applied sciences

MDPI

Article

Dialogue Enhanced Extended Reality: Interactive System for the Operator 4.0

Manex Serras [1,2,*], Laura García-Sardiña [1], Bruno Simões [1], Hugo Álvarez [1] and Jon Arambarri [3]

[1] Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián, Spain ; lgarcias@vicomtech.org (L.G.-S.); bsimoes@vicomtech.org (B.S.); halvarez@vicomtech.org (H.Á.)

[2] SPIN—UPV/EHU, Department of Electrical and Electronics, Faculty of Science and Technology, Campus de Leioa, 48940 Leioa, Bizkaia, Spain

[3] VirtualWare Labs Foundation, Calle Usausuaga 7, 48970 Basauri, Spain; jarambarri@virtuawareco.com

* Correspondence: mserras@vicomtech.org

Received: 20 May 2020; Accepted: 4 June 2020; Published: 7 June 2020

Abstract: The nature of industrial manufacturing processes and the continuous need to adapt production systems to new demands require tools to support workers during transitions to new processes. At the early stage of transitions, human error rate is often high and the impact in quality and production loss can be significant. Over the past years, eXtended Reality (XR) technologies (such as virtual, augmented, immersive, and mixed reality) have become a popular approach to enhance operators' capabilities in the Industry 4.0 paradigm. The purpose of this research is to explore the usability of dialogue-based XR enhancement to ease the cognitive burden associated with manufacturing tasks, through the augmentation of linked multi-modal information available to support operators. The proposed Interactive XR architecture, using the Spoken Dialogue Systems' modular and user-centred architecture as a basis, was tested in two use case scenarios: the maintenance of a robotic gripper and as a shop-floor assistant for electric panel assembly. In both cases, we have confirmed a high user acceptance rate with an efficient knowledge communication and distribution even for operators without prior experience or with cognitive impairments, therefore demonstrating the suitability of the solution for assisting human workers in industrial manufacturing processes. The results endorse an initial validation of the Interactive XR architecture to achieve a multi-device and user-friendly experience to solve industrial processes, which is flexible enough to encompass multiple tasks.

Keywords: extended reality; human–machine interaction; augmented reality; dialogue systems; industrial voicebots; Operator 4.0; Industry 4.0

1. Introduction

Human Augmentation (HA) can be described as a functional extension of the physical body mediated through technology to increase human productivity or enhance body capabilities [1]. In industry, the enhancement of workstations with augmentation technology can reshape the human presence in the process value chain and support the development of self-awareness and new skills, primarily where manual labour is inevitable [2,3].

The major technological factors impacting the industry sector are cyber-physical systems, big data analytics, collaborative and fully connected robots, and interaction devices that exploit the integration of wireless network capabilities with Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (MR) [4]. Industry 4.0 is an umbrella term that encompasses a wide range of concepts belonging to the new industrial revolution moved by interconnectivity, automation, machine learning, and real-time data analytics [4]. Real-time data insights and communication tools can now empower operators, actively and passively, to perceive and learn information about new assembly processes in a timely manner without altering their established work routines.

eXtended Reality (XR) technology (the full spectrum of technologies in the virtual-to-reality continuum) is paving the way to new forms of interaction that disrupt traditional desktop interaction, where the degree of freedom and mobility of users is limited. In this paper, we describe a streamlined approach to integrate spoken and natural interaction to industrial processes as a solution to facilitate the communication between operators and the technology deployed at their workspace. This provides a powerful complement to the main approaches of using visual computing technologies (such as Augmented, Virtual, and Mixed Reality) to enhance the role of the operator 4.0, as suggested in [5]. We propose a decisional process management mechanism to generate XR experiences based on unstructured and user-centred interactions for industrial processes, such as maintenance, repair, and decision making. Our research work assesses two hypotheses: (1) the combination of different HA technologies within a user-centred and interactive workflow reduces the operator′s cognitive barriers when performing industrial tasks, and (2) the proposed decisional process management mechanism can facilitate the processes of managing, distributing, and communicating the domain knowledge in factory-floors as it is easily adopted by the end users. This paper presents an Interactive XR (IXR) architecture using the SDSs modular and user-centred architecture as reference. To validate the proposed approach, two industrial use cases are presented and evaluated with real users in terms of technology usability and adoption parameters. To prove the usefulness of the IXR when it comes to reducing cognitive barriers, users with little knowledge about these processes were selected for the first use case, and workers with cognitive disabilities were selected for the second use case.

This paper is structured as follows: Section 2 is a review of previous studies on XR interaction models and Spoken Dialogue Systems (SDSs). Section 3 formulates the new interactive XR framework built on the concept of natural interaction interfaces. Section 4 describes the prototype's use cases, and Section 5 details the IXR implementation that has been adopted in this paper for the presented use cases. Finally, Section 6 discusses the research findings, and in Section 7, we draw conclusions and discuss possible lines for future work.

2. Background

Industrial control and execution processes are carried out by technical operators, shift leaders, field workers, and engineers. These processes, to be carried out properly, require complex management of the domain knowledge, where the creation, distribution, access, and communication of this knowledge is critical [6]. In addition, the process industry requires specific expert knowledge, thus making the training and education of workers a challenging task, usually ad hoc to each product and factory. As a result, domain knowledge is not distributed at all, and it heavily relies on a few expert operators [7]. As immersive technologies are becoming gradually more robust and affordable, new case studies and applications are being explored to enhance workers with new skills, decreasing the skills gap between untrained and expert operators by providing just-in-operation computer-based task assistance. Along the way, Industrial Augmented Reality (IAR) emerged as a research line focused on how these technologies can cognitively enhance workers in industrial processes. IAR differs from traditional systems in terms of quality and reliability, which must comply with those of manufacturing industries.

The following subsection describes the different Industrial AR applications and how they were used to improve and enhance human capabilities, reduce cognitive barriers and improve workplace efficiency in several scenarios. After this, Spoken Dialogue Systems and their modular architecture are presented, and how they have been used as natural, user-centred process solving systems. This background motivates the conceptual system presented in Section 3, in which different AR mechanisms are encompassed in a modular, natural, and user-centred architecture, similar to the one that SDSs employ.

