
Published: 14 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, there has been a significant demographic shift in the global population and, in particular, population aging and its consequences on society need to be seriously taken into account. Indeed, according to the World Population Prospects provided by the United Nations in 2019 [1], life expectancy reached 72.6 years in 2019 and it is forecast to grow to 77.1 years by 2050. In 2018, the amount of persons with an age of 65 or higher reached for the first time the number of children under 5 years. In addition, the United Nations also declared that by 2050 the number of persons aged 65 years or over would overcome the number of youth aged 15 to 24 years [1]. These projections suggest that population aging may drastically affect the entire society, causing social issues in the organization and cost-management of healthcare systems and family units. Moreover, emergency situations such as the COVID-19 pandemic raises critical issues in monitoring isolated people in their houses, which normally need dedicated assistive operators. Socially assistive robots (SAR) have recently emerged as a possible solution for elderly care and monitoring in the domestic environment [2]. Although the specific role and objectives of a robotic assistant for elderly care need to be concurrently discussed from an ethical perspective, according to

Abdi et al. [3] diverse robotic platforms for social assistance already exist. These studies often brought researchers to limit their study to the human-machine interaction, realizing companion robots with humanoid [4] or pets-like architectures [5,6]. Such robots have been particularly studied for what concerns dementia, aging, and loneliness problems [7,8]. Different studies specifically focus on detailed monitoring tasks, for example, heat strokes [9] and fall detection [10].

Besides the healthcare and elderly monitoring purposes, the potential scope of application of an indoor robot assistant is wide, with the enhancement of domestic welfare as a general goal. Indeed, the awareness of air quality risks is rapidly increasing with the spreading of COVID-19 [11]. Moreover, following the Internet of Things (IoT), the paradigm of the house as we know it is changing with the introduction of multiple connected devices [12]. According to this, recent studies reveal that robots can be identified as complete solutions for future house management [13]. Robotic assistants represent a promising solution also for monitoring and surveillance in diverse environments such as offices and industrial facilities, with the role of constantly supporting workers while checking potential illness conditions, or accomplishing simple service tasks.

However, the success of the service assistant robot has not seen a real peak yet: the adoption of these technologies is still limited by the high research focus on technology in the existing prototypes, while a user-centric perspective should guide the design phase. In Section 2, we discuss the state of the art in assistance robotics and we frame our solution in this context, highlighting its advantages. We believe that the selection of specific target tasks for the robot in a domestic scene is the first necessary step to move toward the spread of robotics assistants as a concrete demonstration of the helpful role of the robot can strongly encourage the user to its adoption. Secondary, we identify a suitable architectural design, flexible and appropriate to the environmental constraints, as the other key factor for a successful final prototype.

Contributions

In this paper, we propose a novel robotic assistive platform: Marvin. The goal of our mobile robot is to provide basic domestic assistance to the user. More in detail, we identify a set of service functions for the Marvin robot within the overall research scope of socially assistive robots: *user monitoring*, *night assistance*, *remote presence, and connectivity*. In the following sections, we present the layered modular design adopted to conceive Marvin, resulting in a system indifferent to small modifications of the domestic environment and features required by the specific application. Differently from previously presented robots for home assistance, discussed in Section 2.1, we chose a tiny omnidirectional base platform [14]. Indeed, Marvin exploits its restricted footprint and four mecanum wheels to autonomously navigate in a cluttered indoor environment such as the domestic one. Omniwheels and mecanum wheels have already been studied in many prototypes [15,16] and they are particularly used in industrial robotics applications [17], where the flexibility and the optimization of trajectories are a priority. Omnidirectional motion offers a competitive advantage compared to the most commonly employed differential-drive system in unstructured environment navigation. In particular, omnidirectional mobility can be exploited to monitor the user while navigating and avoiding obstacles efficiently. The geometrical asymmetry in the platform's footprint, in conjunction with the omnidirectional capability, can also be exploited to navigate in confined spaces, which are very common in domestic environments. Although the study of human–robot interactions does not fall within the specific scope of this project, we design a telescopic positioning device to adjust the height and tilt of Marvin's camera and its potential user interface. This effectively improves its usability for surveillance purposes offering an extended visual range and facilitating access to its screen for a user-centered visual interaction or telepresence. A custom solution for a positioning system is developed because no commercial device, like a lightweight robotic arm, deals with the strict weight and size requirements dictated by the application. A limitation of this first prototype is that it does not provide manipulation

capability, which will be the objective of future studies. In addition to an accurate study of the base platform, the Marvin design has been merged with the adoption of state-of-the-art computer vision and AI methods for perception, person tracking, pose classification, and vocal assistance. Deep Learning lightweight models have been selected from recent literature and optimized for real-time inference with the computational embedded hardware mounted on the robot. Marvin presents an AI-based vocal assistant named PIC4Speech for controlling its actions and selecting the desired task. Differently from commercial solutions such as Google Home or Alexa, the PIC4Speech system described in Section 7 completely runs offline on the onboard computational device of the robot, avoiding privacy risks and issues of an online cloud-based solution.

Overall, Marvin is a novel robotic solution for domestic and, more generally, indoor user assistance. We distinguish our design choices from existing solutions particularly focusing on service functions in which the mobility constraints dictated by realistic cluttered home environments strongly emerge. To this end, a human-comparable footprint and flexible motion planning, combined with effective visual perception and vocal control systems, can drastically increase the adoption of robotic solutions for home assistance. Therefore, the contributions of this work to robotic assistant research are manifold, and can be summarized as follows:

- we conceive a novel modular solution for user monitoring, night assistance, remote presence, and connectivity, prioritizing the agility and flexibility of the platform in complex domestic environments, by adopting a tiny human-comparable footprint and an omnidirectional base platform for the robot (Sections 4 and 5);
- we design a controllable telescopic positioning device (Section 5) for easy access to the visual interface and to extend the visual range of the robot;
- we develop a real-time AI-based vision system (Section 6.3) to constantly check potential critical conditions of the user based on their pose, and automatically set up an emergency call;
- we propose the PIC4Speech vocal control system (Section 7) to provide a reliable, offline vocal interface for the user to express commands to the robot easily.

2. Related Works

In this Section, we present an overview of the state of the art in assistive service robotics, comparing the most popular architectural solutions proposed in the literature so far, and discussing their points of strength and weakness that led us to design our platform and its sub-components. In the Section 2.1, we firstly present the main assistive robotics platforms presented in the literature, highlighting their peculiar design characteristics. Then, in Sections 2.2 and 2.3, we briefly introduce the principal technologies that have been used to develop functionalities of Marvin.

2.1. Assistive Service Robots

In the last years, the robotics research community is focusing its effort on the study of an effective design for an indoor assistant, and different proposals have recently emerged. Diverse researchers based their study on the human-machine interaction, realizing humanoid companion robots such as NAO [4] or pets-like architectures such as Aibo [5] and PARO [6]. These kinds of robots have been particularly studied for research on dementia, aging, and loneliness problems [7,8], although their usage cannot be extended to home assistance without a mobile platform. Different studies specifically focus on detailed monitoring tasks, for example, heat strokes [9] and fall detection [10]. However, although their usual expensive cost, they often result to be unused for a long time horizon due to the complex healthcare task they try to accomplish. Indeed, for the pure purpose of a companion robot, marginal differences exist with the more competitive commercial vocal assistants like Alexa, with a much lower cost. Jibo [18] (Figure 1a) is another example of a social robot for the home which falls in this category. Hence, an assistive domestic robot should go beyond

the conversational skills of common vocal assistants and we decided to choose a mobile platform, trying to identify a clear, helpful role for the robot in a domestic scenario.



Figure 1. Commercial robots developed or suitable for service home-care applications.

There are already a good amount of research prototypes and few commercial mobile platforms for home-care robotics today. Among them, the HOBBIT robot [19] and the Toyota Home Service Robot (HSR) [20] are the results of different research projects and they present a similar architecture composed of a wheeled main body equipped with manipulators for grasping objects. The Pepper robot (Figure 1b) is one of the most popular humanoid robots and it has been also used for nursing and rehabilitative care of the elderly [21]. TIAGo [22] (Figure 1c)is another comparable platform developed for robotics research groups and in general for indoor applications. Even though they are standard differential drive wheeled platforms, all these robots aim at reproducing a human-like overall shape and presence. However, a large footprint and a standard steering system represent strong disadvantages to navigating in a realistic cluttered household environment. The same limitations hold for the SMOOTH robot [23], the resulting prototype of a research study that aims at developing a modular assistant robot for healthcare with a participatory design process. Three use cases for the SMOOTH robot have been identified: laundry and garbage handling, water delivery, and guidance. In agreement with the authors of the SMOOTH robot, we decided to avoid a robotic arm on the robot, due to the higher control complexity it requires and stability issues it causes when mounted on a tiny lightweight mobile platform. Instead, we designed an innovative controllable positioning device to lift the camera point of view to a reasonable adjustable height, and to allow the user to access the robot visual interface easily. The abilities to carry objects without a manipulator and offer physical support to elderly people are the advantages of the SMOOTH robot design. However, we consider the tiny size and the omnidirectional motion of the platform the crucial design choices to enable the introduction of robot assistants in real-world domestic environments on a large scale. We discussed that the current state of the art in mobile assistive robots suggests a wide variety of potential configurations for the platform. Compared to previous works, we take into account the following key considerations to enhance the success of robots for domestic applications:

- a **tiny size** fits better in a cluttered domestic environment and improves the acceptance of the presence of the robot in the house;
- an **agile and flexible mobility** guarantees better performance in navigation tasks with complex obstacles and narrow passages;
- a **positioning device** for camera and tablet is preferred to a robotic arm for this prototype.

