Turk-2, a multi-modal chess player

Levente Sajó*,a, Zsófia Ruttkayb, Attila Fazekasa

a University of Debrecen, Egyetem sq. 1, 4030 Debrecen, Hungary
b University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands

Abstract

With the spreading of computers the demand for user-friendly interfaces has increased. The development of techniques based on multi-modal human-computer interaction makes it possible to create systems which are more natural to use. We built Turk-2, a hybrid multi-modal chess-player with a robot arm and a screen-based talking head. Turk-2 can not only play chess, but can see and hear the opponent, can talk to him and display emotions. We were interested to find out if a simple embodiment with human-like communication capabilities enhances the experience of playing chess against a computer. First, we introduce the architecture of Turk-2, we describe the human experiments and its evaluation. The results justify that multi-modal interaction makes game playing more engaging, enjoyable - and even more effective. These findings for a specific game situation provide yet another evidence of the power of human-like interaction in turning computer systems more attractive and easier to use.

Key words: turk-2, multi-modal chess player, human-computer interaction, HCI, MMHCI

1. Introduction

The development of computer industry has step-by-step introduced computers to our everyday life. Computers become faster and cheaper, which leads to the spreading of computer-based information systems in different
areas, such as general purpose personal computers used in people’s homes but also machines which are built for a specific task, e.g. ATMs at banks or interactive schedules and ticket machines at train stations. Though systems have become increasingly efficient and easier to use, on the part of users some kind of disfavour can be observed (Forrester, 2008; Karlonia, 2007; Wight, 2008). There can be various reasons for this resistance:

- Lack of knowledge for using new technologies.
- As every system should be used differently, on a daily basis one has to learn several different protocols to be able to use the systems. Moreover, the time is getting shorter between the releases of new generation systems.
- Several new systems and services are used internationally, in a multicultural community, but they often reflect, in their design, the culture and communication habits of the developer.

Albeit the use of computer systems is spreading, they pose a big challenge for the user who should learn the “language” of each system and figure out how to interact with it. Hence it has become a hot research topic to make computer-human interaction more natural. A straightforward way to achieve this is to exploit multiple modalities - speech, gesture, facial expressions - common in every day communication for human-computer interaction, too. Since in this paper we will only talk about the communication between humans and computers, the terms ”Multi-modal Human-Computer Interaction” (MMHCI) and ”Multi-modal Interaction” (MMI) will be used as synonyms.

1.1. Multi-modal Human-Computer Interaction

It is difficult to define exactly what makes HCI human-like. A major component is using input and output channels which are more natural for users.

GUIs can be considered as a first step in this direction, introducing visual output in addition to or even, instead of, textual output, including the use of icons referring to every-day concepts (dustbin, letter-box). Multi-modal user interfaces represent the next big step in the direction of natural communication (Ilmonen, 2006; Raisamo, 1999). They allow using multiple, natural modalities, as well as providing input by the user.
The term modality refers to human senses or input/output channels (Raisamo, 1999). Silbernagel (1979) listed six different senses and corresponding modalities: sight (visual), hearing (auditive), touch (tactile), smell (olfactory), taste (gustatory), balance (vestibular). Raisamo (1999) divided human senses into seven groups from a neurobiological point of view: internal chemical, external chemical (taste, smell), somatic senses (touch, pressure, temperature, pain), muscle senses (stretch, tension, joint position), sense of balance, hearing and vision. The goal of multi-modal human-computer interaction is the synergetic use of several (and ultimately, all) of the sensing and expressing abilities that humans have (Jaimes and Sebe, 2005).

In our vision future information systems will be using several input and output modalities in parallel, to interact with the user in a more natural and efficient way. Users can communicate using natural language and gestures, and systems can understand the conversation based not only on the meaning of the utterances, but also the - often additional, but sometimes, contradictory - meaning expressed by the posture, a blush or eye gaze, or by a nervous finger ticking. Ideally, systems should also be able to reply in a similarly rich and natural way, using speech, gestures and facial expressions.

1.2. Related Works

We are still far from the above depicted situation, and the ultimate richness and complexity of human-human communication makes it unrealistic to expect computers passing a kind of multi-modal communication Turing test. However, there is a spectacular increase in research efforts as well as applications paving the road towards human-like multi-modal communication with computers. It is essential to understand the role of the single channels of human-human communication first, and develop technologies for making them available for human-computer interaction. Besides the traditional fields of speech recognition and synthesis, the exploitation of human gestures has become a hot topic. First, Gestural User Interfaces were developed at the beginning of the 1970’s. Myron Krueger’s Videoplaces experiments are influential in this field (Krueger, 1991). Videoplaces used projectors, video cameras and special purpose hardware to create silhouettes of participants and to analyze their movements. The computer responded to the gestures of the participants by interpreting their actions. The participants could “touch” each other’s video-generated silhouettes, or manipulate different graphical objects generated by the system.
One of the first of a series of gesture-based commercial sport applications was Konami’s "Dance Dance Revolution” released in 1998. Nowadays, gestural input is used in several sport, game and training simulation applications, as well as in novel entertainment installations. For instance, in computer aided Tai Chi (Kunze et al., 2006) captured human motion was analyzed and used as the basis for the repertoire of the virtual Tai Chi performer.

