MPEG-4 modell alkalmazása szájmozgás megjelenítésére

TAKÁCS GYÖRGY, TIHANYI ATTILA, BÁRDI TAMÁS,
FELDHOFFER GERGELY, SRANCSIK BÁLINT

Pázmány Péter Katolikus Egyetem, Információs Technológia Kar
{takacs.gyorgy, tihanyia, bardit, flugi, sraba}@itk.ppke.hu

Lektorált

Kulcsszavak: audiovizuális beszédfeldolgozás, fej animáció, multimodális kommunikáció

A cikk áttekinti az MPEG-4 szabványnak a fej és az emberi test alakjának és mozgatásával foglalkozó részének kódolási alapelveit. Bemutatja a nyílt forráskódú LUCIA dekódoló modell jellemzőit és egy speciális alkalmazását. Ebben az alkalmazásban beszédjelből előállított jellemzők vezérlik a fejmodellt, amely siketek számára szolgál kommunikációs segédeszközként. A cikk kitér az alkalmazási kísérletek mérési eredményeire is.

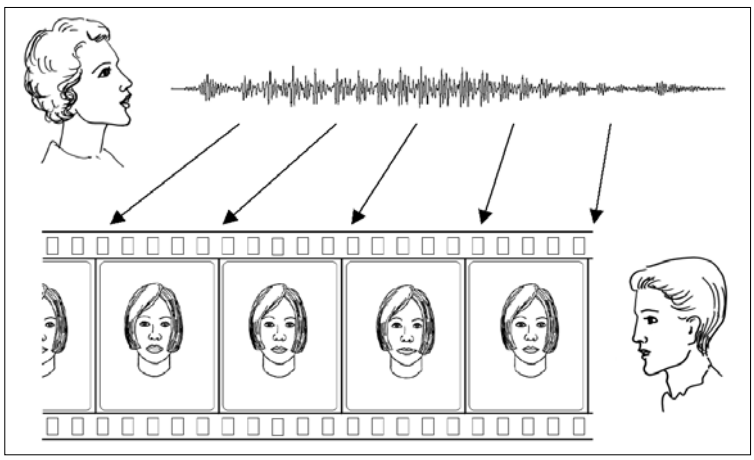
1. Bevezetés

Egyre több szempontból felmerül, hogy az emberi beszédátvitel nem csak hangtani jellemzőkre épülő folyamat. A multimodális kommunikáció megközelítése foglalkozik azzal, hogy a beszédinformáció közlése és felfogása nem csak hallható, hanem látható folyamatok együtteséből áll. A szabványosítás elérte ezt a területet is. Az MPEG-4 szabvány része a fej és az emberi test alakjának és mozgatásának kódolása. Kidolgoztunk egy speciális alkalmazást, amely végletesen használja a beszédfolyamat kettős természetét, azaz magából a hallható beszédjelből származtat látható beszédfolyamatot és ezzel siketek számára ad egy kommunikációs segédeszközt.

Ebben a cikkben nem a segédeszköz felépítését és működésének részleteit ismertetjük, hanem az MPEG-4 szabványon alapuló dekódoló rész technikai részleteit taglaljuk. Ehhez a 3. szakaszban a szabvány érdekesebb részeit ismertetjük és értékeljük, ami előkészíti a következő részt, melyben a nyílt forráskódú MPEG-4 kompatibilis LUCIA modell részletes ismertetése következik. Azok a finom részletek kaptak nagyobb figyelmet, amelyek a szájrólolvasás szempontjából kritikusak. Az alkalmazás hatásosságát mérési eredmények támasztják alá az 5. szakaszban kifejtettek szerint.

2. Előzmények

Egy teljes rendszert dolgoztunk ki, amely alkalmas arra, hogy beszédjelből mozgó száj képét állítsa elő. A mozgó szájról a siketek képesek a beszédet leolvasni. A rendszer ismertetése ugyanezen folyóirat számban megtalálható [3]. Itt azokat a részleteket és általános megfontolásokat taglaljuk, amelyek kifejezetten a megjelenítő egységre vonatkoznak.



1. ábra
Mozgó száj előállítási vázlata

Folyamatos beszédjelből mozgókép folyamot hozunk létre. Ez egy olyan transzformáció, amelynek lényegi részét egy neurális hálózat hajtja végre. A neurális hálózat komplexitását korlátok között kellett tartani, ezért elengedhetetlen volt az emberi beszédfolyamat lényegét jól megragadó, tömör és hatékony leírása a vizuális beszédnek.