Recently, our design considerations have been strictly confirmed by the lastly emerging commercial proposals. Indeed, in addition to research projects, Amazon has recently launched its commercial home assistant Astro [24] (Figure 1d). Even though it is still

in an experimental stage, Astro can surely be considered an enhanced design thought for end-users, which can visually recognize people and interact with them through a visual interface that aims at conveying expressive reactions and thanks to the Alexa vocal assistant. Robotic platforms such as Astro aim at totally managing the house, also providing surveillance and telepresence services. We can notice that Astro presents a reduced size compared to the typical humanoid platforms to guarantee agile movements in the house and does not represent an oppressive figure for the users, at the cost of not being able to carry items. Moreover, it does not present robotic arms for manipulating objects. With our omnidirectional platform, we aim to improve the mobility and obstacle avoidance of a platform like Astro, together with offering an extended visual range and an easy access to the visual interface thanks to our positioning device. Moreover, Amazon designed the robot to integrate it with the home automation system, exploiting Alexa as a vocal interface. However, this choice exposes Astro to high privacy risks and issues, handling both vocal and visual data of domestic private environments. To prevent such risks, we consider the idea of developing a basic **offline vocal assistant**, as discussed more in detail in Sections 2.3 and 7.

2.2. Visual Perception

Visual perception in robotics plays a major role, enabling a thorough and detailed scene understanding with a wide variety of visual tasks. Pose estimation [25,26], object detection [27], and semantic segmentation are the main visual processes that allow the robot to perceive and interpret what surrounds it. The drastic increase of robotic devices both in manufacturing and management facilities and, more recently, in populated environments such as offices, hospitals, and houses, is strictly tied to the breakthrough of Artificial Intelligence (AI) and Deep Neural Networks (DNN) for computer vision applications. From the publication of the ImageNet dataset [28], the escalation of Deep Learning from AlexNet to the most recent DNN architectures [29] is still in progress worldwide. In particular, the real-time detection of humans is a growing computer vision process that provides support in the application field of surveillance and monitoring, and most critically, allows robots to be placed in populated environments. In particular, person detection is the pillar of every visual-based human–robot interaction. An effective perception system of humans is, therefore, a necessary condition to let robots safely plan their activity. According to this, the robotics research community is focusing its effort on person-aware autonomous navigation algorithms [30], and the detection of human figures in the visual stream is the first step towards this goal. Classic deep learning-based one-stage object detectors such as YOLO [31] and SSD [32] networks provide the robot with information about the presence of an object in its field of view. However, these approaches only give such information in the form of a bounding box, namely a region of interest of the image where the object is detected. For robotic tasks, especially for human-aware applications, it is crucial to have some additional knowledge about the person. For this reason, pose estimation models represent a more suitable and powerful choice, and their usage is twofold: to detect the presence of humans and provide information about their pose status. State-of-the-art models for human pose estimation [26,33] provide a skeleton schematic graph of the person. We use PoseNet to estimate human poses as we need a fast inference system and fine-grained information about the human pose. We build our complete pose classification pipeline training a simple DNN, which receives as input key points estimated by PoseNet. Our visual perception pipeline combines both RGB and depth images to autonomously detect emergency conditions of the user, constantly checking if they are standing, sitting, or laying, and to track the person's relative position to enable the human-centered navigation of the robot.

2.3. Vocal Interface

The study of Human–Robot Interface (HRI) has become today a fundamental component for the spreading success of robots in society, allowing the end-users to interact with

the robot in different ways. Visual and vocal interfaces are the most common choices to let humans easily interact with a robotic machine. Joysticks and touch screens are other solutions more diffused in the research world during the development phase. In this work, we focused on a vocal interface, to allow the user to call and interact with the robot also from a suitable distance, without the need to access its screen, and to facilitate access to a high-tech device for the elderly. In the last decade, speech processing has seen huge steps forward [34] according to the progress of robots and AI. However, training Deep Learning models for vocal assistants requires an extremely high amount of data [35], that only giants of the market such as Google and Amazon can collect and exploit easily. Moreover, state-of-the-art models in Natural Language Processing (NLP) [36] provide great performances at the cost of a much higher computational cost, which forbids their usage on embedded devices with constrained hardware resources. Commercial solutions such as Siri, Alexa, or Google Home exploit a cascade activation pipeline of multiple models that transfer the computation from the physical device, when triggered, to the cloud servers to run their NLP algorithms. Indeed, the development of a full pipeline of algorithms for fast-interference low-cost vocal assistance in robots is rare to be found in the research literature, although it is a fundamental aspect of human–robot interaction. According to this, we decided to kick off a research project, the PIC4Speech vocal assistant, with the aim of providing a low-cost, efficient solution to be executed on board the robotic platforms without the need for expensive hardware and, above all, without relying on a stable internet connection. This last consideration is born from the objective to avoid the exposure of private data to the internet, protecting the development of the robot assistant Marvin from privacy issues, which have been faced by other previous prototypes such as the Astro robot. The complete description of PIC4Speech architecture is reported in Section 7. Besides the commercial online assistants, no other similar systems have been identified for a direct, thorough comparison. PIC4Speech is intended to be an offline vocal assistant for human–robot interaction, even though in this primitive shape of development, its goal is principally to allow the user to give commands to the robot vocally and not to carry out a meaningful social conversation. Indeed, Marvin is considered a robot assistant more than a companion robot and social aspects of communication are not treated in this study.

3. Requirements

Against the described state-of-the-art scenario, the researchers at Pic4SeR Center (Interdepartmental Centre for Service Robotics) of the Politecnico di Torino, in association with the researchers at Officine Edison, developed a personal assistant mobile robot called Marvin. The robot has been conceived as a proof of concept to explore the possibilities of autonomous assistive robots in domestic environments, designed for people owning reduced motility, like elderly or people with disabilities. To such an aim, the robot must be able to perform the following service functions:

1. *User Monitoring*: the robot should be able to detect a potentially dangerous situation for the user and call for help. To detect a sudden illness or a drift in the person's behavior, many approaches are possible. For example, it is possible to structure the environment with proper sensors, provide the user himself with wearable devices, or exploit visual and perception sensors directly mounted upon the robotic assistant. The first two solutions require intervention in the users' houses, or the users themselves, which can be perceived as invasive measures in the living environment. Therefore, the monitoring of the person by the robot turns out to be more feasible and immediate. This task requires the robot to track the user as they move within the environment continuously. From a mobility point of view, this also implies the ability to move and reorient the robot sensors at any instant
2. *Night Assistant*: one of the most critical moments in the daily life of elders is the night-time bedroom-to-toilet journey. The robot should cater to assist in all those situations in which, for whatever reason, the user was unable to reach the electric lighting. The Night Assistant service proposes to accompany the user in any desired

location of the domestic environment, enlightening the path and monitoring their movements, giving alarms in the case of need. Again, such a feature requires as-high-as-possible maneuverability to contemporaries providing light and not hampering the user path

3. *Remote presence and connectivity*: the robot must be provided with the ability to access commonly used communication platforms (e.g., Skype, Whatsapp). This task implies the robot should be able to approach the communication interface for the user to answer or perform calls/video calls. Such ability infers with usability requirements: the human–robot interface needs to be positioned and oriented towards the user

These tasks, addressed to provide a service to the user, in turn require a series of robotic capabilities. First of all, to properly monitor the user, the system should be able to perceive them, recognize a laying posture and ask for help if necessary. This comports the implementation of pose detection and classification neural networks and external communication via a mobile connection. The monitoring service would become pretty limited without a continuous track of the user within the different rooms of the domestic environment. To constantly control the user's condition, it is fundamental the implementation of an autonomous *Person Following* functionality, exploiting the information retrieved from the perception system. On the other hand, to efficiently accompany the user, the rover must be able to save a series of points of interest, like rooms or specific locations, towards which it can autonomously navigate. Finally, to facilitate the approach of the user with the remote presence service, a positioning device must be properly designed and deployed on the robotic platform. To this end, it is first necessary to evaluate the workspace requirements. As the application suggests, the tip of the mechanism should reach above common furniture to bring the user interface in a comfortable position. To perform this action, the robot can approach the furniture parking as close as possible to the goal position (Figure 2b,c) or it can go under the furniture if the cabinetry geometry allows it (Figure 2a). Moreover, the device can be mounted near the closer or on the opposite side of the approached entity. Regarding Figure 2, the mounting configuration (Figure 2b) guarantees a better redistribution of the masses to keep the center of mass inside the footprint of the robot, but the device has limited capability at reaching distant points in the longitudinal direction, while the mounting configuration (Figure 2c) enables the device to further reach out in that direction, but it moves the center of gravity away from the center of the platform. The workspace related requirements for the positioning device have been chosen considering the following situations:

- Dinner table: under motion is possible, no longitudinal displacement from the platform border is required, working height approximately 90–100 cm;
- Home bed: under motion is usually not possible, required longitudinal displacement from the platform border of approximately 20 cm, working height approximately 80–90 cm;
- Hospital bed: under motion is usually possible, required longitudinal displacement from the platform border of approximately 20 cm, working height approximately 100–110 cm;
- Standing person: no longitudinal displacement from the platform border is required, working height approximately 120 cm;
- Seated person: required longitudinal displacement from the platform border of approximately 10–20 cm, working height approximately 90–100 cm;
- Person on wheelchair: required longitudinal displacement from the platform border of approximately 10–20 cm, working height approximately 80–90 cm;

To keep the center of gravity low during the motions, the conceived device needs to be retracted as much as possible. Considering the lightweight requirements, the mounting configuration (Figure 2b) seems more suitable for this application because of good weight distribution. The positioning device could also be useful to elevate the RGB camera above most obstacles, to facilitate the tracking of the user in a cluttered environment.