Another example for using body gestures in martial arts games is Kick Ass Kung-Fu reported by Hämäläinen et al. It is a virtual reality martial arts game, which uses a combination of computer vision techniques (background subtraction and optical flow computation) to perceive movements made by the user. In this case hand and leg gestural communication was used to control the human player enabling one to practice martial arts against a virtual opponent (Hämäläinen et al., 2005).

Facial expressions have been used both on the input and the output side. The recognition of a range of facial expressions - often including the ‘6 basic emotional expressions’ identified by Ekman (Ekman, 1982) - is highly developed, producing a good recognition rate, near real-time in near-natural situations, as of lighting conditions, head posture and characteristics of the face (skin color, artifacts) (Littlewort et al., 2002). Expressive talking heads with visual speech - that is, lip sync, facial and head signals accompanying speech - can be found on web pages, chat forums or in SecondLife\(^1\) and commercial services\(^2\)\(^3\)

Besides the study and exploitation of single-modalities of human-human communication, the other major effort is interdisciplinary, focusing on the fused use of different modalities as expressive means for both the user and the computer. After decades of work to develop computational models for the fused use of language and gesture (Bolt, 1998a,b), nowadays, we can use a wide range of interesting applications of the fusion of input modalities. Hämäläinen and Höysniemi (2002) presented a perceptual user interface for controlling a flying cartoon-animated dragon in QuiQui’s Giant Bounce. This is an interactive computer game for 4 to 9 years-old children. The dragon flies according to the user’s movements and breathes fire when the user shouts.

Mitsubishi Electric Research Laboratories presented a digital multiplayer

\(^{1}\)http://secondlife.com
\(^{2}\)http://www.virtualsmartagent.com
\(^{3}\)http://193.108.42.79/ikea-nl/cgi-bin/ikea-nl.cgi
tabletop game using gesture and speech input. They redesigned two commercial single player games (Warcraft III - a command and control strategy game; and The Sims - a simulation game) and implemented multiplayer multi-modal wrappers atop them. Taken together, gestures and speech coupled with gaze awareness, provides a good controlling of these games. For example, units can be directed by voice commands (e.g., move), and spatial gestures (e.g., pointing to a location) disambiguate their meaning (Tse at al., 2006).

For using multiple human-like modalities to generate output by the computer, it is inevitable to have an embodied "virtual human" communicating with the user, often entirely hiding the underlying traditional application (like an information system). A virtual human - also referred to as virtual agent or embodied conversational agent - is a human-like 2d or 3d graphical model residing on the screen, and endowed with multi-modal communicational characteristics of real people. The domain of applications range from tutoring and training via (serious) games to entertainment. Closest to our chess-player application are virtual agents in the role of an opponent of a real human in some mental game, like poker (Kim et al., 2004; Rist et al., 2008; Rehm and André, 2008; Becker et al., 2005).

In the virtual human endowed applications, and particularly, in applications where the VH is a game player, the question arises how engaging and believable the virtual human is, and if the virtual opponent is treated in a similar way as a real opponent. André and Rehm (2005) give convincing evidence that a virtual game partner is taken seriously, and as for verbal and particularly, gaze behaviour, treated similarly as a real player.

Schröder et al. (2008) have presented an interactive poker game in which one human user plays against two animated agents. The application combines several modalities like facial gesture and body animation, emotion modeling and speech synthesis for driving the behavior of virtual characters and this way enhancing the naturalness of interaction in the game.

Robots are not virtual, but physically real humanoids. These - often "emotional" - robots are able to move their body and their hands and they can display human emotions on their faces. Though their embodiment and thus, movement, particularly, their locomotion capabilities are different from those of VHs, their research and development have many topics in common. Focusing on strategic game playing, Leite and Pereira (2008) presents iCat, a cat-like robot developed by Philips, as the opponent of a human player in chess with emotive behavior that is influenced by the state of the game.
Finally, a few words about the origin of our Turk-2. It has some historical inspiration: The Turk or Automaton Chess Player was a chess-playing machine constructed in the late 18th century by Farkas Kempeleln. The mechanism appeared to be able to play a strong game of chess against a human opponent. It was only 50 years later when its secret was finally unveiled that a small sized and skilled human sitting inside the machine was operating the Turk. In spite of the fact that is turned out to be a hoax, it attracted people, both as opponent players and as engineers. In our computer age, the Amazon Mechanical Turk was launched in 2005. It is a web-based software application to coordinate programming tasks with human intelligence, inspired in part by the way Kempeleln’s Turk operated. Ledermann outlined a project where the Turk was realized in augmented reality, as a projected figure sitting and acting opposite the real player (Modern Turk, 2009). REEM-A is an all-round high-technology robot who can also play chess well (REEM-A, 2009).

In the light of the above applications, we set out to develop a chess playing robot which is mechanically simple, but by the addition of a talking head on a computer screen, is capable to communicate by gaze and exhibit facial expressions. Our major motivation was to research how people react to such a "hybrid" humanoid in a game-playing situation.

1.3. Human-Computer Interaction and Turk-2

In general, the basic structure of a system endowed with multi-modal human-computer interaction can be demonstrated by Figure 1.

