Debreceni Egyetem
Informatika Kar
Komputergrafika és Képfeldolgozás Tanszék

MULTIMODÁLIS EMBER-GÉP KAPCSOLATOK

Dr. habil. Fazekas Attila
egyetemi docens

Váradi Péter Zsolt
PTM hallgató

Debrecen
2008

Tartalomjegyzék

1. Bevezetés 3
2. Multimodális ember-gép kapcsolatok 5
2.1. Az ember-gép kapcsolat definíciója 6
2.2. Modalitások 8
2.2.1. Emberközpontú megközelítés 10
2.2.2. Rendszerközpontú megközelítés 11
2.3. A multimodális kapcsolatok mítoszai/tévhitei 12
3. Multi-modális rendszerek 19
3.1. Put – That – There 19
3.1.1. A rendszer képességei 19
3.2. Intelligens tájékoztató- és hirdetőtábla 20
3.2.1. Valós idejű tekintetkövetés 21
3.2.2. Multi-kulcsszavas felismerő 22
3.2.3. Alkalmazás 23
3.3. Multimodális Pool oktató 23
3.3.1. Target Pool 24
3.3.2. Az interfész 25
3.3.3. A rendszer felépítése 25
3.3.4. Virtuális oktató 26
3.3.5. Humántesztek a rendszerrel 27
3.4. Multimodális póker 28
3.4.1. Vezérlés 30
3.4.2. Viselkedés modell 30
3.4.3. Érzelemkifejező szintetikus hangok 32
4. A multimodális sakkozóval végzett humán teszt 33
4.1. A rendszer felépítése 34
4.1.1. A vezérlő 35
4.1.2. Arci érzelmet felismerő modul 36
4.1.3. Beszédfelismerő 37
4.1.4. Beszélő fej 38
4.1.5. Sakkállás felismerő 39

4.1.6. Robotkar 40
5. A multimodális kommunikáció hatásának vizsgálata 41
5.1. Kísérleti összeállítás 41
5.2. Felmerült problémák 42
5.3. A teszt eredményeinek kiértékelése 43
5.3.1. Hatással van a beszélő fej a játékosokra? 44
5.3.2. Személyként kezelték a beszélő fejet? 46
5.3.3. A beszélő fej hatása a játékélményre 47
5.4.4. További megfigyelések 47
6. Összefoglalás 49
Irodalomjegyzék 51
Függelékek 57

Rövidítések

ACM SIGCHI – ACM's Special Interest Group on Computer-Human Interaction

ALMA - A Layered Model of Affect

APT - Automated Pool Trainer

CeBIT - Centrum der Büro- und Informationstechnik

CWI - Centrum Wiskunde & Informatica

FPS - Frame per Sec

GIF - Graphics Interchange Format

GUI – Graphical User Interface

HCI – Human-Computer Interaction

HD - High Definition

iGBBS - intelligent Guiding Bulletin Board System

LCD - Liquid Crystal Display

MARY TTS - Modular Architecture for Research on speech Synthesis Text-to-Speech System

MIT - Massachusetts Institute of Technology

PAD - Pleasure-Arousal-Dominance

PC - Personal Computer

PDA – Personal Digital Assistant

PK - Primary Key

RFID - Radio-Frequency IDentification

SAPI - Speech Application Programming Interface

SDMS - Spatial Data Management System

SK - Secondary Key

UI – User Interface

WIMP - Window, Icon, Menu, Pointing device

SVM - Support Vector Machines

Előszó

A számítógépek, az automatizált rendszerek és a mesterséges intelligencián alapuló technikák manapság az életünk minden területén megtalálhatóak. Nagyon rövid idő alatt olyannyira központi szereplőjévé váltak mindennapjainknak, hogy lassan már el sem tudjuk képzelni az életünket a sokféle hasznos és okos eszköz nélkül. Mikor megszülettem, az emberek java része azt sem tudta, hogy mi az a mobiltelefon, autója is csak a tehetősebbeknek volt. Eltelt 25 év, és talán már nincs is olyan lakosa a fejlett vagy fejlődő országoknak, aki ne használna minden nap számítógépet, bankkártyát vagy mobiltelefont. A járművek többsége beépített fedélzeti számítógépekkel rendelkezik, egyre többen választják a papír térképek helyett a GPS navigációs eszközöket.

A multimodális rendszerek témaköre mindig is őszinte kíváncsisággal és csodálattal töltött el. Nem is olyan rég még csak a tudományos fantasztikus filmekben láthattunk hangvezérelt gépeket vagy személyiséggel rendelkező robotokat. Manapság ez az informatika fejlődésének egyik vezérfonala, és egyre komolyabb eredményekről számolnak be a tudósok. A tradicionális egér-és-billentyűzet interfészek kora lassan leáldozóban van. Amilyen rohamos léptekkel fejlődnek az egyes elektronikai eszközök, úgy kell lépést tartani azok kezelésével és használatával is. Egyre terjednek az új generációs ember-gép kommunikációs felületek, melyek egyaránt könnyítik az oktatást, a navigálást és nem utolsósorban a szórakozási lehetőségeinkbe is új színt visznek.

Mivel a Debreceni Egyetem Informatika karán is megépült egy multimodális kísérleti eszköz, úgy döntöttem, hogy ezzel a témával kapcsolatban szeretném megírni a szakdolgozatomat. Lehetőségem volt ezt a rendszert alaposan megismerni, a dolgozatom egyik alaptémája pedig az ezzel a rendszerrel végzett humánkísérletek ismertetése lett. A szkeptikusok úgy gondolhatják, hogy az emberek úgysem fogják kihasználni a multimodalitás által kínált előnyöket, de az előbb említett kísérletek is bebizonyították, hogy igenis komoly létjogosultsága van az ilyen rendszereknek. Az emberek nyitottak az újdonságokra, és értékelik, ha az adott rendszer a kényelmüket szolgálja.

Ezúton szeretném megköszönni témavezetőmnek, dr. Fazekas Attilának végtelen türelmét és segítőkészségét. Az ő útmutatása és irányítása nélkül bizonyára elvesztem volna a szakterülethez tartozó szakirodalom sűrűjében. Köszönetet szeretnék még mondani Sajó

Levente PhD. hallgatónak a humántesztek előkészítéséért és a tesztek lefolytatása során nyújtott segítségéért és hasznos tanácsaiért.

Ezt a szakdolgozatot Szüleimnek ajánlom. Az ő támogatásuk és bizalmuk nélkül soha nem jutottam volna el ideáig. Remélem, hogy egyszer képes leszek viszonozni azt a megannyi áldozatot, amit értem hoztak. Köszönöm, hogy végig kitartottatok mellettem.

A szerző

1. Bevezetés

Az elmúlt húsz évben az informatika, mint önálló tudományág sokat fejlődött. Ahogy az informatika fejlődött, úgy lett az ember-gép kapcsolatok is (Human-Computer Interaction, HCI), sok más témakörrel együtt, annak egyre fontosabb része. Az elmúlt 10-15 évben ez a terület aktív kutatás tárgya lett, amelyet például az ACM Special Interest Group on Computer-Human Interaction [48], a British Human-Computer Interaction Group [49] vagy a The European Association for Cognitive Ergonomics [57] csoportok művelnek.

Az ember-gép kapcsolatok tudományága - ahogy azt a neve is mutatja - a felhasználók és a számítógépek közötti kapcsolatok természetét kutatja. Gyakran tekintenek úgy erre a területre, mint az informatika, a viselkedéstan, a tervezés és még néhány más tudományág közös vizsgálati területe. Az emberek és számítógépek közötti kapcsolat színtere a felhasználói felület (User Interface, UI), mely egyaránt magába foglalja a szoftvert és a hardvert is.

Az ember-gép kapcsolatok egyszerre vizsgálja az embert és a számítógépet, ezért egyaránt támaszkodik az emberi és a számítógépi ismeretek területeire is. Gépi oldalról a legfontosabb területek a számítógépes grafika, a mesterséges intelligencia, az operációs rendszerek, a programozási nyelvek és fejlesztői környezetek. Emberi oldalról a kommunikáció, nyelvészet, társadalomtudományok, kognitív pszichológia és viselkedéstan van elsősorban hatással. Az ember-gép kapcsolatok multidiszciplináris mivolta miatt nagyon számos terület kutatója kell, hogy hozzájáruljon a kutatáshoz, ami nagy feladatot jelent az együttműködés megvalósításában is.

A felhasználói felületek jövője a multimodalitásban rejlik. A számítógépek életünk szerves részévé váltak. Az asztali számítógépeken túl állandóan használjuk mindennapjaink során a bankautomatákat, mobiltelefonokat, PDA-kat, információs paneleket, intelligens háztartási eszközöket, stb. is. A repülőgépek, hajók és vonatok egyre nagyobb része számítógép vezérelt, a modern autók alapfelszereltségéhez is hozzátartozik már a fedélzeti számítógép. Az ember-gép kapcsolatok tudományterületének elsődleges célja olyan multimodális számítógépes rendszerek és interfészek kialakítása, melyek egyszerűbben kezelhetőek, használatuk szaktudás nélkül is könnyen elsajátítható, ezzel megkönnyítve az emberek munkáját a mindennapokban

A Debreceni Egyetemen dr. Fazekas Attila egyetemi docens vezetésével kialakításra került egy multimodális rendszer. Ez rendszer egy olyan sakkozó gép, mely modalitásokként a beszédet, a vizuális megjelenítést, az arci gesztusokat és használja.

A szakdolgozatom első fele az ember-gép kapcsolatok, mint kutatási terület bemutatását hivatott szolgálni. Ezt követően egy, a sakkozó géppel lefolytatott humánkísérletek kiértékeléséről fog szólni, melynek egyetlen célja azt alátámasztani, hogy a multimodális rendszerek jobbak és emberközelibbek, mint a tradicionális egér-billentyűzet interfészen alapuló rendszerek, ezáltal az emberek is sokkal hatékonyabban használják.

A dolgozat szerkezete a következő szerint épül fel:

A második fejezetben az alapvető definíciók és irodalmi eredmények kerülnek bemutatásra. Sharon Oviatt [31] kutatásai során összegyűjtött néhány mítosza is megtalálható ebben a részben.

A harmadik fejezetben konkrét rendszereket fogok bemutatni, melyek megépítésre és tesztelésre kerültek.

A negyedik fejezetben a Debreceni Egyetem sakkozó gépének felépítését ismertetem részletesen, majd az elvégzett humánteszt kerül bemutatásra.

Az ötödik fejezetben található meg a kapott eredmények kiértékelése, és azok értelmezése.

Az utolsó fejezetben végül egy rövid összefoglalás kap helyet.

2. Multimodális ember-gép kapcsolatok

Ebben a fejezetben áttekintjük a multimodális ember-gép kapcsolatok kutatási területének irodalmát, a fontosabb fogalmakat, definíciókat. Az emberek egymás között tudatosan és tudat alatt sokféle csatornán kommunikálnak egymással. Gondolhatunk itt például a testbeszédre, a különféle gesztusokra, hanghordozásra, stílusra, érintésekre vagy a tekintetre csak hogy néhányat említsünk. A multimodális ember-gép kapcsolatok egyik legfontosabb célja olyan ember-közeli rendszerek, eszközök kifejlesztése, amelyek képesek felismerni az emberi viselkedést, érzelmeket kifejezni, azaz természetes módon tudnak kommunikálni, együttműködni a felhasználókkal.

A felhasználók dolgának megkönnyítésére tett egyik első és korszakalkotó lépés a grafikus felhasználói felületek (Graphical User Interface, GUI) megjelenése volt a ’80-as években. Különböző ikonok, képek jelképezték a számítógépen tárolt adatokat, ezek manipulálására pedig megjelent az első igazán könnyen használható fizikai eszköz is, az egér. Mivel ezeket a felületeket sokkal könnyebb volt használni, mint pusztán szöveges beviteli módban kezelni a gépet, ezért nagyban hozzájárultak a személyi számítógépek gyors terjedéséhez. De akármennyire is volt ez mérföldkő a számítógépek történetében, az ember-gép kapcsolatok szempontjából még mindig csak egy részét használják ki az ember érzékszerveinek. Buxton 1986-ban [1] megfogalmazott egy jövőképet, miszerint a jövő antropológusa milyennek képzelné el az emberi fajt, ha találna egy tökéletesen működőképes állapotban lévő asztali számítógépet az összes korabeli operációs rendszerrel, szoftverrel és kommunikációs interfésszel együtt:

„Azt hiszem leginkább olyannak írná le az emberi fajt, aminek túlfejlett szemei, hosszú jobb karja, csökevényes bal karja, egyforma hosszú ujjai és „low-fi” fülei lennének. A meghatározó jellemzők mindenképp a kiváló vizuális szervek és alulfejlett testi képességek lennének.” (Buxton, 1986)

2.1. Az ember-gép kapcsolat definíciója

Jelenleg nincs egységesen, mindenki által elfogadott pontos definíció az ember-gép kapcsolatok szakterületére. Az egyik legszélesebb körben elfogadott megfogalmazás az ACM SIGCHI-tól származik:

„Az ember-gép kapcsolatok tudományterülete az emberi felhasználásra szánt interaktív számítógép rendszerek tervezésével, kiértékelésével és implementációjával foglalkozik, valamint az ezeket kísérő jelenségeket kutatja."

Roope Raisamo szerint ahhoz, hogy az emberek számára használhatóbb, kényelmesebb rendszereket készíthessünk alapvetően három fogalom megértése szükséges:

- *Felhasználó:* aki a rendszerrel kommunikál,
- *Rendszer:* a technológia és használhatósága,
- *Interakció:* a felhasználó és a rendszer közötti kapcsolat.

Ezek alapján egyértelmű, hogy a multimodális ember-gép kapcsolatok tudománya interdiszciplináris tudomány. Egy ilyen interaktív rendszer megtervezőjének sok témakörben kell jártasnak lennie: a pszichológiában és a kognitív tudományokban, hogy megértse a felhasználó észlelési, kognitív és problémamegoldó képességeit, a szociológiában, hogy általánosságban értse az interakciókat, az ergonómiában, hogy ismerje a felhasználók fizikai képességeit, a grafikus tervezésben, hogy hatékony megjelenítést tudjon létrehozni, az informatikában és a mérnöki tudományokban, hogy fel tudja építeni/programozni a megfelelő technológiákat, stb. [33].

Az ember-gép kommunikáció szempontjából a *bemenet* fogalma nagyon lényeges szerepet játszik. A gyakorlatban az interakciók java része multimodális módon megy végbe, de pontosan tudni kell, hogy mi a különbség az emberi cselekedet és a rendszer által tényleges inputként értelmezett információ között egy-egy interakció során. Például, amikor gombokat nyomunk le a billentyűzeten, az alapvetően az érintés modalitását használja, de vannak emberek, akik a vizuális modalitást is igénybe veszik, mert ránéznek a billentyűzetre vagy a

monitorra, hogy ellenőrízzék, helyesen gépeltek-e [33]. Egy általános asztali számítógép azonban ezekből semmit sem észlel, csupán a begépelt parancsot értelmezi.

Különbséget kell tenni *parancs* és *nem-parancs* interfészek között is. Az előbbi olyan akciókat jelent, melyekkel explicit módon utasításokat hajtatunk végre, pl. menü megnyitása, gombra kattintás, stb. Az utóbbi csoportba az olyan cselekedetek, események tartoznak, amelyekkel indirekt módon lehet hangolni a rendszert a felhasználó igényeihez [33].

