A Novel Modular Architecture for Conversational Robotic Agents

Milan Gnjatović, Dragiša Mišković, Srđan Savić, Branislav Borovac,
Nemanja Maček, Branimir Trenkić

***Abstract*—This paper proposes a novel modal architecture of a conversational robotic agent. Two important characteristics of this architecture are that it is general (i.e., it does not depend on specification requirements for a conversational agent) and modular in a way that it allows for introduction of functionally and structurally different modules in a general manner. The architecture is conceptualized as a bus that serves as a backbone of the system and allows communication between an arbitrary number of modules. All modules use the same plug-in interface to the bus system, so the design of the architecture does not depend on the design of modules. The communication between modules is asynchronous and based on the principle of event bus. Finally, the architecture is illustrated for a particular robotic system (i.e., the assistive humanoid robot MARKO).**

***Index Terms*—Conversational agent, robot, architecture, event bus, asynchronous communication.**

I. Introduction and Motivation

THE design of appropriate architectures that support intelligent and socially believable technical systems still represents a challenging research question [7]. As specification requirements for such systems become more demanding, the underlying architectures become more elaborated and specific, and thus less applicable to different interaction scenarios [11]. This paper considers this research question in a more systematic manner. It introduces a novel modular architecture of conversational robotic agents that is not specific for a particular robotic system or a particular scenario of use, but rather general.

It is fair to say that the modularity of architectures for conversational agents is hardly a new thing. Even typical architectures contain separate modules that implement distinct functionalities (e.g., speech recognition, natural language understanding, image processing, etc.) [9,10]. However, at the heart of a conversational agent, there is usually a central module (i.e., a dialogue manager) that models the interaction context and controls the interaction flow. The specification requirements for such a module depend on the number and types of other modules in a system, and this in turn almost always determines the underlying architecture of a system as a whole. In contrast to such approaches, this paper proposes an architecture of a more general nature that can be applied in different conversational agents.

Milan Gnjatović is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (email: milangnjatovic@uns.ac.rs).

Dragiša Mišković is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (email: dragisa@uns.ac.rs).

Srđan Savić is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (email: savics@uns.ac.rs).

Branislav Borovac is with the Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia (email: borovac@uns.ac.rs).

Nemanja Maček is with the School of Electrical and Computer Engineering of Applied Studies, Vojvode Stepe 283, Belgrade, Serbia (email: nemanja.macek@viser.edu.rs).

Branimir Trenkić is with the School of Electrical and Computer Engineering of Applied Studies, Vojvode Stepe 283, Belgrade, Serbia (email: btrenkic@viser.edu.rs).

II. The Main Idea

The proposed architecture is depicted in Fig. 1. It is conceptualized as a bus that serves as a backbone of the system and allows asynchronous communication between an arbitrary number of modules, each of which implements a functionality relevant to the interaction between the user and the system. An important characteristic of this architecture is that it does not depend on specification requirements for the observed modules.

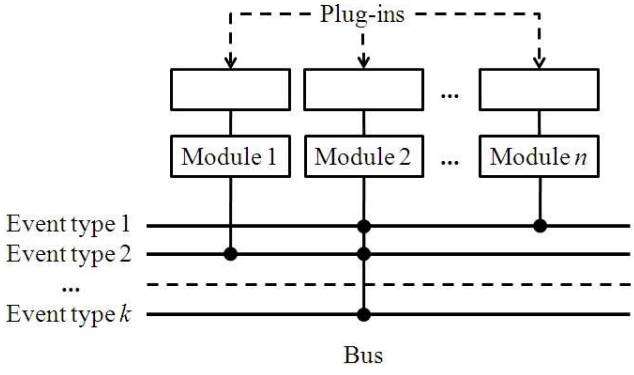

Fig. 1. The proposed architecture for conversational agent. Each module is assigned a set of corresponding interaction event types. The design of a module is separated from the design of the reported architecture.

At the conceptual level, each bus line corresponds to a certain type of interaction event. For example, a designer may decide that one line represents the interaction event of speech recognition (i.e., the line carries information about a speech signal that has been recognized by one of the assigned modules), while other bus line may be assigned to the

interaction event of speech synthesis (i.e., signaling that speech is to be synthesized by one of the assigned modules), etc.