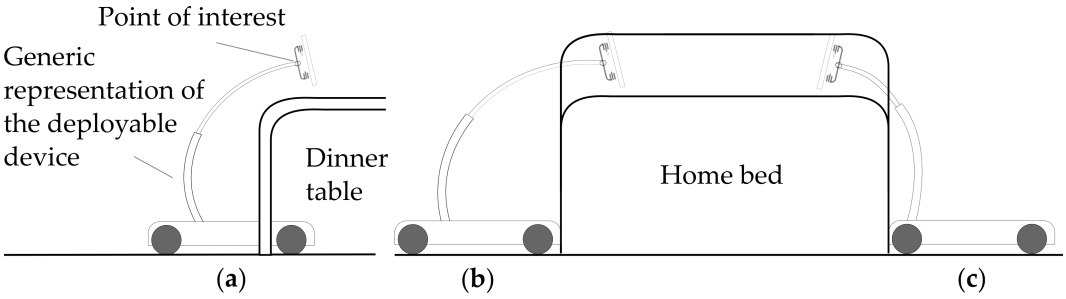

Figure 2. Schematic representation of the system in three situations: (**a**) table approach with undermotion allowed, (**b**) bed approach with the positioning device mounted on the opposite side relative to the deployment direction, (**c**) bed approach with the positioning device mounted on the approach side.

Given the set of tasks to be addressed and the application workspace, some specifications can be worked out for both the robotic base platform and the software design.

- **Performance:** as anticipated, the robot should act as a personal assistant. As a consequence, it should be able to follow the user to provide basic assistance. From a mechanical point of view, this implies the need to perform a velocity similar to that of a human walk ($v \approx 1$–1.5 m/s). The robot should be capable of reaching such cruise velocity in a reasonably short time ($t \approx 1$–1.5 s): it follows an acceptable maximum acceleration range $a_{MAX} \approx 0.7$–1.5 m/s^2.
- **Dimensions:** the environment where the robot should navigate is designed for human needs. To effectively move in this environment, the assistant should have the same footprint as humans have: maximum encumbrance on the ground approximately of 40 cm $\times$ 60 cm.
- **Mobility:** the use case requires the robot to exhibit remarkable mobility to maintain a reduced distance from the user while he is moving within the domestic environment. To such an aim, it turns crucial to provide the mobile platform with full in-plane mobility. Such a feature allows the robot to exhibit velocities in the plane independently from its configuration (orientation).
- **Usability:** the identified users' category suggests that the robot should own an easy-accessible interface to allow efficient interfacing. This feature yields requirements for both software and structural design areas. From the mechanical point of view, the robot layout must allow simple and comfortable access to the interface area. Then, it is interesting to consider the possibility of providing the interface with a proper number of degrees of freedom to make it approach the users' reach when they are unable to.
- **Computational capability:** given the complexity of the software system, a certain degree of computational resources are needed. In particular, one of the most critical components in this regard is represented by the navigation system: to guide the platform in a cluttered, dynamic environment, it needs to re-plan the optimized path very quickly and react appropriately to very different and potentially dangerous situations. In any case, the final implementation needs to be executed completely on board the platform, without the help of external or remote contributions.
- **Modularity:** although the capacities required by the robot and presented before already cover a broad use case, the needs and problems of the domestic environment are constantly evolving. This means that the robot should be able to cope with problems not considered in an early design phase, with the integration of new skills. For this reason, the platform should be designed to be as modular as possible, easily allowing the inclusion or removal of new components, hardware and software. This is necessary also because hardware and software are strictly linked, as an increasing complexity of the application system would require more computational capabilities.

4. Marvin Modular Approach

The architecture of the mobile robot has been developed with a modular approach in order to design a robust architecture for both the physical and non-physical constituent parts such as mechanics, electronics, and software. The goal is to obtain a system where small changes, in either the application environment or the implemented features, do not require structural modifications of the system itself. The overall system can be divided into three main layers, as presented in Figure 3:

- A *Low Layer System* consists of the mechanical structure, the control electronics, and firmware. This layer is responsible for the actuation and control of the system motion given the desired state of the system which is computed by the Upper Layer System.
- A *Upper Layer System* collects the Upper Layer sensors such as lidar, cameras and remote controller, the autonomous navigation stack, and the visual perception subsystem. This module collects data from the sensors and plans the robot response based on the user's commands interpreted by the Human-Machine Interface and on the current state of the robot.
- A *Human-Machine Interface* consists of a vocal control interface and a manual control interface. The implementation of a custom graphic interface has been discussed but postponed because, even if it enriches the user experience, it does not increase the functionality of the robot which is the main goal of this prototype.

The interaction between the different layers is coordinated by predefined communication protocols. In the following sections, the main features of these modules are presented.

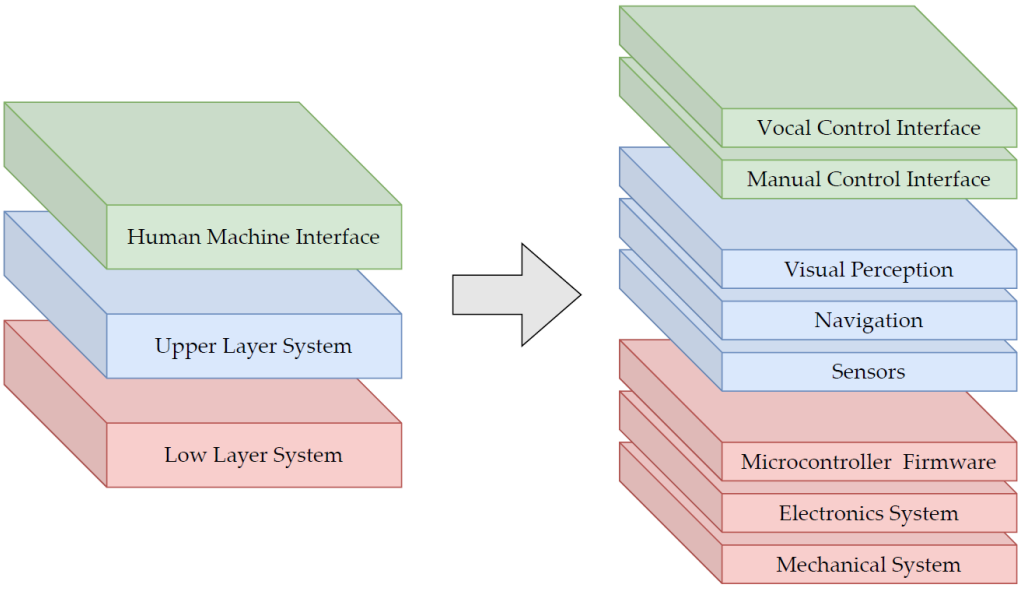

Figure 3. Schematic representation of the modular platform architecture. The three main layers of the architecture (**left**) are decomposed in their respective principal components (**right**).

5. Low Layer System

As the first design step, it is necessary to identify the typology of a suitable base platform on which to build the entire system. The selected platform should fulfill all requirements as best as possible. This preliminary choice is of fundamental importance, as an evaluation error in this phase could affect the following implementations. At first, various differential drive platforms are considered for inspiration, like the Robotis TurtleBot3 Waffle [37] and the TurtleBot2 [38], differential drive robots specially designed taking into account modularity and prototyping. An alternative is the RosBot2 [39], a four-wheel differential drive robot.

To overcome the limitations of differential drive locomotion systems and fulfill mobility requirements, an omnidirectional platform is more suitable for the application. Different solutions can be adopted to achieve this level of maneuverability. For example, specially designed wheels [40], such as Mecanum wheel [41–43], Universal wheels [44], Orthogonal wheels [45], Spherical/Ball wheels [46,47], or conventional steerable wheels [48] can be adopted. In particular, four mecanum wheels and three omniwheels configurations are the most commonly adopted. Reasons for this are the simple control strategy required, omnidirectional mobility with fast response to turn, and simple setup. This improved maneuverability comes with some drawbacks such as discontinuous contact with the ground, a higher sensitivity to floor condition compared to conventional wheels, and payload limitations. Considering the specific environment, ground conditions are quite controlled in indoor applications, even if small obstacles can be found on the floor, while payload limitation is not a problem considering the limited weight of the system. To speed up the prototyping phase, research on commercial solutions has been made. The selected platform, a Nexus 4WD Mecanum robot [49], is characterized by overall dimensions of 400 mm × 360 mm × 100 mm and a limited mass (5.4 kg), with a passive roll joint between the front wheels and the rear wheels to deal with the presence of four contact points with the ground.