Kiváló példa erre az emberek hangulatának kifejezése. Az ember-ember közötti kommunikációban az érzelmek mindig fontos szerepet játszanak. Mivel az érzelmeket gyakran multimodális módon fejezzük ki, ezért ez a multimodális ember-gép kommunikáció tudományának is nagyon fontos része. Az olyan rendszereket, amelyek fel tudják ismerni az emberi érzelmeket, állapotokat (pl. stressz, unalom, figyelmetlenség, harag, nevetés, stb.) és képesek ennek megfelelően reagálni és alkalmazkodni az adott felhasználóhoz, sokkal természetesebbnek, megbízhatóbbnak, hatékonyabbnak fogják látni az emberek [33].

Az ACM SIGCHI szerint informatikai szempontból a hangsúly az interakción van, különösképp az egy vagy több ember és egy vagy több számítógép közötti interakción. Alapesetben az az egyszerű szituáció jelenik meg lelki szemeink előtt, ahogy egy felhasználó ül egy asztali számítógép előtt, és egeret, billentyűzetet és monitort használ a számítógép kezelésére. Azonban ha egy kicsit szabadjára engedjük a képzeletünket, egy nagyon gazdag és kiaknázásra váró interakciók halmazával találjuk szemben magunkat.

Az ember-gép kapcsolatokat mi elsősorban az informatikai szemszögéből fogjuk megvizsgálni majd. Ez azonban nem jelenti azt, hogy nem kellene ugyanolyan figyelmet szentelni ezeknek a tudományoknak.

Az ember-gép kapcsolatok tudományágának további jellemzéseként az ACM SGHCI összeállított egy listát azokról a területekről, melyekkel ez a témakör foglalkozik:

- Az ember és gép együttműködésének hatékonysága.
- Az ember és gép közötti kommunikáció struktúrája.
- Mennyire képes használni az ember a gépet (mennyire tanulható az interfész)?
- Algoritmusok és interfészek programozása.
- Interfészek tervezése és építése.
- Interfészek specifikálásának, tervezésének és implementációjának folyamata.
- Megvalósíthatóság.

Ebből is látszik, hogy az ember-gép kapcsolatok vizsgálata egyaránt merít az elméleti-, mérnöki- és tervezési tudományokból is.

Newell, Perlis és Simon 1967-ben megfogalmazott klasszikus definíciója szerint az informatika a „számítógépek és az azokat körülvevő jelenségek tanulmánya”. Eszerint a számítógépek és az emberek közötti interakciók is szerves részét képezik az informatikának.

Egy másik meghatározás szerint [7] az informatika „az információ leírásának és transzformálásának algoritmikus folyamatainak szisztematikus tanulmánya: elmélet, analízis, tervezés, hatékonyság, implementáció és alkalmazás.”. Ezek az algoritmikus folyamatok egyértelműen magukba foglalják a felhasználókkal való interakciókat is. A mai modern számítógépes alkalmazások tervezésének elkerülhetetlen része egy olyan modul megtervezése, amely a felhasználókkal hivatott kommunikálni. Az ACM SGHCI szerint a valóságban általában a teljes forráskód több mint felét az ehhez a modulhoz kapcsolódó sorok teszik ki.

2.2. Modalitások

A multimodális kapcsolat a tradicionális billentyűzet-egér szinten túl további lehetőségeket biztosít a felhasználó számára az adott rendszer eléréséhez. A leggyakoribb ilyen multimodális kapcsolat a vizuális modalitás (képfelismerés, mint input, és animáció, mint output) és az audio modalitás (beszédfelismerés, mint input, és beszédszintézis, mint output) együttese. Ezen felül természetesen más modalitások is alkalmazhatóak, mint például az

érintés (érintőképernyők, tollak). A multimodális felhasználói felületek az ember-gép kapcsolatok kutatási területéhez tartoznak.

A többféle modalitás használata megnöveli a felismerés rátáját, mivel az egyik modalitás gyengeségét a többi modalitás kompenzálni tudja. Egy mobiltelefonon, amely kis képernyővel és gombokkal rendelkezik, azt a szót, hogy „fantáziadús" elég nehézkes beírni, de nagyon könnyű kimondani. Vagy gondoljunk bele milyen körülményes egy ilyen eszközön átböngészni egy online könyvkatalógust vagy moziműsort. Másik valós példa egy műtét is lehetne, ahol az orvosoknak azonnal meg kell tudni jeleníteni az információt, és nincs idő billentyűzni vagy az egérrel kattintgatni. Ezeket a problémákat sokkal könnyebb lenne hangvezérlés használatával megoldani.

A multimodális felhasználói felületek könnyebben kezelhetőbbek. Egy jól megtervezett multimodális rendszert az emberek sokkal szélesebb köre tudja használni, akár szaktudás nélkül, akár fogyatékkal. A gyengén látók számára nagyon előnyös egy beszédfelismerővel és beszédszintézissel kiegészített rendszer, míg a halláskárosultak inkább a vizuális megjelenítésre koncentrálnak.

Maybury és Wahlster [21] szerint a multimodális ember-gép kapcsolatok előnyei a következők:

- *Hatékonyság:* Minden modalitást arra lehet használni, amire leginkább való.
- *Redundancia:* A kommunikáció nagyobb eséllyel folyik zavartalanul, mivel ugyanarról a dologról több csatornán keresztül több referencia érkezik.
- *Felismerhetőség:* Térbeli kontextus esetén a felismerhetőség növekszik.
- *Természetesség:* Abból fakad, hogy a modalitások használata szabadon választható, és ez olyan ember-gép kommunikációt eredményezhet, amely közel áll az ember-ember kommunikációhoz.
- *Pontosság:* Növekszik, amikor egy másik modalitás pontosabban képes megjelölni egy objektumot, mint a fő modalitás.
- *Szinergia:* Az egyik kommunikációs csatornán érkező információ finomíthatja, módosíthatja, kijavíthatja egy másik csatornán érkező információ tartalmát.

Az ember-gép interakciók esetében modalitáson a következők általános osztályát értjük:

- Érzék, mellyel az ember a számítógép kimenetét észleli
- Egy szenzor vagy eszköz, mellyel a számítógép az ember általi inputot tudja fogadni

Kevésbé formálisan megfogalmazva, a modalitás az ember és a számítógép közötti kommunikációs csatornát jelenti [58].

Roope Raisamo szerint a multimodális interakciókat kétféle szemszögből lehet megközelíteni. Az első a pszichológiában gyökerezik, és az ember oldaláról közelíti meg a témát: észlelés és irányítás. Ebben az esetben a modalitás szó az emberi input és output csatornákra vonatkozik. A második nézet a számítógépek oldaláról közelít, és kettő vagy több számítógépes input vagy output modalitás alapján épít fel olyan rendszereket, melyek az ezeken a csatornákon keresztül érkező információkat együttesen, egymást kiegészítve képesek felhasználni. A következő két szakaszban ezt a két megközelítést tárgyaljuk részletesebben [33].

2.2.1. Emberközpontú megközelítés

A multimodális ember-gép kapcsolatok emberközpontú megközelítése esetén a hangsúly az emberi multimodális észlelésen és irányításon van, vagyis az emberi be- és kimeneti csatornákon. Az észlelés az a folyamat, aminek során az érzékszervi információkat magasabb szinten reprezentálja [36]. A kommunikációs csatorna az érzékszervekből, idegpályákból, az agyból és izmokból épül fel [4].

Ebben a megfogalmazásban a modalitás az emberi érzékekkel van szoros kapcsolatban. Silbernagel 1979-ben az 1. táblázatban bemutatott módon gyűjtötte össze a különféle érzékeket és a hozzájuk tartozó modalitásokat. Ez a táblázat az észlelés egy valamelyest egyszerűsített formáját tartalmazza. Például az érintés érzetét a mélyebb szövetekben és a bőrben található idegvégződésekkel egyaránt érezzük, de ennyire részletesen itt most nem kerül bemutatásra.

Érzet	Érzékszerv	Modalitás
Látás	Szemek	Vizuális
Hallás	Fülek	Auditív
Érintés	Bőr	Tapintás
Szaglás	Orr	Szaglószervi
Ízlelés	Nyelv	Ízlelő
Egyensúly	Egyensúlyi szerv	Vestibuláris

1. táblázat. A különböző érzetek és modalitások [41].

A modalitásokat neurobiológiai szempontból csoportosítva ([14],[39]) hét csoportra oszthatjuk őket:

- Belső kémiai (véroxigén, cukor, pH)
- Külső kémiai (ízlelés, szaglás)
- Fizikai érzetek (érintés, nyomás, hőmérséklet, fájdalom)
- Izomérzetek (nyújtás, feszülés, ízületi pozíció)
- Egyensúlyérzet
- Hallás
- Látás

Mivel a belső kémiai folyamatokat elég nehezen alkalmazhatjuk felhasználó interfészekhez, ezért ettől eltekinthetünk. A többi neurobiológiai modalitás azonban mind megtalálható az 1. táblázatban, kivéve az izomérzeteket (kinesztézia). A kinesztézia fontos szerepet játszik az egyes testrészek pozíciójának meghatározásában.

2.2.2. Rendszerközpontú megközelítés

Az informatika már sokféleképpen definiálta a multimodális felhasználói felületeket. A multimodális kapcsolatra vonatkozó definíciókat Chatty [5] foglalta össze, arra alapozva,

hogy a legtöbb szerző akkoriban (1994 környékén) úgy gondolt a multimodális rendszerekre, mint többféle input eszközzel rendelkező (multi-szenzoros kapcsolat), vagy egy eszközön keresztül érkező input többféle interpretálására alkalmas rendszerekre.

A multimodális kapcsolatok Chatty-féle megfogalmazásával a legtöbb informatikus egyet ért. A multimodális felhasználó felület fogalma alatt olyan rendszert értenek, mely sokféle bemenetet képes elfogadni, melyeket értelmesen kombinál össze. A számítógépes kimenetek már jó ideje sokkal gyorsabban fejlődnek, mint azok az eszközök, amelyekkel irányítani tudjuk a rendszereket. Ha egy pillanatra összehasonlítjuk a napjainkban elterjedt multimédiás kimeneti eszközöket a jelenlegi bemeneti eszközökkel, azonnal látszik, hogy nincs egyensúlyban a két csoport [5]. A kimeneti eszközök (HD Monitorok, Projektorok, 8.1-es hangfalak, LCD és Plazma technológia, robotika) mérföldekkel kifinomultabbak, mint a szokásos bemeneti eszközök (billentyűzet, egér).

Nigay és Coutaz [26] a következőképp fogalmazza meg a multimodalitást:

„A multimodalitás a rendszer azon kapacitását mutatja, ahogy különböző kommunikációs csatornákon képes kommunikálni a felhasználókkal és automatikusan képes a tartalmat kinyerni és értelmezni”

A multimédia és a multimodális rendszerek egyaránt többféle kommunikációs csatornát használnak. Nigay és Coutaz azonban különbséget tesz a kétféle rendszer szerint annak alapján, hogy a multimodális rendszernek képesnek kell lennie automatikusan lemodellezni az információ tartalmát magasabb absztrakciós szinten. Attól még, hogy egy levelezőrendszer képes hang- és videoanyagokat is továbbítani, még nem multimodális, mivel csak egyik postaládától a másikig küldi az információt, de nem értelmezi őket.

2.3. A multimodális kapcsolatok mítoszai/tévhitei

Sharon Oviatt interaktív intelligens térképeivel kapcsolatos humántesztjei során foglalta össze a most következő mítoszokat, tévhiteket a multimodális rendszerekkel kapcsolatban. Az interaktív térképeket beszédparancsokkal, fénytollal, vagy a kettő együttes használatával lehetett irányítani. Az elvégzett humántesztek során az alanyoknak különböző feladatokat

kellett megoldani ezzel az interaktív dinamikus térképrendszerrel. Például helymeghatározásokat, távolságszámításokat. A tesztek során az alanyok előszeretettel választották a multimodális megközelítést, és ösztönszerűen a leghatékonyabb módon irányították a rendszert.

A multimodális interfészekkel és multimodális interakcióval szemben egyre nagyobb elvárásokat támasztanak. Ezek az elvárások gyakran olyan tévhitekhez vezetnek, amelyeknek vajmi kevés köze van az „empirikus valósághoz" (Oviatt, 2002). Oviatt 1999-ben összefoglalta ezeket a tévhiteket, és empirikus tényekkel cáfolta őket. Álljon most itt a lista a multimodális interakciókkal kapcsolatos tévhitekről [31]:

1. mítosz. A multimodális rendszerekkel az emberek multimodális módon kommunikálnak.

Oviatt egy 1997-es tanulmánya szerint a felhasználók 95-100%-a multimodális módon kommunikált, ha szabadon használhattak tollat vagy vokális parancsokat az interaktív térkép környezetben. Ez azonban nem jelenti azt, hogy ha felhasználók minden egyes esetben multimodális módon fogják utasítani a rendszert. A felhasználók szabadon variálták az unimodális és multimodális interakciót attól függően, hogy mik voltak az adott feladat követelményei. Egy tanulmány szerint a felhasználók parancsainak 20%-a volt multimodális, a többi vagy pusztán beszéd- vagy írásos utasítás volt [30]. A multimodális utasításokat leggyakrabban térrel kapcsolatos feladatokban használták (pl. két objektum közötti távolság kiszámításakor, vagy térbeli objektumok helyének, méretének, számának megnevezésekor).

Ha a feladat nem volt igazán térközpontú (térkép nyomtatása), a felhasználók általában unimodálisan viselkedtek. Ez alapján Oviatt azt mondta, hogy a felhasználók szeretnek multimodális módon kommunikálni a gépekkel, de nem minden esetben. A jövő multimodális rendszereinek képesnek kell lennie különbséget tenni, hogy a felhasználók mikor kommunikálnak multimodálisan és mikor nem, így a párhuzamos input folyamokat ennek megfelelően interpretálják egymást kiegészítő vagy egymástól független módon.

2. mítosz. A legdominánsabb multimodális terület a beszéddel egybekötött mutogatás.

Bolt „Put-That-There" rendszere óta a beszéd-és-mutogatás módszerét tekintik a szakemberek a multimodális rendszerek prototípusának. Emiatt a legtöbb multimodális felhasználói felület ezt igyekszik megvalósítani, kiváltképp a deixisek feloldása végett (pl.: „azt" vagy „vele", olyan kifejezések, amelyek feloldásához referenciára van szükség). Ugyanakkor ez a kombináció akár a tradicionális egérnek az „új implementációjaként" is felfogható. A beszéd-és-mutogatás csupán 14%-át teszi ki a multimodális interakcióknak [30]. Oviatt említést tesz McNeill 1992-es kutatási eredményeiről is, miszerint az emberek közötti személyes kommunikációnak kevesebb, mint 20%-a mutogatás [23]. Ezek alapján azok a rendszerek, melyek csupán a beszéd-és-mutogatás elvére épülnek, a felhasználók számára nem fognak igazán sokkal többet nyújtani a használhatóság szempontjából, mint a tradicionális egér-billentyűzet felületek.

3. mítosz. A multimodális input párhuzamos jeleket jelent.

Egy másik gyakori feltételezés, hogy a különböző jelzések, modalitások részben párhuzamosan történnek és ezek az átfedések határozzák meg a rendszer számára, hogy mely jeleket kell együttesen feldolgozni. Például a mutogatással egybekötött beszéd esetén azt gondolhatnánk, hogy az emberek egyidőben beszélnek és mutatnak valamire, amikor azt mondják „ott". Ez a fajta átfedés azonban szintén egy tévhit: mindössze 25%-ban fedik át egymást a különböző deixisek és kéz általi mutatások, Oviatt 1997-es empirikus tanulmánya szerint [30]. A valóságban a gesztus gyakran megelőzi a szóbeli információkat. Emiatt nem szabad félreértelmezni a jelek közötti szinkronizáció meglétét, mivel a szinkronizáció nem törvényszerű.