The cardinality of relationships between bus lines and modules are N:M, i.e., a module can be assigned an arbitrary number of bus lines, while a bus line can be assigned an arbitrary number of modules. It is important to note that the determination of particular relations is a designer's decision, and not an inherent characteristic of the architecture. For example, a designer may decide that a system contains, for some reason, more than one module for automatic speech recognition. In such a case, all plug-ins for speech recognition will be assigned to the same bus line that corresponds to the interaction event of speech recognition. Similarly, a designer may decide that a module should be connected to all bus lines (e.g., a dialogue management module should respond to all interaction events).

Generally, the number and conceptualization of bus lines, the number of modules, and relationships between the bus lines and modules reflect particular specification requirements, and therefore are the subject of a designer's decision. The advantage of the proposed architecture is that communication on the bus is rather general, and not affected by these designer's decisions.

III. Implementation Details

Modules contained in conversational agents are usually very complex and architecture-agnostic. They are thus designed and implemented as separate entities, and their functionality depends on a number of configuration parameters.

The specified modularity of the proposed architecture is achieved by applying the class factory approach [2] that allows for registration of modules in a general manner, i.e., independent of the specification requirements. The applied approach differentiates between plug-ins and modules. The plug-ins encapsulate the implementation of specific functionality (e.g., speech recognition, dialogue management, etc.), while the modules provide a uniform interface between the plug-ins and the bus, i.e., all plug-ins use the same interface to the bus system, irrespective of their implementation details. This encapsulation enables the integration of an arbitrary module into a conversational agent.

The communication between modules is based on the principle of event bus. Modules are connected indirectly, via the bus, which is in line with the above mentioned fact that they are usually independent. During the integration of a module into the system, it is assigned a set of corresponding interaction event types. We recall that each bus line corresponds to a certain type of interaction event. This concept of a bus line is implemented as follows: each interaction event type is defined by a class within the observed module.

The task of a module is twofold:

- to map specific results of the assigned plug-in onto a class representing the given interaction event type, so that they can be detected by other modules that expect interaction events of the given type as their input,
- to detect and collect interaction events of specified types produced by other modules, and represent them in a data-structure that can serve as input to the assigned plug-in.

The plug-in approach allows for asynchronous communication between modules. E.g., one module may generate an interaction event that should be processed by other module. The latter module will process this interaction event at some moment, depending on the current traffic on the bus system, implementation of the module, and (possibly) many other factors. To support asynchronous communication, each interaction event accepted by a module is stored in a FIFO memory within the observed module.

Finally, during the life-time of the system, the registrations of modules can be modified or canceled.

IV. Prototype Conversational Agent

This section briefly reports on a prototypical conversational agent that manages speech-based dialogue between the user and the assistive humanoid robot MARKO (cf. Fig. 2).

Fig. 2. The robot MARKO.