On the left side we have an input interface which contains the input receiver module with the input devices. The inputs from the different devices – on the picture only audio and video inputs are shown, but we can imagine other types of "sense-organs" as input – are processed in relationship with
each other in the input receiver module. The information from the input interface is processed by a core application, which forwards the result to the output interface. The output interface contains two modules, the output generator, which transforms the results to the proper form for the output devices. In this case, besides audio and video devices any other output devices can be added, too.

Our aim was to build a chess playing system with multi-modal interfaces attached to it, and study if the human-like interface provides an added value. As input, Turk-2 monitors the user’s face in order to detect his/her emotional state, listens to the user to interpret his/her speech (of a limited vocabulary), and monitors the chess board to detect moves by the player. The output of the system is conveyed by a talking head, which is capable of directing her gaze, exhibit emotional facial expressions and perform speech with lip-sync as output. The Turk-2 performs chess moves by his single robot arm. (See Figure 2.)

We chose chess as a kernel game application because it is a popular and
wildly known game and therefore suitable to be tested with users. We wanted to put our effort into the part of the system which communicates with the users and not the artificial intelligence part. We implemented the system and performed user studies with two versions, one with multi-modal communication module, including face analyzer, speech recognizer and talking head (referred to in the following as TH) and without it (NoTH). By comparing reactions, we intended to get an insight into the role of multi-modal interfaces.

2. Turk-2 architecture

In this chapter we present Turk-2, our multi-modal chess player system. The structure of Turk-2 matches the general model described above for multi-modal interaction (see Figure 3).

The central part of the system is the controller which consists of three modules: the turn manager, which is responsible for the game flow and orchestrates the communication between the user and the chess engine, the emotion monitor, which keeps track of the emotional state of the user and the Turk-2, and the chess engine which acts as the ”mind” of the Turk-2, deciding on the chess moves. The various components can be grouped into two main modules.

The first module (marked with dots) is the multi-modal communication module, providing multi-modal input and output facilities. This module
has two kinds of input, one from the speech recognizer and one from the player’s face analyzer, which localizes faces and detects facial expressions on the human player’s face. The inputs from these components are processed by the turn manager, which generates the output and forwards it to the talking head.

The second module (marked with dashes) is the chess game module. This module contains the chess state detector which detects the changes on the chess-board, and if there is a move, sends it to the chess-engine. The chess-engine generates the next move, which is executed by the robot-arm on the board.

In the following, the various components of the system will be presented. We will go into details of the modules related to the MMC module.

2.1. The controller

The main function of the controller is to coordinate the various components and establish communication among them. It consists of three main parts: the chess engine, the emotion monitor and the turn manager.

- The **chess engine** is responsible for determining the answers to the human player’s moves. This engine informs the controller about the state of the game and about the prospects related to the outcome of the game. In this system, we used Phalanx as the chess engine, but any other xBooard compatible chess engine could be used \(^4\).

- The facial emotional changes detected by the player’s face analyzer module is received by the **emotion monitor**. There are three emotional states: sad, neutral, happy. These are inverted and forwarded to the talking head by the emotion monitor.

- The **turn manager** is responsible for the game flow. In our case, the game flow is taken care of by a deterministic state transition algorithm (see Figure 4). The following types of turns are maintained:
  - *Introduction.* At the start of the game, the face detector recognizes if someone is sitting in front of the camera, being ready for the game. After that, the talking head greets the human player and introduces himself. Then the controller sets up the chess engine.

\(^4\)http://www.tim-mann.org/engines.html
Figure 4: The deterministic flow of the turn manager. The ellipses represent the input (on the left) and output (on the right) modules. These are respectively: chess-board detector (CBD), speech recognizer (SR), talking head (TH), chess engine (CE), robot arm (RA). The rectangles represent the game phases managed by the turn manager. The arrows show the communication between the turn manager and the different modules.

- **Game playing.** Once, the game is started, the turn manager repeats the following four phases in a cycle. At the beginning of every phase, the turn manager first informs the talking head about the changes. Then in each stage further actions take place:

1. **The player is thinking.** It withdraws the robot arm and switches the chess state detector to active status. Waits until it receives a "hand" message from the chess state detector, which means, the player has started to make a move.

2. **The player is making a move.** If the human player's move has occurred, the information about the step is sent to the chess engine. If the chess engine validates the move then it continues with the next step. Otherwise the talking head informs the player about the illegal move and goes back to "the player is thinking" step. If checkmate occurred, it continues with the
3. *Turk-2 is thinking.* For the chess engine, calculating the next move takes approximately 2 seconds. Because we wanted to make the computer chess player more human-like, in this phase, the generation time of the next move has been extended with a random thinking time. The duration of the computer’s thinking is adapted to the actual game state. The probability of longer thinking time is higher when the computer is in a difficult situation. This is an important state from the point of view of human-computer interaction, because during the time while the computer is thinking, the talking head is supposed to engage the human players. After the turn manager has received the answer move and some information about the actual game status (prospect) from the chess engine and the thinking time has elapsed, the data about the move is sent to the robot arm’s controller which starts to execute the move.