2.1. Industrial Augmented Reality Systems

The first AR prototypes explored the human vision senses and tackled the combination of image registration and content visualisation. The first seminal work can be traced back to Thomas Caudell and David Mizell [8] in 1992. Their prototype allowed for a computer-produced diagram to be superimposed and stabilised on a specific position on a real-world object to help the worker with cable wiring. The contributions of Kollatsch et al. [9] led to an industrial prototype for the visualisation of information from control systems (e.g., Programmable Logic Controller and Computer numerical control machines) directly in situ, enabling real-time digital content to overlap real-world objects. To prevent the drawbacks of AR technology in terms of visual inputs, and mainly their limited field-of-view, projection-based approaches have been broadly presented as an alternative to AR and VR. These solutions are referred to as *projection mapping* or Spatial Augmented Reality (SAR). For example, Sand et al. [10] developed a prototype to project instructions into the physical workspace, which helped end users to find the pieces required to assemble products without prior knowledge. Rodriguez et al. [11] proposed a similar solution in which instructions were directly overlaid with the real world using projection mapping. Similarly, Petersen et al. [12] projected video overlays into the environment at the correct position and time using a piece-wise homographic transform. By displaying a colour overlay of the user's hands, feedback can be given without occluding task-relevant objects. More recently, Álvarez et al. [13] have improved the manufacturing process of a packaging company by integrating an SAR system in a real factory floor to provide assistance to operators during the setup of die cutters. Projecting virtual content directly on physical objects contributed to reducing the mental workload of the operator (no need to interpret the real workspace on a screen) [14] and an improvement of other practical open challenges like the need for the worker to hold a tablet or wear HMD devices. However, there are a number of challenging factors of IAR development that need further research. Some challenges are transversal and related to the necessary interdisciplinary knowledge in areas such as computer graphics, artificial intelligence, object recognition, and human–computer interaction [15]. For example, user intuitive interfaces still remain a challenge, particularly in situations where understanding the user's actions and intentions is required for their adaptation in unexpected conditions. The use of mid-air gestures enable users to interact with virtual objects using their hands and with varying levels of intuitiveness [16,17]. Techniques grounded on gaze [18,19] (based on head and/or eyes' movements), electromyography (by analysing muscular activity) [20], electroencephalogram (brain electrical signals) [18,20], and hands tracking [21,22] were developed to empower users with new forms of interaction that do not require hands to be concurrently used for interaction and to perform the task [23]. This type of interaction is particularly helpful in industrial assembly settings where dexterity of the hands is key to perform the task [3]. Voice-based interfaces' popularity is rising as well, although these solutions are often limited to a small set of fixed commands. Few of them add a natural layer to allow a more comfortable experience. In Section 2.2, we provide a literature review of voice-driven interaction systems and their modelling processes.

2.2. Spoken Dialogue Systems

Spoken Dialogue Systems (SDSs) are voice-enabled Human–Machine Interfaces for natural communication with a computer, robot, and other devices [24,25]. SDSs rely on processing and exploiting domain knowledge to guide users and solve their needs [26–28]. One of their main advantages is that they exploit spoken language, rendering a convenient and frictionless system, which communicates in a natural way.

Additionally, these systems facilitate hands-free access to information, thus enabling online information consumption for processes that require manual work. Also, SDSs can be enhanced with error handling techniques that detect interaction breakdowns during the process due to misunderstanding, channel noise, etc. [29,30], improving the communication between the user and the system. As SDSs provide a suitable response for the knowledge management and expertise distribution challenge, it comes as no surprise that they have been used in multiple sectors. To name a few examples, these include: LARRI, a dialogue-based system for support in maintenance and repair activities for aircraft mechanics [28]; the International Space Station procedural assistant [31]; bus scheduling systems [32,33]; tourism booking systems [34]; resident helping with robotic assistants in nursing homes [35]; elderly assistance and coaching systems [36]; educational dialogue systems for tutoring [37–39]; and retail assistance [40]. Despite their popularity, past limitations in speech recognition engines and in microphones' noise-cancelling mechanisms prevented the application of SDSs in industrial environments.

To carry out a flexible and natural communication with the operator, the proposed framework emulates the modular and user-centred SDS architecture. Traditional SDSs are a pipeline of specialised technological modules (see Figure 1). The first module encapsulates a Speech To Text (STT) service to transform audio signals into textual transcriptions. Audio transcriptions are processed by a Spoken Language Understanding (SLU) module and encoded as semantic actions (i.e., the communication intent of the operator, which can be understood as a simplification of what the user meant to say), so the linguistic variability of the input is reduced (e.g., "Tell me what to do now", "What's the next step?", and "Now what?" can all be represented as the same semantic codification: [*intent=ask, object=next-step*]).

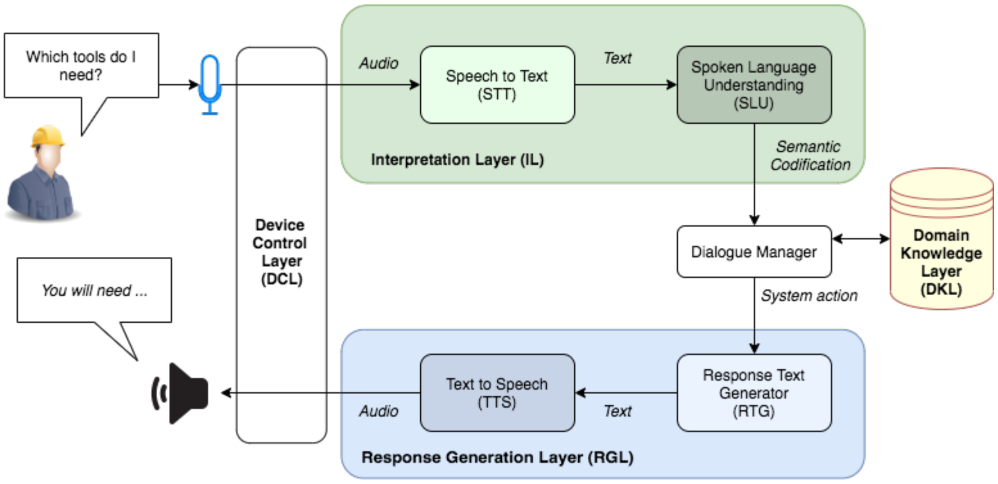
Which tools do I need?
Device Control Layer (DCL)
Audio
Speech to Text (STT)
Text
Spoken Language Understanding (SLU)
Interpretation Layer (IL)
Semantic Codification
Dialogue Manager
Domain Knowledge Layer (DKL)
System action
You will need ...
Text to Speech (TTS)
Response Text Generator (RTG)
Audio
Text
Response Generation Layer (RGL)

Figure 1. Traditional dialogue system architecture.

Semantic actions are passed down to the Dialogue Manager (DM) module, which is in charge of defining the best strategy to give an appropriate answer to the operator. The interaction strategy relies on real-time flow states and existing domain knowledge. In the last process stage, the interaction feedback computed by the DM module is converted into a human-interpretable interactive action by the Response Generator (RG) and synthesised into audio speech by the Text to Speech (TTS) module.

