Ref. Ares(2023)4550332 - 30/06/2023

This project is funded by the European Union
G.A No 101016906

Project Number: 101016906
Start Date of Project: 2021/01/01
Duration: 48 months

Type of document 6.2 – V1.0

Human Robot Communication

Dissemination level	PUBLIC
Submission Date	2023-06-30
Work Package	WP6
Task	T6.3
Type	Report
Version	1.0
Author	UNIROMA1
Approved by	UPC, DTI, PALEBLUE, PAL, AGRIMESSINA

DISCLAIMER:
The sole responsibility for the content of this deliverable rests with the authors. It does not necessarily reflect the opinion of the European Union. Neither the Directorate-General for Communications Networks, Content and Technology, Artificial Intelligence and Digital Industry nor the European Commission are responsible for any use that may be made of the information

This project is funded by the European Union G.A No 101016906

Executive Summary

This report represents the outcome of Task 6.3. The goal of the task is reported in the technical annex of the project as it follows:

"This task will collect the requirements for human robot communication by analyzing the output of human robot collaboration and extracting the instances of communication acts that support the collaboration. The main categories that characterize the communication acts for the chosen scenario will then be identified. Furthermore, a preliminary analysis of the constraints of the operational environment will be carried out to take them into account for the definition of the communication modalities to be investigated in detail. The communication modalities will include spoken language, sound, gestures as well other devices."

Section 1 introduces the rationale followed for this analysis. Section 2 introduces the concepts of actions, activities and goals in vineyard farming, providing for each definition examples from the application scenario. Section 3 introduces the categories of communication acts involved in the human-robot interaction required by the project objective. For each of the communication acts, examples are provided. Section 4 introduces the environmental constraints that can influence communication. Section 5 explains which are the communication modalities that will be used in the project to realize communication acts. Section 6 introduces the architecture defined for Task 6.3. Section 7 and section 8 describe in detail the modules for Human-to-Robot and Robot-to-Human communication respectively. Section 9 concludes the deliverable with final remarks.

As the document is large, we suggest the reader interested only in the technical results to skip Sections 3, 4 and 5, which contain the requirement analysis for the task.

Table of Content

1 Introduction 8
2 Agronomic Actions, Activities and Goals 9
2.1 Actions 10
2.2 Activities and Goals 11
3 Communication act categories 12
3.1 Information 13
3.1.1 Intention 13
3.1.2 Current task 14
3.1.3 Proximity 14
3.1.4 Agronomic element 16
3.1.5 Human presence 17
3.1.6 Environment 17
3.1.7 State 18
3.2 Command 19
3.2.1 Action/Activity 20
3.2.2 Motion 20
3.3 Alert 20
3.3.1 Collision risk 21
3.3.2 Touch 21
3.3.3 Error 21
3.3.4 Velocity 24
3.3.5 Motion 24
3.4 Request 24
3.4.1 Action/Activity 24
3.4.2 Information 25
3.5 Instruction 25
3.6 Greeting 26

3.7 Examples of combined speech-act interaction 26
4 Environmental constraints 27
4.1 Noise 27
4.2 Illumination 28
4.3 Ground 28
4.4 Weather (windy, rainy, humidity) 28
4.5 Branches and stones on the corridor 28
5 Communication modalities 29
5.1 Spoken language 29
5.2 Gestures 30
5.3 Sound 30
5.4 Visual signals 31
5.5 Other devices 32
5.5.1 Smartphone and Tablet 32
5.5.2 Augmented reality technology 32
6 Implemented solution 33
7 Human-to-Robot Communication 35
7.1 Speech based communication 35
7.1.1 Preliminary Investigations 35
7.1.2 Training Data Collection 38
7.1.3 Study Design 40
7.1.4 System Implementation 41
7.2 Gesture based communication 48
7.2.1 Preliminary Investigations 48
7.2.2 Gesture Definitions 50
7.2.3 Data Generation 51
7.2.4 System Implementation 53
7.3 Multi-modal fusion 55

8 Robot-to-Human Communication 56
8.1 Speech based communication 56
8.2 Sound based communication 56
8.2.1 Sound Emission Hardware requirements 56
8.2.2 Sound signals 57
8.2.3 Sound Signal Types 58
8.2.4 Characteristics of audio warning signals 59
8.2.5 Sound Signal Design 60
8.3 Light based communication 62
8.3.1 Light Emission Hardware requirements 62
8.3.2 Robot Light Signal Emission 63
8.3.3 Light signal parameters 64
8.3.4 Signal Design Algorithm 65
8.4 Multi-modal fission 68
9 Wrap-up, future work and final demonstration 68

Abbreviations and Acronyms

3D	Three Dimensional
4G	Fourth generation of broadband cellular network technology
5G	Fifth generation of broadband cellular network technology
AR	Augmented Reality
ARSG	Augmented Reality Smart Glasses
BEM	Box-Exchange Mechanism
CAN	Controller Area Network
DoF	Degrees of Freedom
EtherCAT	Ethernet-based fieldbus system.
F/T	Force Torque Sensor
FOV	Field Of View
GigE	Gigabit Ethernet
GNSS	Global Navigation Satellite System
HRC	Human-Robot Collaboration
HRI	Human-Robot Interaction
HUD	Head-Up Display
IMU	Inertial Measurement Unit
IP	Ingress Protection Code
ISO	Reference to standards from the International Organization for Standardization
KPI	Key Performance Indicator
LED	Light Emitting Diode
LIDAR	LIght Detection And Ranging
LTE	Long-Term Evolution communication standard
NTRIP	Networked Transport of RTCM via Internet Protocol
NUC	Next Unit of Computing
OS	Operating System

PPS	Pulse Per Second
RGB	Red Green Blue (referring to the camera channels)
RGB-D	Red Green Blue + Depth (referring to the camera channels)
ROS	Robot Operating System
RS-485	Serial communication protocol
RTCM	Radio Technical Commission for Maritime Services (GNSS communication protocol)
RTK	Real Time Kinematic
SEE	Serial Elastic Elements
SLAM	Simultaneous Localization And Mapping
UHF	Ultra High Frequency
USB	Universal Serial Bus
VR	Virtual Reality
WiFi	Wireless network communication protocols
WP	Work Package

1 Introduction

"Human-Robot Communication specifications" is the task of WP6 focusing on the requirements of human robot communication to perform collaborative tasks in precision agriculture. The idea of having a joint effort of people and androids in the agriculture scenario could significantly reduce the time of executing particular tasks, so as to improve the efficiency of the overall system. Indeed, robots can support and relieve human operators, enable versatile automation steps and increase productivity by combining human capabilities with the efficiency and precision of machines.

Human-Robot Communication can rely on single sensory channels such as hearing, sight, speech and gestures or through the combination of two or more of them in order to obtain a more robust architecture. Signals that are exchanged along these channels must be coherent, since a clear understanding of their meanings is required to enable communication and to establish a more solid relationship between people and machines. To this aim, we look at the interaction from an agent communication perspective, by identifying different message types (*speech acts*) and characterizing their meaning to allow for a proper understanding of each other. This will allow us to present a detailed analysis of different communication modalities, from which we expect to find out the interaction channels that are more suitable for our application scenario.

In the following sections, we provide an analysis of the speech act categories based on the information content exchanged between humans and robots. Finally, we define the interaction modalities to be considered in developing Human-Robot Communication.

Figure 1. Representation of the resulting collaboration achieved through Human-Robot Communication.

2 Agronomic Actions, Activities and Goals

Vineyard farming in CANOPIES involves different actors, being them humans and robots. Humans include not only workmen but also experts such as agronomists. Robots are instead supposed to be of two different kinds:

a farming robot (F-R): which is dedicated to the execution of agronomic tasks such as harvesting the fruits or pruning the vines;

a logistics robot (L-R): for logistics tasks such as the transportation of boxes of harvested grapes or removed branches.

The farming robot will be equipped with two arms for executing its agronomic tasks in a human-like manner and an area with removable boxes for output – for example the harvested fruit. The logistics robot will be equipped with several removable boxes, a mechanism for taking a full box from the farming robot and replacing it with an empty one, and a good navigation system.

The collaboration between humans and robots happens in the context of the different phases of vineyard farming, namely pruning and harvesting. Pruning consists of managing the vineyard, whereas harvesting concerns collecting clusters once ripe.

In this section, we will distinguish basic (atomic) operations from composed operations. To this aim, we will call the first *actions*, whereas the second will be called *activities*, using a terminology typical of smart spaces. Activities are made up by actions, and the exact sequence of actions performed depends on contextual conditions. **The specific communication language for actions will be the topic of Task 6.3, whereas how to describe activities will be the topic of Task 6.4.** In particular, activities will be modeled by employing techniques typical of the Business Process Management world, such as Petri Nets, or more declarative formalisms.

The list of actions and activities presented in this section must be considered as an initial list, subject to evolution during the project, due to the possible emergence of new needs in the validation phases.

For the benefit of the reader, we also include in this deliverable Figure 2, depicting the structure of a grape cluster (left) and of a typical trellis (right, also called “Pergola system”).

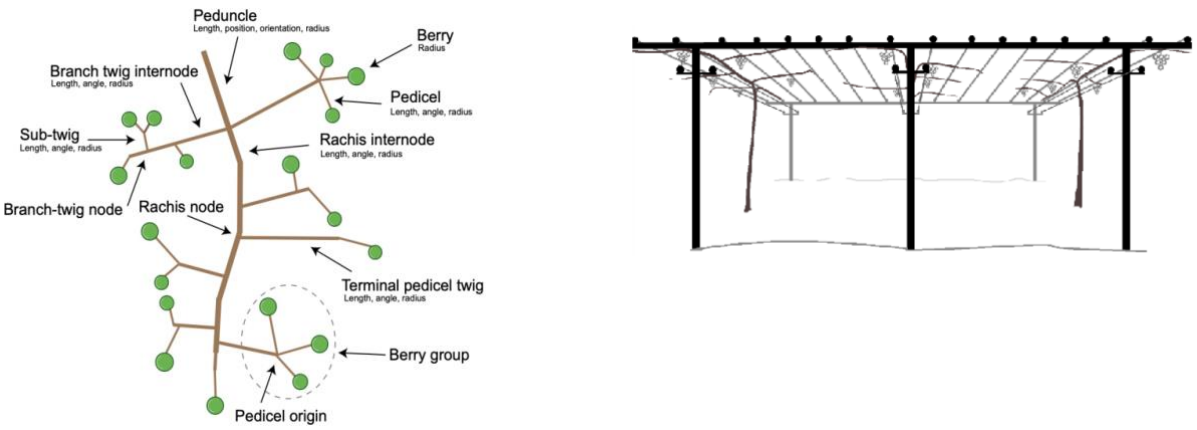

Figure 2. The structure of a grape cluster (left) and of a typical trellis (right, also called “Pergola system”)

2.1 Actions

In this subsection, we introduce an incomplete (more can be added in the following months) list of actions (not including communication actions) to be considered for communication purposes. Table 1 consists of 5 columns. The first column contains a brief description of the action, the second one specifies to which agronomic task (Pruning, Harvesting or Any) the action belongs. The next three columns specify whether a certain action can be performed, in the context of the project by logistic robots, farming robots and/or humans. However, some actions can be executed by multiple categories of actors. The particular conFiguretion of the smart vineyard will specify which actor performs which action.

Action	Agronomic Task	Logistic Robot	Farming Robot	Human
Movement in space	Any	✓	✓	✓
Tool (e.g., cutting device) placement	Any		✓	
Following	Any	✓	✓	✓
Pointing	Any		✓	✓
Removing leaves, branches and stones from the ground	Any	✓	✓	✓
Put leaves, branches and stones in a box	Any		✓	✓
Take a box to destination	Any	✓		
Take a box close to the worker	Any	✓		
Removing leaves	Pruning		✓	✓
Removing branches	Pruning		✓	✓
Sprout binding	Pruning			✓
Small berries removal	Pruning			✓
Mouldy berries removal	Pruning			✓
Stained berries removal	Pruning			✓
Dry berries removal	Pruning			✓
Identify hidden clusters	Harvesting		✓	✓

Check grapes cluster color	Harvesting		✓	✓
Compute cluster Brix index	Harvesting		✓	✓
Identify bad cluster	Harvesting		✓	✓
Identify good cluster	Harvesting		✓	✓
Peduncle detection	Harvesting		✓	✓
Harvest a cluster	Harvesting		✓	✓
Put cluster in box	Harvesting		✓	✓
Evaluating number of remaining clusters	Harvesting		✓	✓

Table 1. List of agronomic actions

2.2 Activities and Goals

In this subsection, we introduce a not complete list of activities (the full list will be defined by Task 6.4). Table 2, in particular, includes two columns. The first column briefly describes the activity, whereas the second one introduces examples of actions (taken from Table 1) involved in the execution of the specific activity.

Each activity has a specific goal and describes which actions are involved and in which order they are executed according to conditions presented over the context. In particular, activities may be specified using Petri Nets or other representation formalisms from the field of BPM. Representing activities using formal languages, allows for reasoning and automation, which are among the main challenges of the project. **The exact structure of activities is left to Task 6.4, together with the choice of the specific formalism to be employed.**

Activity	Examples of involved actions
Main fruiting cane pruning	Removing leaves, identification of a main fruiting cane, ...
Identifying a ripe field or lane	Compute cluster Brix index (several times), Check grapes cluster color
Cleaning the vineyard	Removing leaves, removing branches, ...
Quality assurance	Small berries removal, stained berries removal, moulded berries removal, dry berries removal, ...
Harvesting a ripe field of lane	Peduncle detection, harvest a cluster, put

Activity	Examples of involved actions
	cluster in box, take a box to destination, take a box close to the worker, pointing, …
Cleaning up a lane	Movement, tool placement, ….

Table 2. Preliminary list of activities

3 Communication act categories

Starting from the definition of "speech act", we will describe the proposed communication act categories. Generally, the term *speech act* is used to identify a set of classes that differ according to **what** is the informative content exchanged. However, the theory concerning speech acts was firstly introduced by Oxford philosopher John Langshaw Austin, who provided a clear and exhaustive definition of this concept, by identifying the speech act theory as "a subfield of pragmatics that studies how words are used not only to present information, but also to carry out actions". The introduction of this completely innovative and revolutionary theory, led researchers and experts to carry out various studies during the years to better investigate and deeply understand human communication. An important aspect of the speech act theory concerns the conducted analysis, mostly focused on human-human interaction that led to identify specific categories. Moreover, speech act theory has been the basis for the development of protocols for agent communication, from KQML[1] to the current FIPA standard[2].

Choosing the best interaction modalities and protocols in HRI is a challenging task, as it strictly depends on the application domain, the tasks to be executed, the types of robots and sensors involved. In the last years, a growing number of HRI researchers exploited VR as a means to evaluate proposed solutions, focusing in particular on safety and correctness of collaborative tasks. This allows to prove the effectiveness and robustness of a certain approach in a simulated environment, thus permitting to converge more easily to the best solution, also avoiding to experiment potentially harmful actions in a real scenario.

Starting from this background and based on the informative contents that emerge from the existing works concerning HRI studies in VR, we propose the categories depicted in Figure 3, as the ones that should be considered in the development of Human-Robot Communication in CANOPIES. Each of the presented speech acts is described in a dedicated subsection. In addition, a further subdivision, based on the specific message type, is performed and some examples are also provided in order to clarify the differences among the subcategories.

Results of this analysis have been also presented at the 8th International Conference on Augmented Reality, Virtual Reality and Computer Graphics (AVR 2021).

[1] FININ, Tim, et al. KQML as an agent communication language. In: Proceedings of the third international conference on Information and knowledge management. 1994. p. 456-463

[2] FIPA "FIPA 2000 Specifications", available at http://www.fipa.org.

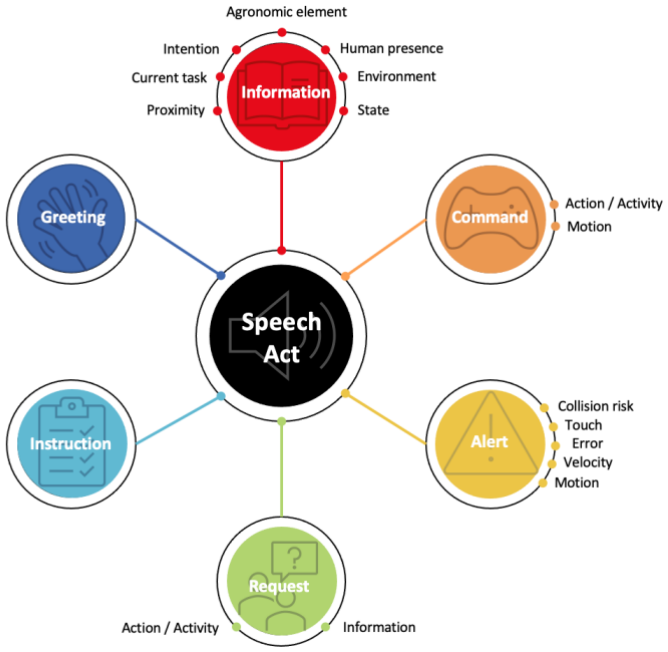

Figure 3. Identified speech act categories

3.1 Information

One of the most frequently adopted speech acts concerning VR experiments in HRI is *Information*, allowing both humans and robots to notify the teammate(s) about their status, particular events and situations on the field. In this case, bidirectional communication (from androids to people and vice versa) is considered for the majority of the proposed subcategories.

3.1.1 Intention

This class allows humans and robots to anticipate their next action/activity to the teammate, in order to facilitate the collaboration in the shared environment. Such subcategory would particularly increase the trustiness between people and machines, and finally the efficiency of the overall system. Knowing in advance the intention of the teammate, would also be beneficial from the point of view of awareness and safety. In particular, this would avoid the possibility of collision between humans and robots. Another interesting aspect to examine that would surely contribute in increasing human awareness and trustiness would be the projection of the robot's trajectory on the ground (either for farming or logistic robot). Such information is necessary for both people and machines, in order to plan their motion without interfering with the teammate's task and in particular, to avoid an unexpected collision with a person and with any table-grape vineyard component. For this reason, the trajectory of the farming and logistic robots should be always available to human workers, so to have under control machines motion and ensure the integrity of the agricultural field (by avoiding the damage of the crop and more generally, of the vineyard).

However, to highlight the importance of this class, we are providing some examples of an interaction concerning the exchange of intention information on the field. Each of the presented scenarios specifies the type of the speech act involved, followed by the related subcategory. In the round brackets, the parameters (from, to, what) are provided.

