Multimodal Dialogue System with NAO and VoiceXML Dialogue Manager

Stanislav Ondáš, Jozef Juhár, Matúš Pleva, Peter Ferčák, Rastislav Husovský
Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics,
Technical University of Košice,
Kosice, Slovakia
stanislav.ondas@tuke.sk, jozef.juhar@tuke.sk, matus.pleva@tuke.sk

***Abstract*— The proposed paper describes a multimodal interactive system based on NAO humanoid robot with an external dialogue management module VoiceON. System can be controlled by voice. NAO produces speech and gestures as a response to the user inputs. Presented multimodal system uses built-in speech recognition and speech synthesis modules adapted to Slovak language, which was originally not supported. Moreover, system accepts VoiceXML dialogue applications for all supported languages. To manage dialogue interaction, a previously developed VoiceXML dialogue manager was adopted. A new module for generation of multimodal output (speech and gestures) was designed, which enables NAO to produce gestures together with speech in several modes.**

Keywords— Multimodal interaction; dialogue management; speech recognition; speech synthesis

I. INTRODUCTION

Humanoid robotics is a very attractive research area. The result of undeniable effort to develop artificial human is several humanoid robots (e.g. Honda Asimo or Aldebaran Pepper) with different capabilities and with a different degree of human-like appearance [1]. With an improved appearance of humanoid robots, people tend to expect a human-like behavior, including speech production [2] and hearing capabilities.

To improve the robot capability of being a dialogue participant requires the involvement of several technologies as: speech recognition and text-to-speech synthesis, spoken language processing, interpretation of multimodal inputs and multimodal fusion, dialogue management, vision, gesture processing and production etc.

In proposed work, we extended a humanoid robot NAO (Fig.1) with our develpoed external dialogue management module VoiceON, which enables to write dialogue interactions for NAO robot in standard VoiceXML language [3]. The novelty of the paper lies in the fact that, according to our knowledge, there is no other VoiceXML-based dialogue manager implementation evaluation for NAO robot.

Fig. 1. Humanoid robot NAO (Aldebaran Robotics)

VoiceXML language is defined in W3C recommendation [4] to enable design spoken dialogue interactions for human-machine interfaces. It offers a simple, XML-based, language that enables to define dialogue content and flow more effective. VoiceXML language can be considered as an industry standard for voice platforms, which serve for providing automatic voice services through the telecommunication network. VoiceXML enables to write system-directed dialogues, which imagine a comfortable and reliable form for many voice applications [5]. Though system-directed dialogues are often seen as less flexible and too restrictive, they ensure a right direction of the interaction and achievement of user's goal.

NAO robot is a great tool to prepare a multimodal interactive system due to its support of vision, hearing, gesture production and body language. Moreover, an autonomous mode of the robot enhances a human-like behaviour.

Designed multimodal dialogue system for NAO robot is a multimodal human-robot interface (HRI) as a special kind of human-machine interfaces (HMI). The areas of human-machine interfaces as well as area of multimodal interaction have common interests with Cognitive info-communications (CogInfoCom), as is noted by authors in [15] and [16]. From CogInfoCom point of view, the system serves for inter-cognitive communication.

Proposed work can be also located in the area of *affective computing*, because designed system is able to exhibit human-like qualities, namely body gestures and vocal qualities [15].

Designed multimodal dialogue system can be seen as *artificially cognitive system*, which can communicate information to the user through synthesized speech. Such speech-based communication is grounded in *cognitive linguistics* research and thus, according [15], is clearly related to both CogInfoCom and cognitive linguistics.

The paper is organized as follows: The second section describes designed multimodal dialogue system based on NAO robot and VoiceON dialogue manager. This section also describes a newly developed wrapper between robot and VoiceON dialogue manager and a newly developed module for generation of multimodal gesture output. Section 3 describes a pilot voice application and results of evaluation experiment.

II. Multimodal Dialogue System

The designed interactive system based on NAO robot (Fig.2) is an asymmetric multimodal dialogue system with unimodal input (speech) and multimodal output (speech and gestures). It can be easily extended to fully symmetric multimodal system. The interaction in the system is managed by the external dialogue manager, which interprets VoiceXML language.