The main peculiarity of the robot, aside from its ability to exhibit full planar mobility, is its capability to deploy its sensors and user interface, exploiting the integrated positioning device (Figure 4). Such aspect is crucial for different reasons:

- it allows improving the perception of the robot of the external environment improving the range of view of the sensors;
- a re-orientable and deployable head enhances the usability of the touch interface for the users, giving a chance also to bedridden or handicapped people to easily interact with the robot.

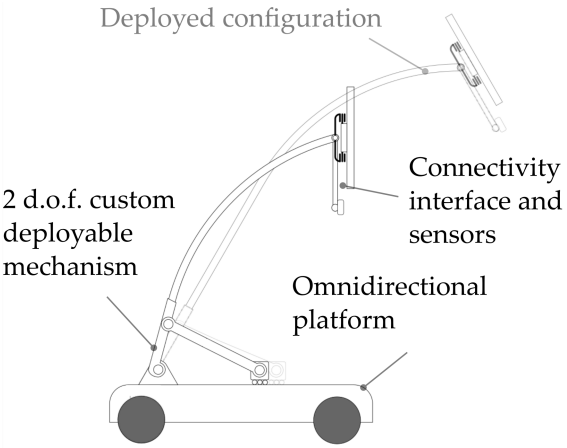

Figure 4. Concept representation of the mobile robot.

In Figure 5, the mobile assistive robot is represented in two configurations: on the left, the telescopic mechanism is deployed for better standing usage, while on the right, the custom mechanism is retracted and inclined forward for better-seated usage. The retracted configuration is also very effective at keeping the center of gravity low during the motion of the robot.

The electronics system firmware is running on the MCU, a PJRC Teensy 4.1 microcontroller (https://www.pjrc.com/store/teensy41.html (accessed on 1 May 2022)), which is responsible for receiving instructions from the computing unit and acting on the actuators to control the motion of the system.



(**a**) (**b**)

Figure 5. Final prototype of the mobile assistive robot in two different working configurations: (**a**) Deployed configuration for standing usage, HMI height = 1.1 m, mechanism tilting angle = 0°, (**b**) Retracted end angled configuration for better-seated usage, HMI height = 0.80 m, mechanism tilting angle = 20°.

6. Upper Layer System

The Upper Layer System contains all the algorithms for the execution of the various task performed by the robot. Each piece of software dedicated to a specific action of the rover needs to exchange information with the rest of the system to fulfill its assignment properly. To this end, the most widespread solution in literature requires the use of a Middleware [50], an abstraction layer that resides between the operating system and software applications. In this project, we decided to adopt the Robot Operating System 2 (Open Source Robotics Foundation, Inc. (https://www.openrobotics.org/ (accessed on 1 May 2022))) (ROS2) [51] due to the variety of compatible algorithms and the very active community supporting it. We preferred it over the original ROS [52] as it is more suitable for real-time systems and has access to more advanced applications [53,54]. ROS2 is based on a Data Distribution Service (DDS) structure, with nodes able to publish and subscribe to different topics.

In our system, all the nodes listen to (or publish on) a specific topic, called *Actions topic*, which contains information about the actual state of the robot and receives requests to perform a new action. This topic is essential to ensure a certain degree of synchronization between all the software components, with consequent robustness of the entire system. To control the platform, the user can use two different human-machine interfaces. The first is the Vocal Command (which will be presented in the next chapter), and the second is a wireless gamepad. The latter allows manual control of the platform, as well as the execution of all the tasks. The manual control interface also provides the possibility to send an emergency signal which immediately disables the platform's current action, guaranteeing safety conditions and risk prevention.

6.1. Sensors

For the robot to effectively work in the domestic environment, a whole series of sensors are required to perceive the surroundings adequately. Alongside classic devices like RGB cameras and Lidar sensors, technology has introduced more powerful tools able to achieve advanced tasks, like self-localization and depth estimation autonomously. On our robotic platform, some of these state-of-the-art devices are employed (Figure 6). In particular, the following sensors are used:

- Intel RealSense T265 Tracking Camera (https://www.intelrealsense.com/tracking-camera-t265/ (accessed on 1 May 2022)) , with VIO technology for self-localization of the platform. It is placed in the front of the rover, to better exploit its capability

- Intel RealSense D435i Depth Camera (https://www.intelrealsense.com/depth-camera-d435i/ (accessed on 1 May 2022)), able to provide color and depth images of the environment. It is mounted on the appropriate support, on the positioning device, which provides a convenient elevated position for the camera
- RPLIDAR A1 (https://www.slamtec.com/en/Lidar/A1 (accessed on 1 May 2022)), exploited for its precision in obstacle detection, a fundamental aspect for obstacle avoidance navigation and mapping of the environment. It is mounted on the platform with a specific structure, capable of elevating it above the other components of the robot. In this way, the only blind spot of the sensor is constituted by the rod of the positioning device, which, however, occupies a very limited area and does not compromise the correct functioning of the sensor
- Jabra 710 (https://www.jabra.com/business/speakerphones/jabra-speak-series/jabra-speak-710#7710-409 (accessed on 1 May 2022)), with a panoramic microphone and speaker. It is particularly useful for voice command. Can be placed on the rover or used wireless from a distance
- Furthermore, a wireless gamepad is employed for manual control operations.

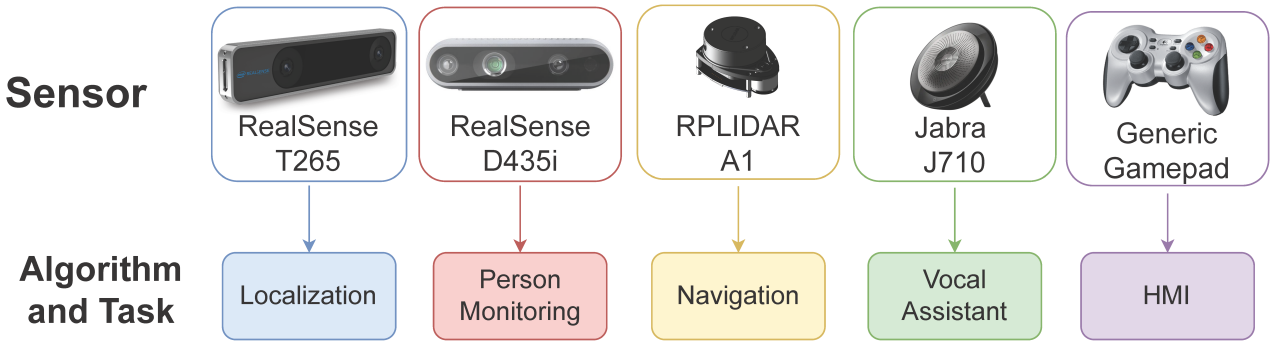

Figure 6. Sensors employed on the Marvin robotic platform, associated with the corresponding task they serve.

6.2. Computational Resources

With technological advancement, software algorithms have exponentially grown in complexity and computational requirements, leading to the abandonment of limited integrated systems in favor of more powerful hardware. Fortunately, these systems are increasingly widespread and are easy to find, allowing researchers to focus on the development of new applications without worrying about hardware limits.

Our system relies on two fundamental components: the *microcontroller unit* (MCU), which manages the low layer system software and a computing unit that executes all Upper Layer system applications. The selected microcontroller, a Teensy 4.1, is chosen for its high clock frequency and excellent general performance. It allows communication between the Upper Layer algorithms and the mechanical and electronic system, and vice versa. On the high-level system, an Intel NUC11TNHv5 is selected as a computing unit, as it represents a good trade-off between high computational power and low energy consumption. A Coral Edge TPU Accelerator is employed alongside the computing unit to run optimized neural network models without the necessity of a full-size graphics processing unit.

6.3. Visual Perception for Person Monitoring

Computer vision is a fundamental component of most recent service robotics platforms. In the last decade, Deep Neural Networks (DNN) have largely been demonstrated to be meaningful solutions for a wide variety of visual perception tasks such as real-time object detection [27], semantic segmentation [55] or pose estimation [26]. Robots can exploit the vision of the surrounding environment to extrapolate information and plan their actions accordingly. Nonetheless, visual perception is an extremely effective method for monitoring

a person in a domestic scenario. We developed a visual perception system for Marvin that contextually detect and track a person from color images. Such information is translated into an effective method for constantly monitoring sudden emergency health conditions of the assisted individual. The RealSense D435i API is used to retrieve aligned color and depth images.

