GALA 2007 submission document

Title: **Relational approach towards the synchronization of head, lip and eye movements of an intelligent virtual agent reading a text with different tempo and different sound volume**

Track: Student
Category: Application
Authors:

Dilyana Valkova Budakova – students' team leader
Technical University of Sofia – Branch Plovdiv,
Department of Computer Sciences and Technologies
PhD in Computer Sciences, Assistant Professor.
Bulgaria

Anastasia Iordanis Nalpantidou
2nd year, Master Student
Computer Sciences and Technologies,
Technical University of Sofia – Branch Plovdiv
Bulgaria

Elena Stoyanova Gesheva
Five months ago she graduated with a Master's degree from the
Technical Unuversity of Sofia – Branch Plovdiv,
Computer Sciences and Technologies.
Bulgaria

Valentin Veselinov Vachev
3rd year Bachelor Student,
Informatics
Plovdiv University "Paisii Hilendarski".
Bulgaria

Contact data:

name: Dilyana Valkova Budakova
e-mail: dilyana_budakova@yahoo.com

url: www.geocities.com/dilyana_budakova;

www.tu-plovdiv.bg;

phone: mobile: +359 878 28 16 16 ;
postal address: *Dilyana Valkova Budakova*
Technical University of Sofia – city Plovdiv,
Department of Computer Sciences,
61, Sankt Peterburg Blvd.
4000 Plovdiv
Bulgaria .

URL: www.geocities.com/dilyana_budakova;

Movie file submitted: *GALA07VideoTUBulgaria.avi*

GALA 2007 submission document

Reference teacher: Dilyana Valkova Budakova
affiliation: Technical University of Sofia – city Plovdiv,
Department of Computer Sciences,
61, Sankt Peterburg Blvd.
4000 Plovdiv
Bulgaria.
e-mail: dilyana_budakova@yahoo.com

Process of development:

An intelligent virtual agent reading a text was developed between January and May 2007, as an application. The team was formed with the only aim to develop the currently presented application especially for participation in GALA.

All members of the team together with their leader have worked for the realization of the programme system. They filled the database with words for pronouncing, with information about the paths to the corresponding wav files, and with the numbers of the bitmap images for the realization of lip, eye and head movements of the agent. (This part was done more specially from 1st student) They separated each word from the wav file in which the texts had been pronounced in Bulgarian and recorded beforehand by the team leader (This part was done more specially from 1st and 2nd student). They realized by means of Photoshop CS the collection of pictures for the two main poses of the agent's head and accomplished the synchronization between the agent's image and the speech. (This part was done more specially from 2nd student and 3rd student) They prepared together the documentation and the video clip for presentation of their work. (This part was done more specially from 3rd student).

The idea about the development of this programme system, as well as the analysis of the task, its division into main sub-tasks and the definition of methods and algorithms for their realization belong to the leader of the project.

All team members show great interest in the undertaken work and are going to continue developing both new IVA and their own technical skills.

Resources used:

1. Visual Basic .NET, and MS Access XP. The programme system has been realized in Visual Basic.NET, and the databases – in MS Access XP.

For accessing the database control system the approach with using objects for accessing data is used (DAO). The model DAO involves a great number of objects having relevant methods and properties for creating databases, defining tables, fields, indices and establishing relations between the tables etc.

In order for the text to be pronounced, a command "Sound" from the Microsoft Multimedia Control enabling reproducing of sound files is executed.

MMCntrlWav.FileName = Name and path to wav file
MMCntrlWav.Command = "Sound"

2. Photoshop CS. The collection of bitmap images presenting all the different head, eye and lip positions is realized by means of Photoshop CS.

3. WavePad. Each word had been separated in WavePad.

Resources required: An intelligent virtual agent reading a text was developed and tested on a PC with Windows XP Prof OS, 512 kb RAM, CPU Celeron 1.00GHz FCPGA2. The PC's microphone and speakers attached were used.

GALA 2007 submission document

Title Relational approach towards the synchronization of head, lip and eye movements of an intelligent virtual agent reading a text with different tempo and different sound volume.