Extension of NAO robot with VoiceXML-based dialogue manager imagines a very simple and comfortable way how to write dialogue interactions for NAO robot in comparison with NAO built-in dialogue module. Such solution enables system-directed dialogues, which offers a reliable way to achieve user's goals. The main drawback of using VoiceXML is lower flexibility and naturalness.

There exist also several other conversational systems for NAO robot, e.g. NAO WikiTalk (see [14]), which offers the complex conversational system to access information from Wikipedia through human-robot interface. While NAO WikiTalk system focuses mainly on multimodal feedback, our system only provides a framework, which can be extended with advanced feedback generation and other functionalities. In the work presented in the paper, the main attention was paid on the integration of VoiceXML dialogue manager into NAO robot. However, VoiceXML was originally developed for unimodal dialogue interaction, We examine the VoiceXML adoption for multimodal interaction, to show its potential.

A. Architecture

The multimodal system consists of following components: external VoiceXML dialogue manager, wrapper between dialogue manager and NAO, block for generation of multimodal output and built-in automatic speech recognition (ASR) module and text-to-speech (TTS) module.

Expect of dialogue manager all modules are integrated directly inside the robot. Wrapper and output generation modules are written in Python and will be described separately in next sections.

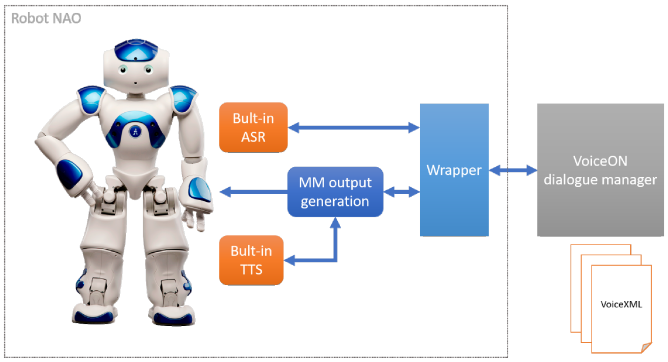

Fig. 2. Multimodal dialogue system based on NAO

Designed multimodal system uses built-in NAO modules for speech recognition and speech synthesis, which enables to use the system with all supported languages. Designed system was tested with English, Czech and Slovak language. To support Slovak language, we decided to use built-in Czech ASR and TTS systems, what is possible thanks to their similarity. Similar experiments with such crosslingual solution were described in [6], where Slovak phonemes were mapped on Czech phonetical set.

B. Dialogue manager VoiceON

VoiceON dialogue manager controls dialogue by interpretation of VoiceXML 2.0 scripts. Its development started during the national project IRKR, which focused on design and development of the Slovak spoken dialogue system (see [7] or [8]). The reason to develop our own VoiceXML-based dialogue manager was a lack of reliable free VoiceXML managers. This situation did not change up to now. There are only few open-source free VoiceXML interpreters, often based on OpenVXI project [9], as are e.g. commercial BladeWareVXi VXML interpreter. In case of Voxy interpreter [10], it is fixed to Asterix solution. JVoiceXML interpreter [11] run on Java, which can bring some limitations. When we start to develop VoiceON in 2003, there was practically only one freely available interpreter – OpenVXI. We did not select it due to two main reasons. The first one was that we wanted to connect the interpreter into Galaxy HUB architecture, where a very detailed view of source code is required. The second reason was that the situation around VoiceXML was changing everyday up to March 2004, when VoiceXML 2.0 became the W3C Recommendation. Building of own VoiceXML interpreter gives us a great opportunity to update the interpreter according a new recommendation.

VoiceON has been split into a VoiceON server and a Galaxy wrapper, which communicates together through TCP/IP socket. This solution has proved to be very beneficial considering further use, especially for usage in server-client or cloud-based solutions. Such arrangement has shown as very suitable also for NAO robot, where we decided to use VoiceON server as an external module, without need to perform any changes inside its code.

The architecture of the VoiceON dialogue manager is shown in Fig. 3.

