TRANSFORMING IBNSINA INTO AN ADVANCED MULTILINGUAL INTERACTIVE ANDROID ROBOT

Nikolaos Mavridis, Alia AlDhaheri, Latifa AlDhaheri, Maitha Khanji, Noura AlDarmaki

Interactive Robots and Media Lab, United Arab Emirates University

ABSTRACT

IbnSina is the world's first Arabic-language conversational android robot, and is also part of an interactive theatre with multiple possibilities for human teleparticipation. In this paper, we describe extensions carried out to IbnSina's software architecture in order to enrich its capabilities in multiple ways, so that it can become an exciting educational / persuasive robot in the future. The main axis for extension were: access to online (Wikipedia) and stored (Koran database) content for dialogue generation, basic multilingual capability exploration (English and Arabic, also utilizing Google Translate), basic read-aloud-text capability (through OCR), and systematization of motor control (with higher-level API for real-time lip syncing, eye blinking, natural looking random face movements, and interpolation between facial expressions including an affective state subsystem). With such capabilities, IbnSina becomes closer to an attractive robot that can find real-world application in malls, schools, as a receptionist etc.

***Index Terms*—** Intelligent Robots, Android, Multilingual, Interactive, Natural Language Interfaces

1. INTRODUCTION

IbnSina [1] is the world's first Arabic-language conversational android robot, which already has capabilities for generating pre-scripted facial expressions as well as head movements, and for utilizing Arabic text-to-speech and speech recognition for the creation of simple dialogues, as described in [2][3]. Furthermore, IbnSina is part of an interactive theatre with multiple possibilities for human teleparticipation, through motion capture, virtual worlds, etc. [4] [5]. However interesting the current results might be, there are still very many open possibilities for extension.

The purpose of this project is to extend IbnSina's capabilities in multiple ways, so that it can become an exciting educational / persuasive robot in the future. The main axis for extension were: access to *online content* (Wikipedia) and *stored content* (Koran database) for dialogue generation, basic *multilingual* capability exploration (English and Arabic, also utilizing Google Translate), basic *read-aloud-written-text* capability (through OCR), and systematization of *motor control* (with higher-level API for real-time lip syncing, eye blinking, natural-looking random face movements, and interpolation between facial expressions). With such capabilities, IbnSina becomes closer to an attractive robot that can find real-world round-the-clock application in shopping malls, schools, as a receptionist etc.

Many technologies and areas are implicated in our endeavor. The main fields are: humanoid robotics, speech and language technologies, computer vision, pattern recognition, and human-robot interaction. Also, in terms of software engineering, this was a very interesting experience because we had to deal with a complex heterogeneous system, and we had to hack around with existing software as well as build our extensions, and to co-ordinate among the different parts. In short, it was a very important educational experience, giving us a firm grounding in real-world advanced systems development as well as cutting-edge research.

2. PHASE I: REQUIREMENT AND ANALYSIS

The following Requirements were set for our system:

R1) Access to online and stored content for dialogues (Wikipedia, Quran)

The conversational system is one of the main subsystems of Ibn Sina's cognitive architecture, and the following requirements were set for it: facilitating simple conversations in Arabic and English, detecting the language that the human is talking in, and responding in the same or the other language. Responses to various forms of queries should be supported: first, replying to *simple questions* like: "hello, how are you?", second, *translating* words; third, answering questions by giving *online info* like: "what is Artificial Intelligence?", fourth relating the asked questions with *stored sourcebooks* (such as the Holy Quran) by mentioning the verse and its translation in English and in which verse and chapter it is mentioned. Ideally, the system should classify the topic that the user is talking about and bring other terms of the same corpora, for example: if the user talks about neural networks and evolutionary computing, it should understand that the topic is about Artificial Intelligence and then respond in sentences of the same topic. In terms of other requirements, the robot should respond to the user quickly and react sequentially so that the user does not feel interruption while the computer is executing the various algorithms. The robot also should give feedback

978-1-61284-119-9/11/$26.00 ©2011 IEEE

to the user in all cases like missing information or incorrect spelling, so that the user be aware of what is happening. In addition, the robot should be able to speaks and also display appropriate images for illustration during the conversation, creating a richer presentation.