The modules that constitute an SDS can be implemented in many different ways and they can integrate a vast range of technologies [41,42]. The in-detail description of these modules exceeds the scope of this paper. Withal, it is noteworthy that module stack strategies facilitate the deployment of SDSs both in scenarios where enough labelled data are available, where Machine Learning can be successfully applied, and in data scarce scenarios, where mechanisms such as encoding expert rules, designing dialogue flows, and others can be implemented.

3. Conceptual System

For HA technology to be effective it has to complement existing cognitive processes rather than creating competing or shared channels of information for them. It must be adapted for the underlying mechanisms and processes of human perception and biology, and avoid cognitive barriers triggered by decisional processes. Most cognitive barriers can be removed when proper information processing methods are applied. For example, the momentary pause that is required for the worker to process the need and execute a specific command is a cognitive barrier, which is removed only when workers intuitively understand what to do to overcome the barrier. That being said, invisible interfaces and unclear interactions still represent a potential abandonment or frustration point to workers who cannot figure out the interaction mechanics.

Spoken language is a natural form of communication for humans to interact and share information. A system that is capable of conveying and managing information in real time using natural language as an interface is expected to be an intuitive and frictionless solution against cognitive barriers that may arise due to lack of knowledge about some process to solve. Recent advances in speech recognition and language processing technologies enabled SDSs to gain popularity as voice-guided task solvers in multiple domains [28,40,43], but their combination with and role within XR scenarios is still unclear.

Conceptual Workflow

In this section, we describe an Interactive eXtended Reality (IXR) system that helps operators to carry out a certain task or process through the combination of XR technologies with the SDS process control logic (see Figure 2). The proposed work streamlines multiple input and output XR devices into the logical scheme of SDSs. As a result, we describe a framework that enhances both classical SDSs and XR devices as Human–Machine communication interfaces.

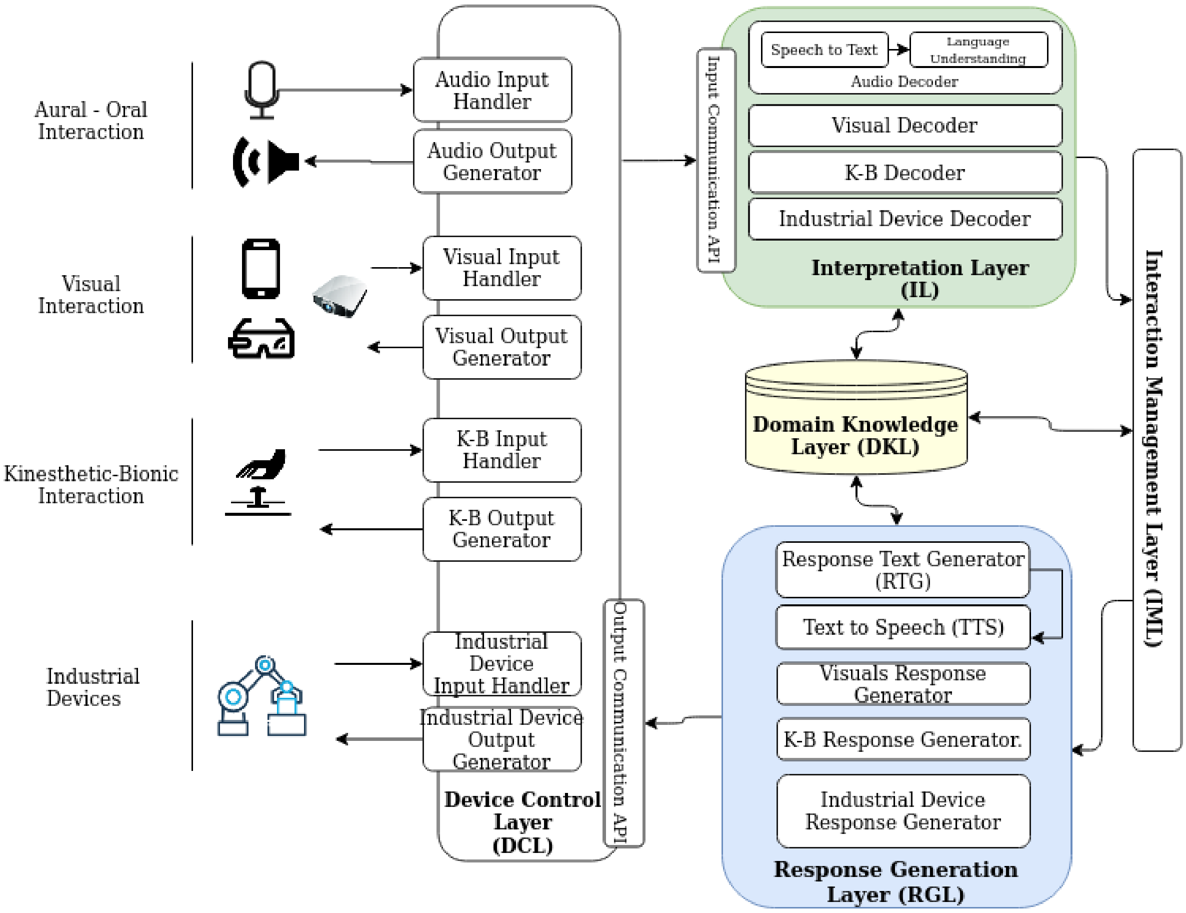
Aural - Oral Interaction
Visual Interaction
Kinesthetic-Bionic Interaction
Industrial Devices
Audio Input Handler
Audio Output Generator
Visual Input Handler
Visual Output Generator
K-B Input Handler
K-B Output Generator
Industrial Device Input Handler
Industrial Device Output Generator
Device Control Layer (DCL)
Input Communication API
Output Communication API
Speech to Text
Language Understanding
Audio Decoder
Visual Decoder
K-B Decoder
Industrial Device Decoder
Interpretation Layer (IL)
Domain Knowledge Layer (DKL)
Response Text Generator (RTG)
Text to Speech (TTS)
Visuals Response Generator
K-B Response Generator.
Industrial Device Response Generator
Response Generation Layer (RGL)
Interaction Management Layer (IML)

Figure 2. Architecture for the Interactive eXtended Reality (XR) framework.

A Device Control Layer (DCL) handles machine-dependent data transformations and communication protocols between devices. The IO API is a modular interface for the use of peripheral devices in industrial environments, for example, collaborative robots. Analogous APIs are available for, for example, aural and visual interaction devices. The DCL copes with the Interpretation

Layer (IL) and the Response Generation Layer (RGL) to gather and send the exchanged data content and format from and to each involved device.

Data processed by the DCL is accessed by the IL and transformed into a normalised semantic encoding, similar to the SLU module in Figure 1. The objective is to discretise the highly-variable data collected by multiple input devices and extract meaning out of it. For example, if the operator asks for the robot's malfunctioning parts, the semantic interpretation of both the operator's audio transcription and the robot sensors' data are extracted in this module.