The monitoring task is carried out through a double-step computing pipeline. Firstly, the person-detection is obtained with PoseNet [33], a lightweight neural network able to detect humans in images and videos. As output, it gives 17 key joints (like elbows, shoulders, or feet) of each person present in the scene. At this point, a second simple convolutional neural network (CNN) receives the key points to classify the pose of the person as standing, sitting, or laying. As already explained, a persisting laying condition can automatically activate an emergency call to an external agent (a relative or a healthcare operator). A custom labeled dataset of images has been collected in a house environment to train the CNN for the pose classification. The total number of images used for this dataset is 25,009. The images are divided into three classes: standing, sitting, and lying, containing 7849, 11,400, and 5760 images, respectively. The classification model reaches an accuracy of almost 99% on the test set, obtained retaining the 20% of the original dataset. The performance of the model is definitely high, probably due to the common background scene of the collected images. A randomized background with scenes of diverse domestic environments may allow for a more challenging testing condition, leading also to improving the generalization performance of the model. Besides the accuracy results, an important remark is that the whole pose estimation algorithm has been drastically optimized to guarantee a real-time monitoring system. Moreover, the lightweight model runs on the Google Coral Edge TPU device for a faster inference: the CNN is able to run at 30 frame-per-second (FPS), which is the maximum frame rate allowed with the Realsense D435i camera.

Moreover, as shown in Figure 7, the key points predicted by PoseNet for the detected person can be exploited for a different assistive task: the person following. Indeed, once a person is recognized within the color image, it is possible to derive the coordinates of such person with respect to the robot from the aligned depth image at any instant. This constitutes an effective method to generate a dynamic goal, corresponding to the position of the user, to be reached by the robot. The person-following navigation system is fed with such information and allows the robot to follow the user around the house. However, more than a single person is usually present in a family house, dramatically increasing the difficulty of an automatic system to recognize the person to follow. For this reason, we combine in our visual perception software a filter called Sort [56], which is used within the same node to track the people recognized. In particular, it assigns an ID to each person in the image and tracks them during their motion. The person with the lowest ID is chosen to be followed: the robot focuses always on a single person and the computational complexity of the task is considerably reduced. Nonetheless, a further improvement of the person-following task can be achieved by adopting a neural network for person re-identification, allowing the robot to discard undesired detected persons and reduce the interference with its monitoring activity.

6.4. Navigation and Mapping System

In any navigation system, the primary necessity consists in localizing the robot within the operating scenario. To achieve this, the localization node exploits the Intel RealSense API to communicate with the T265 camera and get the pose of the rover at any time instant, through visual-inertial odometry technology [57]. The navigation system is based on the Navigation2 stack [58], which has been highly modified to suit the needs of the platform. Details on the entire development and optimization process of the navigation behavioral tree are out of the scope of this paper. When the rover is asked to reach a specific goal or to follow the use, the navigation system retrieves the pose of the rover and exploits the 2D LiDAR points to perceive surrounding obstacles and create a local cost map. From such a cost map, the navigation apparatus plans an optimal path for the platform and guide it

towards the desired destination. Similarly, the mapping system, based on Slam Toolbox [59], uses the pose of the rover and the laser scan to generate a grid map of the environment. Although the navigation system perfectly adapts to mapless circumstances, the generated map of the domestic environment can be saved by the robot.

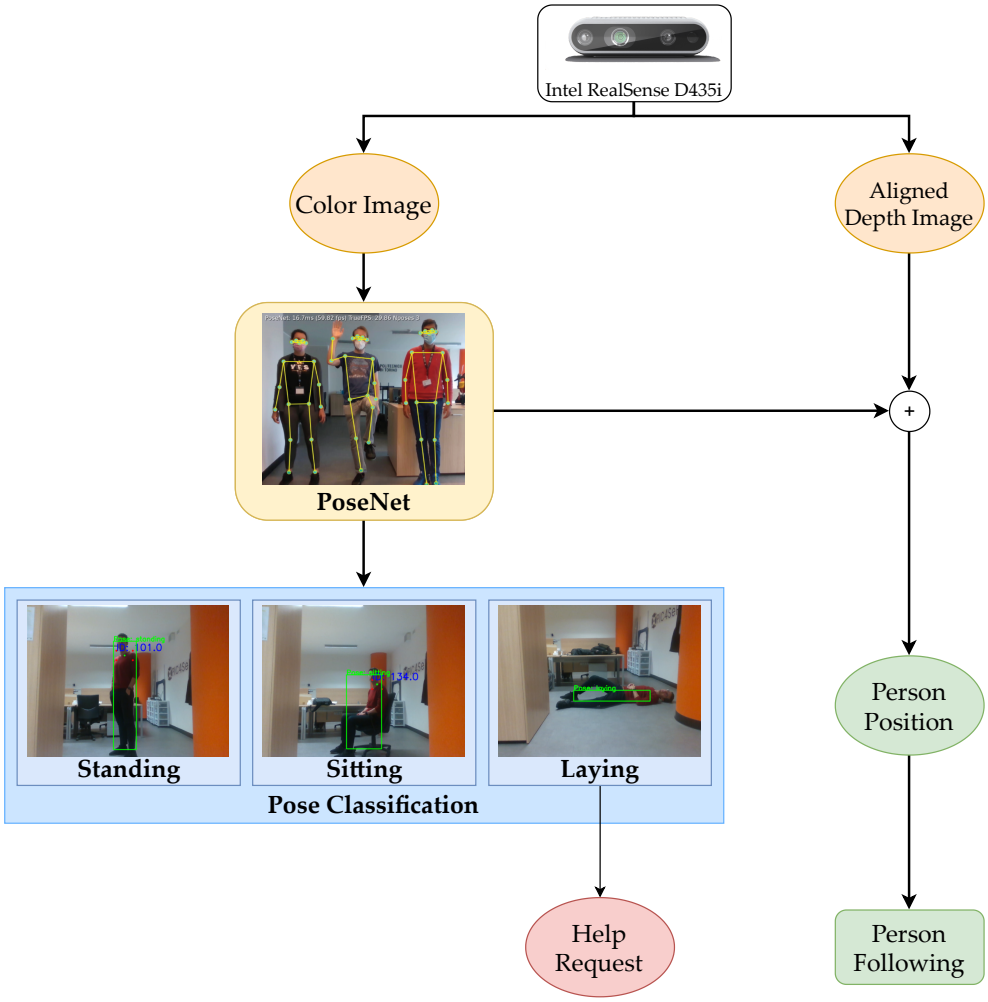

Figure 7. Representation diagram of the person identification system: the estimated pose of the person is continuously classified as standing, sitting, or laying, generating a help emergency request if necessary. Moreover, it is used to extract the dynamic goal coordinates for the person following task.

7. Vocal Human–Robot Interface

The principal user's communication interface with the platform is represented by an offline vocal assistant. We build our vocal assistant system, called PIC4Speech, exploiting the combination of state-of-the-art Deep Neural Networks (DNN) for speech-to-text translation and a simple rule-based model for Natural Language Processing (NLP) taken from literature with the aim of minimizing the computational cost of the pipeline and preserving a flexible interaction. The overall structure of the system is inspired by the most notable products: Siri, Alexa, and Google Assistant. It exploits a cascade of models that are progressively activated when the previous one is enabled. In Figure 8, an overview of the overall architecture of the PIC4Speech vocal assistant is represented, showing the subsequent activation of each block. Although PIC4Speech is designed as an offline vocal assistant, its usage in this primitive version is mainly devoted to allowing the user to give commands to the robot vocally and not to hold a complete conversation. In particular, the system aims at matching a vocal instruction expressed by the user to the corresponding

required task to successively start the correct control process by publishing a ROS message on the Actions topic.

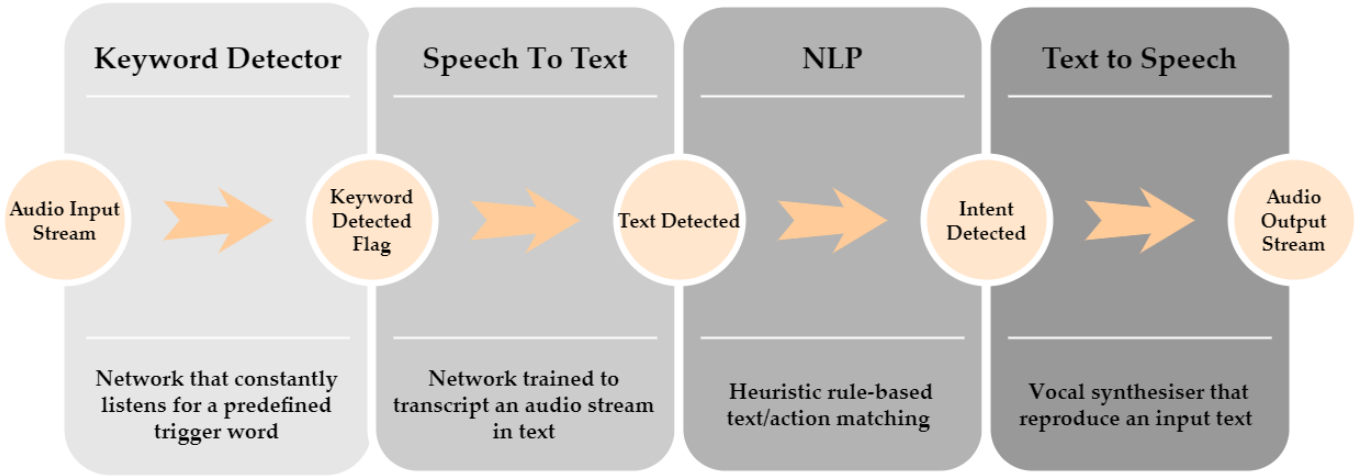

Figure 8. Overview of PIC4Speech vocal assistant architecture. The scheme describes the successive cascade activation of the different components of the vocal assistant pipeline.