4. mítosz. A verbális input minden olyan rendszerben elsőrendű, amelynek része.

Az informatikusok és nyelvészek eleinte úgy gondolták, hogy a beszéd minden esetben az elsődleges kommunikációs forma, minden más input modalitás (gesztusok, mozdulatok,

tekintet, arckifejezés, stb.) annak csak egyfajta kiegészítése lehet Ez a gondolat határozta meg a kezdeti multimodális rendszerek fejlődését is, hiszen először a beszéd inputtal vagy primitív beszéd-és-mutogatás (a mutogatás másodlagos szereppel rendelkezett) inputtal látták el őket. A valóságban vannak modalitások, amelyek bizonyos információkat sokkal hatékonyabban tudnak közölni, mint a verbális csatorna (könnyebben érthető egy térkép vizuális úton, mint szóban elmondani egy útvonalat). Ezen felül, ahogy erről már az előbb is szó volt, bizonyos modalitások nagyon jól artikuláltak lehetnek, és nem is feltétlenül redundánsak: például, nagyon sok gesztusjelzés megelőzi a tényleges beszédet, és egyértelmű információkat közöl a rendszer számára.

5. mítosz. Nyelvészet szempontjából a multimodális nyelv nem különbözik az unimodális nyelvtől.

Gyakran feltételezik azt, hogy „a nyelv az csak egy nyelv", így aztán miért is különbözne a multimodális nyelv alapvetően az unimodális nyelvtől? A valóság azonban az, hogy bizonyos multimodális nyelvek (mint a toll-és-beszéd input) sokkal rövidebbek, szintaktikailag egyszerűbbek, és kevésbé akadozóak, mint a tradicionális unimodális beszéd [29]. Például Oviatt interaktív térképrendszerének használatakor, a felhasználó vagy azt mondta, hogy „Helyezz egy kikötőt a Reward Lake keleti, nem, nyugati oldalára." Vagy egyszerűen csak rajzolt a megfelelő helyre egy négyzetet és azt mondta „Hozzáad kikötő".

Mikor a felhasználók szabad kezet kapnak, hogy mely modalitásokat használhatják, általában a nyelvészetileg egyszerűbb megoldást fogják választani. A multimodális nyelv egyszerűen más, mint a beszélt vagy írott nyelv. Az előbb említett példában is látszik, hogy multimodális beviteli mód esetén a felhasználók a beszélt nyelv nyelvtani szabályait is elhanyagolják, eltérnek az „alany-állítmány-tárgy" konstrukciókról. Tőszavakban, lényegretörően beszélnek. Ez azonban azt is jelenti, hogy a beszédfeldolgozást nem szükségszerű a lehető legfinomabbra csiszolni, emiatt hatékonyabb, kevésbé hibaérzékeny multimodális rendszereket is létre lehet hozni.

6. mítosz. A multimodális integráció magával vonja a tartalom redundanciáját.

A szakemberek gyakran állítják, hogy a multimodális kommunikáció különböző csatornáin érkező információ magas fokú redundanciát mutat. A multimodális kommunikációt azonban úgy is lehet értelmezni, hogy a tartalmat nem redundánsan, hanem kiegészítő módon „rakja össze". A különböző modalitások különböző, de egymást kiegészítő információszeleteket fejeznek ki. Az interaktív térkép esetén a helymeghatározó információt gyakran a tollal jelölték meg a felhasználók, míg az alany, állítmány, tárgy típusú információk általában verbális úton közölték. A jövő multimodális rendszereinek tervezésekor ezért nem érdemes arra támaszkodni, hogy a különböző csatornákon érkező információ redundáns lesz.

7. mítosz. A különböző felismerő technikák ötvözete még nagyobb hibalehetőséggel terhelt rendszert eredményez.

Általában úgy gondolják, hogy több, hibára hajlamos felismerő technológia együttes alkalmazásával (beszéd és fénytoll felismerő) sok „összetett" hiba is keletkezhet, bizonytalanabbá válik a rendszer teljesítménye. A valóságban azonban a multimodális rendszerek sokkal robusztusabbak az unimodális rendszerekkel szemben, mivel az egyes felismerő technológiák hibáinak feloldását más modalitásokon keresztül tudják megoldani. Ezen robusztusság egyrészről onnan ered, hogy a felhasználók tudják, hogy mikor és hogyan használják az adott beviteli módot a leghatékonyabban. Amennyiben lehetőségük van többféle beviteli módra, ösztönösen a lehető legkisebb hibalehetőséggel rendelkező utat választják. Oviatt arra is felhívja a figyelmet, hogy több forrás esetén egy forrás kijavíthatja egy másik forrás hibáját. Például, ha a beszédfelismerő csak az „üzenet" szót ismeri fel, de a felhasználó több objektumot is kijelölt egy másik beviteli módon, akkor a rendszer megfelelően, „üzenetek"-ként fogja interpretálni a parancsot.

8. mítosz. A felhasználók multimodális parancsai uniform módon kerülnek integrálásra.

Minden felhasználó egyéni, saját preferenciái és taktikája alapján létesít kapcsolatot az adott rendszerrel. Oviatt interaktív térkép rendszere esetén a felhasználók vagy egyszerre beszéltek és használták a tollat, vagy egymás után alkalmazták ezeket a modalitásokat. A felhasználási módszer már az elején eldőlt az egyes felhasználóknál, és a rendszerrel való kommunikáció ideje alatt végig ugyanaz maradt. Emiatt a felismerés miatt azok a multimodális rendszerek, melyek képesek igazodni a felhasználó domináns viselkedésmintáihoz, sokkal jobb felismerési és hatékonysági mutatókkal rendelkeznek.

9. mítosz. A különböző beviteli módok összehasonlítható tartalmakat közvetítenek.

A multimodális rendszerek beszédközpontú elvének ellentéteként terjedt el az „alternatív-modalitás" elve. Eszerint az egyes beviteli módok által közvetített információ összehasonlítható, így a kívánt modalitás tetszőlegesen variálható az információtartalom sérülése nélkül. Bár technológiai szempontból nézve a különböző módok egymással alapvetően felcserélhetőnek, ez még nem jelenti azt, hogy minden modalitás egyforma.

A különböző modalitások adatátviteli szempontból sem egyformák. Bár egyesek (például a beszéd és az írás) könnyebben összekapcsolhatók, mások már kevésbé (beszéd és tekintet). Különböző szinten kifejezők és pontosak. Például komplex térbeli objektumokat szóban leírni sokkal nehezebb és nehezebben is értelmezhető, mint megfelelő jelölésekkel leírni. Különbség van az egyes modalitások alkalmazhatósága között is. A beszéd és az írás sokkal direktebb kommunikációs forma, mint például a testbeszéd vagy a tekintet, amely inkább passzív, akár tudatalatti információt közöl a felhasználóról. Egy adott multimodális rendszernek ennek megfelelően kell tudnia értelmeznie a különböző csatornákon érkező információt. Ebből következik, hogy nem lehet minden csatornán pontosan ugyanazt kifejezni.

10. mítosz. A multimodális rendszerek legfontosabb előnye a megnövekedett hatékonyság.

Gyakran feltételezzük, hogy a párhuzamos inputcsatornák miatt a gyorsaság- és hatékonyságnövekedés az elsődleges teljesítménybeli növekedés a multimodális rendszerek esetében. Például Oviatt intelligens térképrendszere esetében az útvonalak lekérése multimodális beszéd-és-fénytoll segítségével 10%-os gyorsaságnövekedést eredményezett az unimodális beszéd-alapú interfésszel szemben. De lehet, hogy ez a növekedés csupán a térképhasználat bizonyos részeinél jön ki [29]. Eddig senki sem tudta demonstrálni, hogy a hatékonyság egy általános és lényeges előny, csak bizonyos rendszerek esetében. A hatékonyságon kívül azonban más lényeges előnyök is jellemezhetnek egy-egy multimodális rendszert. Például rugalmasság (a felhasználók váltogathatnak a beviteli módok között) vagy adaptációs készség (az interfészek a felhasználók szélesebb körében alkalmazhatók, mint az unimodális interfészek), stb.

3. Multi-modális rendszerek

3.1. Put – That – There

A multimodális felhasználói interfész fogalmát 1980-ban Richard Bolt vezette be Put-That-There (Tedd-Azt-Oda) rendszerével [33]. A rendszer leírása Rainer Wasinger könyvéből [44] származik:

A Massachusetts Institute of Technology (MIT) két tudósa, Richard Bolt és Chris Schmandt által leírt rendszer hangfelismerés és mutogatás alapján volt képes eseményeket létrehozni egy nagyméretű grafikus kijelzőn. Ez a rendszer a Térbeli Adatkezelő Rendszer (Spatial Data Management System, SDMS) kutatásához tartozott, melynek célja térben indexelni mindennapi élethez kapcsolódó objektumokat, adatokat, például egy irodában. A rendszer egy úgynevezett 'média szobában' kapott helyet, amiben egy középen elhelyezett karosszék, a szék két oldalán 1-1 képernyő, és a székkel szemben egy nagy kivetítő helyezkedett el. Az egyik oldalsó képernyő az SDMS egészéről szolgáltatott információkat, beleértve a felhasználó pozícióját is. A másik oldalsó képernyőn ennek a pozíciónak egy részletesebb megjelenítése található, lényegében egy nagyítónak fogható fel. A két képernyő navigálására 1-1 joystickot helyeztek el. Az egyik joystickkal a koordinátatengelyeken lehetett mozogni, míg a másikkal az virtuális szoba multimédia objektumaira lehetett ráközelíteni, mint például térképek, könyvek, videók. A képernyőn való megjelenítéséért és a többi eszköz meghajtásáért miniszámítógépek voltak felelősek, illetve helyet kapott még a szobában 4 hangfal is, 1-1 a kivetítő két oldalán, 1-1 pedig a felhasználó székének bal és jobb oldalán.

3.1.1. A rendszer képességei

A Put-That-There rendszerrel való interakció a központi kivetítőn látható egyszerű geometriai formák kezelésére korlátozódott. Ezeket létre lehetett hozni, törölni, mozgatni, másolni, vagy a tulajdonságaikat (színük és méretük) befolyásolni. Az ötlet szerint ezek a formák jelképezték az egyes fizikai tárgyakat a való világban. A rendszer multimodális jellegét az adta, hogy a felhasználó egyaránt tudott kizárólag beszéd útján, vagy beszéd és mutogatás

útján kommunikálni a rendszerrel. A kombinált esetben a mutogatást mindig beszéd kísérte, mivel a mutogatás csak a beszéd-inputban szereplő deixisek feloldását szolgálta.

A rendszer képességeit nagyon jól példázza a „Mozgasd a kék háromszöget a zöld négyzet jobb oldalára” parancs. Ez ugyanazt a hatást eredményezné, ha a „Tedd azt oda” parancsot mondanánk, amennyiben az „azt” a „kék háromszöget”, az „oda” pedig a „zöld négyzet jobb oldalára”-t jelenti. A rendszer komoly erőssége, hogy képes viszonyítási kifejezéseket értelmezni, például egy objektum paramétereinek módosításakor („nagyobb”, „kisebb”) vagy mozgatáskor („objektumtól balra”).

A rendszer beszédfelismerője 120 szót volt képes felismerni, a mutogatáshoz pedig kettő darab szenzor volt szükséges. Az első szenzor a szöget és az irányt ismerte fel, aminek segítségével a központi kivetítő egyes koordinátáira lehetett mutatni (ez volt a felhasználó csuklójára téve), a másik szenzor pedig az első szenzor fizikai pozícióját volt hivatott kiszámolni, miközben a felhasználó a székben ült.

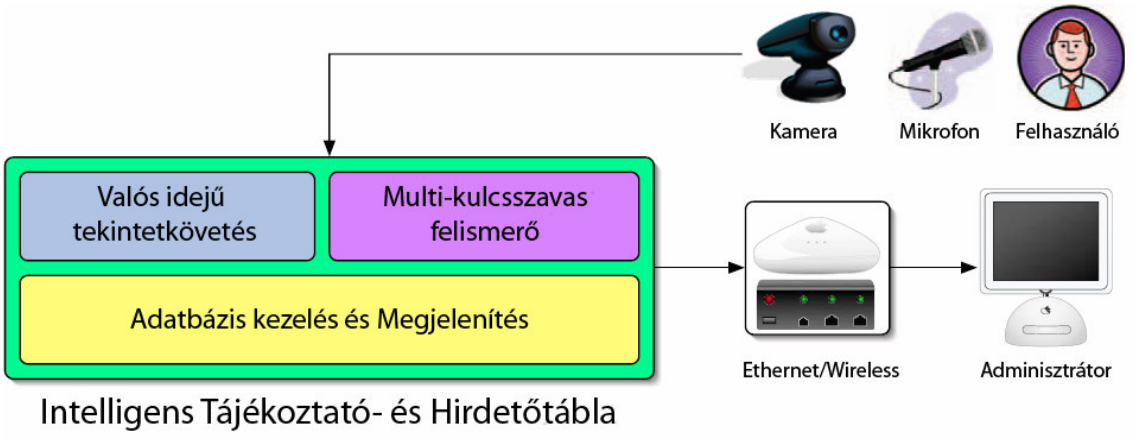
A YouTube-on található egy másfél perces videó [60] ami az eredeti 1979-es felvételből származik.

3.2. Intelligens tájékoztató- és hirdetőtábla

Ebben a fejezetben egy, a taiwani National Cheng Kung University-n kifejlesztett intelligens tájékoztató- és hirdetőtáblát szeretnék bemutatni. A leírás alapjául a [3] szolgált.

A különböző tájékoztató- és hirdetőtáblák napjainkban sok fontos és frekventált helyen megtalálhatóak. Ezeknek a táblák sok helyet foglalnak, rugalmatlanok, és nem viselkednek interaktívan a felhasználókkal. Ezeket a hiányosságokat próbálja pótolni az intelligent Guiding Bulletin Board System – továbbiakban iGBBS – melynek alapját a tekintetkövetés és beszédfelismerés adja. Az iGBBS célja egy kényelmes, könnyen használható multimodális felület biztosítása a felhasználók számára. A felhasználók alapvetően tekintetükkel és fejük mozgásával képesek irányítani a rendszert, míg aki inkább a verbális megoldásokat preferálja, bizonyos kulcsszavak segítségével is képes interakcióba lépni a rendszerrel.

A rendszer felépítése az 1. ábrán látható:



1. ábra. Az iGBBS rendszer felépítése.

Az iGBBS három fő modulja:

- *Valós idejű tekintetkövetés:* Ennek a modulnak a feladata kezelni, hogy mikor ébredjen fel a rendszer és álljon a felhasználó rendelkezésére.
- *Több-kulcsszavas felismerő:* A felhasználó beszédéből képes a megfelelő kulcsszavakat felismerni, és megfelelően reagálni azokra.
- *Adatmenedzsment és Megjelenítés:* az adminisztrátoroknak biztosít kényelmes kezelési felületet.

Ezekből az első kettőt ismertetjük részletesebben.

3.2.1. Valós idejű tekintetkövetés

Az ember alapvetően arra néz, amerre az érdeklődése, figyelme irányul. Ez alapján, ha a rendszer azt érzékeli, hogy egy felhasználó ránéz, az azt jelenti, hogy használni is akarja a rendszert. Az iGBBS esetében a fejlesztők egy megjelenés-alapú és statisztikai megközelítést alkalmaztak, amit Viola és Jones [42] fejlesztettek ki, majd Lienhart [18] módosította azt.