The Interaction Manager Layer (IML) processes the encoded information of every input device to decide the next system response and communicates it to the RGL. In order to decide the appropriate response to give to the operator, the IML takes into account the semantic intention of the operator (in a similar way to [44]), but also the status of the contextual knowledge stored in the Domain Knowledge Layer, such as data from the involved physical industrial devices, ontologies, operator profiles, and so on. The system response is represented as a set of semantic actions, which is then sent to the RGL. With each user turn, the IML updates the interaction state according to the processed input and the selected system responses.

Finally, the RLG transforms the set of semantic actions in the system response to data formats usable by the target output devices so they can be communicated to the operator, for example, rendering commands, natural language audio or text, and device movement commands. These device-usable responses are then sent to the DCL to close the communication loop with the operator.

4. Use Cases

In this section, we describe two use case scenarios that have been validated by 20 operators of different characteristics not familiarised with the project, 10 for each use case. These operators received minimal instructions on how to use the system and about their given task.

4.1. Use Case 1: Universal Robot's Gripper Maintenance

We chose the maintenance of a Universal Robot's gripper as our first use case. Participants were assigned the task to perform a periodic inspection for the robot's gripper (the actual instruction manual can be consulted online in the following address: https://assets.robotiq.com/website-assets/support_documents/document/3-Finger_PDF_20190322.pdf?_ga=2.105997637.750885948.1563187207-1298495031.1563187207assets.robotiq.com/website-assets/support_documents/document/3-Finger_PDF_20190322.pdf, last accessed 29/05/2020—see section *7.3 Periodic Inspection*), which is necessary to ensure its good condition and safe operation. In the experimental setup, the robot arm was placed over a table, fixed on a pre-set position, with the gripper coupled. The gripper was located about the chest-height of an average person and facing up, with its fingers closed, as shown in Figure 3. The tools needed for the task were placed on the working-table.

Before each participant's session would begin, they were informed that they would have to complete a form for the system's assessment afterwards. At the beginning of the session, participants were invited to try and get familiarised with the hardware (in this use case, HoloLens Glasses were used for outputting visualisations). While doing this, a researcher would present the participant with the robot, the system, and the task at hand that they had to complete. Once the explanation was done, the task would begin.

The completion of the task requires participants to operate with both hands, making it an appropriate task for the integration and testing of the proposed XR architecture. The set of instructions described in the manual is equivalent to a series of processes that require a single action to be performed by the user in each step. The compiled set of steps works as a base dialogue flow where each of the steps corresponds to a dialogue state. We added some extra steps to the flow to accommodate the system's contextual needs: an introductory step to define the state prior to the beginning of the task, and a final step to mark the conclusion of the interaction. Figure 4 shows the sequence of steps required to complete this use case's task. In addition, as the system can exploit the domain knowledge to guide the operator, knowledge about multiple question types that could arise during the task was included—what to do in the current step, how to do it, where are the involved parts or items, and so on.

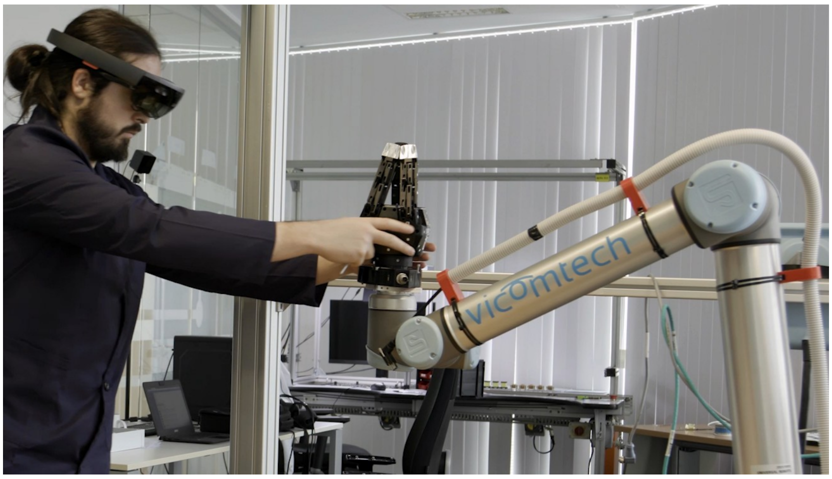

Figure 3. An operator uncoupling the Universal Robot's Gripper.

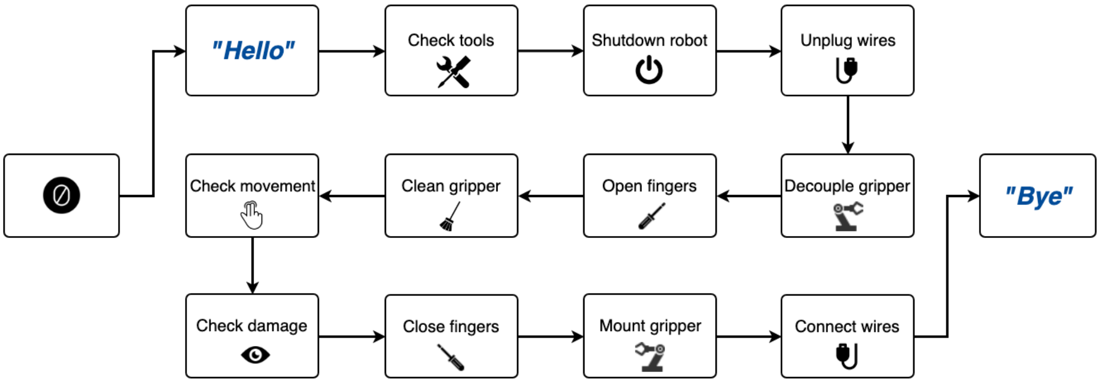

Figure 4. Sequence of steps required to complete the gripper's maintenance task.

During the maintenance task, the IM decides which AR animation to render (pointing the location of bolts to unscrew or cables to connect or disconnect, showing the movement to perform to check the gripper's fingers, and so on). These animations are accompanied with aural responses and contextualised according to the current step in the maintenance process, for example, the bolts to unscrew are different in the decoupling step and the finger-opening step. Furthermore, the operator can explicitly ask to repeat an AR animation as many times as needed.

Ten people (50% male, 40% female, 10% would rather not say their gender) participated in the testing of use case 1. A total of 50% were in the age range between 25 and 29, 30% between 35 and 39, 10% between 30 and 34, and 10% between 18 and 24. All the participants had European Spanish as one of their mother tongues. Participants were all volunteers working at the same centre where the prototype was developed and their data were treated anonymously (the operator appearing in Figure 3 gave consent to its use in this publication).