Scenario 1

Information_Intention (human, robot, moving left)

3.1.2 Current task

The "current task" subcategory is necessary to have a clear understanding of what the human and the robot are doing at a certain time. This is particularly important in order to evaluate whether the teammate requires help in completing a specific task. Such a situation leads both the robot and the human to redefine their planning, by giving higher priority to the actions/activities that need to be completed first. Less priority will be given to other assignments, in order to reduce the overall time required to complete the task involving the worker (either the robot or the person) and allowing them to proceed forward together. A fundamental aspect coming out from this subcategory is the concept of *support*, that emerges between people and machines in the shared environment. Their collaboration allows to establish a more solid relationship among them, from which the entire system will benefit from.

An example of a communication between a human and a farming robot about activity information is provided below:

Scenario 2

Information_CurrentTask (robot, human, leaves removal)

3.1.3 Proximity

Proximity information is also an extremely relevant concept in CANOPIES. Human-Robot distance to perform tasks in a shared scenario is required in order to ensure safety and trustiness. Avoiding dangerous situations is one of the main aspects to take care of when developing a communication system. Both humans and robots need to be informed about their distance from each other and notifications about the depth between the cutting device and the cluster must be provided, allowing the robot to approach or move away from the grape bunches. In this case, the interpersonal distance has a key role in defining the human-robot distance when executing a shared task, in order to avoid robots entering human's private areas with a consequence reduction in trustiness and safety. Figure 4(a) shows 4 zones emerging from proxemics studies conducted on social robots, which are: Intimate, Personal, Social and Public Distances. Considering that we would like to identify the suitable distance to ensure an efficient collaboration between humans and robots, the proxemics representation in Figure 4(b) better describes our goal.

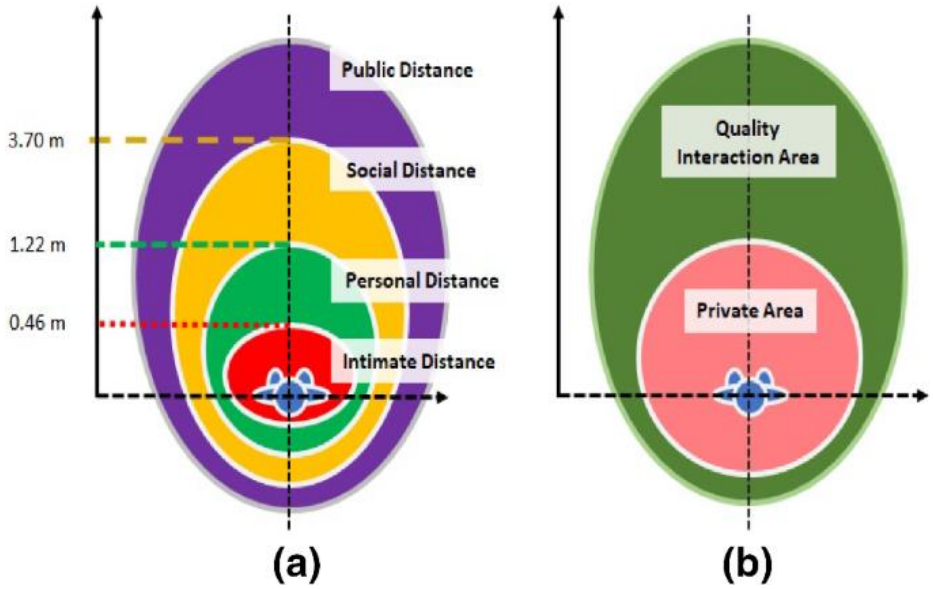

Figure 4. In (a) human interaction area according to proxemics[3], while in (b) interaction area based on the quality of interaction and human privacy is presented[4].

Moreover, the box-filling process is one of the tasks that is present in both harvesting and pruning agronomic activities. The position of the logistic robots on the field should facilitate the insertion of the harvested grapes or removed branches into the boxes. For this reason, the worker's proximity to the logistic robot should also be known, so as to allow people to provide commands and change the robot's position by considering their needs and preferences. However, an analysis concerning the position of the peduncle, grape clusters, logistic and farming robots with respect to the human worker, needs also to be introduced in CANOPIES to ensure a more precise interaction. The peduncle is the part of the grape cluster's architecture that requires particular attention. Indeed, the worker (either human or farming robot) should carefully move the harvested grape cluster into the box by holding it from the peduncle. The identification of clusters hidden by the leaves is one of the tasks that require human support and for this reason, the farming robot is helped in detecting their location and in finding the optimal cutting pose by considering multiple clusters. Particularly relevant is also the position of the two kinds of robots with respect to the human. Examining the case of a logistic robot, its location with respect to human position needs an in depth analysis. For example, for a right-handed person that would cut branches or grape clusters by using the right hand, it would be desirable to have the logistic robot on the left, so as to easily place the crop in the box; vice versa for a left-handed person. A similar consideration can be done with the position of the farming robot with respect to the human. In fact, there is a higher probability that a right-handed person will rotate the body counterclockwise to reach a hidden grape cluster in order to cut it, while it is more common for a left-handed person to rotate his/her body clockwise. Taking this into account, it would be

[3] Edward H (1969) The hidden dimension: man's use of space in public and in private. The Bodley Head Ltd, London.

[4] Patompak, P., Jeong, S., Nilkhamhang, I. *et al.* Learning Proxemics for Personalized Human–Robot Social Interaction. *Int J of Soc Robotics* 12, 267–280 (2020).

interesting to better investigate human behaviour and preferences in such situations. Summarizing, the distances that would be analysed to develop a robust interaction are presented below:

- farming robot - human
- logistic robot - human
- logistic robot - farming robot
- farming robot - logistic robot
- farming robot - grape clusters
- farming robot - branches
- farming robot - leaves
- farming robot - vineyard components (e.g. pergola system)

However, the angular distances would also be essential in defining a robust system. Indeed, the angular distance would be helpful for the robot to align its body or the arm with a specific grape cluster or branch. For this reason, rotation of a particular joint on the robotic arm, the torso and the head will be analysed.

An example of a communication between a human and a farming robot about proximity is provided below:

Scenario 3

Information_Proximity (human, robot, moving close to the grape bunches)

Scenario 4

Information_Proximity (human, robot, moving on human's right side)

Scenario 5

Information_Proximity (human, robot, approaching the peduncle behind the leaves)

3.1.4 Agronomic element

Details about the peduncle, branches to be removed, grape clusters (ripe, good quality and the specific type of grape to harvest) are needed to notify the teammate about additional information concerning a certain agronomic element that could determine the harvesting (grapes) or pruning (branches). The colour, the shape and the size of the clusters that are obtained through image evaluation by the farming robot are extremely important. In this case, the android is supporting the human by providing further information to those available to the person. The specific type of grape to harvest is a concept that must be shared by humans and robots, because on the field, it is common to find corridors where two different types of grapes are present. Multiple clusters are extremely important to be detected, so to find the optimal cutting position for the majority of them, with the aim to reduce the time required for grape harvesting. Necessary information for grape harvesting and branch pruning concern boxes description that are transported by logistic robots. In particular,

it is of interest to know the capacity of the boxes and the possible number of grapes/branches to insert before a full state is reached.

An example of a communication between a human and a farming robot about an agronomic element is provided below:

Scenario 6

Information_AgronomicElement (robot, human, specific grape cluster)

Scenario 7

Information_AgronomicElement (human, robot, pizzutello grapes)

3.1.5 Human presence

Robot's awareness of the human is a central part that needs to be considered in developing the entire system. In all the tasks, the presence of the human has a key role. Indeed, the aim of the person is, in some cases, to support and to guide the robot in the execution of agronomic assignments. Human-Robot collaboration on the field is the main concept that requires people and machines to work together on a specific task in a shared workspace and on which CANOPIES is based. Considering that multiple robots (both farming and logistic ones) and many people are present in the environment at the same time, the aspect of human presence becomes particularly important. For this reason, it would be useful to develop a system in which the farming robot is able to provide a notification about the presence of a human to its teammate when another person is approaching them.

An example of a communication between a human and a farming robot concerning human presence is provided below:

Scenario 8

Information_HumanPresence (human, robot, X needing help)

Scenario 9

Information_HumanPresence (human, robot, human approaching from the right)

3.1.6 Environment

Environmental information subcategory is particularly useful on the field to make some considerations about the state of the grape clusters, the ground, that would influence the harvesting or pruning activity. For example, the humidity percentage and the weather conditions could affect the color of the berries, therefore the workers (both humans and robots) should take them into account to determine if a grape cluster is ripe and ready to be harvested or if it still needs some days/weeks to be completely ripe. Another interesting aspect concerning environmental information, is the possibility that a human worker can notify the robot about the presence of leaves, stones or other obstacles. The necessity of exchanging weather information between people and machines is relevant, in particular when a sudden change in meteorological conditions is expected. Indeed, it is suggested to stop the harvesting and/or pruning activity before a worsening of weather. In addition, the temperature information is also extremely important in this case. For example, in the early

morning, there is higher probability to find dew drops on leaves and in particular on berries, so in the identification of high quality clusters, the farming robot must also consider the temperature and the moment of the day (morning, afternoon, evening, night). The time is a further relevant notification for the human worker that could be obtained by information exchange with the robot. Hence, to help the human worker and to avoid the person wearing multiple devices, we believe that the information exchanged between people and androids could provide many advantages to the entire system.

Scenario 10

Information_Environment (robot, human, time of ripeness for a specific grape)

Scenario 11

Information_Environment (human, robot, presence of stones on the raw)

Scenario 12

Information_Environment (human, robot, presence of a heap of leaves)

Scenario 13

Information_Environment (human, robot, presence of dust on a ripe grape cluster)

Scenario 14

Information_Environment (robot, human, current temperature)

3.1.7 State

3.1.7.1 Robot Activity-related State

Information that could be of help to humans concerns the state of the robots. In this case it could be useful to investigate the situations associated with both, the farming and logistic androids. This category of communication acts is strongly related with the "Current task" category. In particular, while performing a specific activity (as defined in Section 2), it can be useful for the human or the robot to know additional information about the current activity carried on by the other to obtain a clearer understanding of the current condition of the counterpart.

An example of a communication between a human and a farming robot concerning the robot's state is provided below:

Scenario 15

Information_State (robot, human, listening state)

An example of a communication between a human and a logistic robot concerning the robot's activity-related state is provided below:

Scenario 16

Information_State (robot, human, full box)

3.1.7.2 Robot and Human Internal State

It is important for the human to know what is the internal state of the robot. In particular, if any internal problem (e.g., a hardware failure is detected), it must be immediately communicated to the human to decide about stopping action to avoid further potential damages to the robot or harmful behaviour and fix it.

Information about the state of the battery is required, so to allow the human to always have under control the battery percentage. It should be useful for the person to obtain a notification from the robot (both farming and logistic), whether any kind of problem arises (e.g. the battery is damaged). The robots state battery charging is another relevant concept that has the necessity to be communicated by androids to humans. Indeed, it would be useful for people to be notified when a robot is completely charged or when it has low battery, needing to be charged soon.

The same holds for the human who could communicate with the robot about his health status (e.g., I do not feel good).

An example of a communication between a human and a robot (either farming or logistic) concerning the state of the battery is provided below:

Scenario 17

Information_State (robot, human, battery completely recharged)

3.1.7.3 Robot Understanding

The state of robot interaction understanding could be of interest to know if the robot understood the information exchanged by the human. Percentage of the recognized information would help people in communicating with robots. Especially, in case of a vocal interaction, the person is aware whether the robot recognized what he/she said or not. A similar consideration can be applied on a human communication through gesture, for which the android would provide a notification concerning what it understood from the human pointing. However, it would be useful for the person to receive information about words not correctly recognized by the robot in case a vocal interaction is adopted (presenting higher uncertainty).

An example of a communication between a human and a farming robot concerning the state of robot understanding is provided below:

Scenario 18

Information_State (robot, human, certain word not correctly understand)

3.2 Command

An unidirectional speech act category, expressing the commands given by the human to the robot, is represented by the command speech act class. This group of information contains all those messages that **must** be executed by the robot, differently from “Request” (presented below), in which the android has the possibility to avoid executing the request of the human.

3.2.1 Action/Activity

Commands are mainly used to let a robot perform specific actions or activities as defined in Section 2. Examples include commanding the robot to follow a person, to harvest a specific grape cluster or to prune a certain branch. The "follow" command would lead the robot to observe a person and move after him/her to a certain location. Indeed, it could be necessary in case the human is not able to precisely describe the destination to the robot and he/she prefers to guide it to the goal. Command information concerning the specific grape cluster to harvest or the particular branch to prune allows the person to support the android in its tasks.

An example of a communication between a human and a farming robot concerning action command is provided below:

Scenario 19

Command_ActionActivity (human, robot, pruning a specific branch)

3.2.2 Motion

Even though movement has been denoted as an action in Section 2, motion deserves a specific category in our classification, as several kinds of motions exist, and the communication modalities can be different from other actions. Reaching a target or simply a new position, changing the velocity or stopping the current operation are some of the motions that can be executed by the robot (both farming and logistic). As the term suggests, this subcategory contains all the messages that could lead the android to modify its current state, both in position and velocity. An increase of velocity instead, would be necessary to move from one corridor to another, in order to reduce the time of reaching the new grape clusters to harvest or set of branches to prune. Reduction in velocity could be necessary in case the robot is approaching a grape cluster or a branch before cutting it or its hand (in case of a farming robot) is moving between multiple clusters. However, in case an anomalous situation occurs, it should be possible for the human to provide a stopping command to the robot, so as to avoid damaging the field. Anyway, a new position to reach should be given by the human to the robot in order to facilitate its task or guide it to obtain better detection results by looking from a different point of view.

An example of a communication between a human and a farming robot concerning motion command is provided below:

Scenario 20

Command_Motion (human, robot, moving forward)

3.3 Alert

A notification of a dangerous situation is a fundamental aspect that needs to be analysed in Human-Robot Communication in CANOPIES. For this reason, the bidirectional alert speech act category is provided.

3.3.1 Collision risk

Information about collision risk is required in systems where humans and robots must collaborate in executing tasks together, to increase human safety and to avoid the robot damaging the working environment. Hence, notification of higher probability of human-robot collision, branch or grape cluster collision, and pergola collision is needed in order to alert the human about dangerous situations. On the other hand, it would be interesting to collaborate with a robot that can alert the person about a collision risk. The entire system would surely benefit from this subcategory, in particular considering a bidirectional communication, that would ensure a more robust and safe architecture, where humans and robots can trust each other.

An example of a communication between a human and a farming robot concerning a collision risk alert is provided below:

Scenario 21

Alert_CollisionRisk (human, robot, probable collision with the grape cluster)

3.3.2 Touch

The touch subcategory is particularly useful to inform the human worker when a collision with an agronomic element (grape cluster, branch or pergola) is verified. Such conditions should be associated with an immediate notification, so the person could intervene by stopping the robot's activity in the worst case (so as to avoid more serious damage). In this case, higher probability of accidental collisions could be associated with the robot than with humans, so it is key for the person to react quickly to unpredicted touches.

An example of a communication between a human and a farming robot concerning a touch alert is provided below:

Scenario 22

Alert_Touch (robot, human, impacting with the grape cluster)

3.3.3 Error

A further relevant subclass is represented by "error". As the name can suggest, it provides information about errors encountered during a task execution or in human understanding. For this reason, errors can be divided into two subcategories: execution and interpretation errors. Generally, the first class identifies issues associated with a correct understanding of the request, but a wrong execution of the action; while the second one is strictly related to wrong understanding of a sentence, a gesture, a visual or audio signal that will negatively affect the outcome of the action execution. For these reasons, Human-Robot Communication must allow the exchange of all kinds of notifications, in particular the ones concerning sudden errors that could increase the time of executing an assignment and consequently reduce the overall efficiency of the system. Mainly, such information is associated with androids and it is necessary for humans to be aware of them when problems arise, by supporting the robot to overcome the encountered issue as soon as possible. For this reason, errors in human

understanding could be possible and it is important to have a robot with the ability to notify the person about it, in order to avoid uncertainty in the communication, which could lead to dangerous situations. Considering all the agronomic actions in Table 1, it is clearly noticeable that all of these actions can be subject to execution or interpretation errors (mainly by the robot). More details about possible errors are provided in Table 3.

Action	Interpretation Error	Execution Error
Movement in space	Error in understanding the position/direction of motion	A motion to a wrong position or in a different direction is performed.
Tool (e.g., cutting device) placement	Error in understanding the position/direction of the cutting device	The cutting device is positioned wrongly
Following	Error in understanding the person to follow	Motion stopped or a wrong person is followed
Pointing	Error in understanding which is the target element	A wrong element is pointed
Removing leaves, branches and stones from the ground	Error in understanding what elements must be removed	Not all the leaves, branches and stones are removed
Put leaves, branches and stones in a box	Error in understanding what are the elements to consider and where locate them	Error in selecting and positioning the agronomic element
Take a box to destination	Error in understanding what to take and where is the destination	Error in taking the box and reaching a wrong destination
Take a box close to the worker	Error in understanding what to take and where is the destination	Error in taking the box and reaching a wrong destination
Removing leaves	Error in understanding what to remove	Missed leaves or wrongly cut
Removing branches	Error in understanding what to remove	Missed branched or wrongly cut
Sprout binding	Error in understanding what	The sprout binding is wrongly

	to bind	executed
Small berries removal	Error in understanding what to remove	Missed small berries or wrongly removed
Mouldy berries removal	Error in understanding what to remove	Missed mouldy berries or wrongly removed
Stained berries removal	Error in understanding what to remove	Missed stained berries or wrongly removed
Dry berries removal	Error in understanding what to remove	Missed dry berries or wrongly removed
Identify hidden clusters	Error in understanding what to identify	Not detected or missed hidden clusters
Check grapes cluster color	Error in understanding what to check	The cluster color is wrongly detected or the action is not performed
Compute cluster Brix index	Error in understanding what to compute	The brix index is wrongly computed or the cluster is not correctly selected
Identify bad cluster	Error in understanding what to identify	Not detected or missed bad clusters
Identify good cluster	Error in understanding what to identify	Not detected or missed good clusters
Peduncle detection	Error in understanding what to detect	Problem with the cameras in detecting the peduncle or a wrong grape part has been considered as the peduncle
Harvest a cluster	Error in understanding what to harvest	A wrong cluster is harvested rather than the requested one or the cluster is not harvested
Put a cluster in the box	Error in understanding what and where locate the agronomic element	Error in selecting and positioning the agronomic element
Evaluating number of remaining clusters	Error in understanding what to evaluate	The number of remaining clusters is wrongly computed

Table 3. List of agronomic actions with corresponding probable errors

An example of a communication between a human and a farming robot concerning an error alert is provided below:

Scenario 23

Alert_Error (robot, human, wrongly understood a word)

Scenario 24

Alert_Error (robot, human, throwing away a good grape cluster)

3.3.4 Velocity

Robot's notifications about speed increase or decrease are necessary to inform the person about changes in velocities. This subclass is particularly useful to improve safety and trustiness in human-robot collaboration tasks. Indeed, the worker on the field must always be aware of what the robot is doing and what is happening, so to stop the android's action, if necessary.

