AI & Soc (2009) 24:225–235
DOI 10.1007/s00146-009-0213-6

ORIGINAL ARTICLE

Toward a multi-culture adaptive virtual tour guide agent with a modular approach

Hung-Hsuan Huang · Aleksandra Cerekovic · Igor S. Pandzic · Yukiko Nakano · Toyoaki Nishida

Received: 15 October 2008 / Accepted: 27 May 2009 / Published online: 23 June 2009

Abstract Embodied conversational agents (ECAs) are computer-generated, human-like characters that interact with human users in face-to-face conversations. ECA is a powerful tool for representing cultural differences and is suitable for interactive training or edutainment systems. This article presents preliminary results from the development of a culture-adaptive virtual tour guide agent for serving Japanese, Croatian, and general Western users by displaying appropriate verbal and non-verbal behaviors. It is being implemented in Generic ECA Framework, a modular framework for developing ECAs. Dividing the ECA functions into reusable and loosely coupled modules minimizes the effort required to implement additional behavior and facilitates incremental scale up of the system.

H.-H. Huang (✉) · T. Nishida
Graduate School of Informatics,
Kyoto University, Kyoto, Japan
e-mail: huang@ii.ist.i.kyoto-u.ac.jp

T. Nishida
e-mail: nishida@i.kyoto-u.ac.jp

A. Cerekovic · I. S. Pandzic
Faculty of Electrical Engineering and Computing,
University of Zagreb, Zagreb, Croatia
e-mail: aleksandra.cerekovic@fer.hr

I. S. Pandzic
e-mail: igor.pandzic@fer.hr

Y. Nakano
Faculty of Science and Technology,
Seikei University, Seikei, Japan
e-mail: y.nakano@st.seikei.ac.jp

1 Introduction

The recent advances in transport and communication technologies have globalized markets and businesses and have changed the way people interact with each other. Enterprises pursue success in overseas markets to maintain their competitiveness, and businessmen have to negotiate with their foreign customers. In the academic world, attending international conferences is the most efficient way for researchers to gather first-hand information. Overseas trips for tourism and other personal reasons are also becoming easier and more popular. The ability to communicate face-to-face with people who come from other cultural backgrounds is gaining importance.

The differences among cultures appear not only in languages and their use, but also in the display of internal emotional state in facial expressions, gestures, the range of movements, interpersonal distance, and so on (Isbister 2004). Computer graphic characters or embodied conversational agents (ECAs) who can speak in the natural language and display rich facial expressions and who have large degrees of freedom in body movements are ideal interfaces for culture-enabled systems.

A number of research groups have studied the use of ECAs in immersive training and pedagogical applications for intercultural communication. Examples include the TLTS systems developed for training US soldiers in foreign languages and culture to smoothen the execution of their missions abroad (Johnson et al. 2005), an attempt to use virtual peers to encourage African American children to switch their language coding to increase school-based literacy (Iacobelli and Cassell 2007), a proposal for modeling cultural differences as computational parameters based on a combination of the analysis of a video corpus

collected in experiments, and a theoretical model (Rehm et al. 2008a).

Interculturally competent ECA system development typically applies the classic "analysis by synthesis" method:

1. Conduct data acquisition experiments and observe human-to-human interactions.
2. Hypothesize the principal requirements for human–agent interactions and implement a prototype system.
3. Analyze the prototype system and verify the hypotheses; if the results are not satisfying, then go back to step 2.

In this development style, the researchers can clearly benefit if the system can be partially replaced and prototyped rapidly. This article presents a way to achieve this by introducing a common and modular development framework for ECAs, using the realization of a multi-culture adaptive agent named Dubravka as an example. The Dubravka agent was developed in an ongoing international collaborative project aiming to build a tour guide agent who is adaptive to users from general Western, Japanese, and Croatian cultures. In this article, Sect. 2 introduces the objectives of the project and the eNTERFACE'06 workshop where Dubravka was created. Section 3 describes the development details of building the Dubravka agent. Section 4 describes potential extensions to the Dubravka agent with pluggable culture modules. Section 5 concludes the article.

2 Multi-culture adaptive tour guide agent, Dubravka

In order to consider the cultural issues in computer–human interfaces, depending on the needs of the application, there are two approaches: internationalization and localization (Young 2008). Internationalized designs exclude culture-dependent features and implement behavior that will be interpreted in the same by people from different cultures and prevent misunderstanding. Localization includes culture-specific designs for the target audience. According to research reports such as that of Nass et al. (2000), people prefer interface agents with the same ethnicities as themselves; they feel more comfortable with and tend to be more trusting of these agents. Baylor et al. (2006) investigated the impact of the appearance of an interface agent in terms of the age, gender, and "coolness," and reported that participants prefer peer-like (similar to the participants) agents. Pickering and Garrod (2004, 2006) reported that people tend to align their use of language to the interlocutor during dialogues. This alignment is the basis of successful communication. Costa et al. (2008) suggested that speaking in a second language could impair the alignment in dialogues.

In the case of an interface agent for users who may come from many cultural areas, such as a tour guide agent for a sightseeing spot, information transfer should be more efficient if the agent speaks the user's native language and shows behaviors familiar to the user.

2.1 eNTERFACE workshop project

This study was started during the eNTERFACE'06 workshop that focused on the topic of multi–modal human–computer interfaces and was held in Dubrovnik, Croatia in the summer of 2006. Contrary to regular workshops where the researchers only present their research results but do not actually work, the principle of this workshop was to invite volunteer student participants to collaboratively develop proposed application projects in a relatively short four-week period and then present their research results.