4. *Turk-2 is making a move.* During this phase, the talking head is following the robot arm with his gaze. The robot arm notifies the turn manager when the move has been completed. Then the turn manager checks the actual game state. If check has occurred, the talking head announces it to the player in words and the game continues with *"the player is thinking"* phase. In the case of checkmate, the talking head reports it and the turn manager goes to the *"closing"* part.

- **Closing.** These four steps described in the *game part* are repeated until one of the players gives checkmate to the other. At the end of the game the player is asked whether he/she wants to start a new game, or to finish it.

### 2.2. Player’s face analyzer

Face has an important role in human-human communication, too. A lot of information can be gathered from the face; it is possible to determine the age or the gender, or determine the mood of a person (Fazekas and Sánta, 2004, 2005). It is also possible to make video based speech recognition (Gordan et al., 2002). In the case of Turk-2, this module is used for localizing faces and recognizing the various facial expressions on them.
The goal of face detection is to determine whether an image does contain any faces or not and if it does, returns the location and extension of each face. The various conditions (pose, illumination, occlusion etc.) make face detection a challenging task, but several techniques have been proposed in recent years (Pantic and Rothkrantz, 2002; Yang et al., 2002; Hjelmas and Low, 2001; King, 2003). The most successful ones were appearance based techniques. A “template” of the whole face is used to search the image. This face model is learnt from a set of positive and negative examples of training images. Finding faces is done by scanning this face model over a target image. At multiple scales with different window sizes (using an image pyramid) sub-regions are extracted from the original image. These sub-regions are then classified as face or non-face. Therefore face detection is reduced to a binary pattern classification problem.

Many different classification techniques (neural networks, SVM, etc.) have demonstrated excellent results. We adopted the method which was first presented by Viola & Jones in 2001 (Viola and Jones, 2001), having similar consistency and accuracy as the others, but over-performs the previous methods in speed approximately 10 times, being able to detect faces on video streams at 15 frame/sec. This method is also implemented in OpenCv.

Facial expression detection should deal with three basic problems: face detection in a facial image or image sequence, data extraction and classification (Pantic and Rothkrantz, 2002). We built our system based on ideas presented in (Littlewort et al., 2002). The expression recognizer receives image patches located by the face detector. For data extraction, these face patches are converted into a Gabor representation using a bank of 40 Gabor filters, in 8 orientations and 5 spatial frequencies. For pairwise classification Support Vector Machines are used.

Usually, face detection systems considers six different facial emotions plus the neutral one. These emotion groups were first introduced by Ekman and they are the following: anger, disgust, fear, joy, sadness, surprise (Ekman, 1982). Since, in chess games not all of these emotions occur, in the case of Turk-2, we decided for the recognition of three basic emotional states: neutral, happy and sad. By default, we considered a face to be in the neutral emotional state, thus only two separate SVM classifiers needed to be trained for classification between the classes happy vs. not happy and sad vs. not sad.

The face analyzer module processes the video taken by the camcoder placed above the chess board (indicated by number 2 in Figure 2). It runs...
in an infinite loop. For every grabbed frame, first it finds the location of the face. Then it determines the actual emotional state. If it has changed since the previous state, it sends the new emotional state to the emotion manager.

2.3. Speech recognition

This module makes Turk-2 understand some of the utterances of the user. In the case of a chess game, speech does not have a central role but there can occur some situations when the services of this module can be useful. To implement the speech recognition component, we used a library called Hidden Markov Model Toolkit. It has been trained for a limited vocabulary and has been used in two cases: at the start of the game when the player chose the opponent with a difficulty level and at the end, when the player was asked whether he/she wanted to play again.

2.4. Talking head

The face of Turk-2 is a talking head shown on a computer display. The talking head is an animated head, which appears on the screen. Our talking head was made with the CharToon (Han Noot and Ruttkay, 2000) 2D facial animation engine, using a face with features defined in terms of the MPEG4 standard (Pandzic and Forchheimer, 2003). Due to the modular architecture, it would be possible to link some other head designs which ”understand” MPEG4 commands for facial expressions and use viseme repertoire to generate lip sync. The spoken utterance signal is generated by a text to speech engine. This engine, the ProfiVox system, was developed at Budapest University of Technology and Economics (Olaszy et al., 2000).

The two output modules are controlled by the cognitive module. This module, corresponding to the inner states, determines the behaviour of the talking head. The inner states are generated from the combination of the basic ”personality” of the talking head with external parameters, which are received from the emotion monitor and turn manager.

The inner (cognitive) state of Turk2 is characterised by a 4-dimensional vector, depending on: the turn state, the game state(prospect), the emotional state and the time elapsed since the start of the game. The combination of these states influences the appearance of the talking head in four different ways: facial expression, gaze, idle movements, speech. Figure 5 shows how the inner states affect the appearance of the talking head.

---

5http://htk.eng.cam.ac.uk
2.4.1. Facial expressions

The talking head is able to express four discrete emotional states. These are: neutral, sad, happy, bored (Figure 6). These emotions are displayed on both the upper part of the head by the eyes and the shape of his eyebrows, and the lower part by the shape of his mouth. During speech, emotions are expressed only on the upper part of the head.