An example of a communication between a human and a farming robot concerning velocity alert is provided below:

Scenario 25

Alert_Velocity (robot, human, increasing velocity to reach the end of the row)

3.3.5 Motion

A further useful subcategory to increase safety is represented by "motion". An immediate change in robot's trajectory is a relevant aspect that should be notified by the android to the human worker. Indeed, it is required for the person to know in advance the intention of the machines and, before a motion execution, he/she is alerted. The human instead, should inform the robot about dangerous movements, so as to allow the android to replan its motion trajectory.

An example of a communication between a human and a farming robot concerning motion alert is provided below:

Scenario 26

Alert_Motion (robot, human, farming robot's body rotation of 180 degree on the right)

3.4 Request

Another relevant unidirectional communication class, which has not to be confused with the already introduced command category, is request. The difference is straightforward: in the last case, the human is asking the teammate to perform a specific action, selection or provide information, but the robot can refuse or postpone it while, when a command is issued, the android must follow the person's requirement.

3.4.1 Action/Activity

This subcategory includes actions and activities as defined in Section 2. In particular, the robot can be asked to execute a particular operation, such as moving behind the person until a location is reached, so as to easily guide the android to the final destination and avoiding uncertainty in

explaining or showing the path. Other relevant requests could concern grape harvesting and branch pruning. In this case, the robot will be asked to concentrate on a specific portion of the crops, so as to parallelize the work and reduce the time needed to complete the operations on a corridor. However, dirt removal (broken branches or fallen leaves) could also be requested to the robot. The person must be able to traverse the corridor in its length without any problems. Hence, for this reason, the android needs to be aware of the aforementioned situations on the field, so as to ensure human safety. Noteworthy the actions/activities included in this category are a subset of the actions/activities for which a robot can instead receive a command communication act (see Section 3.2.1)

Examples of a communication between a human and a farming robot concerning action request are provided below:

Scenario 27

Request_ActionActivity (human, robot, human's follow to reach the other row)

Scenario 28

Request_ ActionActivity (human, robot, selection of all the branches to remove)

3.4.2 Information

The information subclass has not to be confused with the Information speech act category presented at the beginning of Section 2. This paragraph illustrates the concepts of possible human requests concerning notifications about the environment, the robot's state, intention and trajectory, human-robot proximity, properties and pose (position and orientation) of agronomic elements.

An example of a communication between a human and a farming robot concerning information request is provided below:

Scenario 29

Request_Information (human, robot, time)

3.5 Instruction

A further unidirectional speech act class is represented by "instruction", by which the machine provides instructions or hand motion suggestions to be performed by the human worker to complete an assignment, to reach a certain location through the shortest path and motions to perform in order to reach grape clusters hidden by the leaves. Indeed, such messages are useful to support humans in both common and difficult activities. Instruction information can be provided to the person in various ways, through a visual representation, spoken language or by combining multiple channels, with the aim to support the human during his/her work.

An example of a communication between a human and a farming robot concerning instruction information is provided below:

Scenario 30

Instruction (robot, human, approaching the hidden grape cluster from the right)

Scenario 31

Instruction (human, robot, teaching the speed to keep)

Scenario 32

Instruction (human, robot, recognition of ripe bounches)

Scenario 33

Instruction (human, robot, best pose to reach in a specific situation)

Scenario 34

Instruction (human, robot, harvesting the grape clusters on the right)

3.6 Greeting

A category that emerges when interacting with social robots is greeting. Indeed, the robot can welcome the user to increase human's trust, which is fundamental in collaborative tasks. The android should greet the person by considering the current part of the day (morning, afternoon, evening or night) and whether the human is starting the working day or he/she has already finished it. A friendly robot would surely strengthen the relationship between people and machines, by improving the collaborative work.

Scenario 35

Greeting (robot, human, greeting the teammate)

Scenario 36

Greeting (human, robot, greeting the teammate)

3.7 Examples of combined speech-act interaction

In this Section, examples of Human-Robot Interaction through a combination of multiple speech acts are presented. For each scenario, the involved speech act categories are provided in the round brackets.

Example 1 (Request + Information)

Information_CurrentTask (robot, human, current task)

Request_Information (robot, human, finishing harvesting the clusters of the row)

Information_CurrentTask (human, robot, harvesting not completed)

Request_ActionActivity (human, robot, harvesting grape clusters of the row)

Example 2 (Request + Instruction)

Request_Information (robot, human, reaching the hidden cluster)

Instruction (human, robot, cutting the leaves surrounding the cluster)

Example 3 (Command + Alert)

Command_ActionActivity (human, robot, harvesting the grapes at the end of the row)

Alert_Motion (robot, human, problem with the motor)

Example 4 (Instruction + Alert)

Instruction (human, robot, teaching the specific grape type to harvest)

Alert_Error (robot, human, camera sensor not working)

Example 5 (Instruction + Information)

Instruction (human, robot, how to put the harvested grape cluster in the box)

Information (robot, human, learning the movement)

Example 6 (Request + Instruction + Greeting)

Greeting (robot, human, greeting)

Request_ActionActivity (robot, human, teaching how to reach the field)

Greeting (human, robot, greeting)

Instruction (human, robot, exiting the Cooperativa Agricola Corsira, moving on the right and then forward)

4 Environmental constraints

One of the most important aspects to consider when developing the communication in CANOPIES concerns the limitations of the environment. The introduction of autonomous robots in outdoor scenarios, such as in vineyards, is particularly difficult due to multiple constraints: changes in weather conditions, variations of the illumination, acoustic noises, changes in the state of the ground and presence of broken branches and stones on robot's trajectory. All of these odds influence the ability of robots and humans to understand mutual communication. As an example, acoustic noise may reduce voice communication understanding, whereas strong illumination conditions may reduce the effectiveness of light signals employed for communication.

4.1 Noise

In vineyards, apart from humans and robot workers, agricultural machines are also employed. For this reason, our system should be tested in such scenarios to have a clear understanding of the type of communication to adopt, so as to overcome uncertainty or misunderstandings during the interaction. Indeed, the noise of these vehicles and the sound generated by the motor of the robots are particularly loud, causing problems with Human-Robot Communication. The rustle of the leaves originated by the wind, the noise generated by walking either on the fallen dry leaves or on the cutted branches, are all examples of 'natural' sounds that also require an analysis related to how they could influence the interaction. However, from the robot side, a solution for the aforementioned problems

could be to filter only the human voice that is closer to the android (with higher frequency) and discard all other noises in the background (e.g., by using a noise suppression microphone).

4.2 Illumination

Variations of illumination are another type of environmental constraints concerning outdoor scenarios. The luminosity changes are related to the position and intensity of the light source, the season and the weather. Indeed, more shadows and lower intensity sunrays characterize two parts of the day: early morning and evening, while in the late morning/afternoon the intensity of the sunrays is higher and the shadows are almost absent. Despite the presence of the pergola system, which protects the grape clusters from the sunburn and limits the passage of the sunrays, it could happen that some bunches or a portion of a corridor is widely illuminated, while other portions are temporarily occluded, causing detection issues.

4.3 Ground

The typology of the ground is a fundamental aspect to consider as an environmental constraint, because it could influence the robot's velocity of motion. Hence, two kinds of ground limitation should be investigated: sand and wet. In the worst situation, the robot could get stuck, requiring human support and guidance to escape from such stationary condition. Despite the presence of the pergola system, it is common to find a particularly wet ground (such terrain is very difficult for the robot to navigate) after a heavy rain and a sandy terrain when the temperature is particularly high for several days (mainly during the summer season).

4.4 Weather (windy, rainy, humidity)

Changes in weather conditions could lead to a highly variable agricultural operational environment that the robot needs to be aware of. Relevant, in this case, would be the android's ability to know in real time possible changes in the meteorological conditions, by having direct access to the weather site. A rainy situation could be a problem for the robot for different reasons: the first one has been discussed in the paragraph "Ground" of environmental constraints, while the second is mainly related to the mechanical part that could be damaged by the water. The wind and the humidity are also important, because they could slightly change the color of the bunches and consequently, influence the detection of their ripeness. Indeed, the wind could lift the sand, covering the grape clusters with this thin yellow/orange layer. Instead, during the winter season (mainly in the early morning and in the evening), the humidity could provide a brighter color to the bunches, being covered with water droplets. Hence, in both the presented cases, the robot should be aware of the weather conditions, so as to adapt its next actions or modify its task priority based on the current situation.

Weather conditions may also negatively affect the communication ability of the robot. Dust, may for example cover sensors and emitters making it difficult to perform any kind of communication task with humans.

4.5 Branches and stones on the corridor

A fundamental feature of CANOPIES concerns the person assisting any robot (either farming or logistic) when it gets stuck. Cut or broken branches are common to be found on the corridor and for this reason, both the android and the human require to be aware of such obstacles. Higher probability

is given to the robot to get stuck when wide branches or heavy stones are present on its trajectory. Hence, the human worker should help the machine by guiding it to avoid such situations and, if another path is not available, the person should assist the android, by removing stones and branches that could cause motion problems to it, before the robot approaches the possible stuck position.

5 Communication modalities

Starting from the selected speech act categories and their classification based on the informative content, we devote this section to present different communication modalities that could be adopted in HRI. Our analysis is based on a preliminary study conducted on existing works concerning HRI experiments in VR, in order to find out how the informative content should be exchanged when a collaboration between humans and robots is required. For this reason, an in depth examination of the main interaction channels concerning voice, gesture, sound, visual signals and other devices is provided. However, smartphones, tablets and augmented reality technologies, such as smart glasses or head-up displays, could be adopted in CANOPIES in order to guarantee a safer and a stronger way to exchange information. Each of the aforementioned communication modalities will be described in a dedicated subsection, focusing on the strengths and weaknesses of each of them. In most cases, we would prefer to exploit the power of multimodality (combination of multiple interaction channels) in order to obtain a more stable, efficient and natural communication, by overcoming the problems related to a single channel.

5.1 Spoken language

The most natural way of communication in a human-human interaction is through voice, so we would adopt this interaction channel also in vineyards to allow the exchange of information between humans and robots. Indeed, from an analysis of HRI studies in VR, the spoken language emerged to be the most used medium for the information category, by updating the teammate about processes, so as to increase awareness of the system's features. However, the improvement of performances is another relevant aspect that we would achieve with CANOPIES. Such a result could be reached by clarifying ambiguous situations, giving suggestions to correctly complete a task and communicating the outcome of an action. Considering that one of our primary goals is to develop a robust and efficient communication system, a bidirectional vocal interaction is required, so that both the robot (either farming or logistic) and the human will be able to vocally exchange information. Generally, when designing a Human-Robot Communication, different aspects should to be taken into consideration, but in particular shared terms that are common to be found in a conversation concerning vineyard activities, such as: grape clusters, bunches, branches, harvesting, pruning. Indeed, voice is obviously the modality adopted to provide requests. In some situations, a teammate could require either a detailed description or properties information concerning the environment or a specific element of the vineyard. Frequently, requests of performing an action, selecting an object or information about other teammates tasks could be provided (mainly by a human worker). Summarizing, our idea is to develop a robot able to express concepts in a human-like manner, so that people would feel comfortable and safe. For this reason, we also would like to present a friendly and polite robot that greets a person when starting/finishing the working day and thanks the human worker for his/her support, trying to prioritize people's requests.

This notwithstanding, in outdoor scenarios such as in vineyards, voice is not always the best modality to be adopted due to the presence of noise, misunderstandings, high uncertainty, immediacy and, sometimes, naturalness. For these reasons, we would like to investigate the possibility of combining multiple interaction channels to benefit from the advantages that each of them could provide, overcoming all the problems that are related to the usage of a single modality. Usually, spoken language and gestures are combined together in systems that require interaction between people and androids, so to strengthen the communication.

5.2 Gestures

In a human-human conversation, gestures are often employed to clarify a concept, to better explain how to reach a certain place and to provide a command, in particular to little children, for example by pointing to the final destination to reach. Considering the study conducted on HRI experiments in VR, it emerged that gestures are the preferred communication modality for the command speech act category. Hence, when the interaction between humans and robots is required, as in CANOPIES, such channel could be exploited by a person in order to stop machine's activity in an immediate way by showing his/her opened hand to the mobile vehicle, but also to indicate the next position to reach by the android. As anticipated before, pointing could be employed in different tasks, not only to show a precise position to reach, but also to specify an element of the vineyard (a certain grape cluster, branch) or to indicate the person to follow. However, the introduction of gestures in our system requires the robot to be able to, firstly, detect the position of both human's arm and hand, secondly, to understand their meaning. Conversely, the android should also be able to communicate with the human through gestures, by translating a non-verbal behaviour into an understandable action.

Unfortunately, gestures alone would not be sufficient to develop a robust human-robot interaction, because they could cause misunderstandings and uncertainty. Without a vocal explanation, the meaning of some actions would be difficult to understand, as for example, "pointing" is not a self-explanatory gesture, so if a person points a branch without providing any command or information to the android, the robot would not know the action to associate to such interaction (cut, move, discard the branch). For this reason, in order to disambiguate the gestures, we would combine them with the spoken language, preferring a multimodal approach rather than a single communication channel. As an alternative, Sign Languages, as the International Sign Language could be also employed.

5.3 Sound

Sound signals constitute the emerging interaction channel for the alert speech act category from the analysis conducted on HRI studies in VR. In vineyards, it would be necessary to introduce sound signals on the robot side to alert human teammates about a dangerous situation. Different sound intensities could also be adopted, depending on human-robot distance, so when people and androids are particularly close to each other, a higher frequency signal will be emitted in order to alert the person of being in an unsafe area. For this reason, we would like to provide at least 3 zones that are respectively identifying a dangerous, medium and safe place. Considering a farming robot, a risky situation, with a very high frequency beeping, is encountered when the human worker is positioned inside the private area (see Figure 4b, where the private area has a radius of 122 cm, with 87 cm

representing the maximum length of the farming robot's arm) from the android. The medium area, characterized by a lower sound than the one in the red zone, delimites the person-machine distance between 122 cm and 370 cm (the limit of the so called social distance, see Figure 4a), while a greater value (> 370 cm) identifies a safe area, in which no sound signals are emitted by the robot. Different distances for the zone separation regarding logistic robots (they have no hands) may be considered. For example, in the case of a logistic robot, a dangerous position can be assigned to a human-robot distance which is lower than 100 cm, a medium one between 100 cm and 170 cm, while a safe place is determined by a distance greater than 170 cm. However, another relevant introduction in the system could concern a long audio signal that notifies when a contact between a robot and a human occurs. Furthermore, such interaction channel could be also adopted when a reduction/increment of android's speed is verified or a change in motion is performed, so as to avoid a risky situation. Generally, such communication modality is combined with visual signals to increase awareness and to provide to the human an imminent warning associated with a dangerous circumstance.

5.4 Visual signals

Light signals are emerging as visual feedback means from HRI studies in VR as the preferred modality for information and alert speech act categories. In developing an interaction based on visual signals in CANOPIES, the pergola system covers an important role. In particular, such a condition of having both branches and grape clusters in shadow will allow us to consider the introduction of light signals to exchange information between humans and robots. Indeed, this is a communication channel that is very sensitive to light intensity. In case the pergola system, which creates a shaded working area, would not be present, it would be likely trickier to adopt light signals as a medium of interaction. The intensity of the sun as a light source would interfere with the lights emitted by the robot, making them less visible and reducing the efficiency of this modality in such situations. Instead, without the pergola system, visual signals could be adopted only in case the working day would start after the sunset: in the winter season, it would coincide with 4:30 p.m. considering that the duration of the day is shorter; in the summer, instead, the daylight hours are more and the temperature rarely goes below 32°C during the day. For this reason, it would be interesting to investigate the possibility of starting the working day after 8:30 p.m., so as to benefit from the cooler hours of the night.

The introduction of visual signals in CANOPIES would surely improve human awareness of robot's activities and increase trust among them. A red dot that increases in dimension when the android's hand is very close to the goal, could be adopted to provide information about the machine's intention, for example the next grape cluster to harvest or branch to cut. A relevant aspect that could be added to our system is related to human presence detection, through the usage of different colors exploited to categorize three zones: green, yellow and red. The green one would represent a completely safe area, a medium risk would be associated to the yellow region, while a higher hazard would be assigned to the red sector, in which the human and the farming robot are performing the collaborative task very close to each other, increasing probability of encountering a dangerous situation for the person.

Interesting could be the combination of visual and audio signals to exchange information from robots to humans. For instance, the association of a particular beeping frequency to one of the colored rings

(one for each zone), distinguished based on distance between people and androids in collaborative tasks would allow the human workers to understand the machine's state by considering the sound wave together with different shapes and colors of the signs on the robot. In one HRI experiment conducted in VR, advantages related to the simultaneous reproduction of a colored warning light (positioned in the bottom part of each of the two robotic arms) and a sound emerge. A reduction of mental workload of industrial operators, increment of the final performance and awareness have also been observed.

5.5 Other devices

Other devices to adopt for the communication between humans and robots will be another relevant aspect to investigate, considering that our system should benefit from such interaction channels in order to reduce the overall costs and the time required for harvesting and pruning by exploiting people-machines collaboration. For this reason, in the following subsections we discuss some of the possible solutions concerning the usage of external devices, highlighting pros and cons of each of them.

5.5.1 Smartphone and Tablet

Smartphones and tablets could be useful to guarantee a solid communication between humans and robots on the field when Wi-Fi connection seems to be unstable. In this case, we consider two solutions: either the hotspot or the bluetooth connection. However, the first one is based on the coverage of the telephone operator, so only if the phone has a good network coverage, a strong connection would be instantiated. Conversely, the interaction through bluetooth would be much more flexible, without the limitation of the network coverage, requiring only another device to be associated with the one providing the connection.

Smartphones and tablets need the person to stop his/her work and send information to the robot through such devices and it could lead to waste of time. In general, we would like to avoid interruptions of human work and ensure an interaction “hands-free”, allowing the person to continue the execution of his/her task.

5.5.2 Augmented reality technology

In the last few years, new technologies such as Virtual and Augmented Reality have been adopted in different fields. For this reason, it would be interesting to investigate if the introduction of AR devices could particularly improve awareness and grape harvesting in CANOPIES. Indeed, the goal of such technology is to extend the perception of the real world by introducing virtual information on top of it. To this aim, Augmented Reality Smart Glasses (ARSG) and Head-Up Displays (HUD) could be considered for the development of human-robot interaction in vineyards. Apparently ARSG seems identical to a normal pair of glasses, but with the power of showing extra data to the person that is wearing them. Generally, they are lighter than HUDs and it is an advantage considering that they have to be worn the whole day by the farmers. Another relevant aspect that needs attention is the power, that could be of two types: either through a battery pack or a computer. Obviously, the first option is essential in our scenario, as it is impossible for a farmer to carry a computer on the field in order to charge the batteries of the ARSG. In this case, the battery life also plays an important role. A human worker could wear the AR device throughout the working day and a durable battery is necessary to

enable this. For integrated batteries, the battery life must be at least nine hours, or if fast charging is available, four hours (since the battery can be recharged during the lunch break).