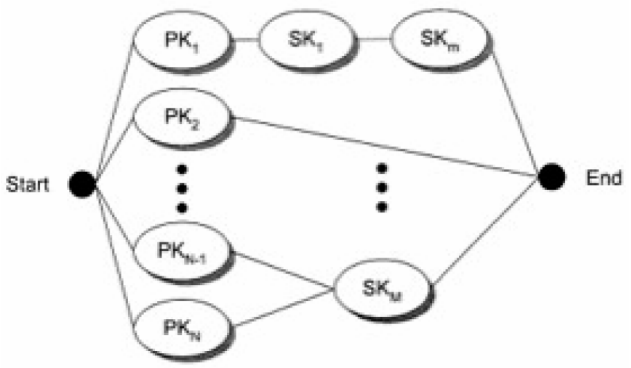
Miután a rendszer felismert egy arcot, már csak meg kellett határoznia annak orientációját, és ezekből eldönthette, hogy aktiválja-e magát vagy sem. Ennek meghatározásának első lépése bizonyos jellemző pontok, sajátságok meghatározása a felismert arc területén belül.

A jellemző pontok megtalálása után a rendszer meghatározta a felhasználó érdeklődésének tárgyát. Ehhez egy módosított, piramis Lucas-Kanade algoritmust használtak ([20],[40]).

3.2.2. Multi-kulcsszavas felismerő

A rendszer kulcsszó felismerője azért különleges, mert egy közlésen belül több kulcsszót is képes felismerni, azokat együttesen kezelni. A szerzők ennek az algoritmusát a [45]-ben publikálták.

Ahhoz, hogy több kulcsszót együttesen tudjon a rendszer használni, létre kellett hozni egy kulcsszó-relációs táblázatot, amely alapján el lehet dönteni, hogy adott kulcsszavak kombinációja mit is jelent. A relációs tábla felépítése (PK, {SK}), ahol PK az elsődleges kulcsszó, {SK} pedig másodlagos kulcsszavak egy halmaza. Ebben a megközelítésben egy relációban csak egy elsődleges kulcsszó engedett, a másodlagos kulcsszavak halmaza pedig üres is lehet, de elemei valamilyen kapcsolatban állnak az elsődleges kulcsszóval. Ehhez egy diagrammot (2. ábra) állítottak össze a készítők, amely alapján és a kulcsszavak mondatbeli elhelyezkedésének figyelembe vételével a rendszer felismerheti a multi-kulcsszavas jelöléseket.



2. ábra. Kulcsszó diagram.

Miután megvannak a kulcsszó tippek, a rendszer végigvizsgálja az elsődleges és másodlagos kulcsszavak lehetséges kombinációit. Ha egy másodlagos kulcsszó túl lazán kapcsolódik az elsődleges kulcsszóhoz (a diagram alapján túl távol van), akkor annak hatását

kisebb mértékben veszi figyelembe. A cikkben bemutatott függvények alapján súlyozza a kulcsszavakat, és így dönt az alkalmazandó kulcsszó kombinációról.

3.2.3. Alkalmazás

A fent bemutatott rendszer prototípusát a National Cheng Kung University Mérnöki Tanszékén üzemelték be, az épület első emeletén. Bárki szabadon használhatta, hogy az adott iroda/laboratórium helyét megtudja. A rendszer egy Pentium IV 1GHz PC-ről futott, 20 képkocka/másodperces sebességgel volt képes a tekintet felismerésre. A több-kulcsszavas felismerő 36,2%-os hibafaktorral működött, tehát a felhasználók átlagosan minden 2,76.-ik kérdésükre a megfelelő választ kapták.

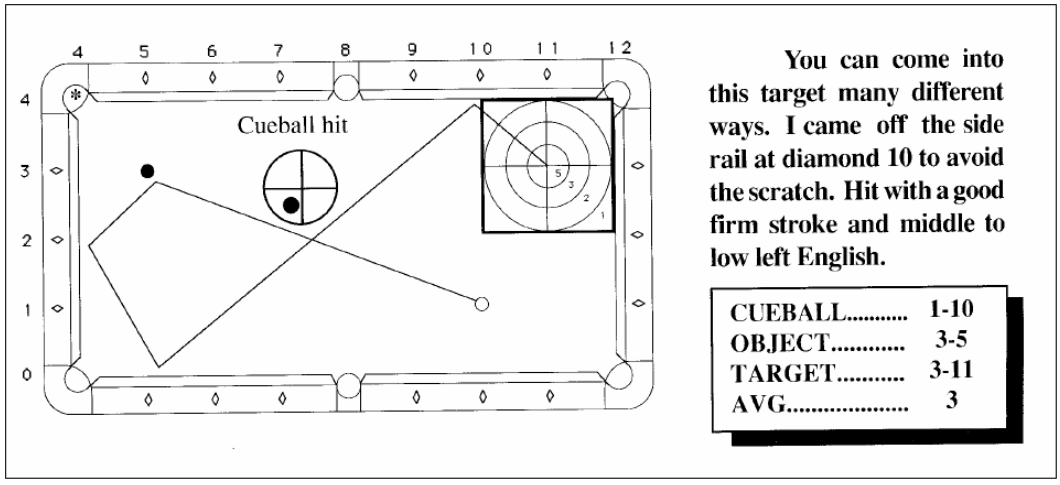
A rendszer alapvetően elérte a célját, hiszen a tradicionális hirdetőtáblák minden hiányosságát kiküszöbölték, hiszen az iGBBS rugalmas, mert könnyen frissíthető volt a tartalma, bárhova telepíthető volt ahova egy sima hirdetőtábla is kihelyezhető, ráadásul multimodális kapcsolatot tudott kialakítani a felhasználókkal.

3.3. Multimodális Pool oktató

Az alábbi bemutatás a [16] cikk alapján készült. Az Aalborgi Egyetemen kifejlesztett automatizált biliárdoktató (konkrétan a pool játék) rendszer célja a tanulási folyamat egyes részeinek automatizálása. Az Automatizált pool oktató (Automated Pool Trainer, APT) multimodális virtuális oktató alapú ember-gép kommunikációt valósít meg. A rendszer a középpontjába az oktatást helyezi, ami egyébként is az informatikai rendszerek egyik központi feladata. Ami külön érdekesség viszont, hogy egy gyakorlati képességet tanít meg, a piacon jelenleg kapható oktató szoftverek pedig szinte kizárólag elméleti anyagokat adnak át. Mivel egy ilyen szituációban a tradicionális WIMP (Window, Icon, Menu, Pointing device) interfész használata kényelmetlen lenne, így remek lehetőség nyílt egy multimodális rendszer megalkotására.

3.3.1. Target Pool

Az APT rendszerének alapját a Kim Davenport profi pool játékos által kifejlesztett Target Pool [6] séma szolgáltatja, melynek segítségével a gyakorolni kívánó játékos előre meghatározott gyakorlatok alapján fejlesztheti és értékelheti ki saját tudását. Minden egyes gyakorlathoz részletes leírás tartozik. A játékos minden egyes lökés után feljegyzi egy papírra a teljesítményét, majd bizonyos számú próbálkozás után kiértékeli az eredményét. Az értékeléstől függően kapja majd meg a következő ajánlott gyakorlatot. A pool asztalon, golyókon és dákón kívül csak a gyakorlatok leírását tartalmazó füzetre, a kiértékelő lapokra és egy célkeresztet ábrázoló vékony szövetlapra van szükség, amit a pool asztal aktuális célterületére kell helyezni. A Target Pool több mint 140 gyakorlatot tartalmaz 10 csoportra osztva. A gyakorlatok túlnyomó többségében csak a fehér és egy másik golyóra van szükség. A 3. ábrán található egy példagyakorlat. Látható rajta a fehér és a másik golyó pozíciója, hogy hol kell meglökni a fehér golyót, illetve egy vastag vonallal megjelölik az ideális vonalat, amit a fehérnek követnie kell.



3. ábra. Target Pool példagyakorlat.

A Target Pool felépítésénél fogva kiválóan passzol egy multimodális rendszerbe, mivel grafikai és szóbeli (szöveges) elemeket egyaránt tartalmaz.

3.3.2. Az interfész

A felhasználó többféle módon is kommunikálhat a rendszerrel. A rendelkezésére álló érintőképernyőn keresztül direkt módon tudja kezelni a GUI objektumait, verbális parancsokat adhat a rendszernek egy vezeték nélküli mikrofonon keresztül, vagy mozgathatja a pool asztalon lévő golyókat is, amiket az asztal felett elhelyezett kamera észlel. Opcionális input eszközként természetesen rendelkezésére áll a szokásos egér + billentyűzet is, melyeket elsősorban komplex utasítások (új felhasználó létrehozása, üzenetek írása) esetén lehet alkalmazni.

A rendszer vizuálisan akár az érintőképernyőn, akár az asztal mellett elhelyezett nagyméretű kivetítőn keresztül is jeleníthet meg információkat. Van egy beépített beszédszintetizátora is, azonban az igazi különlegességét az a lézerfej adja, amely vonalakat fest közvetlenül a pool asztalra, ezzel is segítve a játékost.

Összefoglalva a rendszer a következő módokon kommunikál a felhasználóval:

- *Szintetikus beszéd:* Így a felhasználónak nem kell folyamatosan váltogatnia a figyelmét a képernyő és a pool asztal között.
- *Grafikus megjelenítés:* Az elmondott utasítások illusztrálására szolgál.
- *Video megjelenítés:* Visszajátssza a felhasználó legutóbbi lökését a kiértékelés miatt.
- *Animált virtuális oktató:* Megmutatja az adott feladatot, beszél a felhasználóhoz.
- *Szöveges megjelenítés:* Az elmondott utasítások összefoglalása a képernyőn.
- *Lézersugár:* Közvetlenül a pool asztalra vetített vonalak és körök segítségével jelöli a golyók helyét, az ideális nyomvonalat, stb.

3.3.3. A rendszer felépítése

A rendszer hardveres központját egy csúcsteljesítményű PC képezte, kiegészítve a szükséges eszközökkel és szoftverekkel. A következőkben rövid bemutatásra kerülnek a fontosabb modulok:

A *lézeres modult* egy 'alárendelt' PC által működtetett X-Y szkenner hajtotta. Teljesítményét tekintve 600 pont közé tudott vörös vonalat húzni 50Hz-s sebesség mellett. Az eredmény egy majdnem vibrálás-mentes folyamatos törtvonal megjelenítése a pool asztal

felületén. A cikk írásának idején vörös lézert használtak, amely nem a legszerencsésebb választás a zöld posztóra. Másik hátránya, hogy túl távoli pontok esetén elhalványul a vonal. Részletesebb leírás a ([25],[2],[17]) cikkekben található.

A *beszédfelismerő és -szintetizáló modulokat* az IBM ViaVoice [52] eszközével valósították meg. A felismerés magas színvonalának eléréséhez egy vezeték nélküli mikrofont tettek a felhasználóra, valamint egy szabály alapú grammatikát fejlesztettek ki. Minden új felhasználóhoz szükség volt a rendszert betanítani, hogy a rendszer megfelelően tudjon alkalmazkodni az új kiejtési és hangszínbeli tulajdonságokhoz.

A *képfeldolgozási modul* feladata a golyók helyzetének felismerése a pool asztalon. Ennek segítségével a rendszer tudja, hogy a játékos mikor helyezte le a megfelelő pozícióra a golyókat a gyakorlat elkezdéséhez, illetve a lökés után ki tudja értékelni a golyók relatív helyzetét a célterülethez képest. A rendszer eseményekkel való szinkronizálásán túl még lehetőség van az egyes lökések felvételére, amelyeket a rendszer kiértékel és visszajelzést ad a felhasználónak a teljesítményével kapcsolatban. A cikk írásának idején egy 600MHz Pentium III-as processzor végezte ezt a feladatot, 12 FPS sebességgel. Ez elegendő volt a golyók valós idejű követéséhez. A pozicionálás pontossága 1-1,5 centiméteres hibával volt terhelt, a golyó aktuális helyzetétől függően.

3.3.4. Virtuális oktató

Egyszerű tény, hogy az emberek nem szoktak pool asztalokhoz vagy képernyőkhöz beszélni. Emiatt van szükség valamiféle megszemélyesítésére a rendszernek, hogy a felhasználók kényelmesen érezzék magukat. A felhasználók hajlandóak virtuális személyekkel kommunikálni, amennyiben konzisztens és meggyőző személyiséggel vannak azok felprogramozva [34]. Ellenkező esetben a virtuális személy nem lesz hiteles partner, így a rendszer hamar elveszti a bizalmát az egész rendszerrel szemben.

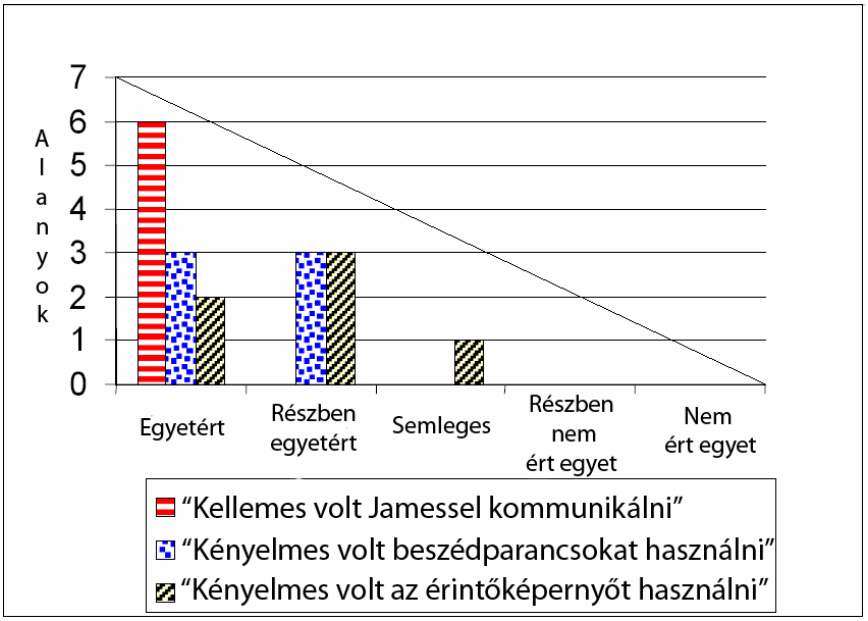
A készítők úgy gondolták, hogy a virtuális oktatónak képesnek kell lennie gesztikulálni és a GUI elemeire rámutatni, ezért olyan animált karaktert kerestek, amely teljes teszt gesztusokat tudott produkálni, illetve mozogni a képernyőn. Végül az egyik elterjedt és könnyen hozzáférhető megoldás mellett döntöttek és az Microsoft MS Agents [55] modulját alkalmazták. Ez DirectX és SAPI kompatibilis, könnyen integrálható bármilyen Windows alkalmazásba. A különböző gesztusokat animált GIF sorozatokkal valósítja meg, ráadásul a

Microsoft egy beépített gesztus könyvtárat is biztosít a modulhoz. A 4. ábrán „James”, a virtuális pool oktató látható. Az oktató elsődleges kommunikációs csatornája a szintetizált beszéd. Elmondja az aktuális feladatot, ötleteket ad, így a felhasználónak nem kell folyamatosan a képernyőre nézni. A gyakorlatok elvégzése közben James hallgat, nem zavarja a játékost, kivéve persze ha maga a felhasználó szól hozzá.

4. ábra. James.

3.3.5. Humántesztek a rendszerrel

Több előzetes teszt után elvégeztek a rendszerrel egy használhatósági tesztet is. Elsősorban James hatását vizsgálták. Hat ember vett részt rajta, fejenként nagyjában másfél órát töltöttek a rendszerrel. Ebbe az időbe beletartozik a beszédfelismerő betanítása is. Az 5. ábrán látható diagramon egyértelműen látszik, hogy a tesztalanyok kényelmesnek érezték a virtuális oktatóval való kommunikációt. A teszt egy másik célja volt kideríteni, hogy a felhasználók mely modalitás használatát preferálták a többivel szemben. A készítők a válaszok alapján úgy vélték, hogy inkább a virtuális oktatót részesítik előnyben, de mivel nagyon kevés alany vett részt, ezért statisztikailag megalapozottak az eredménynek.