The facial expression model of the talking head can be described as a state machine in which the states are the four emotions. In every phase the talking head starts with one of the neutral/happy/sad states which can later switch to bored. Figure 7 shows the transition function between the states. The input parameters of the transition function are: the emotion of the human player (0 is sad, 1 is neutral and 2 is happy), the state of the chess game from the view of the virtual player (0 means the worst state, while 2 is a winning situation), and time parameters. The time parameters are: the current time, including the time passed by in the actual phase of the game, the emotion holding time, including the minimum duration of an emotion shown in the
Figure 7: The emotion function.

<table>
<thead>
<tr>
<th>Phase of the game</th>
<th>Emotion holding time</th>
<th>Bored expression limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human player is thinking</td>
<td>5 s</td>
<td>20 s</td>
</tr>
<tr>
<td>Human player is making a move</td>
<td>4 s</td>
<td>10 s</td>
</tr>
<tr>
<td>Turk-2 is thinking on</td>
<td>10 s</td>
<td>-1 s</td>
</tr>
<tr>
<td>Turk-2 is making a move</td>
<td>10 s</td>
<td>-1 s</td>
</tr>
</tbody>
</table>

Table 1: Emotional settings

face and the bored expression limit, containing the minimal duration before the bored emotion appears on the face. -1 means that the talking head never gets the bored emotion in this phase. These values are chosen based on our own empirical observations during calibration. (See: Table 2.4.1.)

2.4.2. Gaze

The gazing direction of the talking head also depends on the phase of the game. (See Table 2.) The four directions we distinguish:

- looking at the chessboard,
- looking at the opponent,
- following the movement of the robot arm,
- looking around randomly.
Table 2: Turk-2’s gaze direction.

<table>
<thead>
<tr>
<th>Phase of the game</th>
<th>chessboard</th>
<th>opponent</th>
<th>robot arm</th>
<th>around randomly</th>
</tr>
</thead>
<tbody>
<tr>
<td>the player is thinking</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>the player is making a move</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>the computer is thinking</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>the computer is making a move</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0</td>
</tr>
</tbody>
</table>

For all the phases we set the main gazing direction and a numerical value between 0 and 1, which means the probability of looking in that direction. These values are based on own observation. During a phase, the gaze directions are associated with random time durations. A new gaze direction is determined according to the parameters of the phase, but independently from the previous gaze direction.

2.4.3. Idle movements

For life-like behaviour, the talking head should simulate natural motions on real faces like blinking, random eye and mouth movements. These movements are made in the idle time, and in times when emotional changes are to be displayed.

The most frequent facial motion during a chess game is blinking. According to statistical experiments, humans blink in every fifth second. The blinking interval of our talking head is set to five seconds with 1 second deviation generated by the Perlin noise function. Random mouth motions are rare. Random eye movements can only happen, when the probability of the gaze direction is not set to 1.

2.4.4. Speech

Chess players do not usually comment on the game and do not speak to each other. However, to increase engagement, the talking head was prepared to talk in two cases:

- **Commenting while waiting for the opponent.** If the opponent spends too much time without making its move, or the chess engine spends too much time to determine the appropriate step.

- **Announcing its own move.** It happens when the robot arm starts to make the move.
<table>
<thead>
<tr>
<th>Phase of game</th>
<th>Lower time limit</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human player thinking on</td>
<td>20</td>
<td>When do you want to make your move?</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>Do you want to give it up?</td>
</tr>
<tr>
<td>Computer is thinking on</td>
<td>30</td>
<td>This is a tight spot!</td>
</tr>
</tbody>
</table>

Table 3: Examples from the speech repertoire

Moreover, depending on the state and the phase of the game, the talking head can address the opponent, precipitating, or reflecting on its tight spot. In the different phases there are different sentences to say. For each sentence a time parameter is given, that means the lower bound for the time spend in the current phase, before the utterance. The sentences to say are chosen in a random way from the alternatives appropriate for the given situation, (see Table 3).

2.5. Chess-state detector

The chess-state detector processes the video from the first camcoder in Figure 2. It tracks the chess board and detects the changes on it. To minimize the number of preprocessing steps, a stationary physical environment was built. The chessboard and the robot-arm were fixed to a table, and the camcoder was placed above the center of the chessboard. This way we had a top view of the game scene. Because the chess-pieces cannot be easily distinguished from the top, the detector only focuses on detecting the placement and the color of the pieces and does not detect the types. The state of the game is stored by the controller. As the controller constantly receives the computer player’s move generated by the chess engine, it knows the current state of the game. So, it is enough for the chess-state detector to identify the moves made by the human player only. It switches between active and idle modes depending on whether it is the human player’s or the computer’s turn.

In active mode, the chess-state detector performs multiple steps. At the beginning receives a reference game state from the controller and grabs an initial reference frame. For the following frames, a correlation value is calculated for the reference frame. If this value is greater than a predefined threshold then it is interpreted by the detector that there is a hand over the chessboard and the human player is making a move. The detector informs the turn manager about this event and continues to track the board. When
the correlation value decreases to the normal level, it means that the player has completed the move. The next step is to calculate the new game state. For this, the position and the colors of the pieces should be determined. For position detection a difference image of the actual frame and reference frame is used. Color identification is made in HSV color space. After we have calculated the new game state the difference between this and the reference state is analyzed to validate and determine the player’s move. In the final step the calculated move is sent to the controller and the chess-state detector switches to idle mode. The threshold value is chosen empirically in the calibration step.