R2) Basic Multilinguality (English and Arabic)

The subsystem should be able to deal with both languages—Arabic and English, that is; automatically detecting the language of the person communicating with Ibn Sina either by text or speech, and then translating information he is looking for in the other language. The freedom of choice of the language the user chooses, detecting it and responding in the other language facilitates richer human-machine interaction.

R3) Basic read-aloud-text capability (through OCR)

OCR (Optical Character Recognition) is a program that's converts scanned images of the text back to text files. The functional requirement is that we want IbnSina to read what is written in a piece of paper (black ink pixels surrounded by white pixels). Moreover, IbnSina should speak up this text with a loud and clear voice. The technical requireme is that the info on the piece of paper should be sent from a video camera which is located in IbnSina's eyes to the computer that connected with the robot. The audio as well as video hardware should facilitate the resulting paper-camera-speech functionality.

R4) Systematization of motor control (API for real-time lip sync, eye blink, random head movements, and interpolation between facial expressions)

First, in order to achieve higher human-likeness, IbnSina has to *blink his eyes* naturally while talking or listing. The eye blinking should not be purely periodic but should contain stochasticity facilitating apparent naturalness. The Second function required is *lip syncing* with speech. IbnSina has to modulate the opening of his mouth according to what is being said. This step is worked by taking the sound coming from the speech that IbnSina produce, and then analyze it. The outcome of the analyzing step is to be sent to the mouth Servos after converting it to the servo command language in order to make the appropriate movement. The third function is generating *background head movements* while speaking or listing to make IbnSina look more humanly by nodding to the people who speak with. This step involves sending appropriate instruction packets to dynamixel motors. The fourth functional requirement is that IbnSina has to make various facial expressions corresponding to emotional states while he speaking or listing, to look more human.

In terms of technical requirements, the data that is sent to the Hitec and Dynamixels has to *be available fast enough* in order not to lose the required apparent synchronization between the moves, speech, and dialogue state. Another important technical requirement, is to be sure not to push the Servos with values that *exceed the allowable min or max values*; otherwise we might destroy the servos, tear-down mechanical wiring, or even parts of the robot face.

3. PHASE II: DESIGN AND ARCHITECTURE

On the basis of the four requirements R1-R4, three subsystems were designed and implemented, namely S1-S3. Subsystem S1 accounts for requirements R1 and R2, while S2 for R3, and S3 accounts for R4:

S1) Conversational System with access to Online Content and multilinguality

The multi-module conversational system is designed based on object-oriented classes. The chat module consists of two classes: the *client class* and the *server class*, that communicate for sending messages forth and back. The *ChatterBot class* enables making simple conversation and replying the user. The LSA module analyzes the user's speech using Latent Semantic Analysis Algorithms [6] and then finds related topics of the same corpora. The *Wikipedia class* is used for extracting information from Wikipedia about a particular word. The *Translate class* gets a text translation from English to Arabic and reverse. The *Quran class* extracts verses from online dictionary of Holy Quran that is related to a particular word. The *Filter class* is used for filtering and parsing sentences like: client's message, information extracted from Wikipedia and the Quran data. The *Audio class* which is used for running Text To Speech for pronouncing responses, and the Image class that extracts images of a particular topic. In general, we followed the *iterative design approach* [7], that is; each time we got results and feedback from people we enhanced the design in order to get an improved one. In practice the conversational system was tested in two modes: Mode A: in which the input is text, the program automatically detects the input language (English or Arabic text) and the output is text of the same language entered, in addition to speech output while Mode B the input is Arabic speech.