5. ábra. A Pool Oktatóval végzett humánteszt eredménye.

3.4. Multimodális póker

A 2008-as CeBIT kiállításon mutatta be az IDEAS4Games [53] kutatócsapata ezt az interaktív, multimodális pókerjátékot [38]. A játék során egy emberi játékosnak volt lehetősége kettő, a számítógép által vezérelt ellenfél ellen 5 lapos pókert játszania. Az ember vállalta fel az osztó szerepét is, mivel a játék igazi, RFID chipekkel preparált kártyalapokkal folyt. A két mesterséges ellenfél – Sam és Max - kifinomult érzelem-megjelenítéssel és csúcstechnológiának számító beszédszintetizátorral rendelkezett. Sam alapvetően egy barátságos, rajzfilmszerű alak, míg Max egy ellenszenvesebb, Terminátor-szerű robotfigura volt. Mindkettőt a nyílt forráskódú Horde3D [51] grafikus motorral renderelték.

6. ábra. A CeBIT 2008-on bemutatott multimodális pókerjáték.

A 6. ábrán látszik a rendszer kialakítása. Az asztal három részre van osztva, reprezentálva a három pókerjátékos helyét. A kártyák számára kialakított helyen RFID leolvasók vannak telepítve, így ismeri fel a gép a leosztott lapokat. A játékossal szemben lévő asztali monitoron és egeren keresztül tudja az ember kiadni a parancsait, mint például passz, emel, tart, stb. Ezen kívül a játék szempontjából további lényeges információkat jelenít meg, pl. mennyi az aktuális tét, stb. A tábla feletti 42'' monitoron látható Sam és Max, a két virtuális pókerjátékos.

Amikor a felhasználó odalép az asztalhoz és elkezdi a játékot, Sam és Max elmagyarázza a játék szabályait. Ezek után a játékos leosztja az első kört, és kezdődhet a játék. A játék során a két virtuális játékos reagál az eseményekre, az osztásokra, emelésekre. Különböző személyiséggel és különböző pókerstílussal vannak programozva. Sam, aki inkább emberszerű játékos, szabály-alapú algoritmust használ, míg Max nyers erő algoritmusával mind a 2.58 millió lehetséges 5 lapos kombinációra kiszámol egy valószínűséget. A rendszer egyik legnagyobb érdekessége, hogy a két virtuális játékos a játék eseményeitől függően

képes változtatni a személyiségén, hangulatán, hozzáállásán. Ezeket mind párbeszédben, mind testbeszédben közlik az emberi játékossal.

3.4.1. Vezérlés

Sam és Max viselkedését a SceneMaker [10] nevű eszközzel programozták. A dialógusokat ún. jelenetekbe – összefüggő beszédszakaszokba - csoportosították. Egy-egy ilyen jelenet lényegében egy szkript, amelyben meghatározzák, hogy mit mondjon, és hogyan viselkedjen a virtuális szereplő. Ezen felül az egyes mondatokhoz hangulat-befolyásoló jelzéseket (tageket), a szkripthez pedig kiegészítő rendszerparancsokat is hozzá lehet rendelni. Az egyik kiemelkedő jellemzője a rendszernek, hogy a készítők odafigyeltek arra, hogy a virtuális játékosok ne ismételjék ugyanazokat a mondatokat. Egy feketelistás rendszert vezettek be, ami alapján a felhasznált jeleneteket bizonyos időre (pl. 5 percre) blokkolták, és helyette egy másik odaillő jelenetet alkalmaztak az adott pillanatban. Ehhez minden eseményhez jelenetcsoportokat hoztak létre, szám szerint 73 csoportot, összesen 335 jelenettel.

A pókerjáték kötött folyamat, az egyes lehetséges lépések jól meghatározottak. Emiatt volt lehetséges az, hogy a készítők egy jelenetfolyamattal tudták modellezni, hogy mikor mit mondhat, illetve cselekedhet a virtuális játékos. Ezt a jelenetfolyamatot egy hipergráffal írták le. A hipergráf minden csomópontjában algráfok szerepelnek. Minden egyes csomópontot egy vagy több jelenetet vagy jelenetcsoportot rendeltek [10].

Futási időben a rendszer ezeken a csomópontokon halad keresztül a játék aktuális állapota és a három játékos lépései alapján. A csomópontokban kiválasztott dialógusok és testbeszédek határozzák meg a két virtuális játékos multimodális viselkedését.

3.4.2. Viselkedés modell

A valós idejű viselkedés meghatározásához a készítők az ALMA nevű modellt [11] használták. Három viselkedés-tényezővel dolgozik:

- *Érzelmek:* rövid távú viselkedés, általában valamilyen adott eseményhez, cselekedethez, objektumhoz tartozik

- *Hangulat:* középtávú viselkedés, amely általában nem kapcsolódik semmilyen adott eseményhez, cselekedethez, objektumhoz
- *Személyiség:* hosszú távú viselkedés, alapjaiban határozza meg a személy jellemzőit és beállításait

Az ALMA Ortony, Clore és Collins [28] által kifejlesztett kognitív érzelmi modellt implementálja, a PAD [24] hangulat szimulációs modellel és a BigFive [22] személyiség modellel együtt. A viselkedés ezen három szintje között szoros kapcsolat van: a személyiség meghatározza az alapvető hangulatot és befolyásolja a különböző érzelmek erősségét; az érzelmek, mint események hatással vannak a hangulatra; a hangulat pedig erősíti vagy tompítja az egyes események által kiváltott érzelmeket.

Az ALMA 24 érzelem kiszámítására képes. A kiértékelt érzelmek befolyásolják az adott egyén hangulatát. Minél intenzívebb egy érzelem, annál nagyobb a hangulatváltozás. Ami egyedi, hogy az aktuális hangulat viszont befolyásolja az érzelmek intenzitását. Ezzel tudják szimulálni, hogy sokkal nagyobb intenzitású az öröm, és sokkal kisebb intenzitású a bánat, ha az aktuális hangulat épp „kiváló". Egy-egy hangulatot egy (P, A, D) hármassal (Pleasure – Kedv, Arousal – Érdeklődés, Dominance – Dominancia) jelölnek. Például, ha minden tulajdonság pozitív (+P, +A, +D), a hangulat „kiváló".

Az aktuális hangulat és érzelmek befolyásolják a virtuális pókerjátékos viselkedését. A lélegzetvételt befolyásolja a hangulat Kedv és Érdeklődés paramétere. Ha ezek pozitívak, a játékos gyorsabban veszi a levegőt, míg negatív értékek esetén a levegővétel lassú és halvány. A beszéd minősége is változik. Nyugodt hangulat esetén semlegesen beszél, míg ellenséges vagy megvető hangulatban agresszívebb lesz. Ha a virtuális játékos hangulata kiváló, a beszéde érezhetően vidámabb lesz.

A játék kezdetekor Sam és Max viselkedése majdnem teljesen megegyezik, mindketten az alapvető nyugodt hangulatból indulnak. Viszont különböző személyiségük miatt, ahogy halad előre a játék, úgy lesz egyre nagyobb a különbség köztük. Sam kicsit extrovertáltabb, így a hangulata a pozitív, kiváló irányába tendál. Max viszont negatívabb személyiséget kapott, így inkább ellenséges lesz.

3.4.3. Érzelemkifejező szintetikus hangok

A fejlesztők szerint a meggyőző beszéd előfeltétele annak, hogy a virtuális személy hihető legyen. És ez különösen igaz egy olyan rendszer esetén, amely az érzelmek kifejezését tűzte ki egyik céljául.

A megfelelő beszédszintetizátornak megbízható színvonalon, természetes kifejezőképességgel kellene beszédet előállítania. Azonban ezt a két kritériumot nehéz egyszerre megvalósítani. Kétféle beszédszintézis technológiát vizsgáltak meg: Az egység kiválasztás [13] alapú technológia egy előre rögzített mondathalmazt használ fel a szövegszintézishez. Amennyiben az adott beszéd ebből a tartományból kerül ki, akkor szinte emberi természetességgel képes előállítani, de minden egyéb esetben megbízhatatlan a minősége. A statisztikai-parametrikus szintézis [47] egységes minőséget generál, de általában „tompa" hangon szólal meg, a rengeteg simító eljárás miatt.

A fejlesztők végül mindkét módszer módosított változatát alkalmazták, egyiket az egyik virtuális játékosra, másikat a másikra.

Sam hangját egység kiválasztás módszerével készítették el. Az ehhez szükséges beszédtartomány megalkotására a német nyelvű Wikipedia-ról kiválasztott 400 mondatot használtak fel, amelyek lefedték a német nyelvben használt legfontosabb diádokat (angolul diphone). Ezen felül felhasználtak még 200, a pókervilágban gyakran használt mondatot is (kártyákkal, osztással, emeléssel, stb. kapcsolatos kifejezéseket, mondatokat). Ezeket egy profi színész segítségével stúdióban rögzítették. Ezt a hatszáz mondatot 4 stílusban (semleges, boldog, agresszív, szomorú) vették fel, majd ezt a 4 adatbázist használták fel a MARY TTS nevű nyílt forráskódú szövegfelolvasó eszközhöz [37].

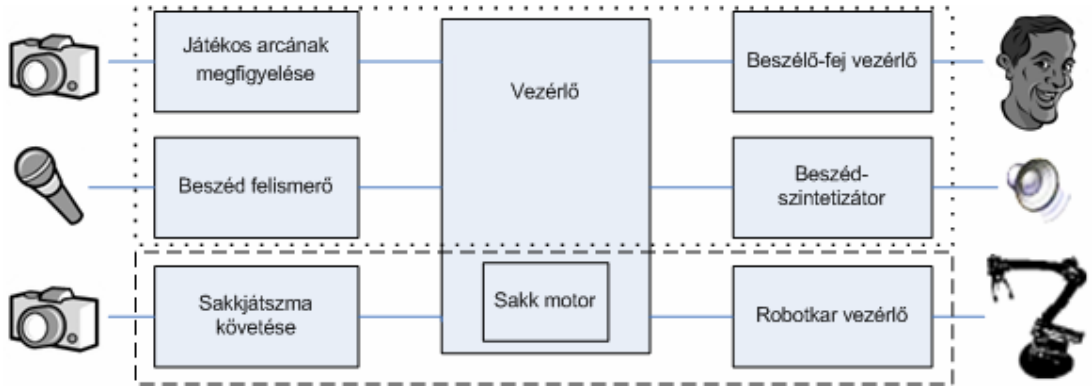
Max hangja a statisztikai-parametrikus megoldást tükrözi. A módszer egy beszéd adatbázis segítségével tanítja be a statisztikai modelleket. Futási időben egy vocoder állítja elő a hangot a statisztikai modellek alapján. A statisztikai jellege miatt az előállított hang általában kissé tompa (a statisztika és a vocoder átlagolása miatt), viszont a minőség egységes az adott szövegtől függetlenül. A kifejezőképesség javítására a fejlesztők különböző audio effekteket alkalmaztak. A vocodernek meg lehet adni a hangmagasságot, frekvenciát és a beszéd gyorsaságot. Ezen felül készítettek még egy kiejtés skálázót, ezzel lehetett nyújtani vagy rövidíteni a beszédet, illetve készült egy suttogás komponens is. Ezek után a különféle hangulatokat tükröző beszédstílust ezen eszközök megfelelő beállításaival érték el.

4. A multimodális sakkozóval végzett humán teszt

A Debreceni Egyetem Informatika Karán épült meg ez a multimodális sakkozó gép. Megépítésében és programozásában graduális és PhD képzésben résztvevő hallgatók vettek részt. A 7. ábrán látható a rendszer általános felépítése. Először részletesen ismertetjük a rendszer moduljait, ezek után a lefolytatott humánkísérlet és annak kiértékelése kerül bemutatásra.

7. ábra. A multimodális sakkozó.

4.1. A rendszer felépítése



8. ábra. A multimodális sakkozó rendszer felépítése.

A rendszer központi eleme a vezérlő és az ebbe integrált sakkozó motor. Ez a komponens felelős az összes többi rendszerelem irányításáért és az elemek közötti megfelelő kommunikációért. Ahogy azt a 8. ábra is mutatja, a komponenseket alapvetően két csoportba lehet sorolni.

Az első csoportban (melyet a pontozott kerettel jelöltünk) helyet foglaló elemek biztosítják az ember-gép közötti kommunikációhoz szükséges interaktív felhasználói felületet. Ennek a csoportnak kétfajta bemenete van, egy beszédfelismerő és egy arci érzelmet felismerő komponens. A beszédfelismerő tanítható (ez feltétlenül szükséges az emberek egyedi hangszíne és kiejtése miatt), egyelőre a játék szempontjából elég az Igen/Nem (a rendszer minden játék után megkérdezi a játékost, hogy akar-e még egy játszmát játszani), illetve az Erikkel/Robottal (a játékos meghatározhatja, hogy melyik beszélő fejjel szeretne játszani) szavak felismerése. Az arci érzelem felismerő feladata lokalizálni a kamera által készített valós idejű felvételen a játékos arcát és azon felismerni a különböző arckifejezéseket. A rendszer kimenetét a beszélő-fej modul valósítja meg.

A második csoportban (melyet a szaggatott kerettel jelöltünk) szereplő elemek magáért a sakkjátékért felelősek. Három komponens tartozik ide, a sakkozó motor, a sakkállás felismerő valamint a robotkar és az azt vezérlő program. A sakkozó motorban megvalósított mesterséges intelligencia dönt a számítógép lépéseiről. A sakkállás felismerő figyeli a sakktáblán történő változásokat, és emberi lépés esetén információt küld a sakkozó motornak. A robotkar fizikailag teszi meg a számítógép lépéseit a sakktáblán.

A teljesség igénye miatt a következőkben részletes ismertetésre kerül a rendszer minden egyes eleme, de a dolgozat szempontjából elsősorban a multimodális interfésznek van fontos szerepe.

4.1.1. A vezérlő

A vezérlő legfontosabb funkciója a többi komponens működésének koordinálása, és a köztük folyó kommunikáció megfelelő biztosítása. A vezérlőben kapott helyet a sakkozó motor is. A sakkozó motor egy xBoard/WinBoard [59] kompatibilis rendszer, a mi esetünkben a Phalanx [56] fantázianevet viselő motor.

A sakkozó motor számítja ki a játékos lépésére teendő válaszlépéseket. A motor tájékoztatja a vezérlőt a játék aktuális állapotáról és a játék várható eredményéről. Ezen információk alapján tudja a vezérlő beállítani a beszélő fej viselkedését. Ha a játékos közel áll a nyeréshez, a beszélő fej visszafogottabb, csendesebb lesz. Az arckifejezése leginkább semleges vagy akár szomorú lesz, ezzel jelezve, hogy a gép nehéz helyzetben van. Amennyiben a gép áll nyerésre, a beszélő fej magabiztosabbá válik, többet beszél. Többet fog mosolyogni, szinte kihívóan fog a játékosra nézni.