The title of our proposed project was "An Agent-Based Multicultural User Interface in a Customer Service Application" (Huang et al. 2006, 2008b). After the announcement of the project proposal in sponsoring universities, we got six student members in our team, three of whom were from our research group. On the basis of the discussions among team members prior to the workshop, the target application was chosen to be a tour guide agent for Dubrovnik city. The entire old town of Dubrovnik has been designated a UNESCO World Heritage Site. Dubrovnik is a famous sightseeing spot and attracts thousands of tourists from all over the world, especially in summer because of the attractive festivals in this period. Since most of the team members come from Japan or Croatia, it was most convenient to gather first-hand Japanese and Croatian cultural information, where the differences are supposed to be fairly obvious. The agent was given a young female appearance and was named Dubravka, which is a regular Croatian female name and can be associated with the city.

Dubravka provides sightseeing information for Dubrovnik to its visitors via verbal and non-verbal interactions. An example usage scenario of the objective system is as follows: when a visitor comes to the system, the system recognizes the visitor as a Western person, Japanese, or Croatian from a combination of the speech recognizer's result and the non-verbal behaviors of the visitor. An example of such obvious cues is bowing, which Japanese people use for greeting. The agent then adapts itself to the Japanese mode, that is, it speaks in Japanese and behaves in Japanese ways to provide the visitors with tour information. At the same time, visitors can interact with the agent not only by speaking in their natural language but also by non-verbal gestures and posture behaviors such as pointing

to an object in the background image or by raising their hand to indicate that they want to ask a question.

Although not all of the ambitious objectives of this project could be achieved during the period of eNTERFACE'06, we continued developing it after the workshop. In eNTERFACE'08, we further explored the possibility of extending the system to allow two users to interact with the agent (Cerekovic et al. 2008).

2.2 Cultural issues involved

Culture is relevant to many aspects of human–human communications. These effects should be also reflected throughout the design of culture-sensitive ECA systems: how the agent interprets its perceptions, how the agent thinks, and how the agent behaves. From the point of view of communication interfaces, the language spoken by the agent directly determines how the user perceives it and is an obvious factor that distinguishes different cultures.

Cultural differences are also displayed in people's non-verbal behaviors. The same gestures may represent different meanings in different cultures and the same meaning may be represented by different gestures. Sometimes the differences are coded culturally, for example, beckoning gestures are displayed in exactly opposite directions by British and Japanese people. The finger gestures representing numbers provide another example; Japanese people use two hands and overlap one of them with the other one while Chinese people use only five fingers of one hand to present numbers from one to eight, even though these two cultures are similar in many aspects. Misuse of these culturally coded emblem gestures may cause misunderstandings and problems in communication.

Handling cultural issues is very relevant to emotion control and the deliberations of the agent (de Rosis et al. 2004). However, in a four-week project, it was not possible to explore these issues in depth. In the case of the Dubravka agent, we were only able to handle the surface of cultural issues, i.e., the perceptions and behaviors of Dubravka including the language that the agent spoke and listened to, and the usage of different culturally coded emblem gestures.

A significant feature that has not yet been achieved is the automatic recognition of the culture class to which a user belongs from her (his) non-verbal behaviors. We realized that it is difficult to find the differences in non-verbal behaviors between users coming from different cultural backgrounds since the beginning of the interaction with the agent. This is an extremely difficult task even for humans and more research is required. Instead of that, the current system is switchable to different culture modes by asking the user to select a cultural mode with a question in English at the beginning of interaction.

2.3 Cultural differences in the non-verbal behaviors of Dubravka

Since our target is a tour guide agent who serves visitors from Japan, Croatia, or somewhere in the Western culture area, the first task was to gather culture-specific behaviors in the tour-guiding context, particularly the culturally coded emblem gestures. The material we used was mainly obtained by taking video data of Japanese tour guides at several famous sightseeing spots in Kyoto and European tour guides in Dubrovnik (Fig. 1). Appropriate non-verbal behaviors of the agent were chosen from observation of the collected video corpus and the ones introduced in (Hamiru.aqui 2004).

While modeling the gesture styles for the character, we aimed to emphasize the diversity of the three cultures. For example, we introduced the "cross hands in front of the chest" gesture in the Japanese mode. This gesture is usually performed with additional head shaking to express negation. It seems to be rather unique and normally draws the attention of Western people who first come to Japan (Fig. 2, left). Another example is the "prohibition" gesture (Fig. 2, right). In Japan, it is expressed by waving with a hand while the arm is extended. Sometimes shaking the head sideways is also added. When asking to wait, Japanese people usually show the palm of one hand to another person. At times, both hands could be used.

Some confusing gestures can make people misunderstand because of different interpretations in different cultures. For example, the beckoning gestures that mean "go away" and "come here" are performed in opposite directions in Western countries and Japan. In Dubravka's Croatian and general Western modes, she gestures "come here" by waving upwards and backwards with one hand and the back of the hand facing downward. However, this

Fig. 1 In one scene collected in the tour guide video corpus, one of the authors is introducing Dubrovnik city to the eNTERFACE'06 workshop participants and is performing a beat gesture

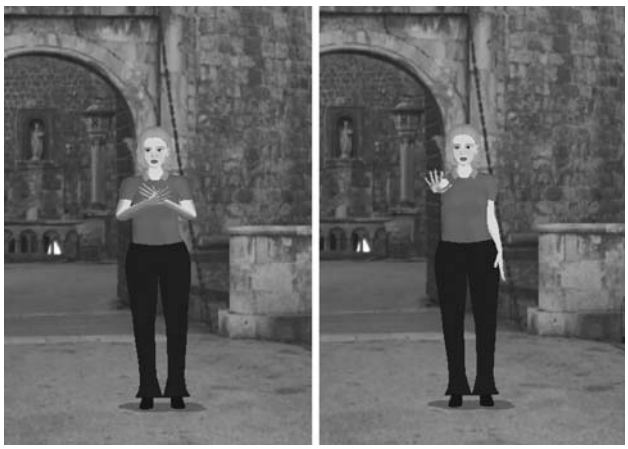

Fig. 2 Dubravka shows Japanese "negation" and "prohibition" gestures

gesture may be interpreted as "go away" in Japan. Therefore, in her Japanese mode, this gesture is performed with the back of the hand facing upward.