The PIC4Speech operative chain can be summarized as follows:

1. The first component is the **keyword detector** [60], which constantly monitors the input audio stream in search of the specific triggering command. In this specific case, that word is the name of the robot: "Marvin."
2. Once the trigger word is detected, a second model performs a **speech-to-text** operation. We exploited the Vosk offline speech recognition API [61] for this block which gives the flexibility to switch language and has ample community support. It continuously analyzes the input audio stream until the volume is below a certain threshold and performs the transcription.
3. The transcribed text is subsequently passed to a **natural language processing (NLP)** algorithm that matches the input with a certain number of predefined intents. The recognized robot action is therefore published on the Actions topic of the ROS framework.
4. The response of the vocal assistant is also given to the user with a **text-to-speech** process. Each OS comes with a default vocal synthesizer that can directly access the speakers.

More in detail, the keyword detection is performed with a DNN based on a Vision Transformer that constantly listens to the audio stream, looking for the target command. First, the mel-scale spectrogram is extracted from each sample of the input stream. These features are treated as visual information and, therefore, they are processed with a Vision Transformer [62], a state-of-the-art model for image classification. We re-trained the keyword detector from scratch on the Speech Commands dataset [63], constituted by 1-second-long audio samples from 36 classes: 35 standard keywords plus a silence/noise class. The re-train model achieved a test accuracy of 97% over the different classes on the 11,005 test samples of the Speech Commands dataset. Specifically, for the current target class 'Marvin', we get the results reported in Table 1, evaluating the performance with standard classification metrics: $Precision = TP/(TP + FP)$, $Recall = TP/(TP + FN)$ and $F1score = 2 \cdot (recall \cdot precision)/(recall + precision)$.

Thanks to the multi-class approach, the keyword can be changed at run-time. Being constantly active, it is of primary importance that this network consumes less energy as possible, but at the same time, it is capable of maintaining a good compromise between false positive and false negative detections. At the same time, the network should deal with different sound environments and noise levels. Further improvements to the keyword detector can be reached by augmenting the training set with newly generated samples to

increase the robustness of the model in crowded, noisy environments. At the moment, a simple rule-based matching mechanism based on keywords is used for the NLP stage. Although it represents a simple approach, a good level of flexibility is guaranteed by the actual solution as the user can introduce new actions for the robotic platform associated with several indicative sentences. Future works may involve the investigation of suitable DNN models for the NLP stage of PIC4Speech, with the aim of semantically matching the encoded query text and the robot actions.

Table 1. Classification results of the target keyword 'Marvin' on the 11,005 test samples of the Speech Commands dataset. Standard classification metrics are used for the evaluation: $Precision = TP/(TP + FP)$, $Recall = TP/(TP + FN)$ and $F1score = 2 \cdot (Recall \cdot Precision)/(Recall + Precision)$.

Keyword Detector Classification Metrics	Results
True Positives (TP)	189
True Negatives (TN)	10,806
False Positives (FP)	4
False Negatives (TN)	6
$F1Score$	0.9742
$Precision$	0.9793
$Recall$	0.9692

Different from commercial vocal assistants, which require a stable internet connection, PIC4Speech works completely offline, running uniquely on the hardware resources of the platform. This competitive advantage derived from the choice of lightweight models in the algorithms pipeline prevents Marvin from exposing the visual data of the domestic environment to internet-derived risks. Diversely, companies prefer to offer a server-client paradigm with the assistant algorithms running on the cloud. That solution presents some computational advantages and enables the connectivity of the platform with the whole house. Furthermore, it dramatically facilitates data collection. This solution may raise privacy issues related to the collection of visual and vocal data collected in domestic private environments currently used on the robot systems. In addition, the dependency on a stable internet connection may weaken the system performance in terms of response time and power consumption. For all those reasons, an offline solution should largely fit most service robotics applications requiring vocal control.

Moreover, it is worth noting that a help request is constantly visually checked by the pose classification node but it can also be called directly through vocal command. In this case, the platform will ask the user for confirmation, avoiding any accidental activation. If confirmed, without any reply within ten seconds, the help request is sent. Otherwise, the platform return to its regular operation state. Further development of the PIC4Speech vocal assistant could see the substitution of the last block for text-to-speech conversion with an additional lightweight neural network, also providing the possibility to choose a more comfortable synthetic voice closer to a real human one.

8. Experimental Demo

We conducted a qualitative demonstration of the platform's capabilities at Officine Edison, Milan, during an institutional presentation specifically organized to test and show Marvin and validate the outcome of its prototyping process. The demonstration took place in an area called Domus (Figure 9), which simulates a real domestic environment made up of a kitchen, bedroom, living room, and bathroom. Like a normal house, the Domus features different obstacles of various heights and dimensions, and rooms are separated by regular size doors.

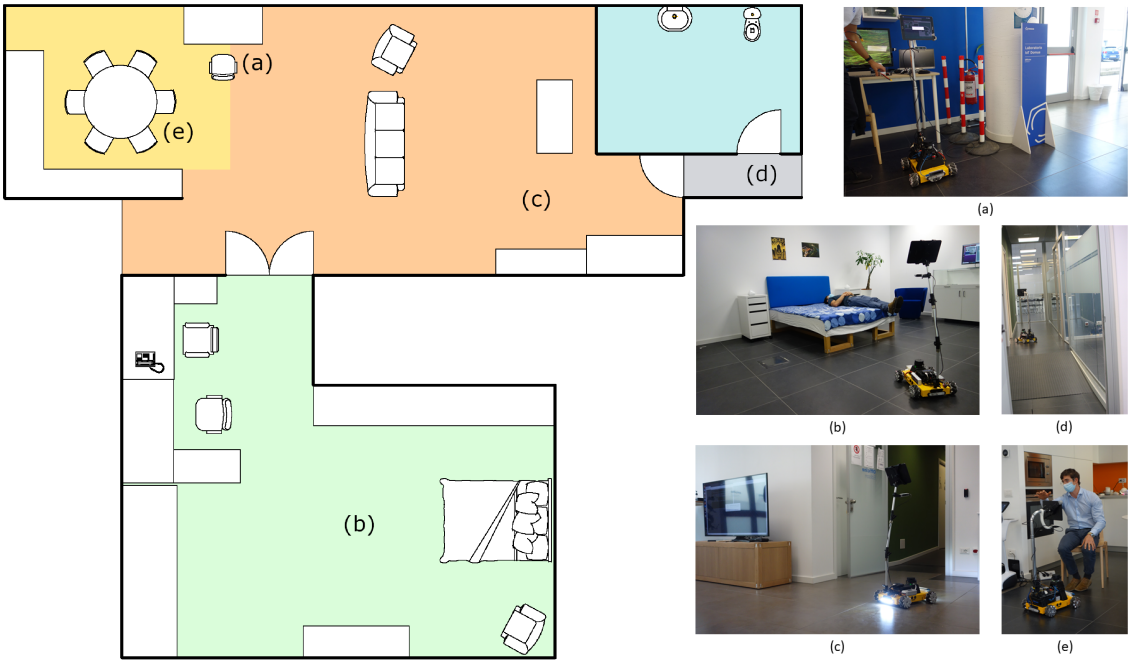

Figure 9. Simplified map of the Domus area at Officine Edison, Milan, with the four rooms, kitchen, living room, bedroom, and bathroom, respectively in yellow, orange, green, and blue. Letters on the map indicate the various goals saved in the environment, associated with the corresponding image: (**a**) starting point, (**b**) bedroom keypoint, with user's pose recognition, (**c**) living room keypoint, with lights turned on, ready for night assistant task (**d**) bathroom keypoint, (**e**) kitchen keypoint, with demonstration of the positioning device capabilities.

In the setup phase, Marvin was guided in each of the different rooms and their relative positions were saved with respect to the starting point, where a docking station for recharging could eventually be placed. Moreover, a telephonic number was memorized for the emergency call task. In the demonstration, all the functionalities of the robot were tested, and a qualitative analysis was conducted. From the starting point (Figure 9a), where a brief introduction was given to attendees, the rover was asked to autonomously reach the bedroom waypoint, passing through the double-leaf door. Here, the user monitor function was demonstrated, showing how Marvin was able to correctly classify the pose of the visualized user, standing, sitting on an armchair, or laying in the bed (Figure 9b). In addition, it was also demonstrated how, after a request from the user, the system was capable of connecting with the pre-configured telephone to call the emergency number. Then, the rover was asked to follow the user from the bedroom to the living room (Figure 9c). There, the user asked Marvin to save a new semantic waypoint, to demonstrate how the various room destinations within the environment can be saved and used by the robot for autonomous navigation tasks. Visual perception algorithms, namely person detection and pose estimation and classification, have been visually validated publicly showing on the television screen what the robot "saw" through cameras. The crowd of attendees received an in-depth explanation of how the various algorithms of person and pose estimation work while looking at the resulting predictions in real-time. Later, the user activated the night assistance task by asking the rover to accompany him to the bathroom, causing the robot to turn on the on board lights (Figure 9d). Finally, Marvin was asked to reach the kitchen and to adjust the inclination and height of the positioning device to adapt to the user, sitting on the chair, so that the mounted tablet could be more easily accessed and operated (Figure 9e).