1. The application and context of the work

Here is presented a relational approach towards the synchronization of head, eye and lip movements of an intelligent virtual agent reading a text. During the pronunciation of the text the user can interrupt the agent and point whether he/she wants the agent to read faster, louder or quieter by clicking on the corresponding button.
The agent follows the mouse movement with eyes and reacts when the mouse goes over her face. When pronouncing the text the agent moves her lips in accordance with each word, blinks, and occasionally changes her head pose.
The agent reads a text in Bulgarian, which is loaded from a text file or introduced by the user in RichTextBox Control. For the purpose of demonstration it has been chosen that the agent reads a film advertisement, recorded in advance as a text file.

2. Novelty/Some of the distinguishing feature of our work are:

1. Relational approach towards the synchronization of the lip and head movements of the agent with the pronounced word. The relational presentation of each gesture animation is convenient, since the code remains the same and the programming of new gestures and mimics can be reduced to adding numbers of pictures in the relational table of the gesture or mimics being modeled.
2. Relational databases for a fast access of the programme to the information required for the animated pronunciation of the text.
3. Buffer in which all the information required for the animated pronouncing of 40 words at the most is recorded.
4. Access to the different sound models of the words.
5. Reading interruption through realization of the work in two threads

3. The architecture

Basic algorithm for reading a text together with a synchronized lip and head movement of the agent. The text to be pronounced is loaded from a text file in RichTextBox control or it is entered by means of the keyboard. From here the text in the form of a string is processed by a sub-programme, which identifies each word in it (paying attention to intervals, numbers or punctuation marks). Another sub-programme extracts all the information about the word from the relational database and stores it in a buffer (the absolute path to the wav file for each word is concatenated this way). When the buffer is full or there are no more words in the text, another sub-programme is called upon, which executes the pronunciation itself and visualizes the images required for the realization of the synchronized lip, head and eye movements of the agent.
Each database has one table with 6 columns, in which there are respectively: the words, starting with the letter mentioned in the name of the database; the names of wav files, in which the pronunciation of these words has been recorded and the names of the folders in which they are placed (the absolute path from the root of the file system to these folders will be defined later); and there are also four columns, comprising numbers of pictures corresponding to the lip position in the pronunciation of each of these words for the first basic agent's head position.
The relational presentation of each gesture animation is convenient, since the code remains the same and the programming of new gestures and mimics can be reduced to adding numbers of pictures in the relational table of the gesture or mimics being modeled.

Relational approach towards the synchronization of the lip and head movements of the agent with the pronounced word. The application continuously follows the current head position (pose) of the agent in order to continue the animation of the agent's lip movement and the eye blinking with this basic position (pose) until the next change of the head position (pose) occurs.

A collection of 20 bitmap files, containing images of the agent's head and realized by means of Photoshop CS is used. We have two basic head positions and all possible eye and lip positions corresponding to them. Each of the images has its own number. For each of the two head positions there are 10 images arranged in the same sequence. Thus the number of each image for the second basic position can be obtained by adding 10 (the number of the images for one of the basic head positions) to the number of the same image for the first basic position.

Now it should be watched which of the two basic head positions is the current one in order to visualize the images for the realization of lip movements for this head position synchronized with the word being pronounced, and also for the realization of the eye blinking once in a while.

A Boolean variable keeps track of which basic head movement is used any moment. It is true when the agent's head is in the first basic position and false when the head is in its second basic position. For the realization of the lip movement in the first basic head position the numbers of the pictures read from the relational database are directly used.

When the Boolean variable is in a false position and the second basic agent's head pose is the current one, then to the number of the picture to be visualized number 10 has to be added (the number of poses for a basic position), and the same picture for the second agent's head position appears.

A counter, counting up to a pre-given number, defines the interval of the change in the agent's head position. Another one, counting up to another pre-given number defines the interval of agent's eye blinking.

When the programme works and the user chooses a text to be pronounced by the agent, the agent starts reading and moves her lips in synchronization with the word being pronounced; she also changes her head position every now and then, and occasionally blinks.