Unlike the Japanese gestures, which are often significantly different from the Western ones, we could not find obvious differences among the Western tour guides, even if they came from different countries, in our observation of the video corpus. Table 1 shows some examples of gestures modeled in the Dubravka agent system.

3 Building the Dubravka virtual tour guide agent

In order to realize an ECA that is capable of carrying out natural multi-modal and face-to-face conversations with humans generally involves the following tasks: Acquire and understand the inputs from human users via sensor devices; choose appropriate behaviors as responses; and realize those behaviors by animating CG characters in multiple modalities.

In acquiring inputs, raw data patterns from sensor devices need not only to be recognized but also to be combined according to timing information for interpretation as semantic meanings (Johnston and Bangalore 2000). In the deliberation process, the agent needs to choose the most appropriate action from its behavior repository based on its perceptions, beliefs, and goals in responding to the current situation. In more advanced systems, the decision may be affected by the internal emotional state of the agent, the modeling of its personality, the setting of the social relationship between it and the users, and cultural issues. In the realization of the agent's behaviors, in addition to the prosody characteristics of the verbal channel, synchronized and precise control of non-verbal behaviors are also required. In order to move the joints of the virtual characters to the designers' intended positions, inverse kinematics computations need to be done. In order to realize the outputs of the agents to the users, its appearance and its virtual environment need to be rendered in 3D computer graphics as life-like, realistic, and believable virtual characters.

In order to realize these functionalities, knowledge and techniques like signal processing, pattern recognition, natural language processing, gesture recognition, artificial intelligence, linguistics, psychology, cognitive science, natural language generation, gesture generation, virtual reality, and computer graphics are required. Due to the involvement of diverse disciplines, it is not easy to start the development of an ECA from scratch. Utilizing available software tools is a typical method in implementing ECAs. For example, SmartBody (Thiebaux et al. 2008) is an open source modular framework for animating ECAs in real time. An EMA emotion framework (Gratch and Marsella 2004), a dialogue movement engine (Traum and Larsson 2003; Larsson and Traum 2000), and an authoring tool for creating dialogue knowledge for tactical questioning (Traum et al. 2007) developed in USC are other such tools.

Table 1 Some examples of the difference of gesture displays in each culture

Action	Culture dependency	Croatia	Japan	West
Bow	In this gesture, we present three types of bowing: shallow bow, using only head; deeper bow (Japanese style) shows respect to the listener	✔	✔	✔
Invite	Croatian and general Western gesture presents waving upwards and backwards with one hand and the back of the hand facing downward. However, this gesture may be misunderstood as "go away" in Japan. In Japanese mode, this gesture is performed with the opposite orientation of the back of the hand	✔	✔	✔
Cross	This is a Japanese emblem gesture, meaning that something is not allowed. The hands are crossed in front of the lower part of the chest		✔	
Extend	This gesture means right arm extended with the palm open and oriented upwards. In the Japanese culture it means "wait please"		✔	
Wave	This gesture presents oscillating right hand waving. Used in combination with the "extend" action as part of the Japanese gesture meaning "No"	✔		✔
Banzai	Throwing both arms up expresses good fortune or happiness		✔	

Because of the nature of the eNTERFACE workshop, there were two general difficulties for each team to achieve its goal.

- The four-week period of the workshop was relatively short to realize significant achievements or start a new project.
- There were some team members who were not directly engaged in this joint research project or not familiar with the fields which the project involved.

Reducing the hurdles for the team members and minimizing the effort of developing new programs were thus essential issues for producing as many results as possible in the limited four-week workshop period. The Dubravka agent was therefore implemented in a modular ECA development framework, generic embodied conversational agent framework (GECA Framework) (Huang et al. 2008a).

3.1 Generic embodied conversational agent framework

In order to facilitate result sharing and rapid prototyping of ECA research, a general purpose ECA programming framework that is meant to seamlessly integrate ECA assemblies is being developed by our group. This framework is composed of a low-level communication platform (GECA Platform), a set of communication API libraries (GECA Plugs), and a high-level protocol (GECA Protocol or GECAP).

GECA Platform is a communication infrastructure based on a blackboard model and XML message exchange in a subscribe-publish mechanism. There is a server that provides common communication services include a naming service, message type subscription, and message forwarding management. For the benefits from the support of two-way communication and an explicit temporal model that are essential in real-time interactive applications, a light weight protocol, OpenAIR (mindmakers.org 2005), was adopted as the low-level routing protocol for communication among the components running on the platform, the GECA server and blackboard managers. *GECA Plugs* are helper libraries that absorb the differences caused by operating systems and programming languages to facilitate the development of the wrappers of individual ECA components while they are plugged into the platform. They are basically AIR Plugs in the OpenAIR context enhanced with additional GECA original routines. *GECAP* is a specification of core XML message types and message formats exchanged among the components connected by the GECA platform. Its syntax is not fixed and can be customized depending on the applications. Because our main interests are human–agent interactions, we currently treat the deliberation process as a black box and focus on the multimodal inputs and outputs of ECA systems. In GECAP, there are three categories of messages: input phase, output phase, and system messages.

The components generating input phase messages acquire sensing data of human users' behaviors and interpret them in multiple channels with modalities such as natural language speech, pointing gestures, nodding, and gazing. In addition to interpreted sensor data, timing information and alternative hypotheses of the interpreted results are transferred. The actuator of software-based ECAs is the character animation player. The components send output phase messages to drive the virtual character in the player to speak and perform non-verbal animations as well as typical controls over the virtual environment such as changing the scenes behind the character. In GECA, verbal information is the master channel and non-verbal behaviors that are configurable at run-time are synchronized with it. Several system message types such as component initialization, operation termination, or status query are also defined.