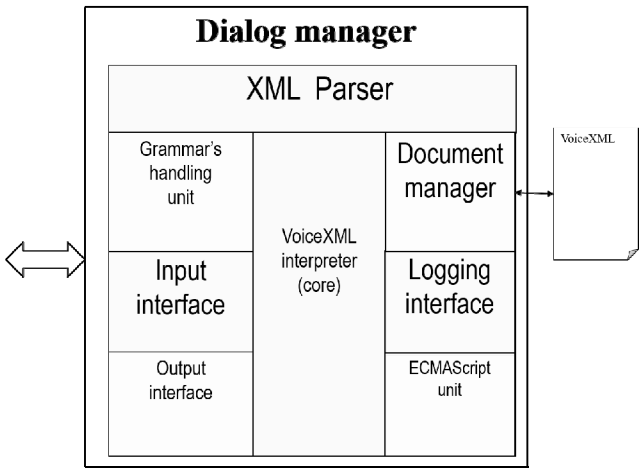

Fig. 3. Architecture of VoiceON dialogue manager

VoiceXML interpreter is a core of the dialogue manager. It performs the Form Interpretation Algorithm, which is a basic algorithm of VoiceXML documents interpretation. It also contains handling functions for each VoiceXML element.

At the beginning of the interpretation process XML Parser parses the interpreted document and creates a hierarchical image of VoiceXML application. This image represents the structure of the dialog. XML Parser is used also for parsing of XML grammars and other XML content.

All variables and expressions as well as logic inside the script elements in the VoiceXML language are represented and performed in ECMAScript language. ECMAScript unit in VoiceON Dialog Manager (DM) process all ECMAScript content.

The next components of the dialog manager are:

- **Input interface block** control processes of taking and processing of user inputs. It manages time intervals and catches input events (Noinput, Nomatch).
- **Grammar's handling unit** implements grammar activation algorithm. It manages scope of grammars, creates support for several grammar formats (for example XML or BNF form of W3C SRGS) and generates grammars when <option> or <menu> elements were used. It can convert grammar formats. Moreover, together with the input interface block enables to use semantic tags according W3C SISR recommendation.
- **Output interface** generates prompts to the user. It performs adding of variable values into the prompts and it makes decision what prompt will be played. Implements prompt selection algorithm.
- **Document manager** is responsible for fetching, storing and downloading files, that are necessary (VoiceXML, audio, grammars and others) for dialog management.
- **Logging interface** serves as a monitoring tool. It logs all events that occur during a dialog into log files. There are also all errors, warnings, diagnostic messages and a textual representation of user-system interaction. This information serves for solving of unexpected states of system.

C. VoiceON Wrapper

The Galaxy wrapper was replaced by a newly designed NAO wrapper, written in Python. The designed NAO wrapper serves as an interface between NAO and VoiceON dialogue manager. Dialogue manager interprets VoiceXML scripts, which results into requirements on speech recognition and speech synthesis services. These requirements are transformed into function calls of built-in ASR and TTS systems in this wrapper. To enable the multimodal interaction in a form of speech + gestures output, the block for multimodal output generation can be inserted between this wrapper and built-in TTS system. It processes a textual input and transforms it into gestures and synthesized speech.

In case of NAO robot, there is a special mechanism for speech grammars, because NAO does not support W3C SRGS recommendation. The VoiceON Wrapper obtains the name of the grammar, which should be active in the actual state of the dialogue, from a VoiceXML document. Then, it looks into a NAO local directory with grammars and search for a grammar with the same name (without extension). This grammar file needs to be in one of the forms supported by NAO ASR system. Content of both grammars (inside the robot and inside VoiceXML) must be the same, but in forms supported by NAO and VoiceON manager.

D. Multimodal output generation

The block of multimodal output generation performs analysis of incoming text, which has to be synthesized and generates the response of the robot as a combination of speech and gestures. The structure of the module is shown in Fig. 4.

Text analysis is performed in two main steps. Firstly, the input text is split into sentences. Next analysis depends on a selected mode, which defines a way of sentence processing. There are three different modes:

- **SpeechOnly** mode. In this mode, system generates only speech without gestures.
- **RandOff** mode. In this mode, input sentences are *analyzed* and searched for *keywords* and *markups*, which are mapped on gestures according to expert manually defined rules in the configuration files. RandOff means that robot random behavior is switched off and robot generates only gestures, which are triggered by performed text analysis.
- **RandOn** mode. The difference between RandOff and RandOn mode is that random robot behavior is here switched on and robot generates also random movements and gestures.