The field of view, denoting the area in which virtual objects can be observed via ARSG, is a crucial parameter to consider when choosing the AR device to adopt in a robotic system. The horizontal field of view is extremely important, as a large horizontal field of view makes it possible to display information on the periphery, keeping the center of view clear for seeing the real world. A human's natural field of view is almost 180 degrees horizontally, so the ideal device should have a field of view as close as possible to this value. Many AR devices implement various kinds of functionalities for controlling and interacting with both ARSG and HUD, as for example, voice control, either through built-in functionality or by integrating a microphone. Voice control support is clearly advantageous, as it makes it possible for the user to control the system while his/her hands are occupied. Some AR devices also support gesture control, in addition to voice, which is also a great advantage, because it is possible for the user to interact with the system in noisy environments where voice control is unsuitable. Supporting both voice and gesture control is doubtless a strength, as the benefits of both can be exploited.

Variations of illumination is another fundamental aspect to examine when choosing the more suitable AR device. Unfortunately, in an outdoor scenario, ARSG are more prone to suffer from problems related to the entrance of sunbeams, limiting their potential. Conversely, HUDs are less exposed to lighting problems, due to their closed structure that avoids sunbeams from entering in the device. Other relevant environmental issues (strictly related to vineyards) that could influence the performance of AR technology are dust and water. For this reason, if we decide to introduce AR devices in CANOPIES, we should consider all these features in choosing the more suitable system.

6 Implemented solution

In this section we describe the main architecture proposed in the context of task 6.3 and shown in Figure 5. The architecture covers both the human-to-robot and the robot-to-human interaction. The *human worker* sends communications acts to the robot exploiting the camera and the microphone mounted on it. These two sensor devices may serve as input for different recognition modules including (i) light signal recognition, (ii) gesture recognition, (iii) motion recognition, (iv) attention recognition, (v) speech recognition and sound recognition. Each of these modules is in charge of handling a specific human-to-robot communication modality. Of these modalities, after involving other partners in the decision, the speech recognition (described in Section 7.1) and the gesture recognition (described in Section 7.2) modules have been implemented, but the framework has been designed for extension with the other modalities in the months ahead.

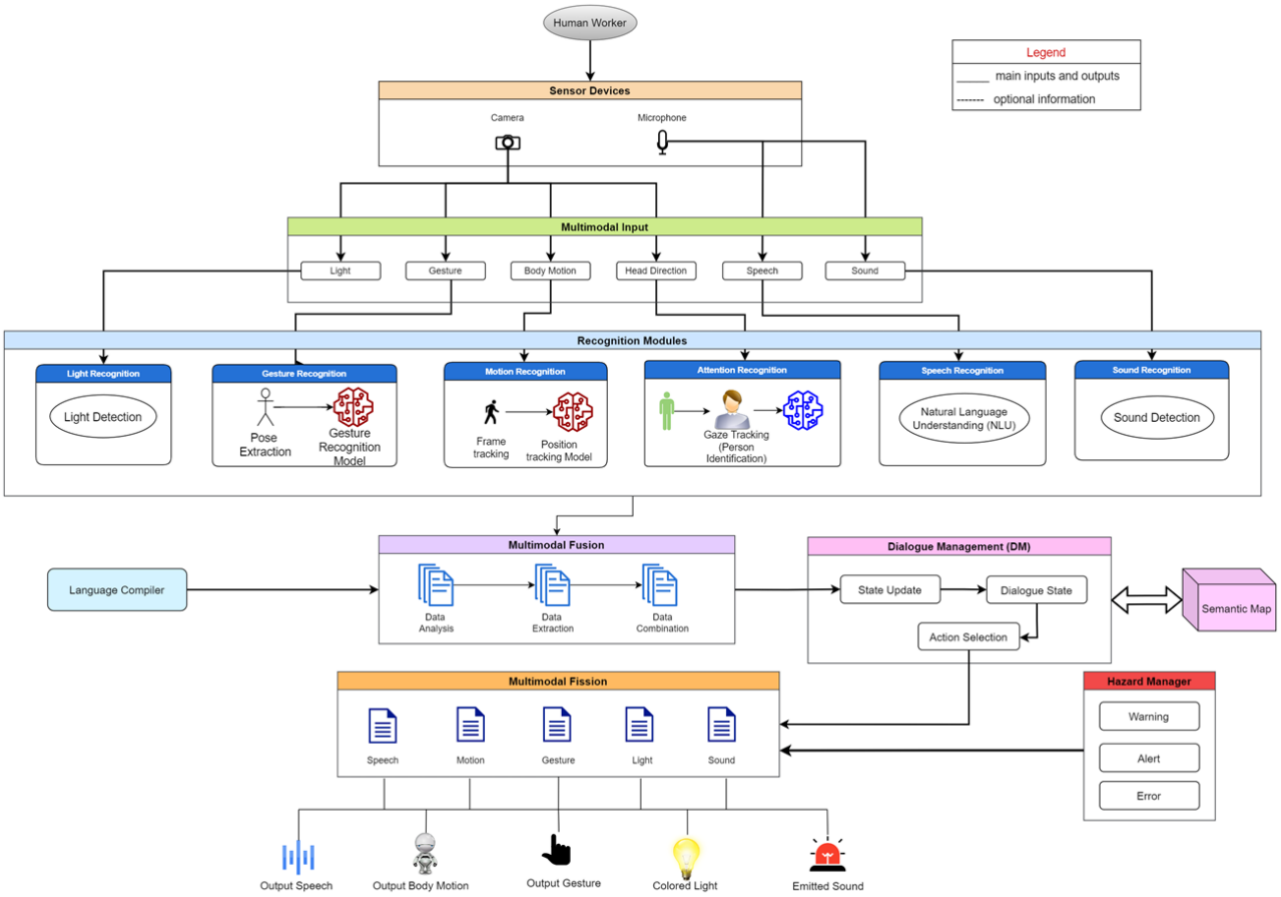

Figure 5. Architecture defined for Task 2.3

The output of each module is fed into the multimodal fusion module. Multimodal fusion can be generally obtained using two kinds of approaches, namely a-priori multimodal fusion and a-posteriori modal fusion. In a-priori multimodal fusion, all input signals are fed into a single module (e.g., a neural network based on multimodal transformer application). In a-posteriori multimodal fusion, fusion is performed after each signal is analyzed specifically. In CANOPIES, the second solution has been applied (see Section 7.3). The output of multimodal fusion is a *fully specified* communicative act. This communicative act must be imagined in the context of a dialog between the human and the robot. For this reason a rule based dialog management system has been implemented. In the context of this dialog, the robot may answer to the human (robot-to-human communication) using different modalities including synthesized speech, body motion, hand gestures, lights and emitted sound. Sentences from the dialog management system are turned into output signals by using the multimodal fission module. At the current stage, only robot-to-human communication based on speech and light signals have been implemented, but further modules can be added to the generality of the proposed architecture.

Finally, the hazard manager module is in charge generating robot-to-human communications which are not part of a dialog but results by the robot monitoring the environment and advising human workers about warnings, alerts and errors.

7 Human-to-Robot Communication

7.1 Speech based communication

Based on the identification of the most suitable speech act classes and communication modalities that could be adopted, in our opinion, in the CANOPIES project, we started conducting some analysis of the main components for the Speech pipeline. Initially, we investigated most of the available libraries and systems for the Speech Recognition module, both online and offline. Once selected the best option for the acquisition part, we proceeded in developing and testing a pipeline for Natural Language Understanding that was exploiting frame semantics.

7.1.1 Preliminary Investigations

7.1.1.1 Speech Recognition

The first analysis for developing the Natural Language Processing (NLP) pipeline for the CANOPIES project was conducted on the already existing Speech Recognition libraries and systems. The identification of the most suitable tool for the Speech-to-Text module, considering the environmental limitations in table-grape vineyards, is of fundamental importance to ensure a stable and robust communication between humans and robots. For this reason, we tested different online and offline Speech-to-Text libraries and systems, such as DeepSpeech (offline), Vosk[5] (offline), Sphinx (offline) and Google Web Speech API[6] (online) in order to determine the best one for reaching our scope.

DeepSpeech is an open-source Speech-to-Text library from Mozilla that exploits recent machine learning technologies and it allows users to create their own training models, to achieve better results. However, we had to discard DeepSpeech for two main reasons: first, due to the company reorganization the future of DeepSpeech is undetermined and it may get shut down and second, the

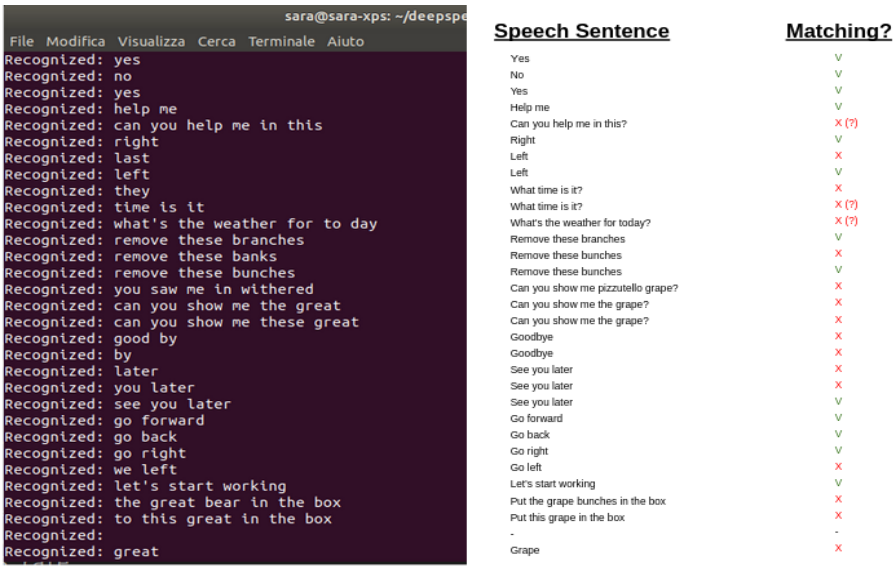

Figure 6. DeepSpeech recognition results. The green "V" represents a utterance correctly understood, while the red "X" identifies a wrongly transcribed sentence.

[5] https://alphacephei.com/vosk/

[6] https://pypi.org/project/SpeechRecognition/

recognition performances were not so good as expected. In Figure 6 it is shown that the percentage of correctly transcribed utterances is very low.

Subsequently, similar tests were conducted with two different Python speech recognizers: Sphinx and Google Web Speech API. The first has the advantage of working offline, but it is not precise in the recognition, while the latter, on the other hand. is very accurate in recognizing the spoken utterances, but the drawback is the requirement of internet connection.

Vosk is an offline open source speech recognition toolkit that supports different languages and dialects. Almost for all languages there are different model sizes available that can be chosen depending on the final platform on which the speech recognizer must work and on the required level of accuracy needed. For this reason, it supplies speech recognition for chatbots, smart home appliances, virtual assistants but also transcriptions. One of the advantages of Vosk is associated to its scalability and precision in the utterance transcription achieved also with small models as demonstrated in Figure 7.

Speech Sentence	Matching?
-	V
Robot take the grape	V
Robot	V
Robot take the grape and put it into the box	X
Can you help me	V
Cut this branch	X
Move on your right	V
Go on your left	X
Follow me	V
Come with me	V
What's the weather for today	V
What's the temperature today	V

Figure 7. Vosk recognition results. The green "V" represents a utterance correctly understood, while the red "X" identifies a wrongly transcribed sentence.

Hence, from the analysis of the Speech-to-Text modules presented above, the most suitable libraries that we selected to be tested on the field were: the online Google Web Speech API and the offline Vosk. However, due to the lack of a stable internet connection on the field, we finally decided to adopt Vosk for the Speech-To-Text module.

7.1.1.2 Natural Language Understanding

For developing our Natural Language Understanding module, we considered as a baseline a neural architecture for Natural Language Understanding (NLU) in Spoken Dialogue Systems, which is: Hermit-NLU. It is a hierarchical multi-task system with multi-layer representation of semantic meaning: dialogue acts, frames and frame elements, respectively associated to the scenario, action and entities of the NLU-Benchmark dataset on which the network was trained. Unfortunately, such a

dataset is not suitable for our purpose, since it contains multiple home assistant task domains and people conversation. For this reason, we acquired a dataset specifically for the vineyard scenario (described subsequently in Section 7.1.2), that was used to conduct some analysis on the frame semantic identification and arguments extraction. In the following, we present a NLU speech pipeline composed of 5 modules, as in Figure 8.

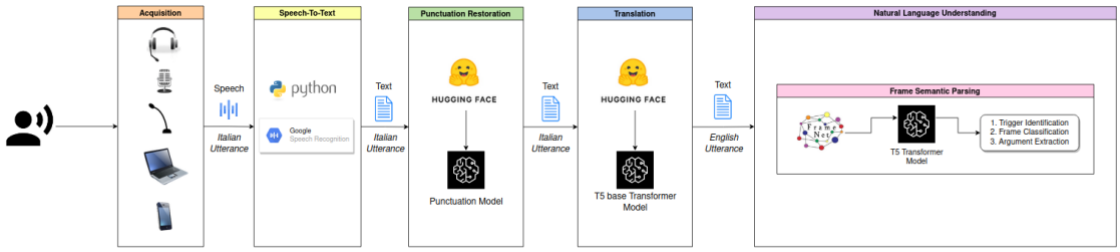

Figure 8. NLU speech pipeline for frame semantic parsing.

Once collected the vocal audio through Jotform, the speech sentences were converted into their textual representation, so to introduce punctuation restoration to obtain a more accurate translation before extracting the semantic frames and frame arguments from each of the translated utterances.

Acquisition

Jotform allows people to access the form through different devices. For this reason, mobile phones, computers, tablets were supported in the process of acquiring user spoken utterances. External devices such as microphones, headsets and headphones were used by most of the participants to reduce the environmental noise.

Speech-To-Text

The choice of the Speech-to-Text recognizer derived from the preliminary analysis conducted on some online and offline state of the art existing systems discussed in the previous sub-section. Finally, we decided to perform some tests with Google Speech API in a first study and then with Vosk, since we noticed that among all the aforementioned systems, these were providing the most accurate transcriptions, with a low percentage of uncertainty and misrecognized words.

Punctuation Restoration

The punctuation restoration applied on the transcribed text was one fundamental step that we decided to introduce in the proposed NLU pipeline. The addition of this intermediate module was essential in order to obtain a more accurate sentence translation. Indeed, punctuation restoration is adopted to separate different utterances, clarify a concept and in some cases, to identify the category a particular

sentence belongs to. For instance, a question mark at the end of a statement has a higher probability to associate the corresponding utterance to the 'Request' class. Initially, we experimented with one model, but we noticed that it was covering only 5 types of punctuation. For this reason, we opted for

the Bert restore punctuation model by Hugging Face[7], being able to manage more English language punctuation than the previously tested system, such as: exclamation mark (!), question mark (?), dot (.), comma (,), hyphen (-), colon (:), semicolon (;), apostrophe (‘).

Translation

The translation module was also required in our pipeline for exploiting the potential of the frame semantic parsing system (last module of the pipeline). Hence, to identify the best translation system, we tested two different Transformer models from Hugging Face: the multilingual Bart for multilingual machine translation, managing the transcription of sentences provided in any of the 50 covered languages, and a particular Helsinki-NLP/opus-mt model available for Italian-to-English translation. Finally, we decided to use the latter one, since our goal is a system robust to any input language, by considering sentence translations from any of the supported languages into English.

Frame Semantic Parsing

The last stage of the proposed NLU pipeline is represented by the frame semantic parser. Such a network works only with English sentences, since it was trained on FrameNet[8]. For this reason, we needed to perform the sentence translation and we could not directly pass the original utterance (in any other language differing from English) to this system. The outcome of the Frame Semantic Transformer[9] is the identification of the trigger, frame and frame arguments associated with a specific sentence provided in input. The network, that is developed on top of T5 Transformer model, is able to assign quite accurately all the aforementioned features from a given utterance.

The outcome is obtained by passing the collected Italian sentences through all the modules of the developed pipeline. Hence, we noticed that such a system was able to extract information about frame semantics and associated frame arguments from most of the analyzed utterances. Since the major limitation of such system was the need to a translation module to permit the frame association, we opted to deeply investigate InVeRo-XL, the state of the art model (based on XLM-RoBERTa) for multilingual semantic role labeling end-to-end that supports 45 languages.

7.1.2 Training Data Collection

Due to the impossibility of physically going to the field multiple times to acquire data or test developed solutions, we decided to exploit VR in all its forms of experience (either immersive or non-immersive) to carry out HRI-related experimental activities in CANOPIES. In other fields of study, it has been demonstrated that acquiring data and evaluating preliminary solutions or complete applications in VR could provide multiple benefits in terms of experimental settings, costs, and safety. To this aim, a larger dataset was obtained when we conducted a user study in immersive and non-immersive VR while testing our approach of speech act classification.

[7] https://huggingface.co/felflare/bert-restore-punctuation

[8] https://framenet.icsi.berkeley.edu/fndrupal/about

[9] https://github.com/chanind/frame-semantic-transformer

7.1.2.1 Acquisition Form

Based on the speech act classes identified along with the suitable modalities for the project,, we initially collected data separately with three different scenarios, one for each category under investigation: Information (22 questions), Command (20 questions), and Request (20 questions). The customized questionnaires were created through the Jotform platform to reach, with our forms, as many participants as possible. Jotform was chosen for its practicality in utterance recording, facilitating the procedure of speech acquisition from the user perspective and, consequently, speeding up the entire data collection process.

In all three forms, we described the general context and some technical terms that users could exploit in generating the vocal responses. Additionally, to avoid uncertainty or misunderstanding in the description of the scenario, a representative image followed each question. Such pictures (some captured from the simulation environment, others in the real field, and a few taken from the web) illustrate tasks, activities, actions, and vineyard elements such as grapes, leaves, branches, and the pergola system.

Given the difficulties in recruiting a sufficient number of vineyard operators for the data collection, we relied upon Bachelor, Master, and PhD students in Engineering in Computer Science at our university. While the features of the subjects need to be properly taken into account, we remark that the kind of system we are developing in CANOPIES and the interaction with robots in the agricultural environment could be of more interest to the next generation of digital farmers, of which young students may represent a reasonable proxy. At the end of the acquisition process, we reached around 40 participants with each form (most of the users provided vocal utterances for all three questionnaires). The subjects who participated in the data acquisition process were mainly males between 18 and 30 years old. Regarding the level of experience with robots, we detected quite balanced robotic skills among the participants, with half of them having good proficiency or interacting several times with a robot. Obviously, the working expertise in the vineyards is generally rather limited.