A játék kezdetekor az arcfelismerő észreveszi, ha valaki a kamera előtt ül, készen a sakkjátszmára. Ekkor a beszélő fej üdvözli a játékost, majd bemutatkozik. Ezek után megkérdezi, hogy szeretne-e a játékos inkább másik ellenfelet választani magának. Esetünkben az ellenfél kétféle lehet, lehet emberarcú vagy robotarcú. A beszédfelismerő komponens feldolgozza a verbális választ, majd annak megfelelően beállítja a megjelenítendő arcot, illetve kiindulási állapotba hozza a sakkozó motort. Ezek után kezdődik a játék, minek során a vezérlő ciklikusan a következő négy állapotban van:

- *A játékos gondolkodik.* A sakkállás felismerő aktív állapotba kerül, és a látóterébe eső mozgást figyeli. A beszélő fej felváltva a játékosra és a sakktáblára néz. A megjelenítendő érzelmet az aktuális állás és a játékos érzelmei is befolyásolják: Ha a játékos szomorúnak tűnik, az jelentheti azt, hogy épp nehéz helyzetben van vagy vesztésre áll. Ilyenkor a beszélő fej boldognak mutatja magát. Ha a játékos sokat mosolyog, a beszélő fej úgy értelmezheti, hogy nyerésre áll, ezért jobban odafigyel a

játékra, akár szomorúvá is válhat. Amennyiben túl sokáig gondolkodik a játékos, a fej türelmetlenné válik, és piszkálni kezdi a játékost, hogy lépjen már.

- *A játékos lép.* Amikor a sakkállás felismerő mozgást érzékel, ebbe az állapotba kerül a rendszer. Ilyenkor a beszélő fej a sakktáblára fordítja a figyelmét. Ha játékos ténylegesen lépett (egy bábu új helyre került a táblán), ezt feldolgozza a felismerő és elküldi a változást a vezérlőnek. Ezek után a sakkállás felismerő passzív állapotba kerül, nem érzékeli látóterében történő változásokat.

- *A rendszer gondolkodik.* A dolgozat témájához kapcsolódóan ez az állapot volt a legfontosabb, mert amíg a gép gondolkodik, addig a játékos könnyen elunhatja magát. A beszélő fej szerepe megnő, fenn kell tartani a játékos érdeklődését. Ilyenkor a fej egyrészt a táblára tekint, másrészt a játékossal próbál szemkontaktust teremteni. Az arckifejezése legtöbbször semleges, de természetesen ebben a játékállapotban is megjeleníthet érzelmet, a sakkállástól függően. Ebben az állapotban beszél a legtöbbet a játékoshoz.

- *A robotkar mozog.* Miután a sakkozó motor döntést hozott a következő lépésről, továbbítja a vezérlőnek, majd a vezérlő utasítást ad a robotkarnak, hogy milyen lépést tegyen meg a sakktáblán. Ezek után a robotkar mozogni kezd és végrehajtja a kívánt lépést. A beszélő fej a táblának arra a területére néz, ahol a lépés történik. Bemondja a játékosnak a lépést, és külön szól érte, ha ütés, sakk vagy matt történt. Az új sakkállás megfelelően fog tükröződni beszélő fej arcán (nyertes/vesztes pozíció, stb.).

4.1.2. Arci érzelmet felismerő modul

Ez a komponens dolgozza fel a játékossal szemben lévő kamera által rögzített videofolyamot. Az arcnak különös jelentősége van az ember-ember kommunikációban is, nagyon sok információt hordoz. Többek között megállapítható az ember neme, életkora, hangulata ([8],[9]). Bár videó alapú beszédfelismerés is lehetséges lenne, ebben az esetben ez a modul az arcok és az azon megjelenő különböző hangulatok felismerésére szolgál.

Az arcdetektáló algoritmusok célja azt eldönteni, hogy az adott képen vannak-e emberi arcok vagy sem. Amennyiben igen, akkor meg kell tudni adni azok pozícióját és méretét. Mivel nagyon sok tényező játszik közre (póz, megvilágítás, árnyékolás, felvétel minősége, stb.) ezért az arcdetektálás már önmagában is egy nagyon nehéz feladat. Az utóbbi években több érdekes és eredményes beszámoló is készült a témával kapcsolatban ([46],[12],[15]). A legsikeresebb technikák manapság ún. megjelenés alapúak (appearence based). Az ilyen megoldások egy arcsablon segítségével keresnek a képen. Ezt az arcsablont a rendszer egy pozitív és negatív képeket (arcot, illetve nem-arcot) tartalmazó halmaz alapján tanulja meg. A képet többféle nagyításban és ablakmérettel darabolják fel alterületekre. Ezután az így kapott területeket osztályozzák arc illetve nem-arc osztályokra. Így az arcdetektálás visszavezethető bináris mintaillesztési problémára.

Több osztályozási technika (neurális hálók, SVM, stb.) is kiváló eredményeket hozott. A multi-modális sakkozógép esetében egy OpenCv-ben implementált módszer került alkalmazásra melyet először Viola és Jones mutatott be 2001-ben [43]. Ugyanolyan pontos és stabil módszer, mint a fent említettek, viszont megközelítőleg 10-szer olyan gyors. Ennek eredményeképp valós időben képes arcokat detektálni videofolyamokon, 15 képkocka / másodperc sebesség mellett.

Az arckifejezést felismerő rendszer három feladatot kell, hogy megoldjon: arcdetektálás, adatkinyerés és osztályozás [32]. A rendszer esetében a [19]-ben közölt ötletek alapján készült el az arckifejezés felismerő. Az arckifejezés felismerő képrészleteket kap az arcdetektortól. Az adatkinyeréshez a képrészletek Gábor reprezentációját vesszük 40 különböző Gábor szűrőre (8 irány és 5 frekvencia) vonatkozólag. A páronkénti osztályozáshoz SVM-eket használ. A sakkozó gép esetében három alapvető érzelem felismerésére volt szükség: semleges, boldog, szomorú. Alapesetben az arcot semleges érzelmi állapotúnak véve elég volt két különböző SVM osztályozó használata: boldog-nem boldog, szomorú-nem szomorú.

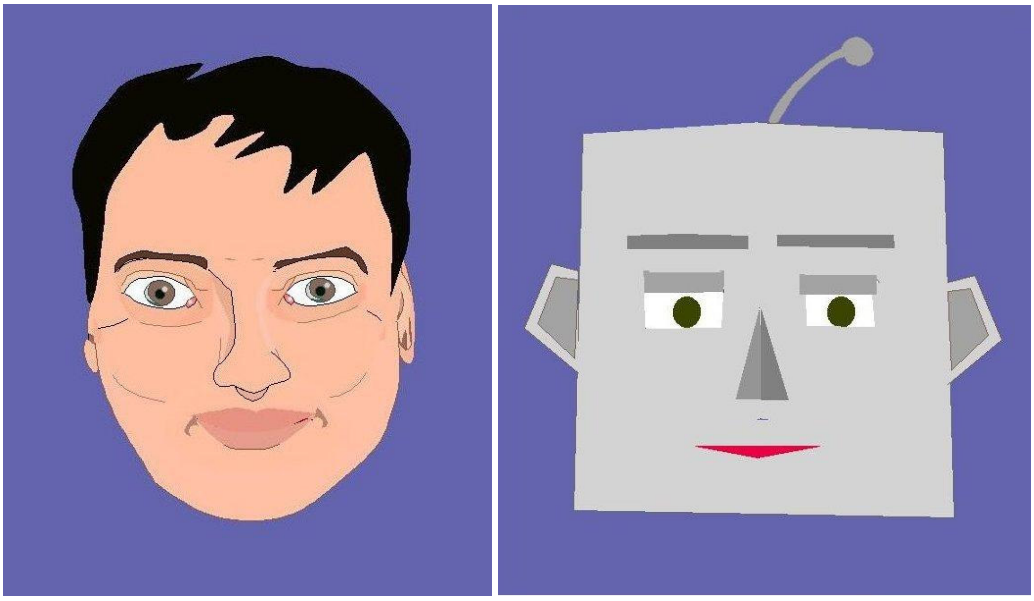
4.1.3. Beszédfelismerő

Ennek segítségével a rendszer képes értelmezni előre megadott hangparancsokat. A sakkjáték esetében a beszédnek nincs kiemelt szerepe, de könnyen elképzelhető néhány olyan szituáció, amikor hasznos lehet. A játék legelején az emberi játékos hangparancsokkal választhat

magának gépi ellenfelet és nehézségi fokozatot, illetve jelezheti a gépnek, hogy milyen bábura cserél le egy gyalogot, ha eljutott vele az ellenfél hátsó vonaláig. A beszédfelismerő komponens implementálásához a Hidden Markov Model Toolkit [50] könyvtárat használtuk fel.

4.1.4. Beszélő fej

A sakkozó rendszer ember-gép kommunikáció vizuális szintjéhez szükséges kimenetét megvalósító modult nevezzük beszélő fejnek. Ez lényegében egy animált emberi arc, amely megjelenik a játékossal szemben elhelyezkedő képernyőn. A rendszerben jelenleg a 9. ábrán látható kétféle arc közül választhat a játékos. A beszélő fej bemenetét a vezérlő egységtől kapott egyszerű szövegformátum képezi, ezeket mondja a fej a játékosnak. Az audio jeleket a ProfiVox [27] rendszer generálja, melyet a Budapesti Műszaki Egyetemen fejlesztettek ki.



9. ábra: Erik és Robot, a multimodális sakkozó két arca.

A beszélő fej képes különböző érzelmek megjelenítésére, mozgatja a száját beszéd közben, és a tábla több különböző pontjára is rá tud nézni. Az aktuális állapotát több tényező is befolyásolja: A játék aktuális állapota, a játékos arckifejezése, a sakkjáték állása. Ezeken felül egy véletlen faktor is szerepet kap, ennek segítségével válik a beszélő fej kevésbé kiszámíthatóvá. Ezeket a paramétereket a modul minden esetben a vezérlőn keresztül kapja.

Jelenleg kétféle arc van integrálva a beszélő fej modulba. Egy emberi arc és egy robot arc. A játékos a játék elején választhat, hogy melyikkel szeretne játszani. Megalkotásukhoz a CWI CharToon ([54],[35]) programját használták.

4.1.5. Sakkállás felismerő

Ez a modul a sakktábla fölé rögzített kamerán keresztül kapja a videofolyamot. Feladata a sakktáblán történő változások felismerése, és azok továbbítása a vezérlőnek. Az előfeldolgozás lépéseinek minimalizálásához egy rögzített környezetet alakítottunk ki. A sakktáblát és a robotkart rögzítettük az asztalhoz. Ezután a sakkállás felismerőhöz használt kamerát elhelyeztük a sakktábla fölé. Mivel felülnézeti képből nagyon nehéz megállapítani az egyes bábuk fajtáját, a modul csupán a bábuk pozíciójának és színének felismeréséért felelős.

A sakkjáték mindig rögzített alapállapotból indul, ezért a vezérlő képes nyomon követni a játék aktuális állását azáltal, hogy a felismerő modul megmondja, hogy a játékos honnan hova helyezett bábut. A modul két állapotban lehet: aktív vagy passzív. Mivel passzív állapotban nem csinál semmit, csak az aktív állapotot érdemes részletesebben tárgyalni.

Aktív állapotban a modul először is kap egy referencia képet a kiinduló állapotról. Az ezt követő képkockákkal egy korrelációs értéket számít ki, amely ha magasabb, mint egy előre definiált határérték, akkor a játékos keze a sakktábla felett van. Ezt úgy veszi, hogy a játékos épp lépést tesz, ezt jelzi a vezérlőnek, majd tovább figyeli a sakktáblát. Miután a korrelációs érték visszacsökken a határérték alá, a modul úgy veszi, hogy a játékos megtette a lépését, és elvette a kezét a sakktábla fölül. A határérték meghatározása a rendszer kalibrálásakor, empirikusan történik.

A következő lépés meghatározni az új játékállást. Ehhez meg kell határozni a bábuk pozícióját és színét. A referenciakép és az új kép közötti különbség adja meg, hogy a játékos

melyik bábuval hova lépett. Ezt az információt küldi tovább a modul a vezérlőnek, majd passzív állapotba lép.

4.1.6. Robotkar

A robotkar külön ehhez a rendszerhez készült. A kar pozicionálása elektromotorokkal és bowdennel történik. Bár zajosabb és pontatlanabb gépezet, mint az a mai korban elvárható, a rendszer kialakításához és a humántesztek lefolytatásához tökéletesen megfelelt. A robotkar irányítani és kalibrálni a vezérlő szoftverjén keresztül lehet. Ez biztosít kommunikációt a rendszer központi vezérlőjének is. A robotkar kalibrációja során meg kell adni az egyes bábuk fogási magasságát, illetve az A1-es és H8-as mezők helyzetét. Ezek alapján a robotkar vezérlője minden szükséges információt kiszámol. Ezek után a robotkar által elvégzendő lépések megadása egyszerű, csupán azt kell megmondani, hogy melyik mezőről milyen bábut melyik mezőre helyezzen.

5. A multimodális kommunikáció hatásának vizsgálata

A rendszerrel lefolytatott humán kísérlet célkitűzése az volt, hogy igazoljuk azt, hogy az ember sokkal jobban érzi magát multi-modális rendszerekkel való interakció során.

5.1. Kísérleti összeállítás

A kísérlet során 16 embert kértünk fel – 8 férfit és 8 nőt -, hogy játsszon a sakkozó robotunkkal. Mindannyian 18 és 25 év közöttiek voltak, és semmilyen előzetes információval nem rendelkeztek a sakkozó robottal kapcsolatban. Sakktudásuk a totális kezdőtől (csak a lépéseket ismeri) a hobbisakkozóig (4-5 lépésig előre gondolkodik, nyitásokat ismer) terjedt. Az alanyoknak két meccset kellett játszaniuk, egyet úgy, hogy csak a robotkar volt aktív, egyet pedig úgy, hogy a teljes rendszer üzemelt. A meccsek sorrendje az adott játékostól független volt, de úgy alakítottuk, hogy 4 férfi és 4 nő a robotkarral kezdjen, míg a másik 8 ember a teljes rendszerrel kezdjen.

A teszt megkezdése előtt röviden bemutattuk a rendszert a játékosnak, felhívtuk a figyelmét az ismert és előforduló hibákra. A játékos helyet foglalt a kialakított tesztkörnyezetben, az operátor pedig a játékos mögött elhelyezett ellenőrző monitorral szemben. A legoptimálisabb eset az lett volna, ha a játékos egyedül van a szobában, de a rendszer kisebb-nagyobb hibái miatt szükség volt rá, hogy jelen legyen az operátor a teszt lefolytatása közben, az esetleges felmerülő akadályok azonnali elhárítása miatt. Mivel a teszt egyik legfontosabb része a játékosról készült videofelvétel volt a két játék alatt, ezért minden esetben a játékos beleegyezését kértük, hogy elkészíthessük a felvételeket.

A játékok során az operátor csendben figyelte a játékost, minden esetben megpróbálta minimalizálni a kommunikációt. A megtett lépésekről jegyzőkönyvet készített, amelyeket az adott időszakban felmerülő megjegyzésekkel, gondolatokkal is kiegészített. Ha a rendszer nem ismert fel egy a játékos által megtett lépést, akkor azt az operátor direkt módon adta meg a vezérlőnek azt. Amennyiben a robotkar nem tudott felvenni egy bábut vagy rosszul rakta le, korrigált.

Hogy minél egységesebbek legyenek a tesztek, a játékos által megteendő lépések számát 15-ben határoztuk meg. Amennyiben a játékos folytatni szerette volna a játékot, akkor megtehette, de a felvételt leállítottuk. Minden esetben felajánlottuk a lehetőséget, hogy hamarabb is abbahagyhatják, ha nem érzik jól magukat, de erre a tesztek során egyszer sem volt példa.

Miután a játékos a második játszmát is lejátszotta, egy kétoldalas kérdőívet kellett kitöltenie. A kérdőív megtalálható a dolgozat függelékében. A 11. – 16. kérdéseire nem volt kötelező válaszolnia.