2.6. The robot-arm

The robot-arm, which was built specially for this system, can be seen in Figure 2. Because of technical and financial reasons, for the robot-arm only electrical and mechanical pieces were used. For positioning the arm, electrical motors and wires have been used. This resulted in a noisier and less precise machine than a hydraulic one, but still usable for us. It is able to position within 1 cm accuracy; an average move takes between 5-10 seconds.

The robot-arm and the chess board were fixed to the table and after that a calibration is done, when the physical positions of the fields and the holding height of the chess-pieces are set for the system. The robot-arm is connected to the system through a robot-controller. For positioning the robot-arm, a controller software is used. The software provides a high level interface for the robot-arm. It is enough to tell where from and where to and which chess piece we would like to move.

3. Experimental study

In recent years, results of academic researches provided examples in different fields of MMHCI which made the comparison of these systems very difficult. Due to the complexity of the problems there do not exist common evaluation and design guidelines for MMHCI systems. (Ruttkay and Pelachaud, 2005) collects the different methodologies for evaluation, from non-perceptional evaluation (“a measured difference” between real and synthesized behaviour) to user-centered evaluation.

In this study, we have performed an empirical evaluation with the aim of finding out whether adding a multi-modal interface to a computer chess game results in better game experience and performance. The Turk-2 chess
player, described in the previous chapter, as well as a version without the Talking Head were used to make tests with humans. We analyzed the game plays in the two situations in terms of subjective and objective measures, and used those to make conclusions. The tests have been evaluated to confirm our observations.

3.1. The scenario

16 people, between the age of 18 and 25 years, participated in the test, 8 males and 8 females. The subjects were everyday computer users. They did not know anything about the Turk-2 system, they were only told that they need to play chess against the computer. The chess playing skills of the subjects varied from novice level (knowledge of the basic rules only) to the level of an average chess player. Advanced chess playing skills were not required, because we wanted to test the system from the point of view of game playing experience, as opposed to an exercise.

Before starting a session, each subject received instructions from an assistant, who also described the process of the experiment and set up the test enviroment. The experimenter remained in the room to take notes occasionally, but he did not further interfere throughout the whole duration of the games. Each test consisted of two chess games, one with a simplified version of Turk-2 - without the multi-modal communication part (NoTH), while the other with the complete system (TH). The order of the games was decided randomly at the beginning. During each game the subject was video recorded and a report was generated from the important (unexpected) events by the assistant. In 6 sessions, the video of the talking head was video recorded in parallel, too. These videos were processed later. Also, at the end of the test the subject was asked to fill in a questionnaire about their first impressions of the experience.

After the tests, each video was annotated with event and timestamp pairs. In the case of videos of humans, the possible events are the following: the player is looking at the chessboard; the player is looking away from the game scene; thinking; the player is making a move; the player is looking at the talking head; the computer is making a move; the player is looking at the robot arm; the player is makeing some facial gesture: smiling, laughing, surprised, sad; the player is mumbling, saying something, talking to the talking head. An example of the annotation file can be seen in the Appendix.

In the case of videos of the talking head, the possible events annotated were: talking, neutral, happy, sad, bored.
The questionnaire consisted of ten yes/no questions regarding the subjects’ impressions about Turk-2 system. The answers to these questions were to be given on a scale of 1 to 10, where 1 represents “not at all” and 10 means “definitely”. The questions are categorized into three main topics: which of the two games did the subjects enjoy more, what is their opinion about the human-likeness of the system provided by the multi-modal techniques and did they consider the talking head as part of the game. The participants were also encouraged to describe their impressions in their own words at the end of the questionnaire. The form is available in the Appendix.

3.2. Evaluation of the experiments

The experiment results were evaluated in both an objective and a subjective manner.

The objective evaluation is based on the data extracted from the annotation files and the questionnaires. The data were then submitted to statistical analysis, using the SPSS v17.0 software. First, the normality of the statistical variables were checked (they varied from p=0.4 to p=0.9), then Student’s t-test was applied. In some cases bivariate correlations and Friedman tests for several related samples were applied.

In the case of the videos recording from the players’ face, the total duration of the events were calculated and these summarized values were used in the statistical analysis. The talking head video and video of the subject pairs were analyzed, in order to determine the effects of the multi-modal components on the subjects.

The operator’s notes and our personal impressions about the videos were the basis of a lot of interesting observation, too. In the following the research questions and their answers are discussed from five different point of views:

3.2.1. How did the players react to the talking head?

Since chess is a turn based game, human players acted differently, depending on whether it was their turn or not. When it was the human player’s turn, most of the time they were looking at the chessboard and thinking on their next moves. During the computer’s turn, there was nothing to engross the players. Usually, they were looking outside of the game scenery. If the robot arm started to move, they followed it with their gaze. The presence of the talking head was important to fill the emptiness while the computer was ”thinking”. Applying t-test, it can be proved that average total durations of looking at the chessboard was statistically equal in both test cases (p=0.822).
Figure 8: The distribution of looking directions in percentage of the whole duration of the game: with TH when the player is on turn, with TH when the computer is on turn, with noTH when the player is on turn, with noTH when the computer is on turn.