4.2. Use Case 2: Industrial Electrical Wiring

The proposed architecture was also validated for cabling tasks, more specifically the assembling of electric panels. This use case was proposed by industrial partners of the project, since it is a process that they carry out on a daily basis. Voice interaction was achieved by headphones with noise-cancelling, which participants could choose to wear on them or leave on the working table. To handle the visual communication, an Optoma projector with a wide angle lens was installed to handle the projection mapping. Additionally, it could also deliver audio output for users that were not wearing headsets. Figure 5 shows an operator in the described setting during the task. Additionally, a third-party optical reader [45] was added to the Device Control Layer to receive visual input. This reader detects the wires grabbed by the operator, so this information can be used for interaction contextualisation.

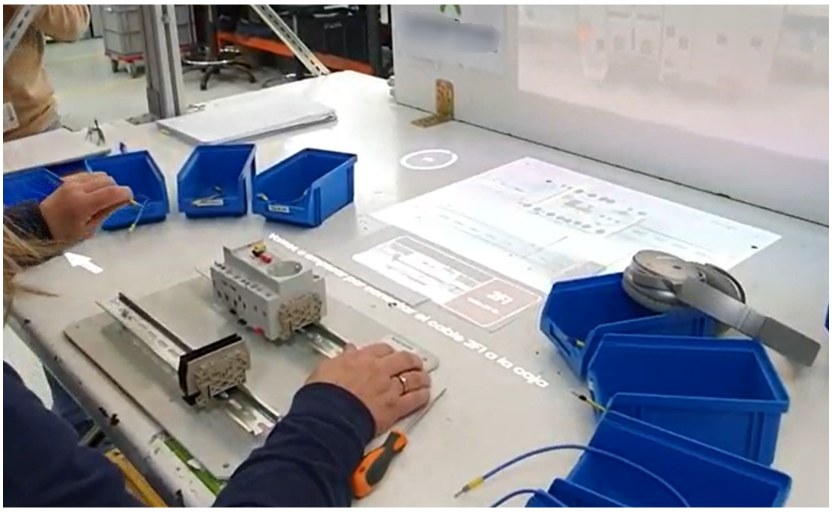

Figure 5. Operator performing a cabling task.

In this scenario, different uses of the projections and voice-based responses were combined to develop a user-centred communication, adjusted to their needs and impairments. For example, for participants with hearing impairments, aural responses were also projected in the form of captions. The completion of this task required the participants to work with both hands, one to manipulate wires and the other to screw the wire connection to the terminal. The best practices for the wiring sequence recommend operators to connect the longest wires first in order to reduce cluttering at early stages. Each wire is identified with a printed label and has its properties, attributes, and process logic stored in the DKL.

A set of ten participants from an external industrial factory was selectively assigned to the testing group of this use case. The inclusion criteria for this group was the level of worker disability. Only workers with a cognitive or physical disability were selected. Due to the sensitive nature of this particular group of testers, we are not providing further information on their age and gender.

5. System Implementation

In this section, the sequence of workflow steps that are involved in the systems implemented for our particular use cases is described in more detail:

1. The DCL is responsible for gathering the operators' input to communicate with the system. For the proposed use cases, the first one uses only voice input while the second one combines additional devices such as optical readers too. For the voice interaction, microphones with noise cancellation capture the audio signals. These signals are streamed using media streaming libraries such as FFmpeg into an energy-based automata that discards any silent or noisy audio segment. The remaining audio segment is sent to the Input Communication API.
2. The IL has to encode the raw and unstructured data into formats that can be interpreted by machines and ontologies. In the presented use cases, the audio segments are transcribed to text by the Speech-to-Text module. The Language Understanding module encodes transcriptions into a semantic structure that can be easily interpreted by the IML. As it is commonly done for Spoken Dialogue Systems, we use a scheme based on act-slot-value sets for representing

symbolic semantic meaning of the operator input. The semantic representation of the user action is dispatched to the IML for processing. For the optical reader input, the detected wires' label is transferred to the IML to contextualise the interaction.

3. The DKL works both as the persistence layer and as semantic ontology for each system. The persistence layer stores the interaction context of each operator to be kept turn-by-turn, as well as the specific domain knowledge encoded in an ontology. In both use cases, this ontology describes the know-how of the steps to perform in the process and their sequence, the physical devices that interact during the task and their properties (e.g., screwdriver size, or wire's location). In addition, it encodes the physical devices tied to each step of the interaction to handle ambiguous questions, for example, if the operator asks "Which size?" in the *Open fingers* step of the first use case, the ontology relations are used by the system's IML to understand that the operator refers to the size of the tool required to this specific step, the precision screwdriver, and not to the hex key used in the *Decouple gripper* step.
4. The IML plans and selects the next action to be performed by the system using both contextual information and the semantic concepts received from the IL. First, it retrieves the task information and the planning rules to complete the task from the Domain Knowledge Layer, and then it defines a strategy to reach the user's objectives. In other words, it consists of a set of expert rules $\mathcal{S} = \{(x,\ c_t) \longrightarrow (y, c_{t+1}) \mid x \in \mathcal{R}, c_t, c_{t+1} \in \mathcal{C}, y \in \mathcal{A}\}$ that evaluate events x and the interaction context c_t into system actions y and the also updates the context c_{t+1} based on the Attributed Probabilistic Finite State Bi-Automata schema as in [33,43,46]. This context, which the IML reads from and writes to the DKL, is maintained and updated through the interaction. In our particular tasks, the input x corresponds to the user semantic representation and the context c, which takes into account the interaction state (e.g., current step), the history of shown AR animations or the selected wire.
5. The RGL translates the IML's output actions into understandable interactions for the users, for example, answering the information in natural language or augmenting the user's surroundings with visuals. For our particular use case, the system uses the Text-to-Speech module to generate synthesised audio when speech modality is required, using the module described in AhoTTS [47]. Additionally, this layer computes suitable visual properties for the action, which include animation selection, as well as feedback duration.
6. The commands given by the RLG are dispatched to the DCL, which interfaces with the output devices. For the presented use cases, visual and aural interfaces are used to communicate with the operators.

Note that the presented XR interaction architecture is flexible enough to encompass different input and output devices, which is the case of our two use cases. This allows to select the most-suitable devices for each task, adapting to the heterogeneous needs of the different industrial processes.