Instead of a complex dialogue management engine, GECA Scenario Markup Language (GSML) describing human–agent interactions and its execution component were developed to supplement GECAP. GSML is an XML-based script language to define a state transition model for a multi-modal dialogue between the user and the agent. The modeled dialogue progress with anticipated verbal or non-verbal inputs from the user causes the transitions between the states. Human–agent interactions are written in AIML-like pattern–template pairs (A.L.I.C.E. AI Foundation 2005). Furthermore, the problem of fusion among multiple modalities is handled in a simplified method and is described in syntax similar to W3C EMMA (W3C 2004). The messages output from it drive the animation player to play multi-modal synchronized character animations and control the scenes in which the character is located.

The development of the first GECA server prototype as well as .Net, C++, and Java versions of GECA Plug have been completed. We have also implemented several general-purpose components such as a Japanese spontaneous gesture generator (Nakano et al. 2004), head orientation tracker, hand shape recognizer, head nodding/shaking recognizer, GSML script executor, speech recognizers, and a character animation player implemented with visage|SDK (Visage Technologies AB 2008). These components are shared by several ongoing projects including an ECA based quiz console (Huang et al. 2007, 2008c) and the Dubravka tour guide agent.

3.2 The development of the Dubravka agent

The functionalities of the Dubravka agent system are divided into standalone GECA components so that each

one of them only supports relatively simple functions and they are loosely coupled with each other. The components then jointly generate the behaviors of the tour guide agent as a single integral system. By this approach, the number of necessary newly developed programs can be decreased and legacy components can be reused without significant modifications.

In the Dubravka agent, some components like the animation player or the sensor devices can be the same in the three different cultures, and some parts like speech I/O or culturally-coded emblem gesture animations are similar but different in the three cultures. The system can benefit from being composed of culture-dependent components which are dynamically switched to the currently appropriate ones according to the cultural mode while culture-independent ones are shared and are always running across different culture modes.

The system was built by reusing as many available components as possible to reduce the efforts required to develop new components. The following is an inventory of the software components and the contents used in the Dubrovnik tour guide application.

The components which can be reused by another ECA system:

Scenario component. This component is an implementation of the GSML script interpreter. The available interactions with the human user in three language modes are defined in a single script.

Japanese spontaneous gesture generator component. This component is a wrapper of the CAST (Nakano et al. 2004) engine which generates the type and timing information of spontaneous gestures from a Japanese utterance input string. This component has been implemented.

Character animation renderer component. This component is a wrapped character animation player that is implemented with visagelSDK. It accepts driving event messages from the animation category and speech synthesizer component and performs the specified character animation. Because the character animations need to be synchronized with voice with a precision of milliseconds, Text-To-Speech (TTS) engines must be tightly bound to the player. In the current implementation, English and Japanese words that the agent speaks are generated by Microsoft SAPI-compatible Pentax VoiceText (Hoya Corp. 2008) TTS engines.

English and Japanese speech recognition components. These components are wrapped recognition engines that recognize Japanese or English spoken by the visitors by matching predefined grammar rules. Because of the lack of a good enough speech recognizer for Croatian, it is recognized by an English speech recognizer with grammar rules, which will be explained later in this section.

Sensor data acquisition components. The non-verbal behaviors of the users are recognized by using the data from data gloves, motion capture, head tracker, and acceleration sensor. In eNTERFACE'08, two new components were introduced. One detects whether there are user movements by using OpenCV (Intel Corp. 2006) and standard image difference techniques are also implemented. The other uses a commercial product, Omron's OkaoVision (Omron Corp. 2008). It is a library that provides accurate face detection and extra functions like face orientation, gaze direction, the positions and openness of eyes and mouth, gender detection, age identification, and face identification from a single image. It has the inherent limitation that when the users turn their heads to the left or right then their faces cannot be detected. These components acquire raw data from the sensor devices, interpret them, represent those events as text strings and send the results to other components for further processing. The configuration of these hardware devices is shown in Fig. 3.

Input interpreter component. This component was introduced to combine the raw data from several sensor components to generate the event messages that can be processed by the scenario component. The task of this component is sensor-dependent but application-independent. In the current system, it combines the raw data from the data glove and from the motion capture to generate user pointing positions and combines data from a motion detecting component and the OkaoVision component to detect the exact number of users present.

The contents need to be specifically created for the Dubravka agent:

GSML scenario script. A GSML script describing the anticipated interactions between the agent and the user in the tour guide context must be created specifically for the application. Currently, the script includes a scenario in

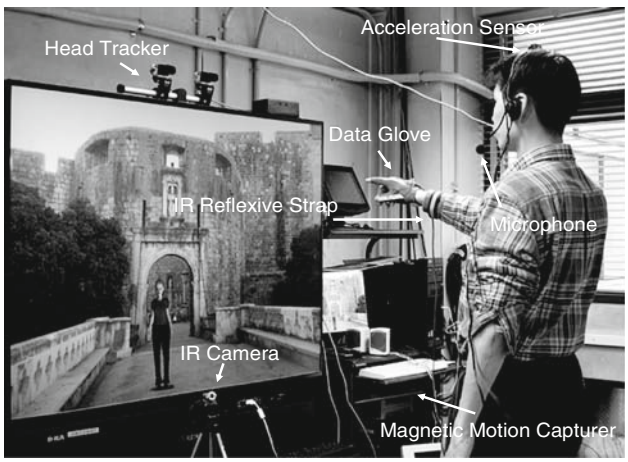

Fig. 3 The hardware configuration of the multi-modal Dubrovnik tour guide agent

three languages (English, Japanese, and Croatian) and possible human–agent interactions in five different scenes: the entrance gate of the Dubrovnik old town, a fountain, a monastery, and two other scenes in Dubrovnik's main street.