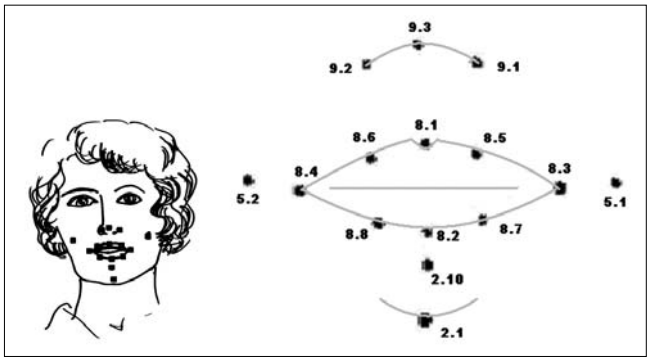
A neurális hálót előfeldolgozott hangadatokkal tanítottunk és képi koordinátákon vártunk a kimeneteken. Főkomponens analízist alkalmaztunk a képi koordináták tömör reprezentálására. Így mindössze 6 kimeneti jellemző kisebb, mint 2% hibával leírta a szükséges képi koordinátákat. A rendszer fejlesztésében külön kezelt probléma volt a mozgókép megjelenítés modellje.

3. Az MPEG-4 szabvány fejmozgások tömörített kódolására

Az MPEG (Moving Picture Expert Group) szabványok fő célja a hang és videó jelek tömörítése. A tömörítés alapvető követelményei a hatékonyság és élethűség. A multimédia-alkalmazásokban elterjedt az MPEG-2 kódolás. Az ezt meghaladó MPEG-4 kódolás is ígéretes jövő előtt áll, ugyanakkor céljainkat közvetlenül támogatja. Az MPEG-4-et nem csak nagy tömörítésre alakí-

tották ki, hanem figyelembe vettek olyan multimédia alkalmazásokat is, mint a 3D-s jelenetek, animációk, szintetizált hangok, képek, szövegek, grafikák külön vagy akár együttes kezelése és élethű megjelenítése.

Az MPEG-4 szabvány egyik legösszetettebb része a fej és az emberi test megjelenítése és mozgatása (Face and Body Animation, FBA). Az FBA-ra vonatkozó szabványrész leírja az arc és a test alakjának és mozgásának kódolási alapelveit. Az FBA egyik legfontosabb tulajdonsága tehát, hogy nem adja meg pontosan a kódolási és a dekódolási eljárást, csak a küldött adat formáját és értelmezését.



2. ábra Felhasznált tartópontok

Az MPEG-4 szabvány az arc modelljét az arc normál állapotával írja le, megad több tartópontot (Feature Point, FP) és az arc mozgását leíró paramétereket (Facial Animation Parameter, FAP), melyek lényegében a normál archoz képesti elmozdulást jellemzik *(2. ábra)*. Az elmozdulások méretét és arányát a szabvány szerint mindig az emberi arcra jellemző alapvető méretek alapján fejezi ki. A szakirodalomban ennek elterjedt rövidítése FAPU (Face Animation Parameter Unit – lásd a *6. ábrát)*. A FAPU-kat az arc olyan jellegzetes távolságaiból kell számolni, mint például a szemgolyók távolsága vagy a száj szélessége.