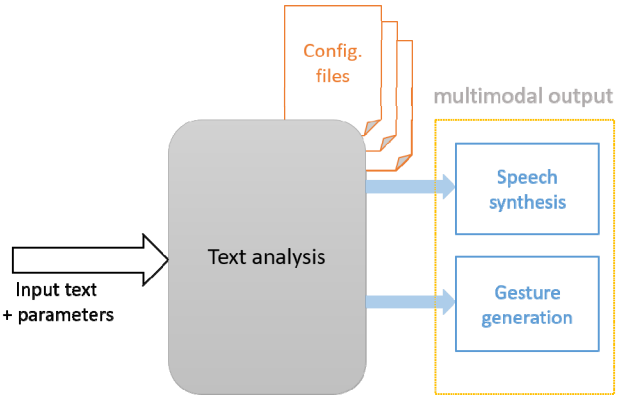

Fig. 4. Multimodal output generation module

Rules for generation of gestures are stored in configuration files. Each rule is associated with a gesture and a group of words and phrases. Then, if a keyword of phrase is found in the input text, then a rule is activated and a gesture marks are included into the output stream.

The block of multimodal output generation enables also to add rule marks directly into the input text. In designed multimodal system with VoiceON dialogue manager, this function enables to control gestures directly from VoiceXML document. Rule marks can be placed into the content of <prompt> element as it is shown in following example:

```
<prompt>
    #hello Hello. How are #you you?
</prompt>
```

As we can see, VoiceXML tag <prompt> contains text and two gesture tags for NAO robot – #hello and #you, which trigger generation of the requested gestures.

III. Pilot Application and Evaluation

Several test dialogues were prepared for the system evaluation. The advantage of designed system lies especially in simplicity of preparing new voice interactions, which consists only from preparing of VoiceXML scripts and modification of rules for gestures generation.

User experiences were evaluated after dialogue interaction with NAO robot. Thirty test subjects interact with NAO with different settings related to gesture generation:a) without gestures b) contextually generated gestures without random behaviour and c) with random behaviour.

After interaction, test subjects filled a questionnaire with eight questions. Due to several conditions, results of questionnaires were very similar in each group. All test subjects assessed performed interaction as enough smooth, natural and interesting. However, slight differences were observed that show importance of gesture generation for perceived naturalness, smoothness and a certain level of user acceptance.

IV. Conclusion

The newly designed multimodal dialogue system with NAO robot has been introduced and described. Dialogue interaction is controlled with the VoiceXML-based dialogue manager (VoiceON) using Python NAO wrapper, which is a result of previous projects. According our knowledge, this is a first work, where VoiceXML-based dialogue manager is integrated with NAO robot. Moreover, the new simple module for generation of multimodal output was designed, which transform incoming output text into a mixture of synthesised speech and robot gestures according synthesized text processing, which fully cooperates with VoiceON dialogue manager and built-in TTS system in NAO.

Designed multimodal system is language-independent; it means that it supports all languages supported by built-in ASR and TTS systems in NAO robot.

The designed multimodal system brings a unique solution for generation of multimodal output in combination with VoiceXML language, where gesture tags can be directly included into VoiceXML prompts or gestures can be automatically generated according keyword spotting analysis of prompts content. Moreover, gesture generation can be combined with random robot behaviour, which results in to natural multimodal output.

Designed solution, including VoiceON dialogue manager, Wrapper in Python and multimodal output generation unit, are available on demand.

In the future, we will focus on the next development of VoiceON wrapper, to enable control of other functions of NAO robot directly from VoiceXML e.g. walking, emotion detection and others. In the frame of project called “Cloud Based Human Robot Interaction” (APPV-15-0731) we plan to work on cloud implementation [1] of the dialogue, integrate emotion recognition [12] and prepare the module for interaction with different platforms [13].

Acknowledgment

This publication was supported partially by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the projects KEGA 055TUKE-4/2016 & partially by the Slovak Research and Development Agency under the contracts No. APVV-15-0517 & APPV-15-0731.