Background images and the positions of the agent. The background images and the coordinates where the agent should stand and can walk to need to be prepared for each scene. The appropriate positions, size, and orientation are computed with ARToolkit (Kato 2006).

Croatian voice tracks. Because of the lack of a Croatian TTS, the agent's Croatian speech is recorded from a native speaker's voice.

Speech recognition grammar. Speech recognition in the current system is keyword-based and the grammar for recognizing those keywords needs to be prepared.

Additional character animations. Additional character animations which are not available in the animation catalog need to be prepared.

The components that are limited to use in this tour guide agent:

None. Although some of the system components were developed in the workshop, they can be used in other applications because of their simple and well-divided functionalities.

The data flow among the components is shown in Fig. 4. The cost of building a tour guide agent that is adaptive to three cultures can be kept low. In the current system, all of the components are culture-independent ones. The scenarios of the three cultures are represented in the same script, but each conversational state in GSML is labeled with a language attribute so appropriate TTS and non-verbal behaviors will picked automatically by the scenario executor. The only exceptions are the speech recognition component; one recognition component is required for each different language and only the results that match the currently valid language will be processed. The following subsections introduce the tasks done for incorporating the three cultures into the tour guide agent.

3.3 Non-verbal user inputs

Because advanced gesture recognition techniques have not been introduced, in the non-verbal input recognition part the system is not recognizing culture-specific non-verbal behaviors from the user but only the following general ones at this moment:

- pointing to the interesting objects shown on the display
- showing the wish to ask a question
- interrupting the agent's utterance
- shaking the head and nodding to express negative and positive answers.

These behaviors are recognized by combining the data from the sensor devices. For example, a pointing gesture is recognized by a pointing shape from the data glove and the pointed positions on the display from the coordinate values of motion capture. The movement detection component and face detection component are used to generate the exact number of available users. Because each type of raw data is not meaningful to the central scenario component, the input interpreter component is responsible for generating the combined information, the position where the user is pointing, for processing by the scenario component.

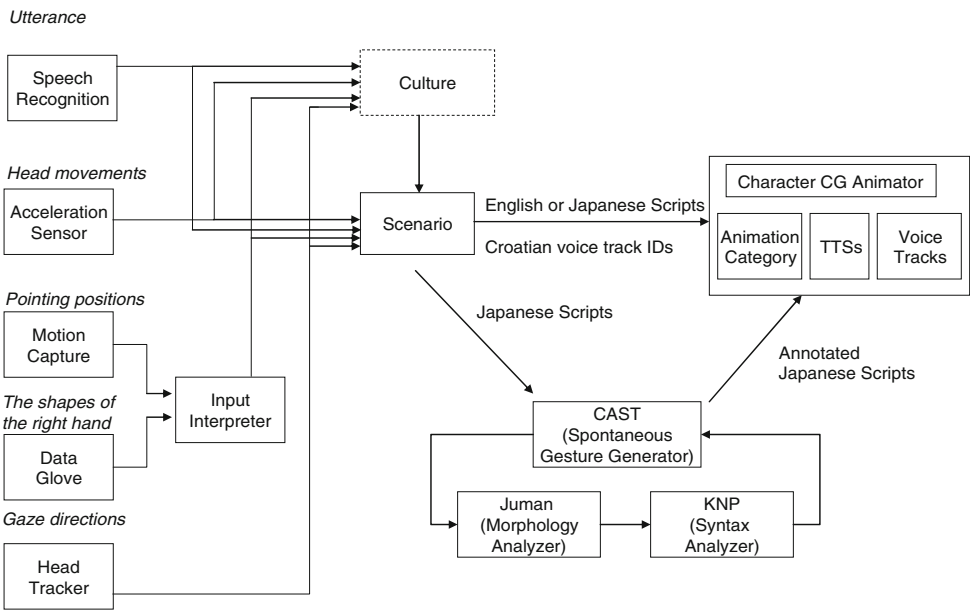

Fig. 4 The data flow and component configuration of the multi-modal tour guide agent. The programs CAST, Juman, and KNP communicate with each other in their original protocols

3.4 Character animations

Some of the gesture animations are created by programming routines that generate joint parameters during runtime. Since we did not have a tool to translate real human gestures into the set of animation parameters in the CG character player, we had to create animations manually. This was a rather time-consuming approach; it took about 5–30 experiments to adjust the parameters for one action, depending on the complexity of the action. Although routine-generated gesture animations have the disadvantage of a relatively unnatural look, they have the advantage that the duration of the animation can be determined at run-time. Locomotion animations have to be implemented by programming. On the other hand, some gestures are modeled in the 3D CG modeling software Autodesk 3ds Max; they look more natural but their duration is fixed. Currently, we have 20 routine-generated gesture animations and 27 animation sequences that are modeled in 3ds Max with fixed lengths. Some of these gestures are shown in Table 1. Since most of the Croatian gestures are also used in many European cultures and in general Western cultures, we did not distinguish them in the current implementation.

3.5 Croatian speech input/output

Although Croatian is spoken by around five million people, the commercial speech and language communities have not yet produced general-purpose recognizers, synthesizers, and translation engines for the Croatian language. This section describes the alternative solutions adopted in the development of Dubravka's Croatian mode.