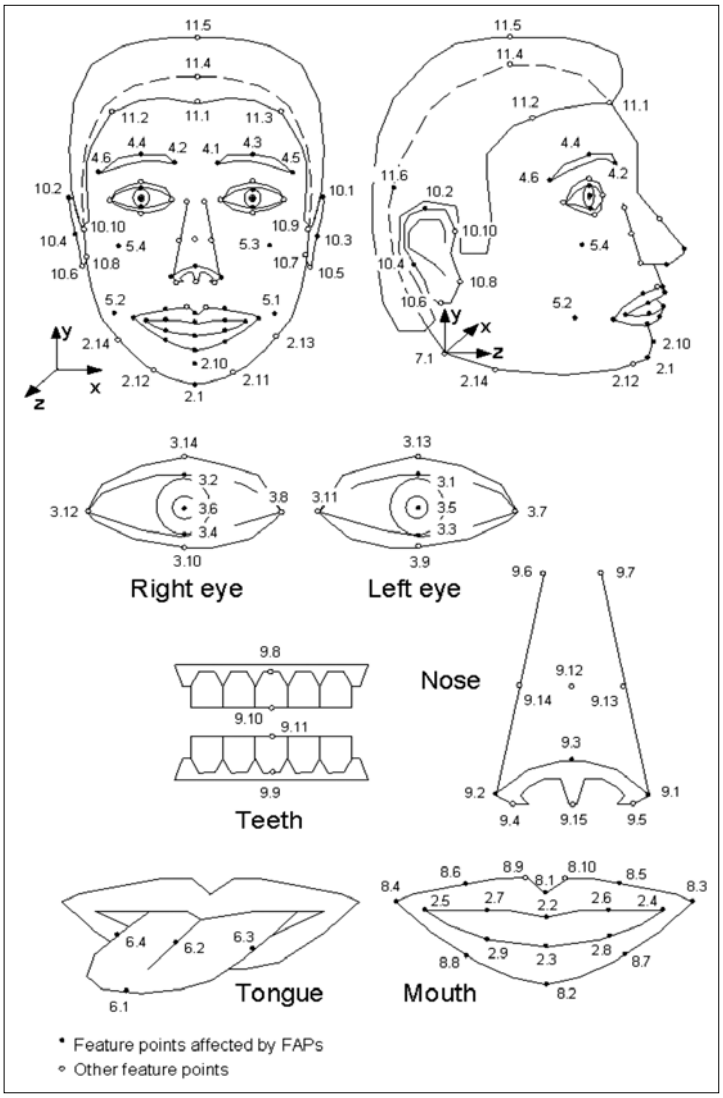
A szabványban 84 tartóponttal írják le az arcot. (Az adatbázisunk összeállítása során mi 15 FP-t használtunk a száj és környékének leírására).

A tartópontok fő feladata, hogy referenciaként szolgáljanak a FAP-ok számára. A FAP-ok által leírt összetett mozgások mindig a normál tartópontok által leírt fejre vonatkoznak. A normál fej csukott szájat és semleges arckifejezést jelent. Vannak olyan FP-k is, melyekre egy FAP sincs közvetlen hatással (például az orr szélei). Ezeket mindössze az arc alakjának meghatározására használják. Az FP-ket minden MPEG-4 kompatibilis modellen a *3. ábra* alapján kell elhelyezni.

FAP-ból a szabvány 68-at különböztet meg, melyet 10 csoportba sorol az alapján, hogy az arc mely részét mozgatja.

Az első két FAP magas szintű paraméter. Ez azt jelenti, hogy ezekkel előre beállított komplexebb mozgást lehet kivitelezni. Az első FAP egy megadott vizéma szerinti megjelenést határoz meg. A vizéma a fonéma képi megfelelője. A második FAP a hat alap érzelem megjelenítésére szolgál, úgy mint öröm, bánat, harag, félelem, undor és meglepetés. Tovább érzelemkifejezéseket az alap érzelmek keveréséből lehet megjeleníteni.

A többi FAP alacsony szintű. Ezek abban különböznek a magas szintű FAP-októl, hogy itt a mozgás irányát és amplitúdóját kell megadni, nem pedig egy összetett feladatra előre összeszerkesztett mozgásvezérlést kell kezdeményezni. Az alacsony szintű FAP-ok általában egy-két tartópontot mozgatnak. Előfordul olyan FAP is, amely az összes FP-t mozgatja, ilyen például a fej forgatása. Az alacsony szintű FAP-oknál a szabvány meghatározza, hogy a mi a hozzá illő FAPU, amiből a mozgás a mérték alapja. A FAP előjele a tartópont mozgásirányára vonatkozó információt hordoz, például a száj nyitására vonatkozó paraméterek pozitív, a zárásra vonatkozóak negatív előjelűek. Ez független attól is, hogy a tartópont a száj alsó vagy felső részéhez tartozik. A mozgatás lehet eltolás, forgatás vagy skálázás.