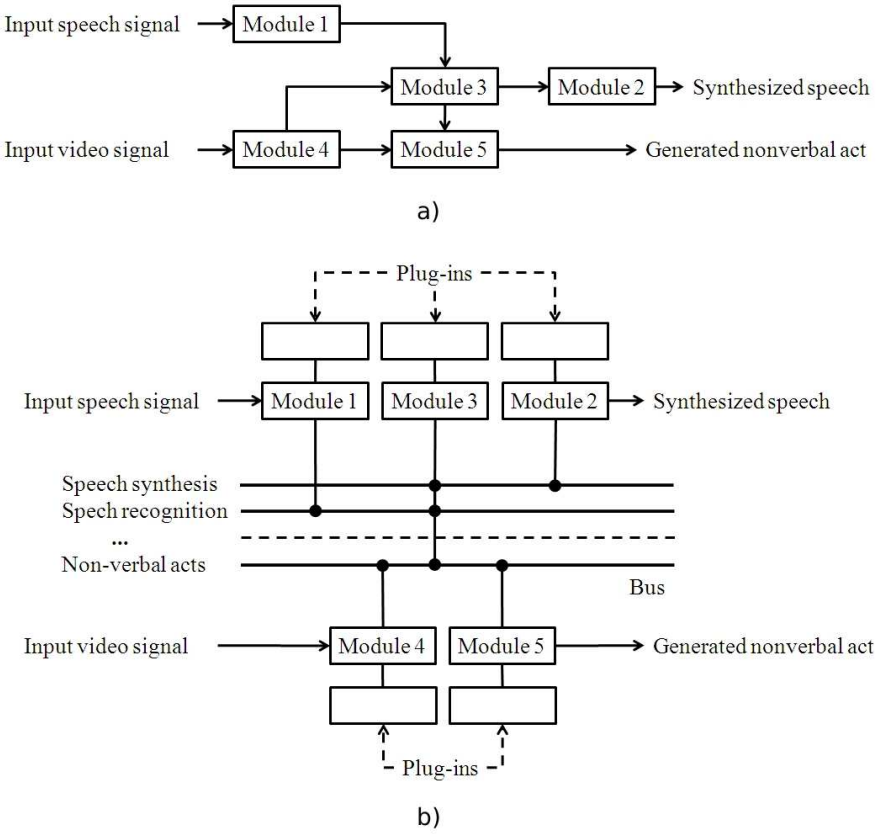

Fig. 3. (a) A task-specific architecture, and (b) the proposed general architecture for conversational agents. Both architectures are illustrated for the same set of modules: speech recognition (Module 1), speech synthesis (Module 2), natural language processing and dialogue management (Module 3), image processing (Module 4), and task management (Module 5).

The specification requirements for this conversational agent may be summarized as follows. The dedicated interaction scenario includes three-party speech-based interaction between the child, the therapist, and the robot MARKO with an integrated conversational agent. The robot takes part in the two-way interaction with the therapist. The therapist verbally instructs the robot to perform nonverbal actions related to facial expressions and gross motor exercises, and it may ask for clarifications and additional information, or comment. In addition, the robot takes part in the one-way interaction with the child, i.e., the robotic system takes initiative. The participants share a verbal context and a spatial context (for more details cf. [3]).

The five following modules are configured in the proposed architecture:

- Module (1) for context-dependent automatic speech recognition in the Serbian language. In contrast to most other automatic speech recognizers, this module integrates aspects of both statistical and symbolic approaches to speech recognition. This allows for accounting for contextual information and using it to improve post-processing of recognition hypotheses [11,8].
- Module (2) for text-to-speech synthesis in the Serbian language [12].
- Module (3) that integrates the functionalities of natural language understanding, dialogue management and natural language generation. The design of this module is based on a cognitively-inspired approach to meaning representation in human-machine interaction, which is intended to contribute to the habitability of a conversational agent across different interaction domains [3,4,5].
- Module (4) for image processing that detects the user's face and supports the robot to simulate eye contact with the user [1].
- Module (5) for task management that serves as an interface to the robotic system [1].

It is important to note that these modules were designed and developed prior to the introduction of the reported architecture. Moreover, they were initially integrated in other conversational agents intended to be used in a different interaction domain and built upon task-specific architectures. For example, the task-specific architecture of the conversational agent introduced in [6] is depicted in Fig. 3a. In contrast to this system, the same modules are integrated in the prototype conversational agent reported in this paper, built upon the proposed general architecture, as depicted in Fig. 3b. The appropriateness of the proposed architecture is reflected through the fact that the reported prototype conversational agent successfully integrates the same modules for a different domain of interaction.

V. DISCUSSION

In the previous section, we briefly noted that the appropriateness of the proposed architecture is reflected through the level of its generalizability. It should be noted that the adoption of the proposed architecture does not change the fact that modules of a conversational agent are more often than not developed outside an architectural context. For example, a more detailed discussion on architecture-agnosticity of the state-of-the-art speech recognition modules is provided in [11]. Thus, in this paper, we have intentionally avoided discussing in more detail the design of the modules of the prototype conversational agent (for these details, the reader may consult the given references). However, it is precisely the feature of architecture-agnosticity that allows us to incorporate these modules in an alternative, more general architecture.

The appropriateness of the proposed architecture can also be considered with respect to its scalability and the possibility to support multimodal human-machine interaction. Generally speaking, the scalability of the architecture and a potential multimodality of a conversational agent are simply based on the facts that the design of the architecture is independent of the design of assigned modules, and that the number of modules is arbitrary.

Finally, practical application of this architecture indicates that it should be optimized for conversational agents that incorporate modules requiring intense real-time communication. For example, a module for image processing or a module that controls movements of a robot may generate data on a millisecond-level, in contrast to a speech recognition module that generates data significantly less intensively, as the spoken natural language human-machine dialogue evolves. On the other hand, the tasks of a module for dialogue management include, inter alia, context-dependent interpretation of the data and adaptation of dialogue strategy. If this module tries to processes all data generated in the scope of intense real-time communication in an unselective manner, it may cause congestion.