Az írásos anyagokat elektronikus formában is eltároltuk, kiegészítve egy dokumentummal, amely az operátor megfigyeléseit tartalmazta. Az elkészült felvételeket annotáltuk, elsősorban az alábbi eseményeket figyelve:

- játékos a sakktáblára néz,
- játékos oldalra néz,
- játékos gondolkodik,
- játékos lép,
- játékos a beszélő fejre néz,
- játékos a robotkarra néz,
- játékos nevet, mosolyog,
- játékos az operátorhoz beszél,
- játékos a beszélő fejhez beszél,
- játékos magában beszél (hangosan gondolkodik).

5.2. Felmerült problémák

Hogy a lehető legteljesebb képet adhassuk a tesztről, a lefolytatás során felmerült problémákról is kell beszélni. A játékosok jól viselték a kísérlet során felmerülő hibákat, azok alapvetően nem befolyásolták a teszt eredményeit.

A leggyakrabban előforduló hiba a robotkar pontatlanságából eredt. Analóg felépítéséből és jellegzetes programozásából adódóan tökéletesen pontosra nem lehetett konfigurálni. Ami érdekes volt, hogy minél tovább volt egyhuzamban használva, annál pontatlanabbá vált, így ha egy nap több teszt is történt, úgy minden játékos után újra kellett kalibrálni a rendszert. Ha

mellényúlt a robotkar, vagy rosszul tett le egy bábut, a játékosok szívesen segítettek neki, ritkán fordult elő, hogy megvárták, amíg az operátor közbeavatkozik. Még olyan eset is előfordult, hogy a játékos egy-egy ilyen hiba után incselkedett a beszélő fejjel, ezzel is alátámasztva a hipotézisünket, miszerint tényleg emberközelibbnek érzik a multimodális rendszereket az emberek.

A második leggyakoribb hibaforrás a sakkállás-felismerő volt. Ez a modul elég érzékeny volt a fényviszonyokra, ezért a legaprólékosabb konfiguráció és előzetes környezet-kialakítás ellenére is előfordult, hogy egy-egy bábut ellenkező színűnek ismert fel a játék folyamán. Attól kezdve, hogy egyszer is átváltott így egy színt, már nem tudta értelmezni a sakkállásokat, ezért az operátornak kellett minden egyes lépést direkt módon megadni a vezérlőnek. A színváltáson kívül még az is előfordult, hogy az aktív üzemmód folyamán a játékos hiába vette el a kezét lépés után a sakktábla fölül, a modul valamiért ezt nem érzékelte, és folyamatosan csak azt az üzenetet küldte a vezérlőnek, hogy a játékos még mindig lép. Ebben az esetben is az operátornak kellett közbeavatkoznia.

A rendszer tökéletesen ismerte a sakk szabályait, ennek ellenére valamiért nem tudta kezelni, ha a játékos sáncolt. Szabályos lépésnek vette, hogy a király 2 (rövid sánc) illetve 3 (hosszú sánc) mezőt lépett, ezt rögzítette is a saját adatbázisában, viszont a bástyát nem helyezte át az új helyére. Emiatt a játékosokat külön meg kellett kérni arra, hogy ne sáncoljanak, mert a rendszer nincs erre felkészítve.

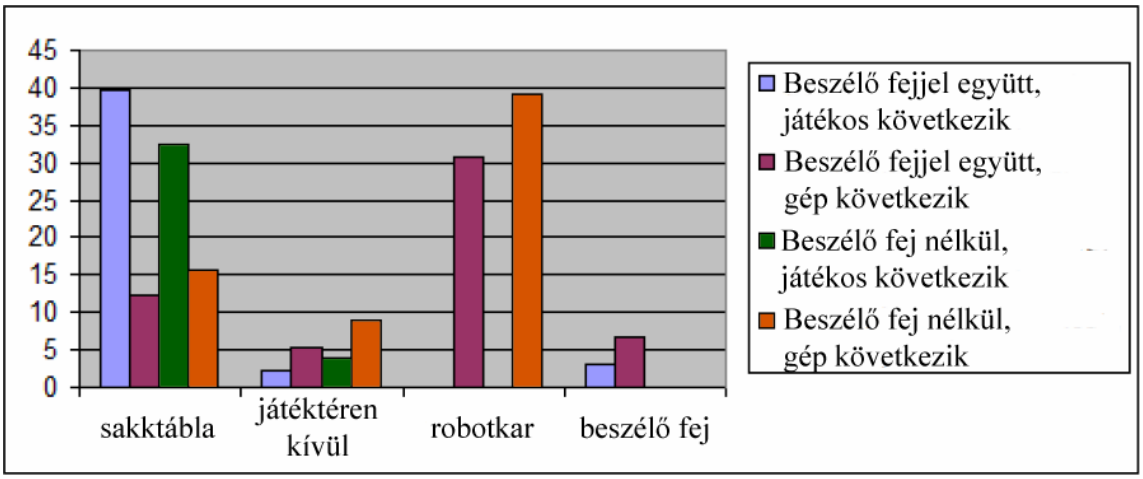
A fenti hibákat leszámítva, egyszer-egyszer az is előfordult, hogy lefagyott a vezérlő vagy a beszélő fej, de ez nem volt jellemző.

5.3. A teszt eredményeinek kiértékelése

A teszt eredményeinek kiértékeléséhez az annotációs fájlokat és a kérdőíveket használtuk fel. Ezek segítségével egyszerű statisztikákat készítettünk, melyeket a tesztek során kialakult szubjektív véleményekkel egészítettünk ki. A kapott eredményeket három különböző szemszögből vizsgáltuk meg:

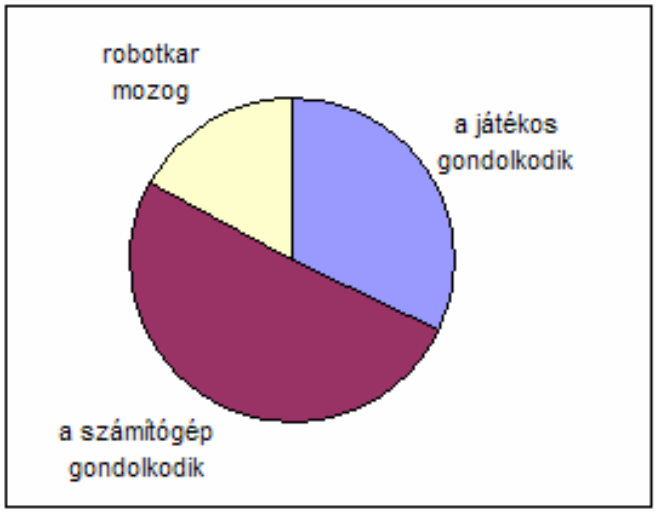
5.3.1. Hatással van a beszélő fej a játékosokra?

A tesztalanyok általános első reakciója a beszélő fejre pozitív volt. Kivétel nélkül megmosolyogták, és köszöntek neki. A sakkjáték jellege miatt az alanyok viselkedése attól függően változott, hogy kinek kellett a következő lépést megtenni. Mikor a játékos következett, az ideje nagy részében a sakktáblát nézte és gondolkodott. Mivel a tesztalanyok egyike sem volt profi sakkozó, amikor a gép következett, semmi sem volt, ami igazán lekötötte volna a figyelmüket. Ilyenkor általában a játéktéren kívülre figyeltek, például az operátorra néztek, az ellenőrző monitort figyelték, vagy egyszerűen csak bámészkodtak. Miután a gép döntött a következő lépéséről, mozgásba hozta a robotkart. Ezt a mozdulatsort már figyelemmel kísérték a játékosok, és újra a játékra koncentráltak. A beszélő fej egyik fontos szerepe az volt, hogy lekösse az emberek figyelmét, amíg a rendszer gondolkodott és semmi nem történt a játéktérben. A 10. ábra mutatja, hogy az alanyok százalékos arányban hova néztek a játékok során, attól függően, hogy ki következett.



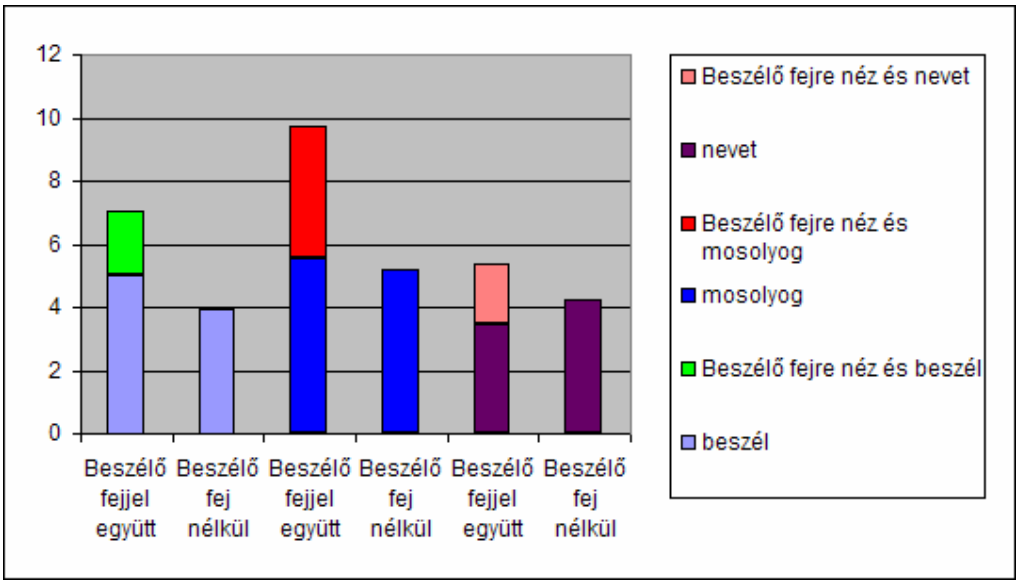
10. ábra. A játékosok figyelme a játékok során, százalékos arányban.

Ha belegondolunk, ember-ember sakkjátszmák esetén a lépések után szinte mindig rápillantunk az ellenfelünkre, egyfajta visszajelzést várva, hogy mit gondol a lépésünkről. Nagyon érdekes volt, hogy a játékosok ösztönösen a beszélő fejre pillantottak miután léptek a játékban. A beszélő fej nélküli játékok során az volt a benyomásunk, hogy a játékosok úgy érezték, mintha magukban játszanának, kevésbé koncentráltak a játékra, a lépések után pedig mivel nem volt kitől visszajelzést kapni, az operátorra vagy az ellenőrző monitorra irányították a figyelmüket. A 11. ábrán azt mutatjuk meg, hogy a játékosok átlagosan mikor néztek rá a beszélő fejre. Jól látható, hogy a „holtidőben" (amikor a rendszer még nem hozott döntést a következő lépésről), foglalta el leginkább a játékosokat a beszélő fej.



11. ábra. Mikor néztek a játékosok a beszélő fejre.

A játékosok hangulata is mindig pozitívabb volt, ha a beszélő fej ellen sakkozhattak. Többet nevettek, mosolyogtak, és beszéltek a játékok során. A jegyzőkönyvek alapján mondhatjuk, hogy jobban is sakkoztak a beszélő fej ellen. A játékosok 80%-ánál a beszélő fej elleni játék tovább tartott (több lépést tettek), mint a sima játék. Igazi ellenfélnek tekintették, komolyabban vették a játékot. A beszélő fej nélküli játékok során egyértelműen látszott, hogy legtöbbször csak túl akartak lenni a játékon, nem igazán kötötte le őket. A 12. ábrán azt szemléltetjük, hogy a játékosok többet beszéltek, mosolyogtak és nevettek a beszélő fejjel együtt történő játszmák során. Külön meg szeretnénk említeni, hogy ez a „többlet" a beszélő fejre nézés közben keletkezett. A beszélő fej hatását külön színnel jelöltük a megfelelő oszlopokon.



12. ábra. A játékosok hangulata a játékok során.

5.3.2. Személyként kezelték a beszélő fejet?

Már láthattuk, hogy a beszélő fej igenis hatással volt a játékosokra. Odafigyeltek arra amiket mondott, a rendszer szerves részének tekintették, nem pedig egy zavaró extrának, ami csupán a figyelem elterelésére jó. A játékosok szóltak az operátornak, ha a rendszer valamilyen apró hibájánál fogva nem mondott be egy-egy lépést, vagy nem szólt, hogy sakkot adott. A beszélő fej megnyilvánulásait a játékosok általában mosollyal, nevetéssel jutalmazták. Válaszoltak a fej kérdéseire, megdicsérték ha jó lépést tett, incselkedtek vele. Sőt, egyes tesztalanyok szabályosan meg is haragudtak, amiért a rendszer gyorsan elverte őket. Ilyenkor mindig a beszélő fejhez beszéltek.

A tesztek során a tesztalanyok fele először a beszélő fejjel játszott, azután a beszélő fej nélkül. Az esetek többségében az alanyok visszaidézték a második játszma során a beszélő fej jellegzetes mondatait: „Jól gondold meg.”, „Nehéz helyzet.”.

5.3.3. A beszélő fej hatása a játékélményre

A kitöltött kérdőívek alapján a játékosok 95%-a azt mondta, hogy jobb volt a beszélő fejjel játszani. Véleményük szerint a beszélő fej érdekesebbé, szórakoztatóbbá tette a játékot. Nagyon érdekes volt, hogy a játékok sorrendje nagyban meghatározta a játékosok hangulatait a teszt során. Ha először a beszélő fejjel játszottak, akkor a második, „csendesebb" játszma során többen is megjegyezték, hogy hiányzik nekik a beszélő fej. Unalmasabbnak, üresebbnek találták a második játszmát. Az alatt az idő alatt, amíg a rendszer a lépésen gondolkodott és semmi sem történt, a játékosok kiestek a ritmusból, elkalandozott a figyelmük. Nem kötötte le őket a játék. A robotkart is lassúnak találták, sokszor hamarabb akartak lépni, mint hogy a kar befejezte volna a mozgását. Egyes alanyok még azt is felajánlották, hogy meglépik a robotkar helyett a lépést, csak hadd haladjon a játék. A beszélő fejjel történő játszmák esetén azonban soha nem volt panasz a sebességre, ráadásul a rendszer gondolkodási ideje alatt is elszórakoztatták magukat a játékosok a beszélő fejjel.

Ha először csak a robotkarral játszottak a játékosok, és csak utána a beszélő fejjel, akkor jobban tűrték a csendes játékot. Lekötötte őket a robotkar újdonsága, de itt is megfigyelhető volt, hogy egy idő után inkább az operátorral foglalkoztak.

A játékok hossza a sorrendtől függetlenül szinte mindig a beszélő fejjel történő játékok javára dőlt.

5.3.4. További megfigyelések

A kérdőív azon kérdésére, hogy a beszélő fej érzelmesebb, vagy csendesebb, pókerarcúbb legyen-e, a résztvevők 70%-a az érzelmesebb mellett voksolt. Általános véleményük volt, hogy „egy csendes, pókerarcú beszélő fejjel játszani olyan, mintha egyedül játszanál". A legkifinomultabb megoldás az lehetne, ha a játékos a játszma elején nem csak a beszélő fej kinézetét választhatná meg, hanem a stílusát is. Így mindenki megtalálhatná a számára legmegfelelőbb robot sakkpartnert.

A fejlesztőknek szánt javaslatok többségében a beszélő fej arckifejezésének, kidolgozottságának javítását említették meg. A szókincs szűkösségével azonban az alanyok nagy többségének volt kifogása. A rendszer sebességére, illetve a robotkar pontatlanságára viszont mindössze egyetlen esetben panaszkodtak a kérdőívekben.