6. Results

In this section, results are reported for each of the use cases where the proposed dialogue-supported XR architecture was applied. Such results are depicted in terms of usability, based on the participants' impressions gathered from post-study usability questionnaires. Note that the goal of these studies is not to measure times nor operational benefits of the systems but to measure the technological adoption of the proposed solutions by the operators. To perform the measurement of the adoption of the described systems, the System's Usability Scale (SUS) [48] was used as a reference. This scale intends to provide a global view of subjective usability assessments over multiple dimensions. User responses are presented as a 1 to 5 agreement scoring scale, where 1 means “strongly disagree“ and 5 means “strongly agree“.

The employed questionnaires contain several questions adapted to each particular task and users' characteristics, with the aim of capturing the operators' responses as detailed and unambiguously as possible. To make the results interpretable within a global scope, each question was clustered within one of the following usability dimensions:

- **Self-confidence:** this dimension involves those questions related to the ability to solve the problems that arise during the industrial processes without the help of any element other than the system itself.
- **Learning Curve:** this dimension measures how hard it was for the operators to adopt the proposed systems and to learn the required concepts and practices to use them.
- **Efficiency:** this dimension involves those questions related to the perceived efficiency by the users, i.e., if they find the system useful and helpful to improve their work processes.
- **Ease-of-use:** this dimension involves questions regarding the systems' difficulty. This is related to the naturalness of the system and whether it was perceived as intuitive by the operators.
- **Consistency:** this dimension measures how predictable the usage of the system is in terms of user experience.

6.1. Usability Results

Figures 6 and 7 depict usability agreement results by dimension for Use Case 1 and Use Case 2, respectively. It is interesting to note that, even if the systems are evaluated independently by different user groups, similar patterns arise in the obtained results.

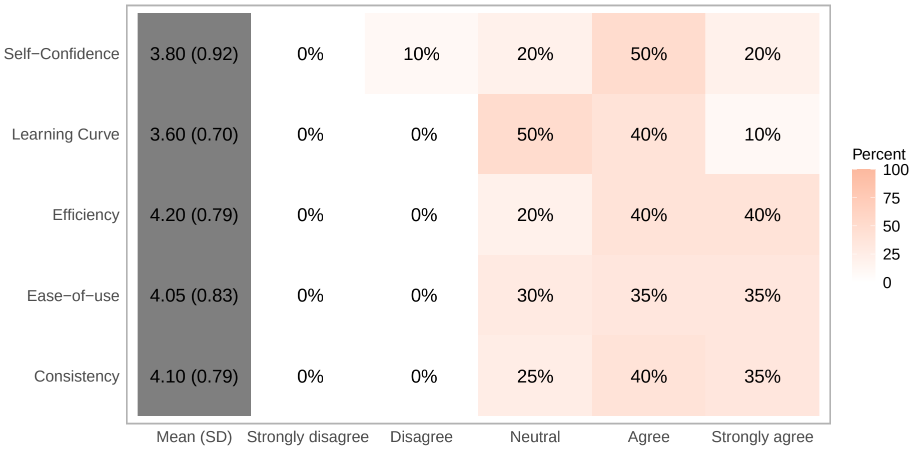

Figure 6. Results showing usability agreement by dimension for Use Case 1.

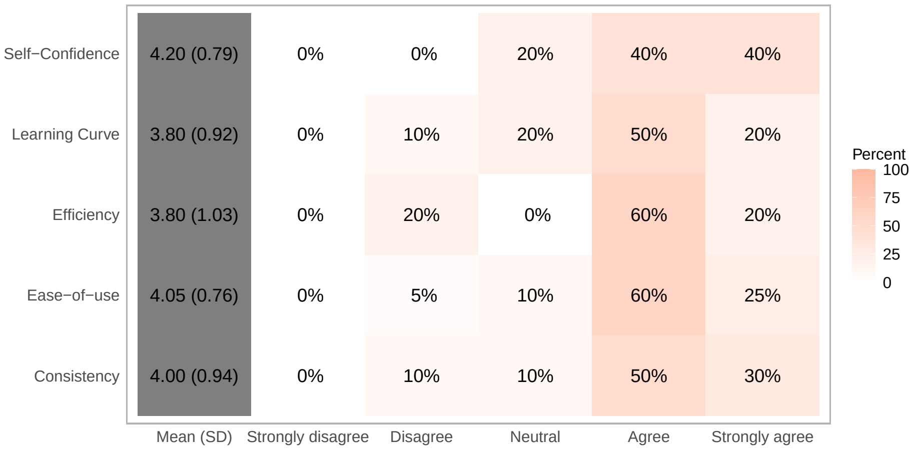

Figure 7. Results showing usability agreement by dimension for Use Case 2.

As can be seen, all the usability dimension scores are bounded between [3.6, 4.2] for both cases. In both situations, the confidence perceived by the operator is incremented due to the multi-modal guidance, and the system is perceived as consistent, even among different operator profiles. Most importantly, both systems are perceived as easy-to-use, one of the main objectives when implementing these systems to overcome cognitive barriers. This score is consistent between both use cases, where the technical profile of the operators differs: note that in the first scenario the operators were non-expert, untrained newcomers to the task, whereas in the second scenario operators were people who had some cognitive impairment. The perceived efficiency fluctuates between one use case and another, yet it still remains high. This fluctuation can be expected as the cognitive barriers that arise may differ between one use case and the other; thus, the perceived efficiency may be lower in tasks that are easier to solve. Finally, the learning curve dimension was the most poorly graded, yet achieving an average score of 3.7 for both use cases. This may be justified due to the experimental setup, where little-to-no instructions were given to the operators on how to use the system.

Participants' feedback corroborates the hypothesis that the combination of aural and visual technologies in an interactive dialogued system helps them to be hands free, which they agreed was an adequate feature given the type of tasks. Apart from the hands-free feature, participants found the system intuitive and natural. According to their observations and post-session informal feedback, the interactive nature of spoken dialogue with task-oriented contextualisation provided helpful real-time insights of the industrial processes to fulfil.

Hands-free systems are appropriate not only for cases like ours were both hands are needed to actually perform the task, but also for other industrial and non-industrial settings where people cannot use their hands for whatever reason (e.g., while driving, or because of having some physical impairment), and for cases where a physical transformation and/or control of the environment has to take place, often implying manual work.

7. Conclusions and Future Work

This paper describes an architecture to develop natural and hands-free human–machine interaction systems for industrial environments, which can combine more classical Human Augmentation technologies (such as Virtual, Augmented, and Mixed Reality) with dialogue based interaction for process solving tasks. The proposed solution facilitates the capture, distribution, and communication of domain specific knowledge to operators in training and production phases. Two systems constructed using the same architecture are presented and initially evaluated in terms of usability and acceptance, one for a maintenance task and the other for assembly scenarios. The implemented systems were hands-free, multi-modal (including visual and speech interaction) and supported multiple technologies. They were well accepted by a group of non-expert operators unfamiliar to the task and a group of operators with cognitive disabilities. The results obtained can serve as a reference starting point in further research for new advances in the industrial sector.