S2) Basic Reading Subsystem

For video-driven read-aloud OCR, we designed an application that is written in Java. First, we take a cropped screen snapshot from the computer, containing the camera video. Then, we take this screen shot as an input image with jpg extension and converts it to pnm extension; for the purpose of feeding the actual OCR code. Finally, the text output is fed to the Acapela text-to-speech for reading aloud.

S3) Novel Motor Control Subsystem

Towards facilitating motor control in our project, we designed an interface to communicate with IbnSina's motors (Hitec and Dynamixel) easily. Our interface is written in Java; to help us use the wide range of libraries and APIs that are available for java users.

There are two types of servos in Ibn Sina: first, Hitec servos, which were used in three functions for the movement parts in our projects, which they are: eye blinking, facial expressions and lips movement. Second, Dynamixel servos were used for head/neck movement.

In order to achieve realistic *Eye blinking*, we used a stochastic method, which produces blinks with time intervals according to a suitably-tuned distribution, by sending appropriate actions to the SSC32 controller of the Hitec servos. *Lip syncing* with speech, is a function that has two steps. First step is capturing sound from the soundcard through microphone device using JavaSound, and then it takes the captured sound wave and converts it to byte array to make the analyzing more easily and produce the amplitude value as an output for the following step. Second step is taking the output from the previous step as an input and convert it to string value that the SSC-32 controller can understand in order for the mouth servos to produce the right movement. *Neck movement*, for which dynamixel motors are activated, uses a stochastic method creating perlin-noise-like movements. In general, when issuing motor commands to Dynamixels, first the specified position value is checked so that it lies on the specified range the order to take a movement, and then it is sent to the USB2Dynamixel which is PC-Based Dynamixel controller that allow us to send command to the Dynamixel motors. The command is sent as an instruction packet in a special format that contains the motor ID, instruction type, address and the values. For *facial expressions*, Hitec servos are used to generate simple face expression such as: happy, exited, sad and angry. Facial expressions can be interpolated and mixed, and arise out of an event-fed affective state subsystem with temporal dynamics.

4. PHASE III: IMPLEMENTATION & TESTING

S1) Conversational System with access to Online Content and multilinguality

The implementation phase took most of time of the project lifecycle. We followed the incremental development method for the implementation, that is; each time we implement a small part of the project we move to another part and so on. For implementing the chat system we started with implementing the Translation module, we used the Google Translate Java API that is available for free use on the Internet [8]. The Wikipedia class was implemented through several phases. At the very beginning we used the URL libraries [9] to pass an input to Wikipedia's URL to get the html data and filter it to get the introduction part. After testing it, it was noticed that the code does not work with all pages because each page in Wikipedia has a different html format. Then, we tried an API for Wikipedia [10] that is written in PHP, we tried to use it but it was not deemed satisfactory. Finally we used the WikiBot Java API [11], which was successfully implemented in our project. The Quran module was implemented using the URL [9, 12] methods mentioned; we used our own databases and PHP codes that are hosted on the Internet for extracting data of the Holy Quran. The chat system took much time to get it done; we started by using a freely-available client-server class, and we spent more than three weeks to combine it with our modules but as this was proving unsatisfactory, we decided to implement our own chat program. Based on basic knowledge, available for example from textbooks such as [13] for implementing the client and server; it was easy to create a simple messenger from the base and then we implemented our own chat from scratch. We then we combined it with the other modules. At the end we added the ChatterBot module [14] that replies the user's messages like "Hello", "How are you doing?", which we also later enhanced it in such a way that it replies to some more advanced questions. After that, we created support the two main operating modes that the conversation is done; for Mode A: in which the input is text that can be Arabic or English; the language character set is detected, for this we use the Arabic encoding windows-1256 [15], the text is then processed through the modules mentioned above and the output is created in a text form in the same language of the input in addition to the speech output (Text-To-Speech) in the same language too. Mode B is implemented so that it takes Arabic speech as input, which is previously processed via the Acapela speech recognition software, to create text for the speech inputs (Speech-To-Text). The text then is processed the same way as Mode A and a speech output is created. Also, we created a module that searches for and displays images that have names related to the terms inquired for. It was also integrated with the chat code. Finally all modules were integrated together and tested, the testing was so important for enhancing results.