The data acquisition process of the Information, Command, and Request speech act categories lasted around two months. At the end of this period, we collected approximately 1,800 vocal utterances with their corresponding textual transcriptions. All the details concerning the data analysis are provided in Table 4; they include the number of sentences before and after removing duplicated utterances and the final number of statements obtained after re-arranging them according to the category of the content information that they represent.

The figures in Table 4 show that the speech act categories that required some post-analysis are Command and Request. Such an outcome is reasonable, as the difference between the two aforementioned classes could be very subtle from the user's perspective. We listed under the Command speech act all the sentences provided by the human that the robot must execute in the short term, such as "Turn left", "Harvest the ripe grape bunches", or "Remove all the dry branches". At the same time, utterances asked by the person that do not require an immediate interruption of the robot's activity have been included in the 'Request' class. For instance, sentences belonging to

Speech Act	Sentences (with duplicates)	Sentences (without duplicates)	Sentences (after category adjustments)
Information	900	804	804
Command	728	517	665
Request	719	534	342

Table 4. Analysis of the acquired data.

this category are "Can you tell me your battery level?", "Could you help me in harvesting these grapes?", "How many damaged grapes have you identified?".

7.1.3 Study Design

Furthermore, in the experiments to evaluate immersive and non-immersive experiences in VR, we recruited students from our university. Again, we argue that they could provide us with good feedback on the adoption of VR technology to test HRI approaches in the CANOPIES project. Since the user studies were conducted for approximately one month in our laboratory, we achieved good student engagement.

Before starting the experiments, a brief description of CANOPIES and the goal of the planned experimental activities were provided to the participants. In presenting the project, we mainly focused on explaining the agronomic robot's ability to enable subjects to interact with it in the virtual environment. Moreover, we put special care into illustrating the different table-grape varieties, such as "white pizzutello", "black pizzutello" and "black magic", to enrich the communication and to compensate for the lack of knowledge of the specific terminology of the table-grape scenario and activities that can be carried out in a vineyard. At the end of the experience, each participant was asked to complete a questionnaire regarding the user study in which he/she was involved.

7.1.3.1 Non-Immersive Experience

The experiments in the NIVE were conducted for two weeks, allowing most Master and PhD students to participate in the user study. Nevertheless, the one-hour duration of the experience in NIVE, we reached a good number of participants (40 students). However, we noticed a higher presence of males (32) than females (8) in our evaluation.

Moreover, we observed that 75% of the users had already experienced VR applications before this study, and 17.5% of them had negative feelings after using such systems in the past. In analyzing the expertise in the vineyard and with robots (in non-immersive and real scenarios), we discovered that 27.5% of people had experience in a vineyard, and 40% worked with robots in the NIVE. At the same time, 52.5% operated real robots multiple times.

For the experiments in the NIVE, we required a computer with high computational power to run the Unity Linux executable of the CANOPIES non-immersive virtual simulator (generated by PaleBlue, our Norwegian partner) and the developed speech system. To this aim, we employed an Alienware Aurora Ryzen Edition R14 computer with 64GB of RAM and an AMD Ryzen 9 5950X 16-Core processor. The study was conducted on a machine running Ubuntu 20.04. In this experiment, to reduce the environmental noise and allow the recording of the vocal utterance in separate files by the system, we opted for a headset with a switcher to activate/deactivate the microphone. Moreover, in the NIVE, the participant could move the avatar (representing himself/herself) through the keyboard and change the horizontal camera view either with the keyboard or the mouse. Hence,

the user could supervise the robot while performing a motion autonomously, chosen based on the recognized speech act category and the match with specific words in its knowledge base.

7.1.3.2 Immersive Experience

Approximately one week was required to perform studies in the IVE of the CANOPIES project. As for the NIVE, also for this experience, most of the involved students were enrolled in a Master's or PhD in Engineering in Computer Science. With such a study, we were able to attract 41 young adults, identifying two groups: people that also participated in the NIVE experience (25) and users that were only included in this experiment (16). Similarly, as in the study in the non-immersive scenario, we observed that most of the participants involved were males (33), with only a few females (8). Then, 61% already used immersive VR applications in the past, and 19.5% of them had a negative experience with such technology. Regarding the expertise in the vineyard, 43.9% of the users had sufficient knowledge of the activities in such a field. However, 48,8% of the participants had experience with real robots, while only 17,1% with androids in immersive environments.

To conduct our studies in the IVE, we employed an additional powerful machine along with the one adopted for the NIVE experiment: an Alienware x17 R2 with 32GB of RAM, a CPU of 12th generation i9 and Nvidia RTX 3080ti GPU running Windows 11 Home. The Alienware Aurora Ryzen ran the developed speech system. At the same time, the Alienware x17, connected to the Oculus Quest 2 headset (through Oculus Link cable), allowed us to analyze what the user was observing during the interaction. The maximum exposure time in the IVE was set to half an hour to reduce the probability of experiencing any discomfort. In such a scenario, the communication device adopted was the Jabra microphone (that participants held in their hands during the interaction) to avoid ambient noise and allow the recording of the vocal utterances in separate files by pressing the microphone button to activate/interrupt the acquisition process.

Since users could change the avatar's position (representing themselves) in the NIVE, we decided to let people also move in the IVE. To this aim, we identified a 2m x 2m squared space in our laboratory to permit users to experience motion in the immersive vineyard while analyzing the robot's responses and activities. However, unlike the experiment in the NIVE, the robot motion in this user study was not directly controlled by the speech system running in ROS since, by the time we conducted the tests, there was not yet a stable connection between ROS and Unity. For this reason, we decided to faithfully reproduce the robot's movements through the Oculus Quest 2 controllers by relying on the robot's comprehension of the utterance pronounced by the person.

7.1.4 System Implementation

The necessity of a modular and robust system that could exploit the speech act classification to simplify the robot's comprehension of the specific information content led us to develop a preliminary ROS framework for spoken interaction between humans and robots. In particular, we implemented our system on ROS Noetic on a computer running Ubuntu 20.04 LTS. The proposed pipeline used for both immersive and non-immersive experiments is presented in Figure 9. The first module is the Speech-to-Text, which is responsible for acquiring human spoken utterances and converting them into their textual representation through Vosk.

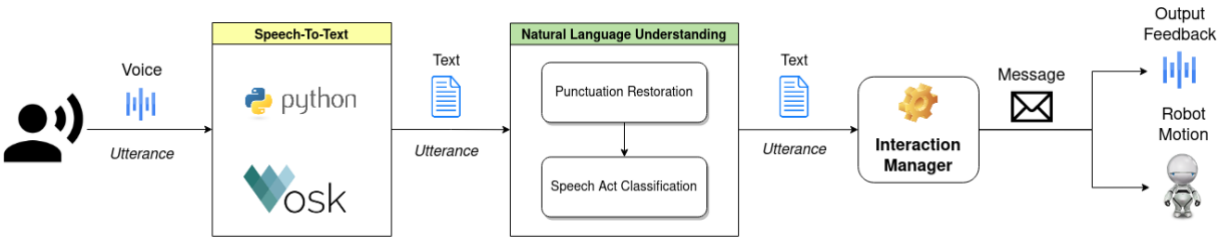

Figure 9. Developed ROS speech pipeline.

Vosk was chosen for its higher accuracy in utterance transcriptions, also with the smallest models, and for supporting different languages. Since our final goal is to develop a robust and reliable system with whom we expect people to interact spontaneously, we believe that also allowing communication in different languages (for foreign human workers) could improve communication with the robot.

Subsequently, the textual transcription of the utterance is processed by the Natural Language Understanding module. Such a component introduces punctuation in the sentences using the Bert restore punctuation model by Hugging Face, to simplify the process of speech act classification. For instance, a question mark at the end of a statement has a higher probability of associating the sentence with the 'Request' class rather than the 'Command' or 'Information' categories.

We trained the speech act classification model on the dataset acquired through Jotform by adopting the AutoGluon library for Natural Language Processing (NLP) tasks. We opted for such an automated machine learning library since it could select the best training parameters by optimizing them with respect to the final goal (in our case, a classification problem). However, another relevant advantage of this library is associated with the existence of a specific predictor for supporting multimodality, which could help us improve our final system.

Finally, based on the sentence prediction class obtained through AutoGluon, the Interaction Manager performs some checks to verify if any of the words of the sentence appears in the system's knowledge and generates a message that is sent to the speakers, so to provide vocal feedback of the robot's speech act understanding ('Information', 'Command', 'Request' or a combination of different categories in case of subordinate sentences) by exploiting PicoTTS 4 for the Text-To-Speech module. Therefore, when the android recognizes an utterance belonging to the 'Command' class, a motion message is sent to the virtual robot's motors, attaining the corresponding motion in the simulated environment (both immersive and non-immersive). Nevertheless, even if a different category is identified, the robot provides feedback to the user by moving its head and looking around to give a feeling of comprehension.

By the end of the NIVE and IVE experiments, we acquired 5,180 vocal recordings and their corresponding textual transcription. However, after removing duplicated sentences (mostly belonging to the Command category), the number of statements we achieved for the final evaluation was 3,364. As anticipated in SectionV, the speech act classification network could assign one of the following classes to the input utterance: 'Information', 'Command', 'Request' or a combination of

these categories in case of a complex subordinate. To this aim, the outcome of our evaluation is presented in Figure 10. From this analysis, we noticed that the sentence misclassification depends on one (or multiple) issues:

1. Misunderstanding of one or more words in the sentence due to ambient noise, words not pronounced clearly or similarities with other terms.
2. Prolonged pause between words within the same sentence, causing the system to separate it into different statements, to each of which one (or more) categories will be assigned.
3. Lack of intonation analysis.

Since our goal is to primarily evaluate the developed speech act classification module in an ideal situation, in which comprehension failures are not detected and propagated among the pipeline by the speech recognition module, we double-checked all the acquired sentences and fixed the ones belonging to the first (346 statements) and second issues (63 statements) listed above, by listening to the recordings collected during the IVE and NIVE user studies. Hence, we ran the speech act classification module on the corrected sentences to identify the issues presented previously from these classification outcomes.

In both cases, the performances of the category classification network increased when working with precise transcribed sentences. In particular, from the 63 statements previously misunderstood, in the recent analysis: 46 were correctly classified, 10 were wrongly recognized due to intonation issues, while an incorrect category was assigned to the remaining 7, as shown in Figure 11. Similar results were also encountered with the 346 statements that previously contained one or more misinterpreted words. Indeed, after sentence correction, 186 were associated with the suitable class, 84 fell under the intonation issue, and 76 were not located adequately by the speech act classification system, as presented in Figure 12. The illustrated results are summarized in Table 5.

# Sentences	Revised	# Sentences correctly classified	# Misrecognised	# Prolonged pauses	# Intonation necessity
3,364	✗	2,746	346	63	209
346	✓	186	-	-	84
63	✓	46	-	-	10

Table 5. Analysis of the acquired data

The outcome highlights some issues the robot can encounter while performing a speech act classification task on spoken utterances. We demonstrated that sentences recognized as separate statements, with erroneous transcriptions due to comprehension errors or word phonetic similarities and the lack of intonation analysis, can cause the android to miscategorize such phrases. To this aim, introducing the spectrum analysis of the vocal utterances would improve the performances of our speech act understanding system, in particular in disambiguating 'Request' from other categories.

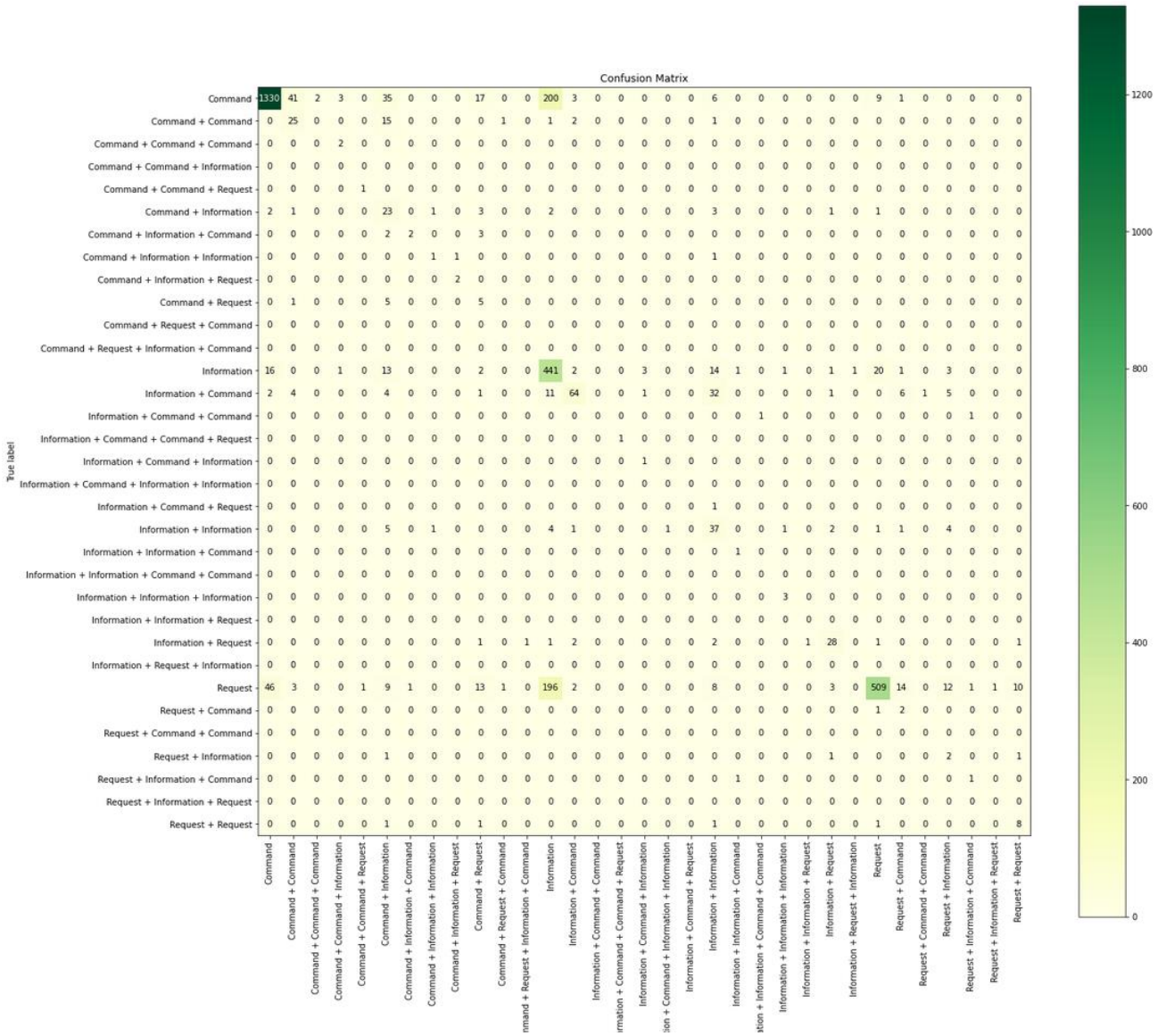

Figure 10. Confusion matrix of all the sentences in the test set (3364).

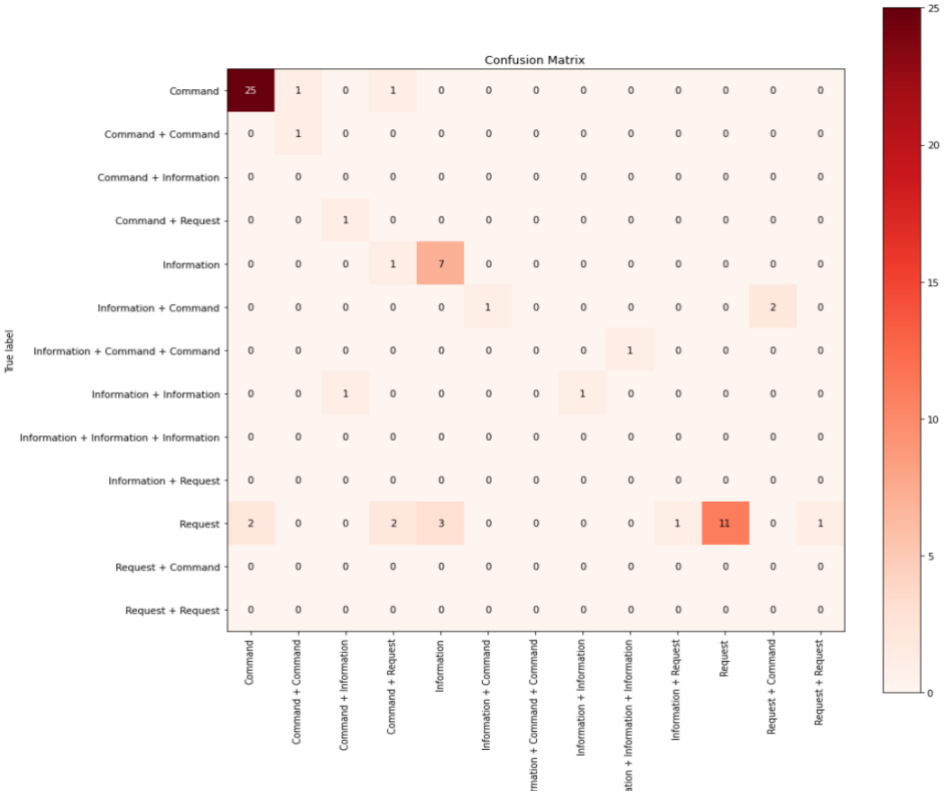

Figure 12. Confusion matrix of the corrected sentences belonging to long pause issue (63).

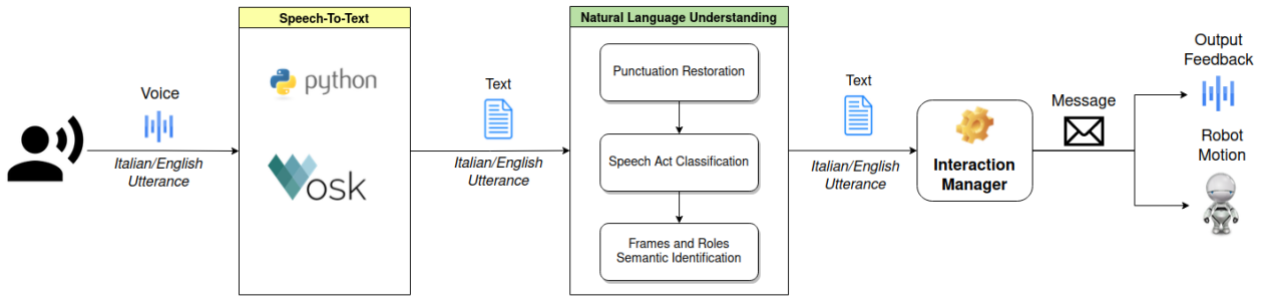

Figure 11. Final ROS speech pipeline including SRL module.