Furthermore, in the case of TH, looking away and looking at the robot arm was less (p=0.006 and p=0.002), and instead the subjects were looking at the talking head. Figure 8 shows the distribution of average looking directions in both cases (TH and noTH).

Applying bivariate correlation, it can be statistically shown that eye contact with the talking head happened when the players were in a passive state – the computer was on turn – (p=0.01, corr=0.836). In most of the cases after finishing their move, to change turns, the players established eye contact with the talking head. Many times they also glanced at the talking head while they were following the movement of the robot-arm. Figure 9 shows the distribution of looking at the talking head in the four game phases.

Since the interaction between the opponents during a chess game is different from an ordinary conversation, the general characteristics of communication is hard to recognize. In a face-to-face conversation it can often be observed that if a smile appears on one of the subjects’ face, the other starts to smile automatically. This was different in the case of Turk-2. Various statistical tests have showed that the talking head’s smile does not induce the human player’s smile: chi-square test (p=0.398), runs test (p=0.164), conditional probability (p=0.281). Even if there is no connection between the talking head’s and the player’s utterances it can be showed that the players were smiling, laughing and talking more in the games with TH (p=0.001) which is illustrated in Figure 10. Furthermore, the rate of these interections
in the games with noTH and in the games with TH, but without looking at the talking head was the same (p=0.794). However, the presence of the talking head made the players perform these utterances in an increased amount while they were also looking at the talking head.

3.2.2. Is the Talking Head treated as a person?

The talking head’s presence influenced the players’ reactions during the game. They treated Turk-2 as a person. Usually human players reacted to the talking head’s manifestations with smile and laughter. Many times the subjects replied to the talking head’s questions and sentences. For example, when the computer finished its move and said ”Check this out” the answer was ”I’m not happy with this”. Sometimes they praised the computer after a good move: ”Clever!” or expressed annoyance ”You are glad to hit, aren’t you?”. They also made different comments about the talking head’s face: ”You are so sly!”, ”Evil!”, ”Is he actually grinning at me!?”, or japed: ”You are so ugly”, ”He has sleepy dust in his eyes”. Some of the subjects were amusing themselves with repeating the talking head’s sentences: ”Take your time”, ”Tough situation”.

By applying Friedman test to the questionarie, it can be statistically proved that the answers connected to the human-likeness of Turk-2 are related to engagement and that the subjects considered the talking head a human-like partner in the game (p=0.06). This shows us the importance of developing more human-like multi-modal interfaces.
Figure 10: Average laughing, smiling and speaking time in percentage of the whole duration of the game. In games with TH the users are laughing, smiling and speaking more than in games with noTH. This happens mainly when they are looking at talking head at the same time.

Figure 11: The expressions appeared on the player’s face during the game: laughing (a), wicked (b) and angry (c).

3.2.3. Effect of the talking head on game experience

Overall, it can be assumed that the subjects enjoyed more playing with TH, more as one can see from the questionnaire, where 95% confirmed that they enjoyed playing with TH more. They spoke highly of the talking head, saying that the talking head made the game more interesting and entertaining. Their concentration during the game also confirms the positive effects of the talking head. Usually subjects played better with TH, 80% of the games were longer, it happened only once that somebody got checkmate in fewer turns when playing with TH. The higher engagement of the subjects was also confirmed, by analyzing the average thinking time in the two test cases. Using t-test it was proved that the subjects were thinking longer in
the games with TH (p=0.017).

The importance of MMI could be particularly well observed in the tests when subjects played with TH first and then with noTH. In these cases, in the second game, the absence of the talking head was more conspicuous and the players expressed their regret: "The first game with the talking head was much better" or "I’m missing the talking head".

The speed of the robot-arm was found to be too slow in some cases, also it usually took the computer a few seconds to generate the next move and repace to execute it. During this time the subjects were confused, they could hardly wait until it finished. In the games with TH, it was easier for the players to realize that it was their turn, from the talking head changing his gaze from the chessboard to the player’s direction, or in some cases an actual verbal warning.

They liked it when the talking head announced his next move or if there was a hit, check or checkmate. In the cases when the talking head did not announce his next move or his hit, usually the participants noticed it: "Oh, he forgot to tell his next step!".

3.2.4. Gender difference

Analyzing the test results for males and females separately, a few differences can be observed. Overall, it can be assumed that males were more interested in playing chess, and winning against the computer. Females enjoyed playing with TH more, they had more interactions with the talking head and gave higher ratings in the questionnaire. By applying t-test, it is
statistically proven that females smiled, laughed and talked more than males (p=0.01) despite the fact that there was no difference in the average thinking time between the two genders (p=0.161).

3.2.5. Further observations

To the question, whether the talking head should talk more and be more emotional or it should be silent with a "poker face", about 70% of the subjects declared that it should be more emotional, because "playing with a silent and poker-faced talking head is like playing alone". Our talking head, we prepared for this system, was somewhere in the middle between the two extremes. Usually, the subjects’ opinion was that it is better to play with a more talkative partner having a large and diverse vocabulary.