Given the number of participants in the groups selected for these initial validations, one of the main goals of future research is to further validate the proposed Interactive XR architecture with a wider sample of participants to ensure the significance of the evaluation. Testing additional systems in new industrial scenarios that require the combination of additional XR interfaces, like Kinesthetic-Bionic devices, Internet of Things, machine-sensors, and multi-operator scenario settings will be addressed to assert the flexibility of the proposed architecture. Future work also includes finding ways to easily adapt the language-specific modules (STT and NLU in the Interpretation Layer, and RTG and TTS in the Response Generation Layer) to multiple languages to overcome possible linguistic barriers for multinational companies.

Author Contributions: Conceptualisation, M.S., L.G.-S., and B.S.; data curation, L.G.-S.; funding acquisition, M.S., H.Á., and J.A.; methodology, M.S., L.G.-S., and B.S.; project administration, M.S. and J.A.; resources, B.S., H.Á., and J.A.; software, M.S., L.G.-S., B.S., and J.A.; writing—original draft, M.S., L.G.-S., B.S., and H.Á.;

writing—review and editing, B.S., L.G.-S., M.S., H.Á., and J.A. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement with ID 723711 (MANUWORK) and by the Department of Economic Development and Infrastructure of the Basque Government under grant KK-2017/00080 (INICIAR).

Acknowledgments: We would like to especially thank all the volunteers who kindly participated in the experimental sessions for their time and valuable feedback, and A. del Pozo for comments that improved the manuscript. Also, we would like to thank J. Posada for his advice on how to approach the manuscript and comments.

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

References

1. Barfield, W.; Williams, A. Cyborgs and enhancement technology. *Philosophies* **2017**, *2*, 4.
2. Aneesh, A. *Virtual Migration: The Programming of Globalization*; Duke University Press: Durham, NC, USA, 2006.
3. Simões, B.; De Amicis, R.; Barandiaran, I.; Posada, J. Cross reality to enhance worker cognition in industrial assembly operations. *Int. J. Adv. Manuf. Technol.* **2019**, *105*, 1–14. .
4. Posada, J.; Zorrilla, M.; Dominguez, A.; Simoes, B.; Eisert, P.; Stricker, D.; Rambach, J.; Döllner, J.; Guevara, M. Graphics and media technologies for operators in industry 4.0. *IEEE Comput. Graph. Appl.* **2018**, *38*, 119–132.
5. Segura, Á.; Diez, H.V.; Barandiaran, I.; Arbelaiz, A.; Álvarez, H.; Simões, B.; Posada, J.; García-Alonso, A.; Ugarte, R. Visual computing technologies to support the Operator 4.0. *Comput. Ind. Eng.* **2018**, *139*, 105550 .
6. Girard, J.; Girard, J. Defining knowledge management: Toward an applied compendium. *Online J. Appl. Knowl. Manag.* **2015**, *3*, 1–20.
7. Schmidt, B.; Borrison, R.; Cohen, A.; Dix, M.; Gärtler, M.; Hollender, M.; Klöpper, B.; Maczey, S.; Siddharthan, S. Industrial Virtual Assistants: Challenges and Opportunities. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, Singapore, 8–12 October 2018; ACM: New York, NY, USA, 2018; pp. 794–801.
8. Caudell, T.P.; Mizell, D.W. Augmented reality: An application of heads-up display technology to manual manufacturing processes. In Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences, Kauai, HI, USA, 7–10 January 1992; IEEE: Piscataway, NJ, USA, 1992; Volume 2, pp. 659–669.
9. Kollatsch, C.; Schumann, M.; Klimant, P.; Wittstock, V.; Putz, M. Mobile augmented reality based monitoring of assembly lines. *Procedia Cirp* **2014**, *23*, 246–251.
10. Sand, O.; Büttner, S.; Paelke, V.; Röcker, C. smart. assembly–projection-based augmented reality for supporting assembly workers. In *International Conference on Virtual, Augmented and Mixed Reality*; Springer: Berlin, Germany, 2016; pp. 643–652.
11. Rodriguez, L.; Quint, F.; Gorecky, D.; Romero, D.; Siller, H.R. Developing a mixed reality assistance system based on projection mapping technology for manual operations at assembly workstations. *Proc. Comput. Sci.* **2015**, *75*, 327–333.
12. Petersen, N.; Pagani, A.; Stricker, D. Real-time modeling and tracking manual workflows from first-person vision. In Proceedings of the 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Adelaide, Australia, 1–4 October 2013; pp. 117–124.
13. Álvarez, H.; Lajas, I.; Larrañaga, A.; Amozarrain, L.; Barandiaran, I. Augmented reality system to guide operators in the setup of die cutters. *Int. J. Adv. Manuf. Technol.* **2019**, *103*, 1543–1553.
14. Baumeister, J.; Ssin, S.Y.; ElSayed, N.A.; Dorrian, J.; Webb, D.P.; Walsh, J.A.; Simon, T.M.; Irlitti, A.; Smith, R.T.; Kohler, M.; others. Cognitive Cost of Using Augmented Reality Displays. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 2378–2388.
15. Industrial Augmented Reality. Industrial Augmented Reality—Wikipedia, The Free Encyclopedia, 2019. Available online: https://en.wikipedia.org/wiki/Industrial_augmented_reality (accessed on 6 June 2012). .
16. Malý, I.; Sedláček, D.; Leitão, P. Augmented reality experiments with industrial robot in industry 4.0 environment. In Proceedings of the 2016 IEEE 14th International Conference on Industrial Informatics (INDIN), Poitiers, France, 18–21 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 176–181.