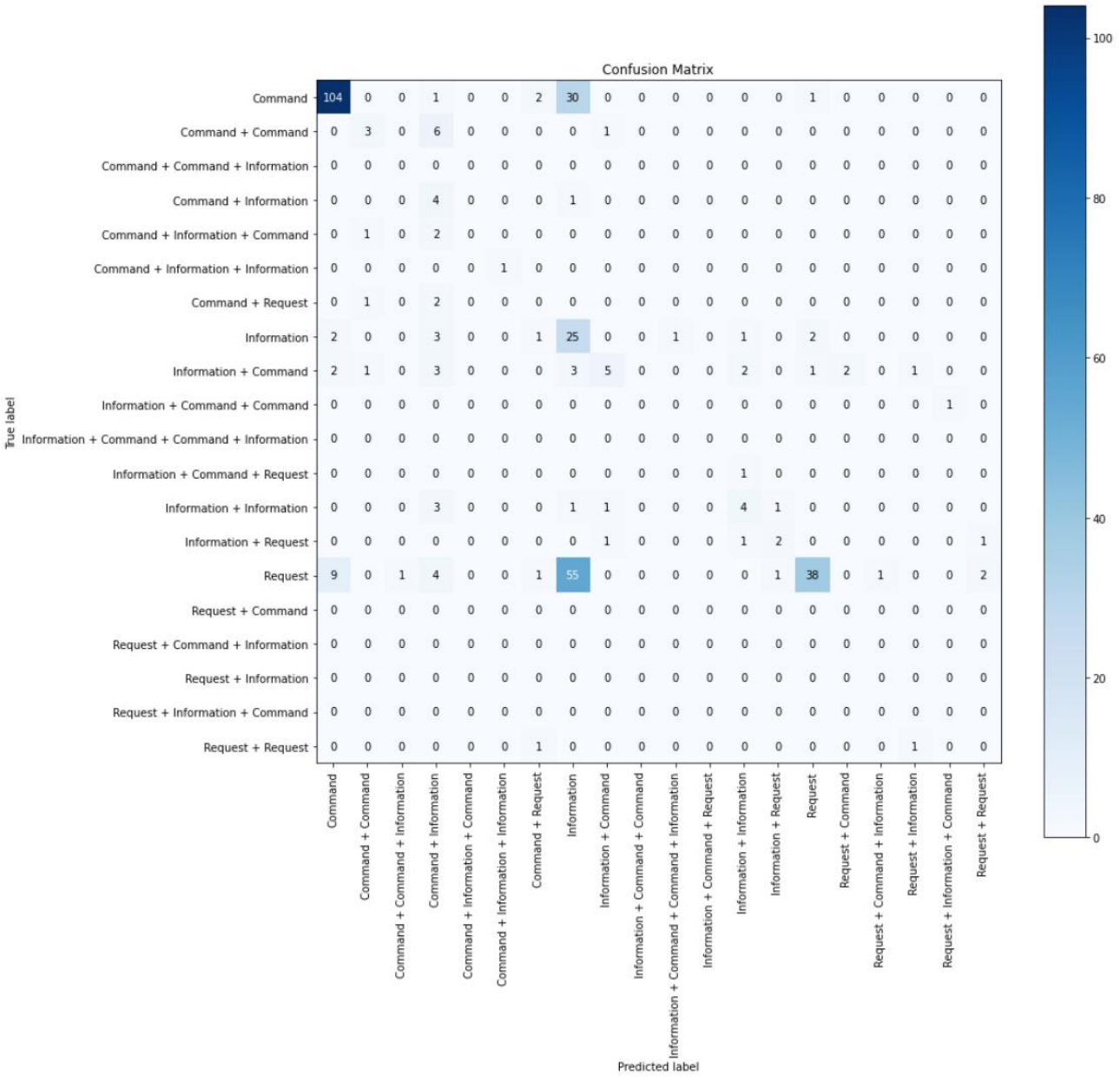

Figure 13. Confusion matrix of the corrected sentences belonging to the misrecognition issue (346).

The proposed approach is supposed to operate as a trigger for specifically trained models, which allow to understand the content of the speech act and consequently perform actions and complex activities. To this aim, Semantic Role Labeling (SRL), whose aim is to automatically address the question: "*Who did what to whom, when, where, and how?*", was included in the final version of our pipeline, which is presented in Figure 13.

Precisely, the identification of the frame would provide information about the type of action to be performed by the robot, together with further details that could help establish with higher accuracy whether a "Command", "Information" or "Request" was given to the android, by analyzing the verb tense. Instead, roles would contribute by enriching robot's comprehension of the human utterances, by specifying relevant features associated to the verb, such as the Direction in which to move, the

Destination to reach, the Time in which an action should be completed, the Theme to check and analyze etc.

All this information would be essential in order to generate the "rules" to identify the provided and missing information into the fusion module to ensure a multimodal communication. Multi-SRL[10] was adopted to generate an initial set of information concerning frames and roles for all the sentences acquired through Jotform and the user studies conducted in NIVE and IVE. For the same sentences, we asked ChatGPT to generate frames and roles (in a similar format to the one provided by Multi-SRL), so to compare their outcomes in terms of assigned frames and common roles. In the end, we double checked around 1,000 Italian sentences, by providing additional information, such as lemmas, part of speech tagging, dependency heads and the associated speech act classification. Such a set of sentences was used to train a network for Semantic Role Labeling that was employing XLM-RoBERTa to support multilinguality, allowing the human workers to communicate with the robot in any of the supported languages.

For instance, considering the utterance "*Posizionati a destra per raccogliere l'uva*" (English translation: "Position yourself to the right for harvesting the grapes"), our system will be able to identify the frames in the sentence ("*Posizionati*" and "*raccogliere*") together with their associated roles, which are respectively "Location" and "Purpose" for the first verb, while "Location" and "Theme" for the second one, as shown in Figure 14.


```
{"sentence_id": 3,
      "words": ["Posizionati", "a", "destra", "per", "raccogliere", "l'", "uva", "."],
      "predictions":
          {"0":
              {"predicate": "PLACE",
               "roles": {"1,3": "Location", "3,7": "Purpose"}},
           "4":
              {"predicate": "HARVEST",
               "roles": {"1,3": "Location", "5,7": "Theme"}}},
      "speech_act": {"0,7": "Command"},
      "lemmas": ["posizionare", "a", "destra", "per", "raccogliere", "la", "uva", "."],
      "pos_tags": ["V", "E", "S", "E", "V", "RD", "S", "FS"],
      "dep_head": [0, 3, 1, 5, 1, 7, 5, 1]
```


Figure 14. Example result of the application of XML-RoBERTa.

[10] https://github.com/SapienzaNLP/multi-srl

7.2 Gesture based communication

Gestures have been explored as a communication channel both in human-computer (HCI) and human-robot (HRI) interaction. In these domains, usually, gestures unlock a new communication channel to send intentional commands to a machine. Indeed, in the preliminary study conducted on HRI experiments in VR, gestures emerged as the preferred communication modality for the command speech act category. Hence, when the interaction between humans and robots is required, as in CANOPIES, such a channel could be exploited by a person in order to immediately stop the machine's activity by showing his/her opened hand to the mobile vehicle, but also to indicate the next position to reach by the robot.

Pointing gestures could be employed in different tasks, not only to show a precise position to reach, but also to specify an element of the vineyard (a certain grape cluster, branch) or to indicate the person to follow. However, the introduction of gestures in CANOPIES requires the robot to be able to: firstly, detect the position of both human's arm and hand, secondly, to understand their meaning. Conversely, the robot should also be able to communicate with the person through gestures, by translating a non-verbal behavior into an understandable action. Human-Robot Interaction system, because they could cause misunderstandings and uncertainty.

Gestures were identified to be of the command category in the speech act classes and after exploring the communication modalities that could be adopted for the CANOPIES project, we conducted preliminary analysis of the key components for the Gesture pipeline. Initially, we investigated most of the available libraries and available systems for the Gesture Recognition module, both online and offline. Once selected the best options for the recognition of gestures, we proceeded in developing and testing a pipeline for Gesture Recognition System (GRS).

7.2.1 Preliminary Investigations

A preliminary analysis was conducted for the development of the Gesture Recognition System (GRS) for the CANOPIES project on the already existing Gesture Recognition libraries and systems. The possibilities for this type of system to give commands was to read the Hand Gestures or Full-Body Gestures Figure 15. The identification of the most suitable joints of the said approaches depends on the context such as the type of conversation between the Human and Robot, Background, camera angle, distance of the human from the recognition hardware, etc. Based on the environmental limitations in table-grape vineyards, it is important to ensure that robots should be able to read and understand the command (Gesture). For this reason, we tested the existing gestural definitions, systems for both hands based and full-body based gesture systems of different architectures in order to determine the best one for reaching our scope.

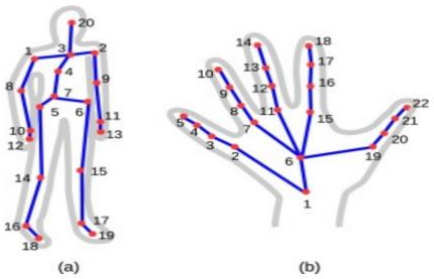

Figure 15. Representation of point key-map of Full-body (a) and Hand (b) gestures.

The state-of-the-art methodologies in gesture recognition techniques for HRI include various techniques broadly classified into Computer Vision-based, Machine Learning-based, Wearable-based, Hybrid, and Virtual Reality-based techniques. Computer Vision-based techniques process images or videos of a person performing a gesture and extract features that can be used to recognize the gesture. Machine Learning-based techniques train models that can recognize gestures from image or video data. Wearable-based methods use sensors such as accelerometers or gyroscopes to detect and track gestures. Hybrid techniques combine multiple input sensor modalities to improve the accuracy and robustness of gesture recognition systems. Virtual Reality-based techniques use VR simulations to generate datasets of gestures that can be used to train and evaluate gesture recognition systems.

However, several challenges need to be addressed in gesture recognition systems for HRI. These challenges include variations in gesture performance, occlusions and self-occlusions, background noise and lighting, real-time processing, privacy-security concerns, scalability and generalizability, and benchmarking and standardization. Ensuring that gesture recognition systems can accurately detect and interpret human gestures in different scenarios and environments while protecting privacy and ensuring compliance with regulations remains a significant challenge for researchers and developers in the field of HRI.

While conducting the preliminary analysis on the available datasets, we have found that there is no generalization across the datasets representing the same gestures. For this reason, this work also aimed to create a dataset generalized of ground vehicles and the ground Robots.

7.2.1.1 Gesture Recognition

We initially created a pipeline for Gesture Recognition System using the readily available mediapose (see Figure 16). Initially we considered the basic navigation commands for this analysis such as move variations, turn variations, start and stop.

The person performs gestures in the view field of Intel Depth enabled Real Sense camera D455 sensor at a prescribed distance range (1mt to 4mts was selected range). These motions were processed into frames and each frame was passed through Mediapipe's Pose Estimation algorithm for extracting the key point map of the respective joints. The key points were verified with the predefined map of the 8 gesture classes corresponding to the navigation of the robot as mentioned above. We used ROS-Python implementation to convert the predicted class into the navigation command and pass these messages to the robot though the connected computer over wireless or wired communication. A

separate log file was created for the predicted gesture, recognised key points, acknowledgement of the command and the command performed by the robot for each of the motions performed by the Human. We also enabled the robot acknowledgement of each command and prompted it back to the user via ROS messages for the user to know which command the robot has really understood from the gesture performed. The figure below shows the implementation of the preliminary pipeline.

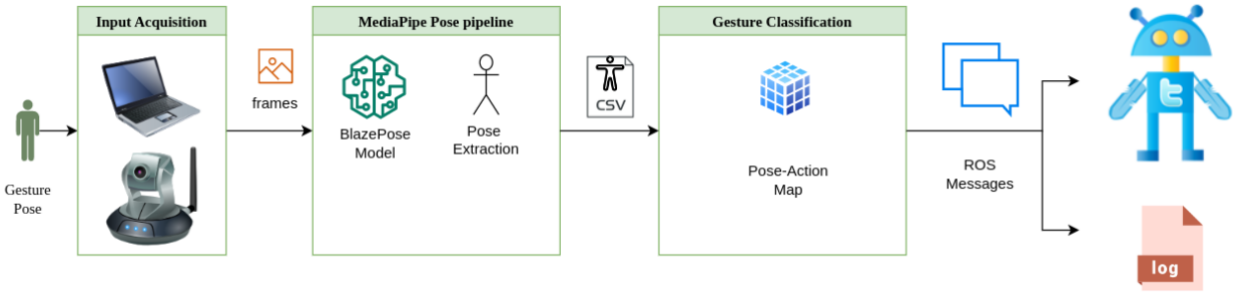

Figure 16. Gesture pipeline using pose estimation and ROS.

7.2.2 Gesture Definitions

There were several definitions of gestures based on the applications and their usage context, yet most still need to be standardized. Regarding industrial applications, gesture definitions change according to the type of industry and context used. General movement and context-specific gestures were defined to generalize and standardize the signs in the industry. These were defined after following some gestures for communicative purposes in HRI and from the gesture language between human-human and human-robot and in the workspace for the CANOPIES project. The primary scientific goals of CANOPIES involve creating innovative approaches to HRI, the collaboration between humans and robots (HRC), and coordination among multiple robots (MRC). These methodologies aim to establish a productive collaborative system between human workers and teams of multiple robots to achieve precision farming of permanent crops. The effectiveness of these approaches will be evaluated in the setting of table-grape vineyards. This is the reason for collecting the data in the table-grapes field to understand the noise and the variations in lighting conditions under the grape canopy.

The defined gestures, as shown in Figure 17 consist of 10 dynamic and 11 static signs, where 9 gestures can be performed using two hands while the rest are with one hand. We initially defined a large set of gestures for acquisition purposes. They can be conveniently reduced to a smaller set based on performance and feedback from experimentation. Some gesture definitions were specific to the Project of CANOPIES [*CUT, CHANGE BOX, RECEIVE, TAKE*], while the rest are generic to the all ground robots. Human workers in the table-grape field will be working along with robots in performing agricultural activities, which require at least one hand in most cases to instruct the robot, without interrupting their activities. The 9 gestures that need two hands can have any other tools in the hands of the human in most scenarios.

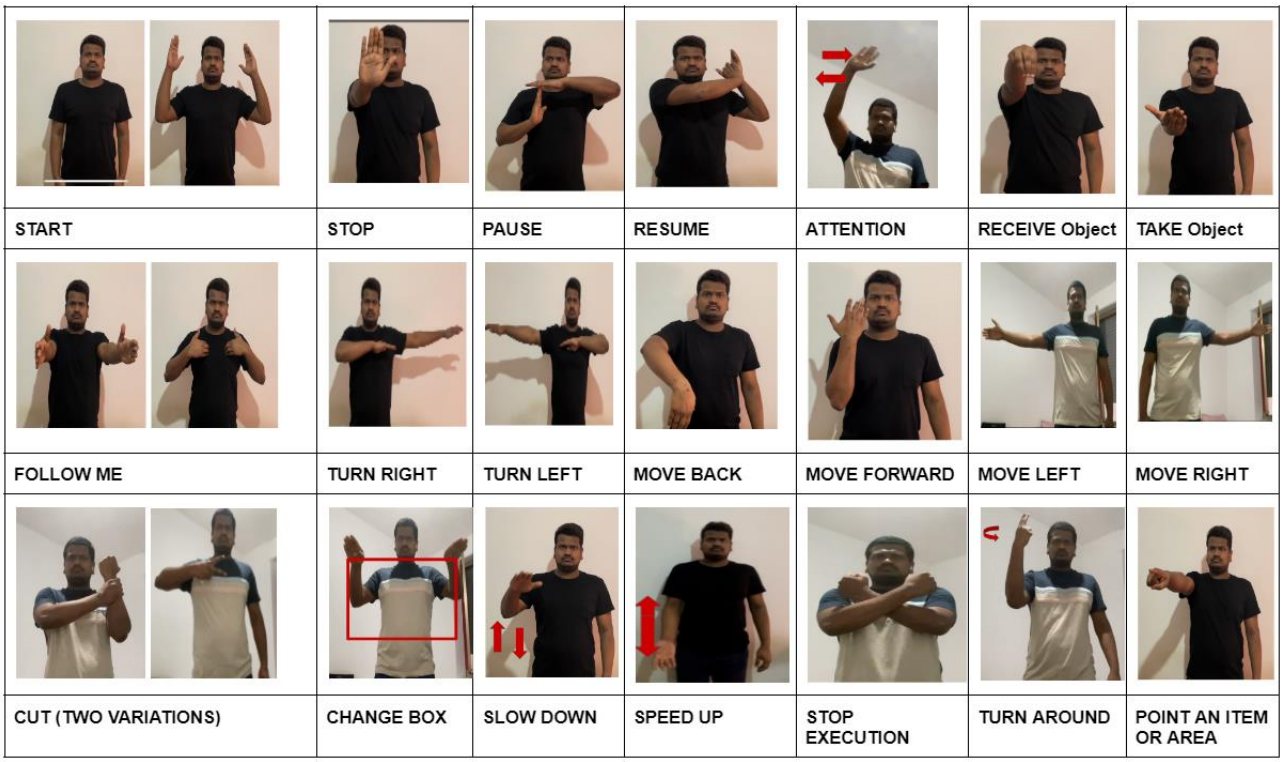

Figure 17. Defined gestures based on the workspace ethic and human communication.

7.2.3 Data Generation

Based on the defined Full- Body Gestures, we needed to generate the gestural motions for the training of the defined network. For this reason , we acquired real-time data from the persons involved in the project based on their consent to use this data across the project.

7.2.3.1 Real-Time Data from the field

The real-time data was captured using Intel realsense Depth camera D435i with IMU and stereo vision in the table-grapes field in Aprilia (Lazio), a worksite for the project. The gesture data was recorded after consent to use and publish it for research purposes. To this aim, 8 persons from different age groups and gender participated in recording the defined gestures, as shown in Figure 18. These gestures were recorded at the resolution 1280x720 at 30 FPS with depth RGB-D information. The gestures were performed and captured at 4 different distance settings, starting at 1.5 meters from the camera and increasing the interval by considering multiples of 1.5, to store such information in RGB-D rosbags.

All 21 gestures were recorded into a single ROS-Bag file for each distance per person. All the files were then post-processed to extract videos and frames of the performed signs based on the type of the gesture(Static/Dynamic). Then, Mediapipe's pose estimation algorithm was applied to verify if all the key points of the poses were being detected.

Figure 18. Real-time Gestures captured in the table-grape field with different persons at different distances.