Interpretation of identical expressions: the same facial expression of the talking head was interpreted differently by the players. At the beginning of the game the smile of the talking head was only an "innocent" smile, but by the end of the game, when the computer was close to win, the same smile was considered to be "malicious".

4. Conclusion

In this paper we present Turk-2, a hybrid with a robot arm and a human-like face displayed on a screen. Turk-2 is able to play chess but also implements various multi-modal channels to communicate with the opponent. We let people play with Turk-2, and also with a faceless and silent variant of it. In the expriment we studied the effect of multi-modal interfaces in the human-computer interaction. Despite the relatively small number of participants in these tests, it can be said that Turk-2 received positive feedbacks. The results suggest that a system which implements MMI solutions appears to be more human-like and provides better game experience. Giving a face to the computer evokes more emotional expressions from the human players – in some cases they even treated Turk-2 as a person by talking to it and arguing with it.

In the future we will perform more dynamic tests with a large number of subjects in a similar game setting. For this, we are going to use another game, called rock-paper-scissors. It will be played by a virtual human (or talking head) using different modalities. Having a larger number of tests statistical measures could give more precise results. But even these results show the positive effects of multi-modal interfaces.
5. Acknowledgement

We would like to thank Sándor Baran from the University of Debrecen for his help in the statistical analysis of the tests.
A. Annotation file

00:00:00 00:00:06:14 The player is making a move.

00:00:06:14 00:00:20:09 Looking at the talking head. Smiling.

00:00:20:09 00:00:41:03 The computer is making a move. Looking at the robot arm.

00:00:41:03 00:00:45:00 Looking at the talking head. Laughing.

00:00:45:00 00:00:53:12 Looking at the chessboard. Thinking.

00:00:53:12 00:00:59:14 The player is making a move.

00:00:59:14 00:01:03:09 Looking at the talking head.

00:01:03:09 00:01:10:07 Looking at the chessboard. Thinking.

00:01:10:07 00:01:13:00 Looking away.

00:01:13:00 00:01:38:13 Looking at the talking head. Smiling. The computer is making a move.

00:01:38:13 00:01:47:05 Looking at the chessboard. Thinking.

00:01:47:05 00:01:50:08 The player is making a move.

00:01:50:08 00:01:52:11 Laughing.

00:01:52:11 00:01:59:12 Looking at the chessboard.

... 00:09:33:04 00:09:44:09 Looking at the chessboard. Thinking.

00:09:44:09 00:09:50:05 The player is making a move. Smiling.

00:09:50:05 00:10:03:11 Saying something. Looking away. Looking at the talking head.

00:10:03:11 00:10:04:06 Looking at the talking head.

00:10:04:06 00:10:10:11 Looking at the chessboard. Saying something.

00:10:10:11 00:10:18:12 Looking at the chessboard. Thinking.

00:10:18:12 00:10:29:10 Looking at the robot arm. Glances out.

00:10:29:10 00:10:34:14 Looking away. Saying something.

00:10:34:14 00:10:42:05 Looking at the chessboard. Saying something.

00:10:42:05 00:10:51:02 END OF THE GAME. CHECKMATE.
B. Questionnaire

1. Do you fully agree that you enjoyed playing with the talking head more than without?
   10 - 9 - 8 - 7 - 6 - 5 - 4 - 3 - 2 - 1

2. Do you fully agree that the talking head is similar to a human player?
   10 - 9 - 8 - 7 - 6 - 5 - 4 - 3 - 2 - 1

3. Do you fully agree that the presence of the talking head was annoying during the game?
   10 - 9 - 8 - 7 - 6 - 5 - 4 - 3 - 2 - 1

4. Do you fully agree that the talking head enjoyed it when you were in a difficult situation?
   10 - 9 - 8 - 7 - 6 - 5 - 4 - 3 - 2 - 1

5. Do you fully agree that the sentences of the talking head were easy to understand?
   10 - 9 - 8 - 7 - 6 - 5 - 4 - 3 - 2 - 1

6. Do you fully agree that the face of the talking head is human-like?
   10 - 9 - 8 - 7 - 6 - 5 - 4 - 3 - 2 - 1

7. Do you fully agree that the face of the talking head expressed his feelings quite well?
   10 - 9 - 8 - 7 - 6 - 5 - 4 - 3 - 2 - 1

8. Do you fully agree that the talking head is more human-like than the computer chess player?
   10 - 9 - 8 - 7 - 6 - 5 - 4 - 3 - 2 - 1

9. Do you fully agree that the sentences of the talking head made the game more exciting?
   10 - 9 - 8 - 7 - 6 - 5 - 4 - 3 - 2 - 1

10. Do you fully agree that many times you felt that the talking head was watching you?
    10 - 9 - 8 - 7 - 6 - 5 - 4 - 3 - 2 - 1

11. What is your opinion about a computer chess player?

12. Do you have some advice for the developers?

13. What do you think about the future of Turk-2?

14. Do you have some other comments connected to the system?

15. Should the talking head be more emotional and talkative or should it be more "poker faced"?
16. Should it even have a face?
References