3.5.1 Croatian speech recognition

In the field of Croatian speech recognition, some research studies have been done, but none of them have produced general-purpose recognizers. Ipsic et al. (2003) and Peic (2003) developed a bilingual database of Slovenian and Croatian weather forecasts. Their recognition results for the two languages are very similar and in the future, they plan to perform bilingual speech recognition system simulation. Nevertheless, a Croatian speech recognition engine is still unavailable to the research community or to industry. Therefore, we decided to configure an English speech recognition software component to recognize Croatian speech by tailoring the recognition grammar. Within the system, classification of the user's utterance is done with limited vocabularies of specific keywords spoken by the user that trigger the scenario component. The pronunciation of Croatian keywords in scenarios is approximated by using the English alphabet. Since some Croatian words in the scenario were impossible to represent in the English alphabet, we had to choose other words instead. If the grammar contained similar words, those words sometimes confused the recognizer, so we were careful to choose words that are not too similar. For example, the pronunciation of the Croatian word "da" (in English: "yes") is approximated in the English alphabet as "ddhaa". Although the speech recognizer works well with the recognition of the word "da" in Croatian, it is often confused by words that contain the syllable "da", like "slobodan" ("free"). We therefore could not choose short words like "da" or "dan" (day) that can appear in longer words, and thus the Croatian scenario is slightly different from the English and Japanese ones. In the end, the following two principles were followed in choosing words to compose the Croatian scenario. The keywords approximated with the English alphabet are not very short and do not contain the syllables of other keywords. Table 2 shows Croatian words used for recognition and the corresponding pronunciations of those words represented in the English alphabet. Because there are only five scenes in the current system, transitions between the scenes and between the states in each scene do not require many keywords from speech input. In the English and Japanese scenarios, we used eight words for transitions and seven of them in Croatian.

3.5.2 Croatian speech output

Since there is still no available Croatian TTS with satisfactory quality, Croatian speech output can only be implemented with a recorded human voice. After the Croatian scenario was composed, a native Croatian speaker's voice was recorded to prepare all the utterances that are supposed to be spoken by Dubravka. The recorded voice tracks are paired with lip animations that are generated automatically by (Zoric and Pandzic 2005). The speech signal is derived from a type of spectral representation of the audio clip and is classified into viseme classes by using neural networks. The visemes are then mapped to MPEG-4 facial animation parameters and are saved as MPEG-4 FBA tracks when the Croatian speech utterances

Table 2 Croatian words and their approximated English alphabets used in speech recognition

No.	Croatian word	Meaning in English	English alphabets
1	bok	hello	bohk
2	grad	city	ghraadh
3	setati	to go for a walk	shetthaati
4	fontana	fountain	fonthaana
5	pitka	drinkable	peethka
6	samostan	monastery	saamostaan
7	super	super	supearh

were being recorded. They are then played by the player with synchronized timings at run-time.

4 Potential extensions

The Dubravka agent built in the eNTERFACE'06 workshop was relatively simple and only addressed the surface issues of multi-culture competent ECA. In this section, we would like to discuss possible extensions to it.

Since multiple users may come together in a real-world application, in the eNTERFACE'08 workshop (Cerekovic et al. 2008), we proposed another project to incorporate basic two-user interaction abilities into Dubravka (Fig. 5). These include behaviors responding to the dynamically changing number of the users currently present, the engagement of the users during the interaction, simultaneous utterances of the users, addressee identification, and gaze direction distribution. It could be a complex but interesting challenge to combine the multi-user and multi-culture tasks. What should the agent do if the users do not belong to the same culture class?

Another possible extension is use for training or pedagogical purposes. Fig. 6 shows another system that we developed for experiencing the differences in gestures between different cultures. There is an avatar that replays the user's hand gestures, such as beckoning, while ten computer-controlled agents react to those gestures differently pretending that they are Japanese or British. The user's actions are captured by a magnetic motion capturing device and interpreted to low-level joint angles to drive the avatar character in real-time. The computer-controlled agents are driven by individual reflexive controlling components and a common BAP catalog component. They are driven by low-level MPEG-4 BAPs in real-time, too. We would like to incorporate this extension into the Dubrovnik tour guide system in the future.

One of the benefits from the modular and distributed design of GECA is that extending the current system to incorporate another culture at the same detail level is straightforward. The developers only need to prepare the speech recognition and TTS engine for that language, additional character animations if required, and the scenario script. In addition, the dashed "Culture" box depicted in Fig. 4 is a potential extension of the current system with a culture module.

In addition to emblem gestures, as suggested in the CUBE-G project (Rehm et al. 2007a, 2008a, b), the cultural class to which the user belongs to can potentially be inferred from the characteristics of the user's non-verbal behavior. The classification criteria can be collected from empirical and statistical results. For example, how frequently the user performs gestures, the strength of the gestures, the distance from the agent chosen by the user, and so on could be informative.

The culture module can then be built to accept the sensor data from the non-verbal input modules, analyze their characteristics, and then classify where the user come from according to a Bayesian network (Rehm et al. 2007b). The results from speech recognizers certainly provide clear evidence of culture. The classification result from the culture component can then be sent to the scenario or deliberation component to affect the characteristics of the agent's behaviors in a parameterized way, for example, done faster or with a larger spatial extent. Solomon et al. (2008) have proposed a language for describing ethnographic data in a pluggable design that could be a candidate of the internal representation of the culture component.