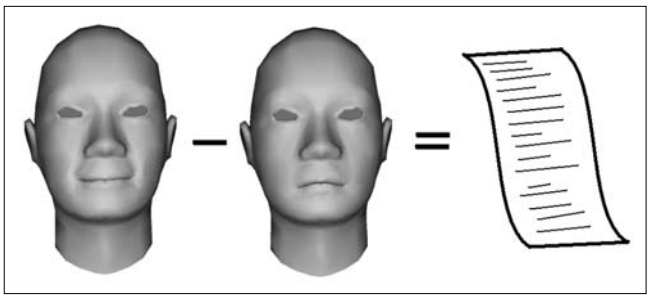
3. ábra
A tartópontok szabványos elhelyezkedése a fejen

4. A LUCIA modell

A legtöbb modell, legyen az két- vagy háromdimenziós, hálókból áll. A háló (mesh) több egymáshoz illeszkedő nem feltétlenül egy síkban levő sokszöget tartalmazó felület. A hálóban a csúcspontok koordinátáin kívül a lapok, az élek és a csúcsok illeszkedési viszonyait is nyilván kell tartani [12].

A modell felületi jellemzői, textúrája erre a rácsra van ráhúzva. Ahogy mozgatjuk a háló csúcspontjait, úgy mozog vele a textúra is. Ám az MPEG-4 szabványban csak az FP-k mozgatására van mód, az egyes hálókéra közvetlenül nincs. Egy modell tetszőleges számú és finomságú hálóból állhat, a szabvány erre nem terjed ki. Minden MPEG-4 kompatibilis fejmodell azonban azonos tartópont rendszerre épül. A hálók mozgatása a tartópontok mozgatásával történik.

A LUCIA modellt Cosi vezetésével olasz kutatók fejlesztették ki [1]. Ez egy nyílt forráskódú mozgó fejmodell. A LUCIA egy MPEG-4 megvalósítás, ami alkalmas vizémák és érzelmi állapotok FAP paraméter alapú közvetlen megjelenítésére. Az MPEG-4 modell tömörítést kifejtő (decompress) része egy grafikus modell mozgatási feladat, alapvetően az *5. ábra* szemléltetése szerinti információk felhasználásával képes átvinni a mozgás jellegzetességeit. A szabványosított eljárás során az alaphelyzetű fej teljes képének meghatározása és vevő oldalra történő átvitele valósul meg, és a továbbiakban csak az alaphelyzettől történő eltérések átvitelére van szükség a tömörített adatközlés során.

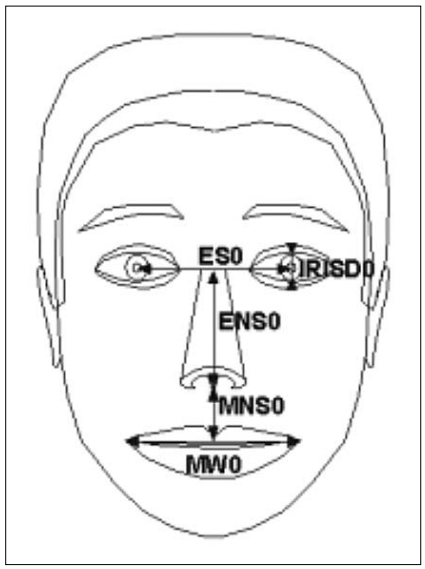


4. ábra
Az MPEG-4 rendszerű tömörítés koncepciója

Az MPEG-4 tömörítési folyamat *(4. ábra)* azon az elven működik, hogy a tömörítendő mosolygós fej lényeges paramétereinek valamint az alaphelyzetű fej paramétereinek különbségéből meghatározza a tömörített jellemzőket. Az MPEG-4 koncepció szerint ez a jellemzősor a fej alakjától és környezetétől független adatokat tartalmaz.

A visszaállítási folyamat *(5. ábra)* során a tömörített jellemzőkhöz, amely jelenleg a mosolygás adatait tartalmazza „hozzáadva" egy tetszőleges alaphelyzetű fej paramétereit, egy mosolygós fej képét kapjuk. Az alaphelyzetű fej meghatározó adatai között kell elhelyezni a felületi jellemzőket valamint az esetleges további adatokat, mint például a modell haja, szeme stb. A visszaállítás során kell létrehozni a felületeket azok megvilágítástól függő színezésével együtt [2].