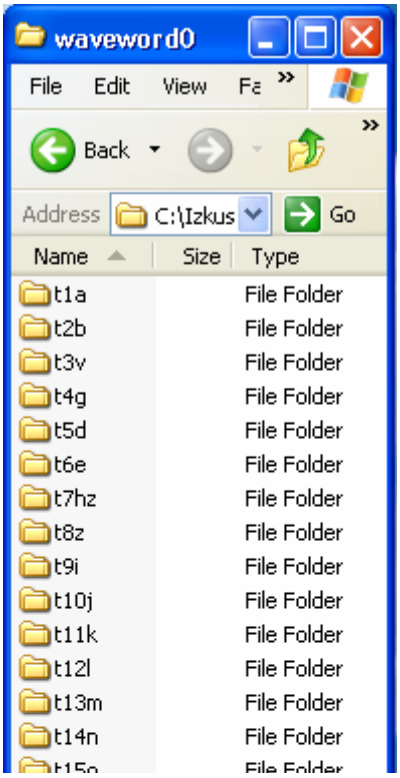


Fig. 1

4. Performance

Location of the wav files with the recorded words.

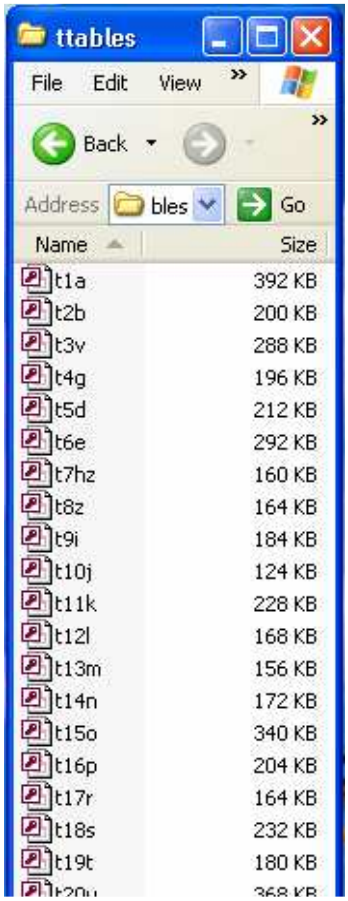
All the words, which the agent could pronounce, have been recorded as separate wav files.

Since the words amount to hundreds thousands (we have recorded a small amount just for the experiment), a number of folders, equivalent to the number of letters in the Bulgarian alphabet has been created. All the words, beginning with the same letter, are recorded in the corresponding to this letter folder. For example, all the words, starting with "a" are recorded in the folder, named t1a, the ones, starting with "b" are recorded in the folder t2b, those starting with "c" – in the folder t3c etc. Fig. 1

As the agent is supposed to be able to pronounce the text slowly and quietly or slowly and loudly, or fast and loudly, or fast and quietly, all the words have been recorded four times for each of the possibilities listed above. Therefore all the folders t1a, t2b,t3v, …., t29ya where we put the words falling into the same model have been placed in a separate folder again – e.g. slowly_quietly folder is for the words to be pronounced slowly and quietly.

All the folders t1a, t2b,t3v, …., t29ya, in which we put the model “fast and quietly” are placed in the folder fast_quietly. All the folders t1a, t2b,t3v, …., t29ya, containing the model “fast and loudly” are put in the folder fast_loudly.

Relational databases for a fast access of the programme to the information required for the animated pronunciation of the text.



The number of the realized relational databases corresponds to the number of letters in the Bulgarian alphabet. The name of the database contains the name of a letter from the alphabet. Fig 2. Each database has one table with 6 columns, in which there are respectively: the words, starting with the letter mentioned in the name of the database; the name of wav files, in which the pronunciation of these words has been recorded and the names of the folders in which they are placed (the absolute path from the root of the file system to these folders will be defined later); and there are also four columns, comprising numbers of pictures corresponding to the lip position in the pronunciation of each of these words for the first basic agent’s head position. (Tables 1a, 1b).

The first two pictures are visualized (they change once and consecutively with a short pause between the visualization of the first and the second one) before the word is pronounced; the second two are visualized after the pronunciation is completed (again in the way described above). Thus an impression is created that the agent moves his/her lips in accordance with the pronounced word.