During the demonstration, we collected some observations about the performance of the system. In particular, we focused our attention on the success of the three proposed service functions, determined by the fulfillment of the required robotic tasks, as reported

in Table 2. Thanks to its mobility and the elevated position of the camera, mounted on the positioning device, the rover managed to efficiently track and follow the person, albeit the several obstacles with diverse heights. This allowed to continuously monitor the user, classifying their pose in any instant, and calling for help during potentially dangerous situations. Moreover, the rover showed no difficulties in autonomously navigating between the various rooms, avoiding static and dynamic obstacles arranged in different configurations, and accompanying the user to the desired destination. In particular, the robot was also able to manage its handling upon less conventional surfaces, such as the skirting board of fire doors and a small ramp placed in the corridor before the bathroom. Finally, the positioning device's flexibility, in conjunction with the platform maneuverability, guaranteed the user an easy operation of the tablet, standing, sitting, or lying down.

Table 2. The three service functions (left) provided by the Marvin robot in the context of home assistance are associated with the robotic tasks (right).

Service Function	Robotic Tasks	Demo Execution
User Monitoring	Person Following Pose Classification Emergency Call	(b) $\Longrightarrow$ (c) always running (b)
Night Assistance	Autonomous Navigation Lighting Control	(c) $\Longrightarrow$ (d)
Remote Presence & Connectivity	Positioning Device Control Autonomous Navigation	(e) (a) $\Longrightarrow$ (b)

Further analyses were conducted to validate the vocal assistant. The vocal command was tested in three different scenarios. The first one took place in an environment characterized by good acoustic conditions and without any relevant noise apart from the tester's voice giving commands. The second test scenario is located in the same environment described before, but with the addition of background noise caused by a group of people talking. The third test scenario was realized in a silent environment presenting not-optimal acoustic conditions, like echo and reverb. To evaluate the performance of the vocal assistant, twenty test commands were given for each scenario and six parameters were taken into consideration. Of these, the results proved only three are scenario dependent:

- Keyword detector success frequency, indicate how many times the Keyword detector is triggered when the trigger word is pronounced
- Keyword detector accuracy indicate the average confidence over the prediction of the trigger word
- Command understanding frequency, indicate how many times the command given after the trigger is correctly understood by the speech-to-text model

The other three parameters returned no significant differences among the various scenarios:

- Trigger delay, indicate the average time delay between the pronunciation of the trigger word and the actual trigger
- Command delay, indicate the average time delay between the given command and the received feedback from the system
- Help request delay, indicate the average time delay between the help request command and the starting of the call towards the emergency number

The final results can be consulted in Table 3.

Table 3. Performances of vocal assistant PIC4Speech during the experimental demo.

	Muffled, Silent Environment	Muffled Environment with People Talking	Environment with Echo and Reverb
Keyword detector Success rate	100.00%	95.00%	95.00%
Keyword detector Accuracy	91.14%	88.24%	89.04%
Command understanding Frequency	100.00%	84.21%	94.74%
Trigger delay		<0.5 s	
Command delay		1.45 s	
Help request delay		<0.5 s	

9. Conclusions and Future Works

In the era of automatic machines, technology is progressively reshaping the domestic environment as we know it. In particular, service robotics is recalling an ever-growing interest of markets, industries, and researchers. Their exploitation in the caregiving sector could relieve the pressure on assistive operators, providing basic assistance which does not require particular dexterity or adaptation capability. In this scenario, we developed Marvin: a modular assistive mobile robot for autonomous applications in the field of home assistance. In this work, Marvin has been initially presented as a robotic assistant solution tailored for the practical use case of monitoring elderly and reduced-mobility subjects in their domestic environment, although its applicability can be easily extended to the alternative person monitoring scenarios in indoor environments. Hence, this paper aims to fully describe Marvin, a four mecanum-wheel robot provided with a custom positioning device for the human-machine interface and state-of-the-art Artificial Intelligence methods for perception and vocal control. The robot has been fully prototyped and qualitatively tested in a domestic-like environment and it proved to be successful in the execution of the target tasks.

Future works will firstly try to enrich the experimentation by providing more task-specific experimental results and subsequently extend the applicability of Marvin to unseen service robotics functions. More in detail, a great focus will be devoted to the application of Marvin to person-centered autonomous navigation tasks. A secondary future direction deals with the upgrade of the human–robot interface of the robot, enhancing its proactive behavior in social domestic environments and its awareness of the context through more sophisticated visual techniques. Furthermore, the combination of vocal and visual inputs can help the robot contextualize its actions better, resulting in higher precision in tasks execution.

Author Contributions: Conceptualization, A.E., M.M., L.T., D.G., M.C., and G.Q.; methodology, A.E., M.M., L.T., D.G., M.C., and G.Q.; software, A.E. and M.M.; validation, A.E., M.M., and L.T.; formal analysis, A.E., M.M., and L.T.; investigation, A.E., M.M., and L.T.; writing—original draft preparation, L.T., M.M., and A.E.; writing—review and editing, L.T., M.M., and A.E.; visualization, L.T., M.M., and A.E.; supervision, D.G., M.C., and G.Q.; project administration, D.G., M.C., and G.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was developed by a collaboration between EDISON Spa, grant number 06722600019 and the interdepartmental research group PIC4SeR of Politecnico di Torino, Italy.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Speech Commands Dataset: https://www.tensorflow.org/datasets/catalog/speech_commands (accessed on 1 May 2022).

Acknowledgments: The work presented in this paper has born from the collaboration between the PIC4SeR Centre for Service Robotics at Politecnico di Torino and Edison S.p.A. In particular, we sincerely thank Riccardo Silvestri and Stefano Ginocchio, as well as the entire team from Officine Edison Milano that fruitfully contributed to the funding and conceptualization of Marvin, and supervised the whole design process. We demonstrated Marvin's capabilities in the Smart Home facility at Officine Edison Milano, simulating a real-case domestic assistance scenario, showing how Marvin successfully fulfill the tasks requirements identified in the design process.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. United Nations. Shifting Demographics. Available online: https://www.un.org/en/un75/shifting-demographics (accessed on 1 May 2022).
2. Vercelli, A.; Rainero, I.; Ciferri, L.; Boido, M.; Pirri, F. Robots in elderly care. *Digit.-Sci. J. Digit. Cult.* **2018**, *2*, 37–50.
3. Abdi, J.; Al-Hindawi, A.; Ng, T.; Vizcaychipi, M.P. Scoping review on the use of socially assistive robot technology in elderly care. *BMJ Open* **2018**, *8*, e018815. [PubMed]
4. Gouaillier, D.; Hugel, V.; Blazevic, P.; Kilner, C.; Monceaux, J.; Lafourcade, P.; Marnier, B.; Serre, J.; Maisonnier, B. Mechatronic design of NAO humanoid. In Proceedings of the 2009 IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May 2009; pp. 769–774.
5. Fujita, M. AIBO: Toward the era of digital creatures. *Int. J. Robot. Res.* **2001**, *20*, 781–794.
6. Šabanović, S.; Bennett, C.C.; Chang, W.L.; Huber, L. PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia. In Proceedings of the 2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR), Seattle, WA, USA, 24–26 June 2013; pp. 1–6.
7. Góngora Alonso, S.; Hamrioui, S.; de la Torre Díez, I.; Motta Cruz, E.; López-Coronado, M.; Franco, M. Social robots for people with aging and dementia: A systematic review of literature. *Telemed. E-Health* **2019**, *25*, 533–540. [CrossRef] [PubMed]
8. Gasteiger, N.; Loveys, K.; Law, M.; Broadbent, E. Friends from the Future: A Scoping Review of Research into Robots and Computer Agents to Combat Loneliness in Older People. *Clin. Interv. Aging* **2021**, *16*, 941–971. [CrossRef]
9. Yatsuda, A.; Haramaki, T.; Nishino, H. A Study on Robot Motions Inducing Awareness for Elderly Care. In Proceedings of the 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), Taichung, Taiwan, 19–21 May 2018; pp. 1–2. [CrossRef]
10. Mundher, Z.A.; Zhong, J. A real-time fall detection system in elderly care using mobile robot and kinect sensor. *Int. J. Mater. Mech. Manuf.* **2014**, *2*, 133–138. [CrossRef]
11. Saini, J.; Dutta, M.; Marques, G. Sensors for indoor air quality monitoring and assessment through Internet of Things: A systematic review. *Environ. Monit. Assess.* **2021**, *193*, 66.
12. Mocrii, D.; Chen, Y.; Musilek, P. IoT-based smart homes: A review of system architecture, software, communications, privacy and security. *Internet Things* **2018**, *1*, 81–98.
13. Marques, G.; Pires, I.M.; Miranda, N.; Pitarma, R. Air quality monitoring using assistive robots for ambient assisted living and enhanced living environments through internet of things. *Electronics* **2019**, *8*, 1375.
14. Doroftei, I.; Grosu, V.; Spinu, V. *Omnidirectional Mobile Robot-Design and Implementation*; INTECH Open Access Publisher: London, UK, 2007.
15. Al Mamun, M.A.; Nasir, M.T.; Khayyat, A. Embedded system for motion control of an omnidirectional mobile robot. *IEEE Access* **2018**, *6*, 6722–6739.
16. Costa, P.J.; Moreira, N.; Campos, D.; Gonçalves, J.; Lima, J.; Costa, P.L. Localization and navigation of an omnidirectional mobile robot: The robot@ factory case study. *IEEE Rev. Iberoam. Tecnol. Del Aprendiz.* **2016**, *11*, 1–9. [CrossRef]
17. Qian, J.; Zi, B.; Wang, D.; Ma, Y.; Zhang, D. The design and development of an omni-directional mobile robot oriented to an intelligent manufacturing system. *Sensors* **2017**, *17*, 2073. [CrossRef] [PubMed]
18. Jibo Robot Website. 2017. Available online: https://jibo.com/ (accessed on 1 May 2022).
19. Fischinger, D.; Einramhof, P.; Papoutsakis, K.; Wohlkinger, W.; Mayer, P.; Panek, P.; Hofmann, S.; Koertner, T.; Weiss, A.; Argyros, A.; et al. Hobbit, a care robot supporting independent living at home: First prototype and lessons learned. *Robot. Auton. Syst.* **2016**, *75*, 60–78. [CrossRef]
20. Hashimoto, K.; Saito, F.; Yamamoto, T.; Ikeda, K. A field study of the human support robot in the home environment. In Proceedings of the 2013 IEEE Workshop on Advanced Robotics and Its Social Impacts, Tokyo, Japan, 7–9 November 2013; pp. 143–150.
21. Tanioka, T. Nursing and rehabilitative care of the elderly using humanoid robots. *J. Med. Investig.* **2019**, *66*, 19–23. [CrossRef] [PubMed]
22. Robotics, P. TIAGo. Available online: https://pal-robotics.com/robots/tiago/ (accessed on 1 May 2022).
23. Juel, W.K.; Haarslev, F.; Ramirez, E.R.; Marchetti, E.; Fischer, K.; Shaikh, D.; Manoonpong, P.; Hauch, C.; Bodenhagen, L.; Krüger, N. SMOOTH Robot: Design for a novel modular welfare robot. *J. Intell. Robot. Syst.* **2020**, *98*, 19–37. [CrossRef]
24. Amazon. Introducing Amazon Astro—Household Robot for Home Monitoring, with Alexa, 2021 Available online: https://www.youtube.com/watch?v=sj1t3msy8dc (accessed on 1 May 2022)..