This problem is not exclusively related to the proposed architecture, but is rather general. Still, in the context of this work, it can be addressed at two levels. At the level of the architecture, this problem can be addressed in such a way that each module can set an acceptance priority for each interaction event type. At the level of a module, these priorities can be dynamically adapted, i.e., according to the current load and interaction context. In this way, a module can dynamically reduce or completely stop the inflow of data. These considerations will be part of future work.

VI. CONCLUSION

This paper introduced a novel modal architecture of a conversational robotic agent and illustrated it for the assistive humanoid robot MARKO. Two important characteristics of this architecture are that (i) it does not depend on specification requirements for a conversational agent, and (ii) its design is separated from the design of modules contained in a conversational agent. Thus, the proposed architecture is general to the extent that it can be applied for a wide range of conversational agents, and modular to the extent that it allows for registration of modules in a general manner, irrespective of their specification requirements and implementation details.

ACKNOWLEDGMENT

The presented study was sponsored by the Ministry of Education, Science and Technological Development of the Republic of Serbia (research grants III44008 and TR32035), and by the intergovernmental network EUREKA (research grant E!9944). The responsibility for the content of this article lies with the authors.

REFERENCES

[1] B. Borovac, M. Gnjatović, S. Savić, M. Raković, M. Nikolić, "Human-like Robot MARKO in the Rehabilitation of Children with Cerebral Palsy", in: *New Trends in Medical and Service Robots. Assistive, Surgical and Educational Robotics*, Vol. 38 of the series Mechanisms and Machine Science, Springer International Publishing, pp. 191-203, 2016.

[2] E. Gamma, R. Helm, R. Johnson, J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley, 1995.

[3] M. Gnjatović, "Therapist-Centered Design of a Robot's Dialogue Behavior", *Cognitive Computation*, Vol. 6, No. 4, pp. 775-788, 2014.

[4] M. Gnjatović, B. Borovac, "Toward Conscious-Like Conversational Agents", In: *Toward Robotic Socially Believable Behaving Systems, Volume II - Modeling Social Signals*, A. Esposito, L.C. Jain (eds.), volume 106 of the series Intelligent Systems Reference Library, Springer, pp. 23-45, 2016.

[5] M. Gnjatović, V. Delić, "Cognitively-inspired representational approach to meaning in machine dialogue", *Knowledge-Based Systems*, Vol. 71, pp. 25-33, 2014.

[6] M. Gnjatović, J. Tasevski, M. Nikolić, D. Mišković, B. Borovac, V. Delić, „Adaptive Multimodal Interaction with Industrial Robot", in: Proceedings of *2012 IEEE 10th Jubilee International Symposium on Inteligent System and Informatics (SISY)*, Subotica, Serbia, pp. 2012.

[7] N. Hawes, J.L. Wyatt, M. Sridharan, H. Jacobsson, R. Dearden, A. Sloman, G.-J.Kruijff (2010) "Architecture and Representations", in H.I. Christensen, G.-J. Kruijff, J.L. Wyatt (eds), *Cognitive Systems*, Springer Berlin Heidelberg, pp 51-93, 2010.

[8] N. Jakovljević, D. Misković, M. Janev, D. Pekar, A Decoder for Large Vocabulary Speech Recognition. Proc. of Int. Conf. on Systems, Signals and Image Processing (IWSSIP), Sarajevo; 2011. p. 1--4.

[9] K. Jokinen, M. McTear, *Spoken Dialogue Systems*, Synthesis Lectures on Human Language Technologies. Morgan and Claypool. 2009.

[10] D. Jurafsky, J.H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, second edition, Prentice-Hall. 2009.

[11] D. Mišković, M. Gnjatović, P. Štrbac, B. Trenkić, N. Jakovljević, V. Delić, "Hybrid Methodological Approach to Context-Dependent Speech Recognition", in *International Journal of Advanced Robotic Systems*, Vol. 14, No. 1, 12 pages, no pagination, 2017.

[12] M. Sečujski, V. Delić, D. Pekar, R. Obradović, D. Knežević, "An Overview of the AlfaNum Text-to-Speech Synthesis System", in Proceeding of the *XII International Conference "Speech and Computer" (SPECOM'2007)*, 5 pages, no pagination, Moscow State Linguistic University, Moscow, Russia, 2007.