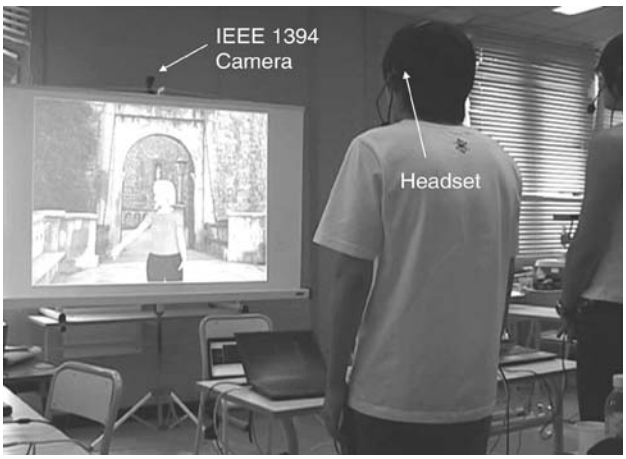

Fig. 5 Dubravka interacting with two visitors in the eNTERFACE'08 workshop

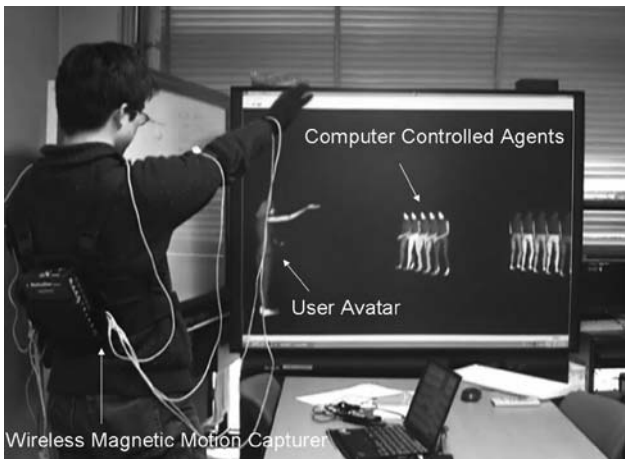

Fig. 6 A culture difference experiencing application with 1 user avatar and 10 computer-controlled agents driven by raw parameters to raw parameters

5 Discussion and conclusion

ECAs are very useful tools for representing cultural differences in training and edutainment applications. In this article, we have presented preliminary results from the development of our culture-adaptive tour guide agent system that is implemented in a modular way with the GECA Framework to minimize the development cost. It can switch its behaviors and speech language to three culture modes: general Western, Japanese, or Croatian. Although both the tour guide agent and GECA itself are still in relatively early stages of development, this very loosely coupled and modular framework can have three possible benefits in handling cultural issues.

- Culture researchers who are not familiar with technical issues can introduce ECA technology more easily because they need only concentrate on culture-dependent issues and implement them as a separate component. The component can then be integrated into a culture-independent skeleton ECA for quick enculturation.
- Collaborative studies with research teams from several countries can separately implement their own culture module more easily.
- Research efforts done in the analysis by synthesis style can be refined incrementally more easily.

This study focuses on the rapid building of ECAs and only features the surface traits of culture, that is, languages, emblem gestures, and probably culture-dependent characteristics of gestures. A more thorough study based on theories of inter-culture communication is necessary in the future. For example, we have noticed that in the case of an interface ECA serving Japanese and Western users, the high-context/low-context differences proposed in (Hall 1992) should cause obvious differences in the behaviors of real humans. Nevertheless, our system models the agent behaviors in a one-to-one mapping sense; the agent always do something in Japanese mode or its counterpart in Croatian mode, even though real Japanese people and Croatian people might make totally different decisions in the same situation.

Finally, by using scripts to describe human–agent interactions, the range of possible interactions will be relatively limited and the quality of the whole system heavily depends on the knowledge and skill of the agent designers. At this moment, we are only showing the feasibility of our modular approach. Obviously, this is not yet a sound solution, but we would like to further develop the deliberative part of the agent with culture modules that affect its outputs with culture-specific differences, and to explore the high level aspects of cultural issues like the use of silence during dialogue, intonation, the choice of words, and so on in the future.

Acknowledgments We thank Kateryna Tarasenko, Vjekoslav Levacic, Goranka Zoric, and Margus Treumuth for their contributions to this project during the eNTERFACE'06 summer workshop, and Takuya Furukawa and Yuji Yamaoka for their contributions to this project during the eNTERFACE'08 summer workshop. We also thank Tsuyoshi Masuda for his contribution in the application for experiencing the cross-cultural differences in gestures.

References

A.L.I.C.E. AI Foundation (2005) Artificial Intelligence Markup Language (AIML). http://www.alicebot.org

Baylor AL, Rosenberg-Kima RB, Plant EA (2006) Interface agents as social models: The impact of appearance on females' attitude toward engineering. In: Conference on human factors in computing systems (CHI'06), Montreal

Cerekovic A, Huang HH, Furukawa T, Yamaoka Y, Pandzic IS, Nishida T, Nakano Y (2008) Implementing a multiparty support in a tour guide system with an embodied conversational agent (ECA). In: The eNTERFACE'08 international workshop on multimodal interfaces. Orsay, France

Costa A, Pickering MJ, Sorace A (2008) Alignment in second language dialogue. Lang Cogn Process 23(4):528–556

de Rosis F, Pelachaud C, Poggi I (2004) Transcultural believability in embodied agents. A matter of consistent adaption. In: Agent culture: human–agent interaction in a multicultural world. Lawrence Erlbaum Associates, London, pp 75–105

Gratch J, Marsella S (2004) A domain-independent framework for modeling emotion. J Cogn Syst Res 5:269–306

Hall ET (1992) Beyond culture. Peter Smith Publisher, Gloucester

Hamiru.aqui (2004) 70 Japanese gestures—no language communication. IBC Publishing, Westminster

Hoya Corp (2008) Pentax VoiceText text-to-speech engine. http://voice.pentax.jp/

Huang HH, Cerekovic A, Tarasenko K, Levacic V, Zoric G, Treumuth M, Pandzic IS, Nakano Y, Nishida T (2006) An agent based multicultural user interface in a customer service application. In: The eNTERFACE'06 international workshop on multimodal interfaces. Dubrovnik, Croatia

Huang HH, Inoue T, Cerekovic A, Nakano Y, Pandzic IS, Nishida T (2007) A quiz game console based on a generic embodied conversational agent framework. In: Seventh international conference on intelligent virtual agents (IVA'07). Paris, France, pp 383–384