A sakkrobot jövőjéről minden alany pozitívan nyilatkozott. Úgy gondolják, hogy még kell fejleszteni és finomítani a rendszert, de mindenképp lesz jövője. Van olyan játékos, aki játéktermekben látná viszont szivesen, míg vannak olyanok, akik otthonra fogadnának el egyet.

A játékosok szerint a beszélő fej kidolgozottságát még finomítani kell, illetve a robotkar pontatlanságára volt még panasz, de egyetlen esetben sem jelezték a résztvevők, hogy ezek bármelyike a játékélmény rovására ment volna.

6. Összefoglalás

A multimodális sakkozóval végzett teszt sikeres és pozitív eredménnyel zárult. Mindenképpen meg kell említenünk, hogy 16 tesztalany számszerűleg kevés egy igazán pontos statisztikai teszt elvégzéséhez, így inkább a szubjektív vélemények domináltak. Ettől függetlenül is azért úgy gondoljuk, hogy sikerült néhány olyan tényt megmutatni, ami alapján kijelenthetjük, hogy a multimodális rendszerek pozitív hatással vannak az ember viselkedésére. Kellemesebbnek, kényelmesebbnek érzik az interakciót, a hangulatukra is pozitív hatással van a multimodalitás.

A Debreceni Egyetem sakkozógépe még nem kiforrott technológia. A szakdolgozat során rámutattunk annak előnyeire és hátrányaira egyaránt. Egy nagyon jó kezdeményezés, de egyértelmű, hogy van még munka vele. Ez a humánteszt ebből a szempontból is pozitívan zárult, mivel megerősített minket abban a hitben, hogy igenis van értelme tovább finomítani, fejleszteni, bővíteni a rendszert. A Pool oktató ötletéből kiindulva a sakkozót is lehetne alkalmazni oktatásra vagy sakk feladványok bemutatására. A beszélő fej funkciójának és képességeinek finomításával egy első osztályú multimodális rendszert lehetne létrehozni.

A jövő multimodális felhasználói felületeké. Egy rendszert nem elég csupán okosabbá, gyorsabbá, nagyobb kapacitásúvá tenni ahhoz, hogy az átlagemberek jobban ki tudják használni annak lehetőségeit. A kifinomultabb, emberközelibb felhasználói interfészek nélkül nem lesz kényelmesebb és barátságos a rendszer, így az emberek sem fogják bátrabban kezelni őket. A multimodalitás egyaránt megkönnyítheti az életünket (hangvezérelt fedélzeti számítógépek, interaktív intelligens térképek, hirdetőtáblák, stb.), szélesebb felhasználói körnek biztosít elérhetőséget (nem szakmabeliek, fogyatékkal élők, stb.), és emberközelibbé teheti az informatika eszközeit a nagyközönség számára.

Irodalomjegyzék

[1] W. Buxton, *There's more to interaction than meets the eye: some issues in manual input.*, User Centered System Design, (1986), 319–337.

[2] T. Brøndsted, P. Dalsgaard, L. B. Larsen, M. Manthey, P. McKevitt, T. B. Moeslund and K. G. Olesen, *A platform for developing Intelligent MultiMedia Applications*, Technical Report, Aalborg University, Denmark, (1998).

[3] C. Chang, C. Yang, Y. Yeh, P. Chung, J. Wang and j. Yang, *An intelligent guiding bulletin board system with real-time vision and multi-keyword spotting multimedia human-computer interaction,* in Proc of IEEE International Conference on Multimedia and Expo, 9–12, July, (2006), Toronto, Canada, 421–424.

[4] H. J. Charwat, *Lexicon der Mensch-Maschine-Kommunikation*, Oldenbourg Verlag, Kirchheim, Deutschland, (1992), ISBN 978-3486209044.

[5] S. Chatty, *Extending a graphical toolkit for two-handed interaction,* In Proc. of ACM UIST '94 Symposium on User Interface Software and Technology, 2–4, November, (1994), Marina del Rey, California, USA, 195–204.

[6] K. Davenport, *Target Pool*, Target Pool Productions, Marysville, Michigan, USA, (1992).

[7] P. J. Denning, *Is computer science science?*, Communications of the ACM, **48/4**, (2005), 27–31.

[8] A. Fazekas and I. Sánta, *Recognition of facial gestures from thumbnail picture*, in Proc. of NOBIM'2004, 27–28, May, (2004), Stavanger, Norway, 54–57.

[9] A. Fazekas and I. Sánta, *Recognition of facial gestures based on support vectore machines*, Lecture Notes in Computer Science, **3522**, (2005) 469–475.

[10] P. Gebhard, M. Kipp, M. Klesen and T. Rist, *Authoring scenes for adaptive, interactive performances*, in Proc. of 2nd International Joint Conference on Autonomous Agents and Multi Agent Systems, 14–18, July, (2003), Melbourne, Australia, 725–732.

[11] P. Gebhard, *ALMA - a layered model of affect,* in Proc. of 4th International Joint Conference on Autonomous Agents and Multi Agent Systems, 25–29, July, (2005), Utrecht, The Netherlands, 29–36.

[12] E. Hjelmas and B. K. Low*, Face detection: A survey*, Computer Vision and Image Understanding, **83/3**, (2001), 236–274.

[13] A. J. Hunt and A. W. Black, *Unit selection in a concatenative speech synthesis system using a large speech database*, in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, 7–10, May, (1996), Atlanta, Georgia, USA, I/373–376.

[14] E. R. Kandel and J. R. Schwartz, *Principles of Neural Sciences,* Elsevier Science Publishers, Amsterdam, The Netherlands, (1981).

[15] A. King, *A Survey of Methods for Face Detection*, kézirat, (2003), 33 pp.

[16] L. B. Larsen, M. D. Jensen and W. K. Vodzi, *Multi Modal User Interaction in an Automatic Pool Trainer*, in Proc of Fourth IEEE International Conference on Multimodal Interfaces, 14–16, October, (2002), Pittsburgh, Pennsylvania, USA, 361–366.

[17] H. Lausen, *LaserXI (2.2), documentation,* dokumentáció, Center for Advanced Technology (CAT), Roskilde, Denmark.

[18] R. Lienhart and J. Maydt, *An extended set of haar-like features for rapid object detection,* in Proc of International Conference of Image Processing, 22–25, September, (2002), Rochester, New York, USA, I/900–903.

[19] G. Littlewort, I. Fasel, M. Stewart Bartlett and J. R. Movellan, *Fully automatic coding of basic expressions from video*, Technical Report, UCSD INC MPLab, (2002).

[20] B. D. Lucas and T. Kanade, *An iterative image registration technique with an application to stereo vision (darpa)*, in Proc. of 1981 DARPA Image Undersatnding Workshop, April, (1981), 121–130.

[21] M.T. Maybury and W. Wahlster, *Readings in Intelligent User Interfaces (Interactive Technologies)*, Morgan Kaufmann, San Francisco, California, USA, (1998), ISBN 978-1558604445.

[22] R. R. McCrae, and O. P. John, *An introduction to the five-factor model and its applications*, Journal of Personality, **60**, (1992) 175–215.

[23] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought,* University of Chicago Press, Chicago, Illinois, USA, (1992).

[24] A. Mehrabian, *Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament*, Current Psychology: Developmental, Learning, Personality, Social, **14**, (1996), 261–292.

[25] T. B. Moeslund, M. Blidegn and L. Bakman, *Controlling a Movable Laser from a PC,* Technical Report, Aalborg University, Denmark, (1998).

[26] L. Nigay and J. Coutaz, *A design space for multimodal systems: concurrent processing and data fusion,* in Proc of INTERCHI '93 – Conference on Human Factors in Computing Systems, Proceedings, 24-29, April, (1993), Amsterdam, The Netherlands, 172-178.

[27] G. Olaszy, G. Németh, P. Olaszi, G. Kiss and G. Gordos, *PROFIVOX - A Hungarian professional TTS system for telecommunications applications*, International Journal of Speech Technology, **3**, (2000) 201–216.

[28] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge, England, (1988), ISBN 052-1353645.

[29] S. Oviatt, *Multimodal interactive maps: Designing for human performance*, Human-Computer Interaction, **12**, (1997), 93–129.

[30] S. Oviatt, A. DeAngeli, and K. Kuhn, *Integration and synchronization of input modes during multimodal human-computer interaction,* in Proc. of CHI '97 – Conference on Human Factors in Computing Systems, 22–27, March, (1997), Atlanta, Georgia, USA, 415–422.

[31] S, Oviatt, *Ten myths of multimodal interaction,* Communications of the ACM, **42/11** (1999), 74–81.

[32] M. Pantic and L. J. M. Rothkrantz, *Automatic analysis of facial expressions: The state of the art*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **22/12**, (2000), 1424–1445.

[33] R. Raisamo, *Multimodal Human-Computer Interaction: a constructive and empirical study,* Academic Dissertation, Tampere, Finland, (1999), 86 pp.

[34] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, Cambridge University Press, Cambridge, England, (1996).

[35] Zs. Ruttkay, A. Fazekas and P. Rigó, *Hungarian talking head according to MPEG-4*, in Proc. of Harmadik Magyar Grafika és Geometria Konferencia, 17–18, November, (2005), Budapest, Magyarország, 16–23.

[36] L. Schomaker, J. Nijtmans, A. Camurri, F. Lavagetto, P. Morasso, C. Benoît, T. Guiard-Marigny, B. Le Goff, J. Robert-Ribes, A. Adjoudani, I. Defée, S. Münch, K. Hartung, and J. Blauert, *A Taxonomy of Multimodal Interaction in the Human Information Processing System.* A Report of the Esprit Basic Research Action 8579 MIAMI, (1995), 195 pp.

[37] M. Schröder and A. Hunecke, *Creating German unit selection voices for the MARY TTS platform from the BITS corpora*, in Proc. of Sixth ISCA Workshop on Speech Synthesis, 22–24, August, (2007), Bonn, Germany, 95–100.

[38] M. Schröder, P. Gebhard, M. Charfuelan, C. Endres, M. Kipp, S. Pammi, M. Rumpler and O. Türk, *Enhancing Animated Agents in an Instrumented Poker Game*, kézirat, (2008), 8 pp.

[39] G. M. Shepherd, *Neurobiology, 2nd edition*. Oxford University Press, Oxford, England, (1988).

[40] J. Shi and C. Tomasi, *Good features to track,* in Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 21-23, June, (1994), Seattle, Washington, USA, 593–600.

[41] D. Silbernagel, *Tatchenatlas der Physiologie*. Thieme, (1979).

[42] P. Viola and M. Jones, *Rapid object detection using a boosted cascade of simple features",* in Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 8–14, December, (2001), Kauai, Hawaii, USA, 511-518.

[43] P. Viola and M. Jones, *Robust real-time object detection*, Technical Report, Cambridge Research Laboratory, (2001).

[44] R. Wasinger, *Multimodal Interaction with Mobile Devices: Fusing a Broad Spectrum of Modality Combinations*, (2006), ISBN 9783898383059.

[45] C. Wu and Y. Chen, *Multi-keyword spotting of telephone speech using a fuzzy search algorithm and keyword-driven two-level cbsm*, Speech Communication, **33/3**, (2001), 197–212.

[46] M.-H. Yang, D. Kriegman and N. Ahuja, *Detecting faces in images: A survey*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **24/1**, (2002), 34–58.

[47] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, *Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis*, in Proc. of Sixth European Conference on Speech Communication and Technology, 5–9, September, (1999), Budapest, Hungary.

[48] ACM Special Interest Group on Computer-Human Interaction, http://www.sigchi.org/, 2008-11-27.

[49] British Human-Computer Interaction Group, http://www.bcs-hci.org.uk/, 2008-11-27.

[50] Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk/, 2008-11-27.

[51] Horde3D Team, *Horde3D - Next-Generation Graphics Engine*, http://www.horde3d.org/home.html, 2008-11-27.

[52] IBM Corporation, *IBM ViaVoice for Windows, Personal Edition,* http://www.nuance.com/viavoice/personal/, 2008-11-27.

[53] IDEAS4GAMES, http://www.dfki.de/lt/iprojects/ideas4games.php, 2008-11-27.

[54] H. Noot and Zs. Ruttkay, *CharToon Software,* http://old-www.cwi.nl/projects/FASE/CharToon/index_netscape.html, 2008-11-27.

[55] Microsoft Corporation, *MS Agents*, http://www.microsoft.com/msagent/, 2008-11-27.

[56] Phalanx Engine, http://wbec-ridderkerk.nl/html/enginesindex.htm, 2008-11-27.

[57] The European Association for Cognitive Ergonomics, http://www.eace.info/, 2008-11-27.

[58] Wikipedia, http://en.wikipedia.org/wiki/Modality_(human-computer_interaction), 2008-11-27.

[59] XBoard, http://www.tim-mann.org/xboard.html, 2008-11-27.

[60] YouTube, http://www.youtube.com/watch?v=RyBEUyEtxQo, 2008-11-27.

Függelékek

1. Jegyzőkönyv részlet

	JÁTÉKOS LÉPÉS	ROBOT LÉPÉS	JÁTÉKOS ESEMÉNY	ROBOT ESEMÉNY
1.	H2H4	D7D5	Neveti az arcot	Nem ismerte fel a lépést!
2.	H1H3	C8H3 ütés	Bosszankodik, Elnézte a saját lépését	Nem ismerte fel a lépést!
3.	G2H3 ütés	E7E5	Válaszol a gép „Te jössz” mondatára	

2. Annotálás során keletkezett file szerkezete

00:00:00:00 00:00:14:10 Robotkart figyeli.

00:00:14:10 00:00:19:10 Beszélő fejre néz. Mosolyog.

00:00:19:10 00:00:20:01 Táblát néz. Gondolkodik.

00:00:20:01 00:00:23:08 Beszélő fejre néz. Beszél.

00:00:23:08 00:00:28:15 Táblát néz. Gondolkodik. Beszélő fejre pillant.

00:00:28:15 00:00:32:02 Játékos lép.

00:00:32:02 00:00:32:09 Táblát néz.

00:00:32:09 00:00:34:01 Beszélő fejre néz.

00:00:34:01 00:00:35:04 Oldalra néz.

00:00:35:04 00:00:49:10 Beszélő fejre néz. Táblára pillant.

00:00:49:10 00:00:55:12 Táblát néz. Gondolkodik.

00:00:55:12 00:01:07:02 Robotkarra néz. Beszél. Mosolyog.

00:01:07:02 00:01:14:10 Táblát néz. Gondolkodik.

00:01:14:10 00:01:20:12 Beszélő fejre néz. Beszél. Mosolyog.

00:01:20:13 00:01:27:05 Táblát néz. Gondolkodik. Robotkarra pillant.

3. A kérdőív

1. **A beszélő robottal élvezetesebb játszani, mint a néma robottal?**

2. **A beszélő robot hasonlít egy emberi játékosra?**

3. **A beszélő robot megnyilvánulásai idegesítettek-e a játék során?**

4. **A beszélő robot élvezte, amikor nehéz helyzetben voltam.**

5**. A beszélő robot beszéde jól érthető.**

6. **A beszélő robot arca jól hasonlít egy emberi arcra.**

7. **A beszélő robot arca jol kifejezte az erzelmeit.**

8**. A beszélő robot emberibb játékos, mint a sima robot?**

9. **A beszélő robot megjegyzései izgalmasabbá tették-e a játékot?**

10. **Gyakran éreztem, hogy a beszélő robot figyelte a játékomat.**

11. **Mi a véleménye a beszélő robotról?**

12. **Mit javasolna a fejlesztőknek?**

13. **Mit gondol a sakkrobot jövőjéről?**

14. **Egyéb megjegyzése, benyomása, javaslata?**

15. **Legyen a robot beszédes, érzelmes, vagy legyen inkább pókerarcú?**

16. **Legyen-e arca egyáltalán?**