References

[1] J. Collinászy, M. Bundzel and I. Zolotová, “Implementation of Intelligent Software using IBM Watson And Bluemix,” *Acta Electrotechnica et Informatica*, 2017, 17(1), pp. 58–63. ISSN 1335-8243. DOI:10.15546/aeei-2017-0008

[2] M. Sulír, and J. Juhár, “Hidden Markov Model Based Speech Synthesis System in Slovak Language with Speaker Interpolation,” *Acta Electrotechnica et Informatica*. 2015, 15(4), pp. 8–12. ISSN 1335-8243. DOI: 10.15546/aeei-2015-0029

[3] S. Ondáš, and J. Juhár, “Dialog manager based on the VoiceXML interpreter,” In: *Proc. 6th Intern. Conference DSP-MCOM*, Košice, Elfa, 2005. pp. 80-83.

[4] J. A. Larson, "W3C speech interface languages: VoiceXML [standards in a nutshell]," *IEEE Signal Processing Magazine*, 2007, 24(3), pp.126-131. DOI: 10.1109/MSP.2007.361612

[5] J. Juhár, S. Ondas, A. Cizmár, M. Rusko, G. Rozinaj and R. Jarina, "Development of Slovak GALAXY/VoiceXML based Spoken Language Dialogue System to Retrieve Information from the Internet," In: *Ninth International Conference on Spoken Language Processing (ICSLP 2006)* Pittsburgh, ISCA, 2006, pp. 485-488.

[6] S. Lihan, J. Juhár and A. Cizmár, "Comparison of Slovak and Czech speech recognition based on grapheme and phoneme acoustic models," In: *Ninth International Conference on Spoken Language Processing (ICSLP 2006)*, Pittsburgh, ISCA, 2006, pp. 149-152.

[7] J. Juhár, A. Čižmár, M. Rusko, M. Trnka, G. Rozinaj, and J. Jarina, "Voice operated information system in Slovak," *Computing and Informatics*, 2012, 26(6), pp. 577-603. ISSN: 1335-9150.

[8] S. Ondáš, J. Juhár, M. Papco, M. Trnka and V. Király, "The integration of the Hungarian language in to the Slovak Spoken dialogue system," In *Proceedings of the 9th WSEAS international conference on signal, speech and image processing, and 9th WSEAS international conference on Multimedia, internet & video technologies,* 2009, Budapest, WSEAS, pp. 102-105.

[9] B. Eberman, J. Carter, D. Meyer and D. Goddeau, "Building VoiceXML Browsers with OpenVXI," In *Proceedings of the 11th international conference on World Wide Web*. 2002, Honolulu, ACM, pp. 713-717. DOI: 10.1145/511446.511538.

[10] L. Lerato, M. Molapo and L. Khoase, "Open Source VoiceXML Interpreter over Asterisk for Use in IVR Applications," In: *Proceedings of the SATNAC,* Southern Africa, SATNAC, 2009, p. 6.

[11] D. Schnelle-Walka, S. Radomski and M. Mühlhäuser, "JVoiceXML as a modality component in the W3C multimodal architecture," *Journal on Multimodal User Interfaces*, 2013, 7(3), pp.183-194. ISSN: 1783-7677. DOI:10.1007/s12193-013-0119-y

[12] P. Takáč, M. Mach and P. Sinčák, "Cloud-based facial emotion recognition for real-time emotional atmosphere assessment during a lecture," In *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*. 2016, Budapest, IEEE, pp. 1696-1701. DOI: 10.1109/SMC.2016.7844481

[13] R. Üveges, F. Ďurovský and D. Lindr, "Open Robotic Controllers," *Acta Electrotechnica et Informatica*. 2016, 16(3), pp. 8–13. ISSN 1335-8243. DOI: 10.15546/aeei-2016-0017

[14] A. Csapo, E. Gilmartin, J. Grizou, J. Han, R. Meena, D. Anastasiou, K. Jokinen, and G. Wilcock, "Multimodal conversational interaction with a humanoid robot," . In: Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on, Kosice, IEEE, pp. 667 - 672.

[15] P. Baranyi, A. Csapo, and G. Sallai, "Cognitive Infocommunications (CogInfoCom)," Springer International, 2015.

[16] P. Baranyi, A. Csapo, "Definition and Synergies of Cognitive Infocommunications", in *Acta Polytechnica Hungarica*, 2012, 9(1), pp. 67 – 83.