25. Toshev, A.; Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
26. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 172–186. [CrossRef] [PubMed]
27. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef]
28. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
29. Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Van Esesn, B.C.; Awwal, A.A.S.; Asari, V.K. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv* **2018**, arXiv:1803.01164.
30. Mateus, A.; Ribeiro, D.; Miraldo, P.; Nascimento, J.C. Efficient and robust pedestrian detection using deep learning for human-aware navigation. *Robot. Auton. Syst.* **2019**, *113*, 23–37. [CrossRef]
31. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
32. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland , 2016; pp. 21–37.
33. Papandreou, G.; Zhu, T.; Chen, L.; Gidaris, S.; Tompson, J.; Murphy, K. PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
34. Moore, R.K. From talking and listening robots to intelligent communicative machines. In *Robots That Talk and Listen*; WALTER DE GRUYTER Incorporated: Boston, MA, USA, 2015 ; pp. 317–335.
35. Skantze, G. Turn-taking in conversational systems and human-robot interaction: A review. *Comput. Speech Lang.* **2021**, *67*, 101178. [CrossRef]
36. Tenney, I.; Das, D.; Pavlick, E. BERT rediscovers the classical NLP pipeline. *arXiv* **2019**, arXiv:1905.05950.
37. TurtleBot3 on Robotis Official Site. Available online: https://emanual.robotis.com/docs/en/platform/turtlebot3/overview/ (accessed on 1 May 2022).
38. TurtleBot2 on TurtleBot Official Site. Available online: https://www.turtlebot.com/turtlebot2/ (accessed on 1 May 2022).
39. RosBot2 Pro on Husarion official Site. Available online: https://store.husarion.com/products/rosbot-pro (accessed on 1 May 2022).
40. Taheri, H.; Zhao, C.X. Omnidirectional mobile robots, mechanisms and navigation approaches. *Mech. Mach. Theory* **2020**, *153*, 103958. [CrossRef]
41. Ilon, B.E. Wheels for a Course Stable Selfpropelling Vehicle Movable in Any Desired Direction on the Ground or Some Other Base. U.S. Patent 3,876,255, 8 April 1975.
42. Pin, F.G.; Killough, S.M. A new family of omnidirectional and holonomic wheeled platforms for mobile robots. *IEEE Trans. Robot. Autom.* **1994**, *10*, 480–489. [CrossRef]
43. Salih, J.E.M.; Rizon, M.; Yaacob, S.; Adom, A.H.; Mamat, M.R. Designing Omni-Directional Mobile Robot with Mecanum Wheel. *Am. J. Appl. Sci.* **2006**, *3*, 1831–1835.
44. Cuevas, F.; Castillo, O.; Cortes-Antonio, P. Towards an adaptive control strategy based on type-2 fuzzy logic for autonomous mobile robots. In Proceedings of the 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 23–26 June 2019; pp. 1–6.
45. Mourioux, G.; Novales, C.; Poisson, G.; Vieyres, P. Omni-directional robot with spherical orthogonal wheels: Concepts and analyses. In Proceedings of the 2006 IEEE International Conference on Robotics and Automation, ICRA 2006, Orlando, FL, USA, 15–19 May 2006; pp. 3374–3379.
46. Tadakuma, K.; Tadakuma, R.; Berengeres, J. Development of holonomic omnidirectional Vehicle with "Omni-Ball": Spherical wheels. In Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, USA, 29 October–2 November 2007; pp. 33–39.
47. Ferrière, L.; Raucent, B. ROLLMOBS, a new universal wheel concept. In Proceedings of the 1998 IEEE International Conference on Robotics and Automation (Cat. No. 98CH36146), Leuven, Belgium, 20 May 1998; Volume 3, pp. 1877–1882.
48. Ferland, F.; Clavien, L.; Frémy, J.; Létourneau, D.; Michaud, F.; Lauria, M. Teleoperation of AZIMUT-3, an omnidirectional non-holonomic platform with steerable wheels. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010; pp. 2515–2516.
49. Nexus 4WD Mecanum Wheel Mobile Robot on Nexus Official Site. Available online: https://www.nexusrobot.com/product/4wd-mecanum-wheel-mobile-arduino-robotics-car-10011.html (accessed on 1 May 2022).
50. Saha, O.; Dasgupta, P. A Comprehensive Survey of Recent Trends in Cloud Robotics Architectures and Applications. *Robotics* **2018**, *7*, 47. [CrossRef]
51. Macenski, S.; Foote, T.; Gerkey, B.; Lalancette, C.; Woodall, W. Robot Operating System 2: Design, architecture, and uses in the wild. *Sci. Robot.* **2022**, *7*, 66. [CrossRef]
52. The Robot Operating System Official Site. Available online: https://www.ros.org/ (accessed on 1 May 2022).

53. Maruyama, Y.; Kato, S.; Azumi, T. Exploring the Performance of ROS2. In Proceedings of the 13th International Conference on Embedded Software, Pittsburgh, PA, USA, 1–7 October 2016. [CrossRef]
54. Changes between ROS2 and ROS1. Available online: https://design.ros2.org/articles/changes.html (accessed on 1 May 2022).
55. Zhang, X.; Chen, Z.; Wu, Q.J.; Cai, L.; Lu, D.; Li, X. Fast semantic segmentation for scene perception. *IEEE Trans. Ind. Inform.* **2018**, *15*, 1183–1192. [CrossRef]
56. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple Online and Realtime Tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016.
57. Debeunne, C.; Vivet, D. A review of visual-LiDAR fusion based simultaneous localization and mapping. *Sensors* **2020**, *20*, 2068. [CrossRef]
58. Macenski, S.; Martín, F.; White, R.; Clavero, J.G. The marathon 2: A navigation system. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24–30 October 2020; pp. 2718–2725.
59. Macenski, S.; Jambrecic, I. SLAM Toolbox: SLAM for the dynamic world. *J. Open Source Softw.* **2021**, *6*, 2783. [CrossRef]
60. Berg, A.; O'Connor, M.; Cruz, M.T. Keyword Transformer: A Self-Attention Model for Keyword Spotting. *arXiv* **2021**, arXiv:2104.00769.
61. Andreev, A.; Chuvilin, K. Speech Recognition for Mobile Linux Distrubitions in the Case of Aurora OS. In Proceedings of the 2021 29th Conference of Open Innovations Association (FRUCT), Tampere, Finland, 12–14 May 2021; pp. 14–21.
62. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
63. Warden, P. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv* **2018**, arXiv:1804.03209.