17. Song, P.; Goh, W.B.; Hutama, W.; Fu, C.W.; Liu, X. A handle bar metaphor for virtual object manipulation with mid-air interaction. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Poitiers, France, 18–21 July 2012; ACM: New York, NY, USA, 2012; pp. 1297–1306.
18. Zander, T.O.; Gaertner, M.; Kothe, C.; Vilimek, R. Combining eye gaze input with a brain–computer interface for touchless human–computer interaction. *Int. J. Hum. Comput. Inter.* **2010**, *27*, 38–51.
19. Parker, C.L.; O'hanlon, M.L.W.; Lovitt, A.; Farmer, J.R. Interaction and Management of Devices Using Gaze Detection. US Patent 9,823,742, 21 November 2017 .
20. Stokic, D.; Kirchhoff, U.; Sundmaeker, H. Ambient intelligence in manufacturing industry: control system point of view. In Proceedings of the 8th IASTED International Conference on Control and Applications, Montreal, QC, Canada, 26 May 2006; pp. 24–26.
21. De Amicis, R.; Ceruti, A.; Francia, D.; Frizziero, L.; Simões, B. Augmented Reality for virtual user manual. *Int. J. Interact. Des. Manuf. IJIDeM* **2018**, *12*, 689–697.
22. Simões, B.; Álvarez, H.; Segura, A.; Barandiaran, I. Unlocking augmented interactions in short-lived assembly tasks. In Proceedings of the 13th International Conference on Soft Computing Models in Industrial and Environmental Applications, San Sebastian, Spain, 3–8 June 2018; Springer: Berlin, Germany, 2018; pp. 270–279.
23. Gupta, L.; Ma, S. Gesture-based interaction and communication: automated classification of hand gesture contours. *IEEE Trans. Syst. Man, Cybern. Part C Appl. Rev.* **2001**, *31*, 114–120.
24. Gorin, A.L.; Riccardi, G.; Wright, J.H. How may I help you? *Speech Commun.* **1997**, *23*, 113–127.
25. Abella, A.; Brown, M.; Buntschuh, B. Developing principles for dialog-based interfaces. In Proceedings of the ECAI Spoken Dialog Systems Workshop, Budapest, Hungary, 13 August 1996.
26. Lemon, O.; Gruenstein, A.; Battle, A.; Peters, S. Multi-tasking and collaborative activities in dialogue systems. In Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue, Philadelphia, PA, USA, 11–12 July 2002; pp. 113–124.
27. Serras, M.; Perez, N.; Torres, M.I.; Del Pozo, A. Entropy-driven dialog for topic classification: detecting and tackling uncertainty. In *Dialogues with Social Robots*; Springer: Berlin, Germany, 2017; pp. 171–182.
28. Bohus, D.; Rudnicky, A.I. LARRI: A language-based maintenance and repair assistant. In *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*; Springer: Berlin, Germany, 2005; pp. 203–218.
29. Bohus, D.; Rudnicky, A.I. Error handling in the RavenClaw dialog management framework. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2005; pp. 225–232.
30. Li, T.J.J.; Labutov, I.; Myers, B.A.; Azaria, A.; Rudnicky, A.I.; Mitchell, T.M. An End User Development Approach for Failure Handling in Goal-oriented Conversational Agents. In *Studies in Conversational UX Design*; Springer: Berlin, Germany, 2018.
31. Dowding, J.; Hockey, B.; Rayner, M.; Hieronymus, J.; Bohus, D.; Boven, B.; Blaylock, N.; Campana, E.; Early, S.; Gorrell, G.; et al. Talking through procedures: An intelligent Space Station procedure assistant. In *Demonstrations*; 2003. Available online: https://www.aclweb.org/anthology/E03-2001/ (accessed on 6 June 2020).
32. Raux, A.; Bohus, D.; Langner, B.; Black, A.W.; Eskenazi, M. Doing research on a deployed spoken dialogue system: One year of Let's Go! experience. In Proceedings of the 9th International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.
33. Serras, M.; Torres, M.I.; Del Pozo, A. Online learning of attributed bi-automata for dialogue management in spoken dialogue systems. In *Iberian Conference on Pattern Recognition and Image Analysis*; Springer: Berlin, Germany, 2017; pp. 22–31.
34. Crook, P.A.; Keizer, S.; Wang, Z.; Tang, W.; Lemon, O. Real user evaluation of a POMDP spoken dialogue system using automatic belief compression. *Comput. Speech Lang.* **2014**, *28*, 873–887.
35. Pineau, J.; Montemerlo, M.; Pollack, M.; Roy, N.; Thrun, S. Towards robotic assistants in nursing homes: Challenges and results. *Rob. Autonom. Syst.* **2003**, *42*, 271–281.
36. López Zorrilla, A.; Velasco Vázquez, M.d.; Irastorza, J.; Olaso Fernández, J.M.; Justo Blanco, R.; Torres Barañano, M.I. EMPATHIC: Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly. *Procesamiento del Lenguaje Natural* **2018**, *61*, 167–170. .

37. Lubold, N.; Walker, E.; Pon-Barry, H. Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand, 7–10 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 255–262.
38. Reidsma, D.; Charisi, V.; Davison, D.; Wijnen, F.; van der Meij, J.; Evers, V.; Cameron, D.; Fernando, S.; Moore, R.; Prescott, T.; others. The EASEL project: Towards educational human-robot symbiotic interaction. In *Conference on Biomimetic and Biohybrid Systems*; Springer: Berlin, Germany, 2016; pp. 297–306.
39. Graesser, A.C.; VanLehn, K.; Rosé, C.P.; Jordan, P.W.; Harter, D. Intelligent tutoring systems with conversational dialogue. *AI Mag.* **2001**, *22*, 39–39.
40. Agarwal, S.; Dusek, O.; Konstas, I.; Rieser, V. A Knowledge-Grounded Multimodal Search-Based Conversational Agent. *arXiv preprint* **2018**, arXiv:1810.11954 .
41. Young, S.J. Probabilistic methods in spoken–dialogue systems. *Philos. Tran. R. Soc. Lond. Ser. A Mathem. Phys. Engi. Sci.* **2000**, *358*, 1389–1402.
42. Chen, H.; Liu, X.; Yin, D.; Tang, J. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explor. Newslett.* **2017**, *19*, 25–35.
43. Serras, M.; Torres, M.I.; del Pozo, A. User-aware dialogue management policies over attributed bi-automata. *Patt. Anal. Appl.* **2018**.
44. Posada, J.; Wundrak, S.; Stork, A.; Toro, C. Semantically controlled LMV techniques for plant Design review. In Proceedings of the ASME 2004 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Salt Lake City, UT, USA, 28 September–2 October 2004; American Society of Mechanical Engineers Digital Collection: New York NY, USA, 2004; pp. 329–335.
45. Kildal, J.; Martín, M.; Ipiña, I.; Maurtua, I. Empowering assembly workers with cognitive disabilities by working with collaborative robots: A study to capture design requirements. *Procedia CIRP* **2019**, *81*, 797–802.
46. Serras., M.; Torres., M.I.; del Pozo., A. Goal-conditioned User Modeling for Dialogue Systems using Stochastic Bi-Automata. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods–Volume 1: ICPRAM, INSTICC*; SciTePress: Setúbal, Portugal, 2019; pp. 128–134.
47. Hernaez, I.; Navas, E.; Murugarren, J.L.; Etxebarria, B. Description of the AhoTTS system for the Basque language. In Proceedings of the 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, Perthshire, UK, 29 August–1 September 2001.
48. Brooke, J. *System Usability Scale (SUS): A Quick-and-Dirty Method of System Evaluation User Information*; Digital Equipment Co Ltd.: Reading, UK, 1986; pp. 1–7 .