Huang HH, Cerekovic A, Nakano Y, Pandzic IS, Nishida T (2008a) The design of a generic framework for integrating ECA components. In: Padgham L, Parkes D, Muller JP (eds) The 7th international conference of autonomous agents and multiagent systems (AAMAS'08), Inesc-Id, Estoril, Portugal, pp 128–135

Huang HH, Cerekovic A, Tarasenko K, Levacic V, Zoric G, Pandzic IS, Nakano Y, Nishida T (2008b) An agent based multicultural tour guide system with nonverbal user interface. Int J Multimodal User Interfaces 1(1):41–48

Huang HH, Furukawa T, Ohashi H, Ohmoto Y, Nishida T (2008c) Toward a virtual quiz agent who interacts with user groups. In: The 7th international workshop on social intelligence design (SID'08). Puerto Rico

Iacobelli F, Cassell J (2007) Ethnic identity and engagement in embodied conversational agents. In: Proceedings of the 7th

international conference on intelligent virtual agents (IVA'07). Springer, Paris, pp 57–63
Intel Corp (2006) Open computer vision library (OpenCV) 1.0. http://sourceforge.net/projects/opencvlibrary/
Ipsic S, Zanert J, Ipsic I (2003) Speech recognition of Croatian and Slovenian weather forecast. In: Proceedings of 4th EURASIP conference, France, pp 637–642
Isbister K (2004) Building bridges through the unspoken: embodied agents to facilitate intercultural communication. In: Agent culture: human–agent interaction in a multicultural world. Lawrence Erlbaum Associates, London, pp 233–244
Johnson WL, Vilhjalmsson H, Marsella S (2005) Serious games for language learning: How much game, how much AI? In: Proceedings of the 12th international conference on artificial intelligence in education. Amsterdam, The Netherlands
Johnston M, Bangalore S (2000) Finite-state multimodal parsing and understanding. In: Proceedings of the 18th conference on computational linguistics. Saarbrucken, Germany
Kato H (2006) Artoolkit. http://artoolkit.sourceforge.net/
Larsson S, Traum DR (2000) Information state and dialogue management in the trindi dialogue move engine toolkit. Natural Language Engineering, Cambridge University Press 6(3–4):323–340
mind.makersorg (2005) OpenAIR protocol specification 1.0. http://www.mindmakers.org/openair/airPage.jsp
Nakano Y, Okamoto M, Kawahara D, Li Q, Nishida T (2004) Converting text into agent animations: assigning gestures to text. In: Proceedings of the human language technology conference (HLT-NAACL'04). ACL Press, Prague
Nass C, Isbister K, Lee EJ (2000) Truth is beauty, researching embodied conversational agents. In: Embodied conversational agents. The MIT Press, Cambridge, pp 374–402
Omron Corp (2008) OKAO vision. http://www.omron.com/rd/coretech/vision/okao.html
Peic R (2003) A speech recognition algorithm based on the features of Croatian language. In: Proceedings of the 4th EURASIP conference. Dubrovnik, Croatia, pp 613–618
Pickering MJ, Garrod S (2004) Toward a mechanistic psychology of dialogue. Behav Brain Sci 27:169–226
Pickering MJ, Garrod S (2006) Alignment as the basis for successful communication. Res Lang Comput 4:203–228
Rehm M, Andre E, Bee N, Endrass B, Wissner M, Nakano Y, Nishida T, Huang HH (2007a) The CUBE-G approach—coaching culture-specific nonverbal behavior by virtual agents. In: The 38th conference of the international simulation and gaming association (ISAGA). Nijmegen, New Zealand
Rehm M, Bee N, Endrass B, Wissner M, Andre E (2007b) Too close for comfort? In: Proceedings of the international workshop on human-centered multimedia, ACM Multimedia
Rehm M, Nakano Y, Andre E, Nishida T (2008a) Culture-specific first meeting encounters between virtual agents. In: Prendinger H, Lester J, Ishizuka M (eds) Proceedings of the 8th international conference on intelligent virtual agents (IVA'08). Tokyo, Japan, pp 223–236
Rehm M, Gruneberg F, Nakano Y, Lipi AA, Yamaoka Y, Huang HH (2008b) Creating a standardized corpus of multimodal interactions for enculturating conversational interfaces. In: Workshop on enculturating conversational interfaces by socio-cultural aspects of communication, 2008 international conference on intelligent user interfaces (IUI2008). Canary Islands, Spain
Solomon S, van Lent M, Core M, Carpenter P, Rosenberg M (2008) A language for modeling cultural norms, biases and stereotypes for human behavior models. In: Proceedings of the 17th conference on behavior representation in modeling and simulation (BRIMS'08)
Thiebaux M, Marshall AN, Marsella S, Kallmann M (2008) Smartbody: Behavior realization for embodied conversational agents. In: The 7th international conference of autonomous agents and multiagent systems (AAMAS'08), Estoril, Portugal
Traum D, Larsson S (2003) The information state approach to dialogue management. In: Smith R, van Kuppevelt J (eds) Current and new directions in discourse and dialogue. Kluwer, Dordrecht, pp 325–353
Traum D, Roque A, Georgiou ALP, Gerten J, Martinovski B, Narayanan S, Robinson S, Vaswani A (2007) Hassan: a virtual human for tactical questioning. In: The 8th SIGdial workshop on discourse and dialogue, Antwerp, Belgium
Visage Technologies AB (2008) VisagelSDK. http://www.visagetechnologies.com
W3C (2004) Emma: extensible multimodal annotation markup language. http://www.w3.org/TR/emma/
Young PA (2008) Integrating culture in the design of ICTS. Br J Educational Technol 39(1):6–17
Zoric G, Pandzic IS (2005) A real-time language independent lip synchronization method using a genetic algorithm. In: the Proceedings of ICME'05, Amsterdam, The Netherlands