Az MPEG-4-ben a tömörítés során meghatározott és felhasznált távolság mértékrendszer *(6. ábra)* lehetőséget biztosít arra, hogy a tömörített információ felhasználásával tetszőleges más alaphelyzetű fejre lehessen alkalmazni a visszaállítást, és így lehessen változtatni a visszaállítás folyamatát.



6. ábra
Az emberi arcra jellemző méretek

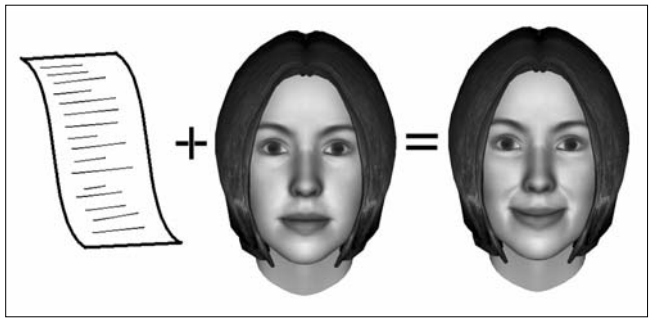
ES0= *szemgolyók távolsága;*
IRISD0= *az írisz átmérője;*
ENS0= *az orr hossza;*
MNS0= *orr és a száj távolsága;*
MW0= *a száj szélessége*

Az ES0; IRISD0; ENS0; MNS0; MW0; távolságok határozzák meg az adott arcberendezésen alkalmazandó távolságegységek halmazát. A távolságmérésnek ez a módszere biztosítja azt a lehetőséget, hogy a visszaállítás során az eredetitől jelentősen eltérő felépítésű alaphelyzetben álló fejre is visszaállíthatók legyenek a tömörített információk.

Az MPEG-4 szabványnak köszönhetően az arc mozgásához nem kell képkockáról képkockára megadni a videó minden egyes pixelét, mindössze a mozgatott FP-khez tartozó FAP-okat kell továbbítani. Ennek köszönhetően igen alacsony sávszélességen keresztül is elérhető a real-time arcanimáció.

Az MPEG-4 szabvány előnyeit leginkább internetes alkalmazásokban használják. Találkozhatunk olyan rendszerrel, mely az e-mail-eket alakítja át olyan videóvá, ahol az általunk kiválasztott személy mondja el az üzenetet. Léteznek olyan alkalmazások, melyek internetes áruházakban „eladókat" alkalmaznak, vagyis egy MPEG-4 szabványú modell ad segítséget az árakról, a minőségről vagy éppen a készletről.

5. ábra
Az MPEG-4 rendszerű visszaállítás koncepciója

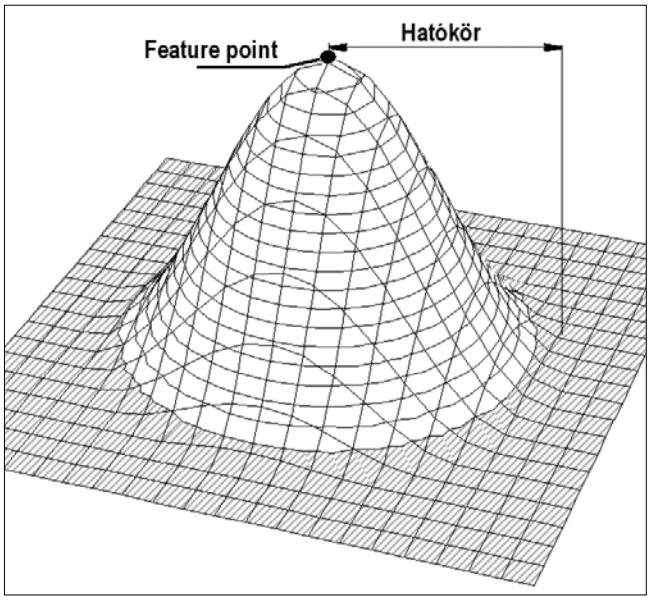


A szintetikusan létrehozott szájmozgás megjelenítésére felhasznált LUCIA modell egy szokásos 3D grafikus modell, amely animálható és így a céljaink megvalósítására alkalmas.

Az adatbázis felvételnél az arcra felfestett pontok kis mértékben eltérhetnek a szabványban előírt tartópontok helyétől. Ezt a hibát úgy korrigáltuk, hogy a tartópontokat ráillesztettük a felfestett pontokra úgy, hogy szintetizáláskor egybeessenek.