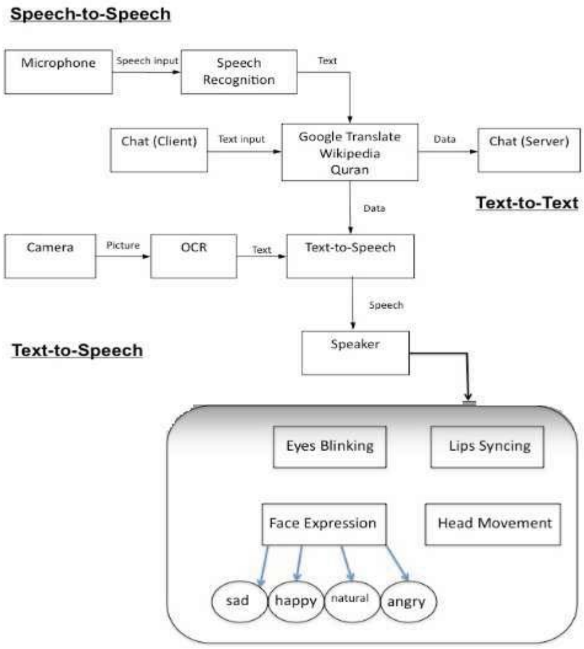

Figure 1. .Modular Subsystems of Ibn Sina

S2) Basic Reading Subsystem

In implementing the OCR we first connected IbnSina with the control computer. IbnSina's eye cameras provide us with video, so we can put a piece of paper of white background with black text in front of him to be read out. The code takes a screenshot from the computer from what IbnSina sees through camera capture. The screenshot is then converted from jpg to pnm. The OCR code outputs the text in this picture and saves it in a text document through the OCR module [16]. Acapela text-to-speech speaks up the text from the output of the OCR code.

S3) Novel Motor Control Subsystem

First, after basic experimentation, we succeeded on moving motors on both SSC32 [17] and Dynamixel [18]. Then, we created code implementing the methods which end up sending appropriate commands to the SSC32 controller for eye blinking and lip movement. Later, we implemented code, which creates background neck movement using the Dynamixel RX motors. The eye blinking parameters were tuned in order to achieve realisticity on the basis of subjective viewer opinions. We also tested the subjective realisiticity of the lips movement, but initially delay was noticeable, the movement was slow and some samples were ignored. The solution was increasing the range of samples, and we went through a sequence of one-time open/close viewing tests. Regarding facial expressions, we started by hand-coding a set of end-positions of "frozen" basic expressions, including natural, happy, angry & sad. Then, code was devised that is able to create weighted interpolations of the above extreme expressions; and finally, code for issuing an apparent facial state on the basis of an underlying emotional state, and whose components are increased / descreased on the basis of incoming events of a given affective significance, while decaying slowly back to the neutral state in case no events take place.

Finally, all the subsystems were integrated in a demo which illustrates the new capabilities of the robot.

Multiple future steps are planned for the robot; currently, work toward providing full-body embodied telepresence through motion capture is underway, extending along the lines of [4]. Also, capabilities for supporting multi-conversational partners are being created. Most importantly, a repackaging and fusion with existing modules of the long-term-relationship-targetting FaceBots social robots [21] is underway, enabling Ibn Sina to also capitalize on as well as deposit social information available on the internet. Furthemore, multiple HRI experiments are at the planning and design stage.