7.2.3.2 VR- Simulated Data

The way we choose to generate datasets of gestures is through VR simulations. Here, a character (Human-Avatar) can perform gestural movements in a virtual environment, and the system can record the gestures and their corresponding annotations. This allows for the generation of a large number of examples of different gestural signs, which can be used to train and evaluate gesture recognition systems. Indeed, a single character can be mapped to many animations and the virtual environment can be configured accordingly to the right parameters.

There are a few advantages of using VR to generate a dataset:

- VR can provide a controlled environment where the lighting, background noise, and other factors can be standardized to minimize the issues that would be faced in real-world datasets, such as occlusions and background noise.
- VR simulations can provide a safe and repeatable way to generate a large number of examples of different gestures, without the need for expensive or dangerous equipment.
- VR simulations can provide a way to generate examples of gestures that would be difficult or impossible to perform in the real world, such as dangerous or demanding ones.

Autodesk character creator tool was used to generate 10 characters representing the varied height, gender, ethnicity, and age. These characters were imported to the Unity Hub tool . The collected videos were processed using DeepMotion tool to extract animations (in .fbx file format) and then were mapped to the generated characters in the Unity Hub.

The simulation background of the table-grape CANOPIES environment was provided by PaleBlue. We have verified if the animations were working as expected by manually playing characters with respective motions in the simulation. By doing so, we have estimated that we can generate extensive data for gesture recognition. To this aim, we successfully deployed a deep learning algorithm trained on the hybrid data of virtually simulated and real-time data acquired. The virtual characters performing some gestures are shown in Figure 19.

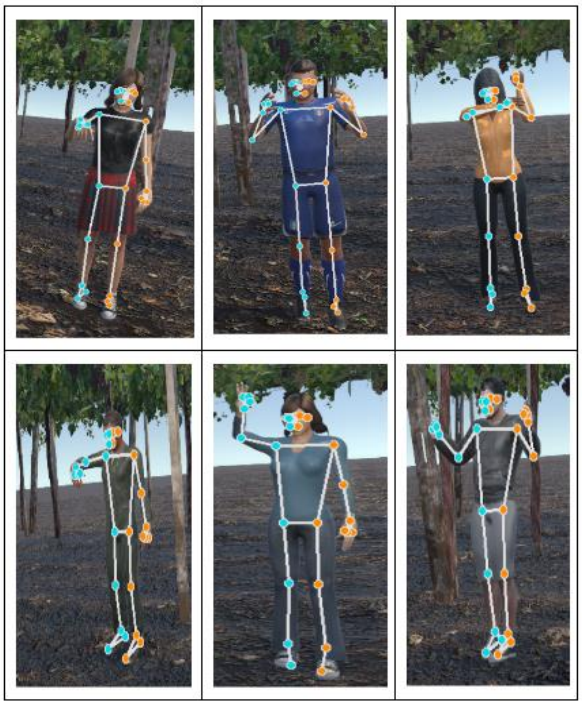

Figure 19. The virtual characters performing some gestures.

7.2.3.3 Data Augmentation

We acquired 3,380 frames of gestures from all 30 characters for 8 (7+1) gestures. However, after post-processing the frames for inconsistencies, we were left with 3,096 frames for these classes. All these frames were passed through Mediapipe's pose estimation algorithm for estimating the key-points.

7.2.4 System Implementation

The necessity of a modular and robust system that could exploit the multi-modality in HRI for the robot's comprehension of the specific information pathway led us to develop a preliminary ROS framework for gesture interaction between humans and robots. Several existing works identify gesture as the secondary communication modality.

We implemented our preliminary framework version on ROS Noetic on a computer running Ubuntu 20.04 LTS. Alienware x17 R2 with 32GB of RAM, a CPU of 12th generation i9 and a Nvidia RTX 3080ti GPU notebook running Windows 11 Home and Ubuntu 20.04 LTS was used for developing the system and its evaluation. Figure 20 presents the pipeline adopted for both Virtual and Real-Time data for training the gesture recognition model and improving the existing model if necessary. The first pathway is represented in a continuous black line for the virtual simulated data and the evaluation, whereas the blue dotted line represents the Real-Time data experiments for the model training and optimisation.

The experimentation was conducted in three phases to answer the research question, *"How efficiently the model trained on simulated data can perform on real-world data?"*. In the first experimentation phase, we generated more virtually simulated data by creating additional virtual human characters. These characters were mapped to 7 of the extracted gesture animations (21

gestures defined for the use-case). The system recorded the gestures performed by these characters through multiple cameras in the Unity tool with various distances and angles. We repeated the data collection with one character performing gestures at a time and multiple characters performing the same or different gestures in a single frame.

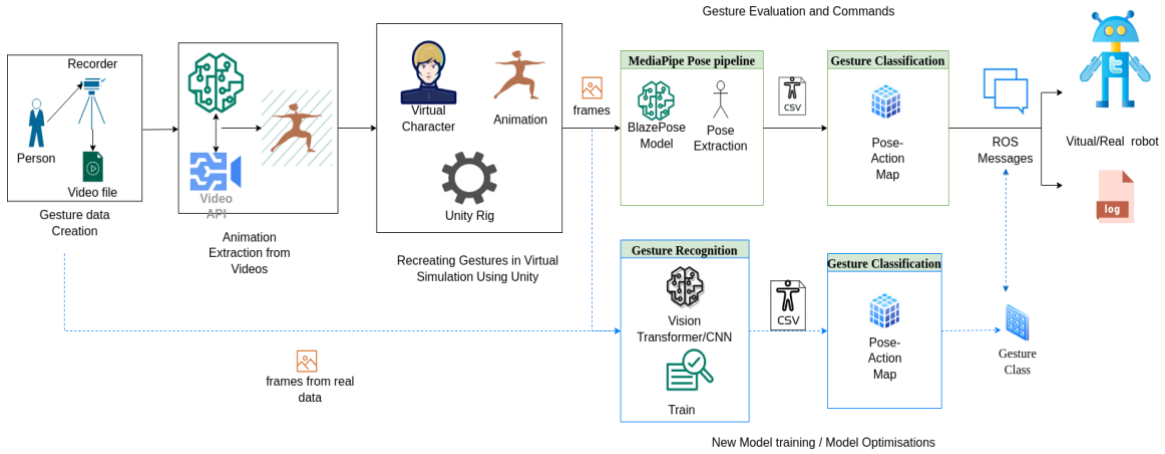

Figure 20. Final pipeline

In the second phase of the experimentation, we used Mediapipe's pose estimation to evaluate the generated data. The recorded frames were evaluated to check if the key points were being recognized and at what threshold. This will be taken as a base score. In the virtual simulated data, the identified key-points were evaluated against Unity-generated joint coordinates as the ground truth. Our experiments and evaluations are based on the hypothesis that a model trained on the hybrid data can be used to identify the gestures both in simulation and the real-world. We employed a CNN to classify the obtained poses accordingly with the gesture classes. This CNN is trained on the key-point data acquired by running the pose estimation algorithm to classify the pose clusters into the defined gesture classes.

In the Third Phase of experimentation, a suitable deep-learning approach was adopted to train a network on such data. The network was trained only on Real-data, Virtual-simulated data, and hybrid data. The Mediapipe's model and the trained models were compared against each other with variations of training data. After evaluations, we found the possible data combinations (virtual-real) for improving the network's performance. Hence, the network with the best performances was deployed to a virtual robot in the simulation, being able to see the virtual human avatar performing gestures so that the virtual robot communicated or executed the corresponding actions based on the gesture commands given by the virtual character. The users in the simulation had the control of choosing gestures via a keyboard (input device) in a {*non-immersive*} experience and with the hand controllers in an {*immersive*} experience based on the device. All these experiments will further be cataloged for system enhancements, and the model's performance evaluation will be conducted on real-world data by deploying such an approach into the real robot.

The gesture classification network on the corrected frames to identify the issues presented previously from these classification outcomes. In future, we will run additional pose estimation networks to

TARGET / OUTPUT	Start	Stop	Move_left	Move_right	Turn_left	Turn_right	Pause	Unknown
Start	387 12.50%	3 0.10%	2 0.06%	0 0.00%	3 0.10%	1 0.03%	1 0.03%	24 0.78%
Stop	5 0.16%	381 12.31%	6 0.19%	2 0.06%	1 0.03%	1 0.03%	0 0.00%	12 0.39%
Move_left	2 0.06%	1 0.03%	336 10.85%	1 0.03%	1 0.03%	2 0.06%	1 0.03%	10 0.32%
Move_right	1 0.03%	3 0.10%	5 0.16%	352 11.37%	2 0.06%	0 0.00%	0 0.00%	7 0.23%
Turn_left	2 0.06%	1 0.03%	1 0.03%	2 0.06%	348 11.24%	2 0.06%	1 0.03%	5 0.16%
Turn_right	6 0.19%	2 0.06%	1 0.03%	1 0.03%	1 0.03%	365 11.79%	0 0.00%	15 0.48%
Pause	2 0.06%	1 0.03%	1 0.03%	1 0.03%	1 0.03%	0 0.00%	371 11.98%	3 0.10%
Unknown	22 0.71%	1 0.03%	2 0.06%	1 0.03%	2 0.06%	1 0.03%	2 0.06%	379 12.24%

Figure 21. Confusion Matrix for 7 Gestures and an Unknown gesture class.

draw comparisons between the networks and evaluate their performances in hybrid data settings. The classification results were shown in the below Figure 21.

7.3 Multi-modal fusion

As anticipated in Section 6, multi-modal fusion is performed in the proposed architecture a-posteriori. This means that fusion is performed after each recognition module completed its analysis.

Our multi-modal fusion approach is strongly rooted in frame semantics. Frames are linguistic predicates providing a semantic description of real world situations. A frame is evoked in a sentence through specific lexical units (lu), i.e., words (such as nouns or verbs). A frame characterizes the set of semantic roles, i.e. semantic arguments called Frame Elements (FE), describing all participants to the event. An example collection of frames has been defined by FrameNet https://framenet.icsi.berkeley.edu/fndrupal/frameIndex. Anyway FrameNet, as well as other collections, do not include frames specific enough for the precision agriculture scenario. For this reason, specific frames have been defined for CANOPIES according to the functionalities implemented by other partners.

Fusion is performed with two different aims:

- Complementarity: each recognition module provides a specific piece of information (e.g., a frame element) which is combined with the others to fill the frame with all the needed frame elements. E.g., the speech module takes as input the sentence “Robot, move there” from which the “move” frame is extracted but no information about the “direction” frame element

is provided. The gesture recognition module recognizes the human worker indicating “left”, thus completing the instance of the frame.

- Redundancy: similarly to the previous case, each recognition module provides a specific piece of information, but two different modules can provide the same part of information.

The developed solution supports both the aims. If, after all recognition modules completed their work, the information received is not sufficient or it is inconsistent (e.g., the human worker said left with the speech but indicated right with the gestures), the dialog management module is employed to ask the human to finalize or to repeat the multimodal communicative act.

The multimodal fusion module (i) gathers all the information provided by the different recognition modules in a specified time window (e.g., 10 seconds), (ii) identify all the frames explicitly or implicitly (e.g., with gestures) indicated by the human users, (iii) associate the frame elements to all the information provided by the gestures, (iv) verify the consistency of the associations, (v) employ the dialog management module to solve inconsistencies if needed.

Few words more can be spent about the consistency check. In particular, in some cases multiple, discording, information can be associated with a specific frame element. In this case, if any of the recognition modules does not agree with the results, dialog management is used to ask the human worker to resolve the inconsistency. From the point of view of frames instead, specific recognition modules could provide a set of possible options in an implicit way (e.g., the “left” gesture may imply “move” as a frame or “pointing”. In this case we expect another modality (e.g., speech explicitly reports the correct frame).

Once a frame has been fully filled in, the corresponding service is called, through ROS queues, to perform a specific action or to retrieve or store a specific information.

8 Robot-to-Human Communication

8.1 Speech based communication

Speech based communication is obtained by using text-to-speech in order to transform frames generated by the dialog management module into speech which is synthesized using speakers.

8.2 Sound based communication

8.2.1 Sound Emission Hardware requirements

The areas of application of audible signaling devices are highly versatile. Multi-tone alarm sounders, piezo buzzers, and alarm horns are used in various industries. Application examples can be found in mechanical engineering, automation, and special vehicles. Audible signals are used, especially when there is no permanent visual contact between workers and the signaling device. They are generally used to signal faults or dangerous situations acoustically. The signal is given by one tone, several tones, or a particular tone sequence. There is an extensive range of alarm horns or sounders suitable for many applications. The efficiency of any audible signaling device depends on different parameters such as the size and/or distance of the area to be covered, the background noise, the pattern and frequency of a tone. In the following we are going to explain these parameters.

8.2.1.1 Distance

Loudness decreases as the listener gets further from the sound source, mainly due to "divergence." The intensity decreases because the energy is spread over a larger area, and it decreases inversely with the square of the distance from the source at a rate of 6dB for each doubling of the distance. So the sound output from an alarm rated at 106dB(A) will travel twice as far as a sounder rated at 100dB(A). If a sounder is rated at 100dB(A) at 1 meter, at 2 meters, it will be 94dB(A), at 4 meters, it will be 88dB(A), and so on.

8.2.1.2 Ambient background noise

The effective distance of a sounder is when the calculated dB(A) reaches at least 5dB(A) above the known ambient background noise. For example, the effective distance of a 100dB(A) at a 1-meter sounder in an ambient noise of 65dB(A) is the distance at which the sounder output level reduces to 70 dB(A).

A 120dB(A) at 1-meter sounder has a 70dB distance of approximately 300 meters i.e., 10 times the effective distance and, more importantly, 100 times the coverage area.

8.2.1.3 Pattern and Frequency of a tone

Sound output is also affected by the frequency of the sound. Lower frequencies tend to travel further, penetrate structures better, and are less likely to be attenuated by obstructions. However, Perception of a tone is not entirely dependent on frequency and sound level, and output with differing frequencies and/or temporal patterns will have a more distinct sound. This can be useful in areas of background noise where hearing protection may be worn. Usually, two-tone, intermittent, ramp-up, or ramp-down frequencies are the most effective.

Some companies offer state-of-the-art signaling devices, which we can find models for all applications. Audible signaling devices are divided into the following two categories:

Electronic sirens (multi-tone alarm sounders, surface, and panel mount buzzers):

Also named electronic sounders, are perfect for giving audible signals in case of emergency or for warning about imminent critical situations.

Electromechanical horns (mini horns and signal horns):

Electromechanical alarm horns are suitable for industrial and general applications and are used in various industries. With electromechanical signaling devices, the signal tone is generated mechanically.

8.2.2 Sound signals

Machines and appliances often use audio and visual cues, in the form of sound and light, to signal various states (e.g., error, idle, ongoing, or completed task). Sound signals can be used in different cases such as below:

1. For *warning* to call attention to imminent or potential danger
2. For presenting *information*
3. For situations when *darkness limits vision* or makes seeing impossible

Generally, auditory signals are better suited when the message must attract attention, even in a busy environment. Therefore, auditory signals are predominantly used for warning, especially when the message is short or simple, often with a flashing light. Such auditory signals may be single tones, sounds (mixture of tones), or spoken messages. In particular, tones and sounds may be continuous, periodic, or unevenly timed. They usually come from horns, bells, sirens, whistles, buzzers, or loudspeakers.

Generally, sound signals are provided under the following conditions:

1. The information to be processed is short, simple, and transitory, requiring an immediate or time-based response.
2. The criticality of the event makes supplementary or redundant notification desirable.
3. It is desirable to warn, alert, or cue the operator to the subsequent additional response.
4. Custom or usage has created anticipation of a sound signal.
5. Voice communication is necessary or desirable.

8.2.3 Sound Signal Types

The attention-gaining characteristics of the sound are based on different parameters such as rapidity of pulse pattern, frequency, intensity, and they should match the relative priority of the signal.

When an audio emission is required, the optimum signal type should be presented and the audio signals should not interfere with other sound sources, including verbal communication.

Each audio signal must have only one meaning. The ear acts as an effective detector of periodic signals in noise. Even when it is considerably weaker than the background noise, if the signal is a sinusoid (pure tone) or a combination of sinusoids (complex tone), the ear can detect it. The ear also efficiently detects periodic modulation in the very-low-frequency range and responds to variations in intensity or frequency. In addition, when speech supplements are used, the length of the initial alerting and the actual message must not interfere with other auditory inputs, including interpersonal voice communication unless the message is critical.

Like light signals, we can define three distinct sound signal types to convey different meanings in the desired condition. As light signals, sound signals can also convey Notification, Warning, or an Error message to indicate the status or warning. To emit sound signals, tonal sounds are recommended, and they should have the characteristics listed below to be used to notify the people around in the vineyard scenario.

1. Tonal signals should be at least 10 dB louder than the ambient noise.
2. Signal frequencies should differ from those dominating background noise.
3. Signal frequency range should be within 200 and 5000 Hz, best between 500 and 1500 Hz. If the signal undulates or warbles, 500–1000 Hz are advised.
4. Frequencies should be below 1000 Hz when signals must travel long distances (more than 300 m).

5. Frequencies should stay below 500 Hz when signals must “bend around” sound barriers or penetrate them.

The conspicuousness of a tonal signal can be elevated by increasing its intensity, by interrupting it repeatedly, and by changing its frequency. An example: increase from 700–1700 Hz in 0.85 s, be silent for 0.15 s (cycle time 1 second), then start over.

Audio warning signals should normally consist of two sequential elements: an alerting signal and an identifying or action signal.

Two element signals. When reaction time is critical and a two-element signal is necessary, an alerting signal of 0.5 second duration must be provided. All essential information must be transmitted in the first 2.0 seconds of the identifying or action signal.

Single element signal. When reaction time is critical, signals must be short. If a single element signal is permissible, all essential information must be transmitted in the first 0.5 second.

Caution signals. Caution signals must be readily distinguishable from warning signals and must be used to indicate conditions requiring awareness, but not necessarily immediate actions.

8.2.4 Characteristics of audio warning signals

Range: The frequency range shall be between 200 and 5,000 Hz and, if possible, between 500 and 3,000 Hz. When signals must travel over 300 m (985 ft), sounds with frequencies below 1,000 Hz should be used. Frequencies below 500 Hz should be adopted when signals must bend around obstacles. These obstacles in a vineyard scenario can be the walls or some partitions that the sound should pass through. The selected frequency band should differ from the most intense background frequencies and should be in accordance with other criteria in this section.

Intensity: The intensity, duration and source location of audio alarms and signals shall be compatible with the acoustical environment of the intended receiver as well as the requirements of other personnel in the signal areas.

Compatibility with clothing and equipment: Audio signals must be loud enough to be heard and understood through equipment or garments (e.g., parka hood, hearing protective devices) covering the ears of the listener.

Discomfort: Audio warning signals should not be of such intensity as to cause discomfort or "ringing" in the ears. Levels should not exceed 115 dB at the ear of the listener.