When the relation to the information about a particular word is to be found, it is first defined which is the initial letter of the word. Then an access to the database containing the records of the relations concerning the words starting with this letter is realized. From the table of this database the line containing the recorded information about a word can be found and read. Thus the process of finding the information required for pronouncing each word from the text file is facilitated.

Fig. 2

Table 1a

Word	Wav file road	Number figure lips1	Number figure lips2	Number figure lips3	Number figure lips4
казано	t11k\kazano.wav	1	1	7	8
кандидатства	t11k\kandidatstva.wav	1	7	0	1
като	t11k\kato.wav	1	2	7	8
крис	t11k\kris.wav	1	2	2	1
корпорация	t11k\korporatciya.wav	1	6	0	1
кубче	t11k\kubche.wav	1	6	2	1

Table 1b

Word	Wav file road	Number figure lips1	Number figure lips2	Number figure lips3	Number figure lips4
работа	t17r\rabota.wav	1	0	7	1
реди	t17r\redi.wav	7	6	6	6
рекордно	T17r\rekordno.wav	1	6	7	6
решава	t17r\reshava.wav	1	2	0	1
рубик	t17r\rubik.wav	1	6	0	1
ръба	t17r\ryba.wav	1	1	0	1

Pros and cons of your system.

- Relational approach towards the synchronization of the lip and head movements of the agent with the pronounced word. The relational presentation of each gesture animation is convenient, since the code remains the same and the programming of new gestures and mimics can be reduced to adding numbers of pictures in the relational table of the gesture or mimics being modeled.

- Relational databases for a fast access of the programme to the information required for the animated pronunciation of the text. When the relation to the information about a particular word is to be found, it is first defined which is the initial letter of the word. Then an access to the database containing the records of the relations concerning the words starting with this letter is realized. From the table of this database the line containing the recorded information about a word can be found and read. Thus the process of finding the information required for pronouncing each word from the text file is facilitated.

- **Buffer in which all the information required for the animated pronouncing of 40 words at the most is recorded**. In a buffer from the operative memory, comprising six arrays (the same number as the number of columns in the database) 40 consecutive words from the text and the information about them, derived from the database, can be saved. When the buffer is full, the sub-programme for pronouncing the speech and for synchronizing the head, eye and lip movements of the agent gets started. After running low the buffer is ready to store the information about the next 40 words to the exhaustion of the text to be read. The use of the buffer guarantees the independence of the processes of pronouncing the speech and synchronization and looking for the information about the words in the database.

- Access to the different sound models of the words. In dependence with the user's desire for the agent to speak fast, slowly, loudly or quietly, the first portion of the absolute path to the file, where the word to be pronounced has been recorded, will be different. The user clicks on one of the four buttons offering an option of the way of reading the text and thus he/she actually defines the first portion of the path to the corresponding sound model.

- Reading interruption through realization of the work in two threads. The possibility to interrupt the agent after she has begun reading the text and has not completed it is realized by starting two threads. The first thread accomplishes the process of reading. The second thread can interrupt the first one in accordance with the user's desire. When the first thread calls upon the functions for realization of synchronized with the head and lip movements speech pronouncing, the second one generates a message, visualized in the message box and informing the user that the process of reading could be

interrupted by clicking on the OK button in this message box. If the user clicks on this button, the second thread interrupts the work of the first one and the agent stops reading the text.

- Following the mouse movement with eyes and shaking head when the mouse passes over the agent's face. Agent's eyes follow the cursor movements before the reading has begun. On the purpose the event mouse move is processed and, in accordance with the current coordinates of the mouse position, a collection of images with various agent's eye positions is visualized again, which creates the impression of eye movement following the movement of the cursor.
When the cursor goes over the agent's face, he/she reacts by shaking his/her head. A change between the two agent's head positions is realized again.

- After adding the realization method for more smooth and more affective reading complying with the text being read, this program could be applied in everyday practice in Bulgaria.

Fig. 3 Running form

Fig. 4 Design view