5. CONCLUSION

In this paper, latest efforts towards transforming the IbnSina Arabic-language conversational robot, into an advanced multilingual interactive android robot, were presented. We described multiple extensions carried out to IbnSina's software architecture in order to enrich its capabilities in multiple ways, so that it can become an exciting educational / persuasive robot in the future. The main axis for these extensions were: access to online (Wikipedia) and stored (Koran database) content for dialogue generation, basic multilingual capability exploration (English and Arabic, also utilizing Google Translate), basic read-aloud-text capability (through OCR), and systematization of motor control, with the creation of a higher-level API for real-time lip syncing, eye blinking, natural looking random face movements, and interpolation between facial expressions, including an event-based affective state subsystem with temporal dynamics. With such capabilities, IbnSina becomes closer to an attractive robot that can find real-world application in malls, schools, as a receptionist etc. And most significantly, this work provides important research insights, and a working real-world proof-of-concept of corpus- as well as online-info mining conversational androids, with multi-lingual as well as reading and display capabilities, and which in the future, might well support multiple conversational partners as well as sustainable longer-term relations with humans, and thus eventually become important daily assistants and friends to us.

6. REFERENCES

[1] N. Mavridis and D. Hanson, "The IbnSina Center: An Augmented Reality Theater with Intelligent Robotic and Virtual Characters", IEEE RoMAN09

[2] Ahmed Daheri, Saeed Merri, Muhamed Kuwaiti, Salem Katheri, "Conversing with Ibn Sina: Towards Dialogic Robots with Arabic Language, Gestural and Expression Capabilities", Senior Project Report, UAEU CIT 2009

[3] L. Riek, N. Mavridis, et al. "Ibn Sina Steps Out: Exploring Arabic Attitudes Toward Humanoid Robots", AISB 2010

[4] N. Mavridis, E. Machado et al.,"Real-time Tele-operation of an Industrial Robotic Arm Through Human Arm Movement Imitation", IRIS 2010

[5] C. Christoforou, N. Mavridis et al.,"Android tele-operation through Brain-Computer Interfacing: A real-world demo with non-expert users", IRIS 2010

[6] "Latent Semantic Analysis problem". Jul 21, 2008. [online]. Available at: old.nabble.com/Latent-Semantic-Analysis-problem-td18562801.html . Feb 2010

[7] Hesham Kamel. ITBP440. Class lecture: "Design, Prototyping and Construction". College of Information Technology, UAE University, Al Ain, UAE. March, 2010

[8] google-api-translate-java. (2010). [online]. Available: from code.google.com/p/google-api-translate-java/

[9] Reading from and Writing to a URLConnection, (Dec 1, 2010). [online]. Available from: java.sun.com/docs/books/tutorial/networking/urls/readingWriting.html

[10] MediaWiki API, (2010). [online]. Available at: www.mediawiki.org/wiki/API. March, 2010.

[11] Java Wiki Bot Framework. [online]. Available at: jwbf.sourceforge.net/pw/index.php April 2010.

[12] How can I execute a PHP script from Java? [online]. Available from: stackoverflow.com/questions/655620/how-can-i-execute-a-php-script-from-java . Feb, 2010.

[13] H. Deitel. "Java: How to Program". Prentice Hall, 2010

[14] "Tutorial on making an Artificial Intelligence Chatbot". Available from: ai-programming.com/java_bot_tutorial.htm

[15] Ahmas Hammad. Java and Arabic Support. Dec 16, 2006. ahm507.blogspot.com/2006/12/java-and-arabicsupport.htm

[16] «Joerg Schulenburg, "open-source character recognition", Retrieved from: jocr.sourceforge.net/index.html

[17] SSC-32 Servo Controller Board, Hardware Manual

[18] ROBOTIS, Dynamixel RX-28 Motor Hardware Manual

[19] T. Kanungo, G. Marton, O. Bulbul, "OmniPage vs. Sakhr: Paired Model Evaluation of Two Arabic OCR Products": citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.9496&rep=rep1&type=pdf

[20] B. J. Fogg, "Persuasive technology: using computers to change what we think and do", Morgan Kaufmann, 2003

[21] N. Mavridis, W. Kazmi, P. Toulis, C. Ben-AbdelKader, "On the synergies between online social networking, Face Recognition, and Interactive Robotics", CaSoN 2009