Audibility: A signal-to-noise ratio of at least 10 dB should be provided in at least one octave band between 200 and 5,000 Hz at the operating position of the intended receiver (it is usually sufficient). Signal to noise ratios can be greater as long as the levels do not exceed 115 dB at the ear of the listener.

Attention and avoidance of startle reaction: Signals with high alerting capacity should be provided when the system or equipment requires the worker to provide attention. Such signals should not, however, be so startling as to preclude appropriate responses or interfere with other functions by

holding attention away from other critical signals. To minimize startle reactions, the increase in sound level during any 0.5 sec period should be not greater than 30 dB. In addition, the first 0.2 sec of a signal should not be presented at maximum intensity and for this reason, the use of square topped waveforms, or present abruptly rising waveforms is suggested.

Onset and sound pressure level: The onset of critical alerting signals should be sudden, and a relatively high sound pressure level should be provided as specified above.

Headset: When the operator is wearing earphones covering both ears during normal equipment operation, the audio warning signal should be directed to the operator's headset as well as to the work area. Binaural headsets should not be used in any operational environment below 85 dBA where sounds that provide the operator with useful information cannot be directed to the operator's headset. Such sounds may include voices, machine noise that indicates wear or malfunction and other audible indications of system performance/mission status.

Use of different characteristics: Audio signals should be discriminable by using different characteristics for them. When several different audio signals are used to alert an operator about different types of conditions, discriminable differences in intensity, pitch, beats and harmonics, or temporal patterns should be provided. Signal intensity should not be used alone as a means of discriminating between signals. Warnings, instead, should differ on two or more parameters.

Coding. Where discrimination of warning signals from each other will be critical to personnel safety or system performance, audio signals should be appropriately coded. Alarms that are perceptibly different should correlate with different conditions requiring critically different operator responses. Such signals should be sufficiently different to help the worker to distinguish the signal meaning. Harmonically related frequencies should not be used to code different signals; they may, however, be used within a single signal.

Critical signals: The first 0.5 second of an audio signal requiring fast reaction shall be discriminable from the first 0.5 second of any other signal that may occur. Familiar signals with established names or associations shall be selected. Speech should be used whenever feasible.

In conclusion, all sounds provide only supplemental information that users should never need to rely on sound alone for feedback. Thus, all sounds, including those for errors and warnings, should be brief, subtle, and calm. In general, higher pitched sounds about the 1KHz range are more unpleasant and therefore draw more attention and are more used for alerts that need immediate attention (e.g. a fire alarm)

Lower pitched tones about the 100Hz range are more pleasant and more suitable for notification signals. Then, the faster the beep, the more annoying and hence alerting. It rarely makes sense to use a beeping sound for a notification, so they are usually used more for attention grabbers.

8.2.5 Sound Signal Design

Auditory warnings are functionally defined as audible stimuli that capture attention and provide listeners with information, prompting behavior. Based on Patterson's hierarchical framework, a

warning sound is comprised of a set of sound bursts that are themselves made up of a set of sound pulses. A pulse is the fundamental unit of sound containing one tone or multiple harmonics.

Technical characteristics of warning sounds include elements of the signal's sound spectrum and its temporal dimension; both characteristics can affect perceptions of the sound and consequently impact behavioral responses. Key features of an auditory warning's sound spectrum include:

- the length of the sound pulse,
- the frequency of each harmonic within the pulse,
- the length of any onset and/or offset envelopes, and
- the sound intensity (often measured in decibels).

The temporal dimension includes the number of times each pulse repeats, the time interval(s) between repeated pulses (speed), the total duration of the warning sound, and any changes in the pitch or fundamental frequency of pulses over the duration of the warning sound.

Psychoacoustic experiments identified speed and fundamental frequency (the lowest frequency among harmonics within a warning sound) as having the most notable impacts on perceived urgency and reaction times. Higher speed and higher fundamental frequencies evoke shorter reaction times, and both are associated with higher perceived urgency. However, speed had the most significant effect.

The total duration and loudness level of auditory warnings can impact behavioral response. Aversive aural stimuli with short durations (e.g., 5000 milliseconds [ms]) reduce unwanted behaviors more reliably than stimuli with longer durations. When sounds are very short (i.e., ≤ 200ms), the human ear fails to discriminate acoustic properties typical of warning signals. Thus, auditory warnings expressly designed to modify behavior might best fit the 200 ms to 5000 ms range. Loudness level is another crucial characteristic to consider when designing a signal to modify behavior. Sounds below the masking threshold of a given environment are challenging to discriminate and could fail to command attention or provide information. Sounds that are too loud can irritate users, mainly when applied in environments with noise-sensitive persons, interfere with interpersonal communications, and lead users to switch the sounds off.

Based on previous studies, auditory warning selection criteria are as follows:

- Be aversive but not sufficient to elicit irritation or evoke counter aggression (e.g., damaging the instruments).
- Be perceived as urgent (i.e., with a fast pace and high frequency)
- Be composed of sound bursts with either constant or descending frequencies.
- Be sufficiently loud to be heard over ambient noise at a "normal" volume.
- Consist of synthetic sounds that did not resemble musical instruments or a human voice.
- Have a duration greater than 200 ms but less than 5000 ms.
- Not to signal previously learned responses (e.g., an alarm clock).
- Not be commonly used in homes or workplace settings.

Sound signal number	Sound spectrum	Temporal description
1	Pulse 500 milliseconds(ms) long: 1 tone at 500 hertz (Hz); onset envelope 10 ms; 62 decibels (dB)a	Pulse repeats every 270 ms for 3 cycles
2	Pulse 500 ms long: 1 tone at 800 Hz; onset envelope 10 ms; 84 dB	Pulse repeats every 270 ms for 3 cycles
3	Pulse 18ms long: 1 tone at 500 Hz; 70 dB	Pulse repeats every 45 ms for 3 cycles with one pause of 348 ms; Burst repeats every 600 ms for 3 cycles
4	Pulse 18 ms long: 1 tone at 800 Hz; 72 dB	Pulse repeats every 45 ms for 3 cycles with one pause of 348 ms; Burst repeats every 600 ms for 3 cycles

Table 6. Proposed sounds.

Table 6 is an example of four distinct sounds proposed. Results from empirical sound tests in previous studies are used to select the auditory warnings for the behavioral intervention. Findings indicate that warning sounds 1 and 2 are more aversive, more often associated with a warning or alert and with the color red, and more likely to elicit a behavioral response. Sounds 2 and 3 are more pleasant to hear with respect to others which are suitable to be used as notification sounds.

8.3 Light based communication

8.3.1 Light Emission Hardware requirements

Light signals are used in a wide variety of industrial applications. Therefore, there are different pieces of equipment to emit light. A few of them are Signal beacons, signal transmitters, and signal towers and they are not limited to specific sectors or industries.

Certain signaling equipment is more suitable for certain industries due to its technical specifications and the characteristics of the industry and its application. Hardware requirements are met in different projects in collaboration with consumers based on the difficulty and specificity of the application.

An example of signal lights and signal lighting products are strobe lights. Strobe lights are one of four types of beacons in signaling technology, and they generate the highest level of attention. A wide range of strobe lights exists in various sizes, forms, panel or surface mount, and light technology. Strobe lights are available in two different technologies:

- LED strobe lights
- Xenon strobe lights

The table below illustrates a comparison between these two technologies.

LED Technology	Xenon Technology
Accurate light direction with a sharply defined radiation angle	Omnidirectional radiation characteristics
Different modes (permanent, blinking or flashing)	Flash mode with selectable frequency
Optional brightness adjustment	Maximum brightness
Long flash sequences	Short flash duration
Life cycle approx. 50,000 operating hours	Life cycle approx. 10 million flashes

Table 7. Comparison between LED strobe lights and Xenon strobe lights

LED or Xenon strobe lights achieve the best signaling effect and attract maximum attention. Both LED and xenon technology have their place in the range of visual signaling devices.

However, some advantages make LEDs suitable for outdoor scenarios and, more specifically, the vineyard. These advantages are described as followed:

1. The Ability to Operate in Cold Conditions: Traditional lighting sources do not like cold weather. When the temperature drops, lighting sources, particularly fluorescent lamps, require a higher voltage to start, and their light intensity diminishes. On the other hand, LED lights perform better in cold temperatures by about 5%. This is why LED lights are a better choice for lights used in outdoor scenarios.
2. Design Flexibility: LEDs are very small (about the size of a speck of pepper), which means that they can be used in almost any application.
3. Low Voltage Operation: LEDs are perfect to use in a location where flooding may occur because they operate on very low voltage. Using a low-voltage system in areas that may be prone to flooding, It is possible to protect the staff and others from potentially harmful or fatal shocks. This also makes them extremely useful to use outdoors where other lighting solutions might not meet with local codes.
4. Instant Lighting and the Ability to Withstand Frequent Switching: LED lights are not affected by frequent switching, and it does not cause any reduction in their lifespan or efficiency. Therefore, they can be rapidly cycled for flashing light displays or applications that require sensors that frequently switch from on to off and back again.

8.3.2 Robot Light Signal Emission

Robots are complex machines that require a high degree of expressivity to convey the many facets of their state. Recent works have suggested using light signals as a possible robot communication modality since authors of recent studies discovered that light signals, specifically LEDs, are affordable, reliable, salient, and versatile. Modulating light signal parameters, such as color, position, duration, and frequency, can be used by robots to communicate a wide range of state information. Although many devices have employed light signals, there are few universal guidelines or principles to drive their design. As robots take on different forms and purposes, their signaling needs will continue to

grow. To this aim, recent studies proposed a model for generating robot light signals for Human-Robot Interaction (HRI). The key insight of these works is to consider human users' existing perceptions or biases to design more intuitive light signals for robots.

The signal types are defined as a Notification, Warning, or Error. Notification is defined as an information provided by the robot that may be beneficial for the human. Warning signals should be provided to warn personnel of impending danger, to alert an operator about a critical change in the system or equipment status, and to inform an operator about a critical action or actions that must be taken. An alerting/warning signal shall provide the worker with a greater probability of detecting the triggering condition than normal observation would provide in the absence of the signal. Error signals, on the other hand, are used when the successful operation of the robot is prevented.

Generating the robot light signals can be done by manipulating the robot's light signal and base motion parameters. In the following section we are going to define these variables.

8.3.3 Light signal parameters

Three common and fundamental light signal parameters are color, pattern, and pattern frequency which are described below:

1. Color: Color is a common differentiator for light signals on individual devices. The common colors that usually are adopted are green, red, orange, blue, and white.
2. Pattern: Four basic and frequently utilized patterns are Steady, Blink, Beacon, and Wipe. In a static light signal, all lights in the array are on. Blink consists of the entire array turning on and off at a constant frequency. The beacon is made up of three lights moving across the array. Wipe is visually similar to a progress bar, in which subsequent lights are turned on to fill the array. Table 8 illustrates each light pattern.
3. Pattern Frequency: The steady pattern is static and does not have a pattern frequency. For the other three patterns, we can define the frequency by varying the speed of the pattern to repeat itself, either fast or slow.

Steady	
Blink	
Beacon	
Wipe	

Table 8. Light patterns

Light Signal Parameters																					
Color	Blue							Orange							Red						
Pattern	***Std***	***Blk***		***Bcn***		***Wp***		***Std***	***BLK***		***Bcn***		***Wp***		***Std***	***Blk***		***Bcn***		***Wp***	
Pattern Frequency		**S**	**F**	**S**	**F**	**S**	**F**		**S**	**F**	**S**	**F**	**S**	**F**		**S**	**F**	**S**	**F**	**S**	**F**

Table 9. Pattern frequencies

Moreover, Two well-known motion parameters are path and speed, which significantly affect light signal perception.

1. Motion Path: Motion path is defined as the robot's base motion direction and has three levels: no path, approaching, and receding.
2. Motion Speed: The robot's motion speed can be either slow or fast.

The light and motion parameters affect human attribution of desired task-related properties.

There are also three dependent measures: urgency, error, and difficulty to ignore which are defined as follow:

1. Urgency: Urgency is defined as the need for immediate attention.
2. Error: Since robots are agents that can physically interact with the world, the robot's ability to operate or function correctly should be considered for human safety and task success.
3. Difficulty to Ignore: This property measures the salience of a signal and is shown to affect the success and speed of eliciting a human response.

Dependent parameters are used as attributes to determine the signal types in designing light signals.

8.3.4 Signal Design Algorithm

As a systematic approach for signal design, a decision tree for light signal generation is proposed based on light signal parameters. The decision tree depicted in Figure 22 is employed for each of the three notification attributes: urgency, error, and difficulty to ignore. The decision tree includes two signal design parameters: information type and urgency. Light signals are better suited for conveying broader categories of information. Therefore, the goal of this decision tree is to generate a set of generalizable signals well suited for many HRI scenarios.

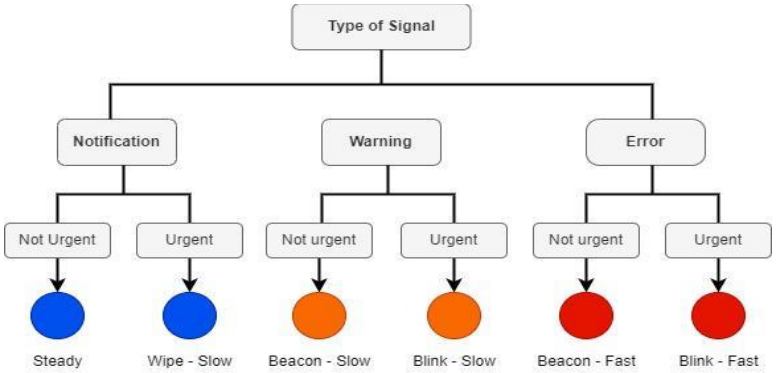

Figure 22. Representation of the resulting decision tree for light signal generation.

8.3.4.1 Generating Light Signals

This section will discuss each of the three types of designs and the different parameters we should choose to better define each light signal. Regarding the robot base (Alitrack), we would like to consider the light signals as in cars, for a total of 4, in particular: 2 horizontal strip LED lights (Figure 23) on the back and 2 horizontal strip LED lights on the front of the Alitrack. Moreover, we are investigating the introduction of 2 lights on the torso of the Tiago++ robot, precisely one horizontal strip LED on the back and one flat LED (Figure 2.2) on its head. Hence, by selecting a feasible length for the aforementioned devices, we believe that each horizontal strip LED light can be of 15 cm length, while the flat one can be equipped with 2 small LED lights emitting signals with different colors. The final selection of strip leds will not anyway influence the decisions taken.

Figure 23. Horizontal strip LED lights that can emit different colors (white, yellow, orange, red, green, blue).

Figure 24. Flat LED light that can emit different colors (white, yellow, orange, red, green, blue).

8.3.4.2 Notification Signal Design

In general, there are two distinct color groupings for light signals:

- red, orange
- green, blue, white

For light signals indicating the notification, both blue and green signals can be employed. However, blue signals are preferred since green signals are rated more salient than blue or white, and the green color is more suitable to express a positive state (e.g., operating correctly).

Generally, non-urgent notification signals are used for information that is not critical to the human or constrained by time; hence, they require low levels of awareness or attention from the human worker. The steady light pattern is rated the lowest in salience, urgency, and error, making it suitable for non-urgent signals.

Considering the user studies, we can conclude that the wipe pattern is rated the next lowest for salience and error; therefore, it can be chosen to represent urgent notifications.

8.3.4.3 Warning Signal Design

In designing light signals, color provides the highest information gain for the error measure, with red and orange indicating the highest error levels. Since orange (or yellow) is often employed for warning signals in other applications, we suppose that orange signals can be employed for the warning signal type also in CANOPIES. Moreover, warnings require a more significant level of human attention, suggesting the need for higher saliency signals. Beacon and blink signal patterns are rated highest in saliency and error. The slower pattern frequency can be employed for warning signals that are less critical than error signals.

8.3.4.4 Error Signal Design

As red signals are rated significantly higher for urgency, error, and salience, red can be employed for both types of error signals. Like the warning signal design, beacon and blink can be chosen for their higher ratings of error and saliency. Like the warning signal design, beacon and blink can be chosen for their higher ratings of error and saliency. All errors require higher levels of human awareness and, consequently, we think that only the faster pattern frequency can be employed for such signals.

8.4 Multi-modal fission

Multi modal fission has been implemented using a rule based system that, according to the message generated by the dialog manager or by the hazard manager, selects the best option for implementing the specific communication pattern.

9 Wrap-up, future work and final demonstration

In this deliverable, we presented the results obtained in the Task 2.3, concerning the specification of the human-robot interaction to be implemented in the context of the CANOPIES project. In particular, the document introduces the concepts of actions and activities, whose representation formalisms will be the goal of Tasks 6.3 and 6.4 respectively. In addition, we introduce the different communication acts that will be implemented and how they will be available in the form of communication modalities. We also introduce which are the environmental constraints that may tamper the communication between humans and robots. The document also defines the multimodal architecture defined for Task 2.3 and how Human-to-Robot and Robot-to-Human communication have been implemented. The work for Task 2.3 also resulted in the following scientific publications:

[1] Sara Kaszuba, Francesco Leotta and Daniele Nardi. "A preliminary study on virtual reality tools in human-robot interaction." Augmented Reality, Virtual Reality, and Computer Graphics: 8th International Conference, AVR 2021, Virtual Event, September 7–10, 2021, Proceedings 8. Springer International Publishing, 2021.

[2] Sara Kaszuba, Sandeep Reddy Sabbella, Francesco Leotta and Daniele Nardi. "Testing Human-Robot Interaction in Virtual Reality: Experience from a Study on Speech Act Classification." 6th International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions at HRI 2023. IEEE (To appear).

[3] Sara Kaszuba, Sandeep Reddy Sabbella, Francesco Leotta and Daniele Nardi. “Speech Act Classification in Collaborative Robotics”. IEEE Roman 2023, Hybrid Event, August 28-31 2023. IEEE (To appear)

[4] Sandeep Reddy Sabbella, Sara Kaszuba, Francesco Leotta and Daniele Nardi. Virtual Reality Applications for Enhancing Human-Robot Interaction: A Gesture Recognition Perspective. 23rd ACM International Conference on Intelligent Virtual Agents (Extended abstract). ACM (To appear)

In the following months, the activities of Task 6.3 will continue as part of Task 6.4, with the refinement of the implemented modules and the final version of supported actions and activities.

The final demo to demonstrate the outcome of the task must still be defined with the other partners, but possible demonstration could include:

- Demonstration of the human-to-robot communication based on speech only
- Demonstration of the human-to-robot communication based on the gestures only
- Demonstration of multimodal fusion including the effect of incomplete frame specification through the dialog manager

This project is funded by the European Union G.A No 101016906

- Demonstration of the robot-to-human communication with the different modalities, also including the hazard manager