Az *7. ábra* bemutatja egy eredetileg vízszintesen elhelyezkedő négyszögháló felhasználásával készített animáló eljárás hatását abban az egyszerű esetben, ha az eredeti helyzetből függőleges irányban felfelé kívánjuk elmozdítani a síknak egyetlen pontját.

7. ábra
Az FP függőleges elmozdításának hatása a vizszintes felületre



A hatókörükben összeérő, egymás mellé eső pontok egymásra hatását megfelelő súlyozással kell kiküszöbölni. Elképzelhető, hogy egy hálórészt több tartópont is mozgatni akar. Ilyenkor természetesen súlyozottan összegződnek az elmozdulások. A súlyozás meghatározásánál az elmozdítást eredményező pont hatását annak távolságával fordított arányban határozzuk meg, ez a módszer azt eredményezi, hogy a modell rácspontjainak elmozdulását a FP-hez közeli rácspontok esetén nagymértékben az FP helyzete határozza meg. A vázolt eljárással lehetséges kijelölt pontok és hozzájuk tartozó területek rögzítése. Ilyen technikával oldottuk meg a 3D-s LUCIA fej állának mozgatását.

Annak érdekében, hogy az állcsont a megfelelő forgáspont körül elforduljon, az állcsúcsot (2.1-es FP) mozgattuk. Az állcsont miatt nagy hatókörrel kell a 2.1-es FP-t mozgatni, aminek az a hatása, hogy szemből nézve úgy tűnik, mintha a fej egész álla leesne. A jelenséget meg lehet szüntetni oly módon, hogy az arc körvonalához tartozó 2.13 és 2.14-es FP-t minden irányban 0-val mozdítjuk el, ennek hatására a 2.13 és 2.14-ös FP-k nagy súllyal helyben tartják az arc körvonalát és csak elenyésző mértékben mozdul a környezetük a 2.1 és 2.10 pontok mozgatásának hatására. Az alkalmazott technika teljesen kiküszöböli az áll leesésének a jelenségét.

A LUCIA modell tartalmazza az alsó és felső fogsort valamint a nyelvet is. Az alsó fogsor mozgatását kizárólagosan az állcsúcs mozgása határozza meg, a felső fogsor mozgatását az orr megfelelő pontjaihoz kötöttük, így annak elmozdulása minimális, hiszen az orr középpontját tekintettük a munka során referenciának. A nyelv mozgatásával a projekt nem foglalkozott.

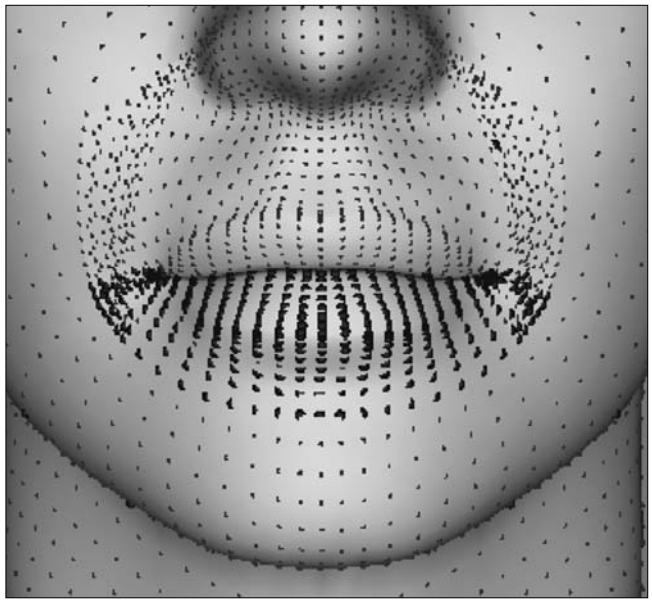
A mozgatandó felületen a háló törése, szakadása (például szem, száj) azt a problémát jelenti, hogy a szakadási vonalnál tovább azon átnyúlva nem alkalmazhatjuk az előzőekben vázolt módszert. Például az alsó ajak mozgatása nem húzza magával a felső ajkak hálórészét, pedig azok a hatókörön belül esnek. Ezzel a módszerrel kezelhető a száj, a szemek természetes nyitása.

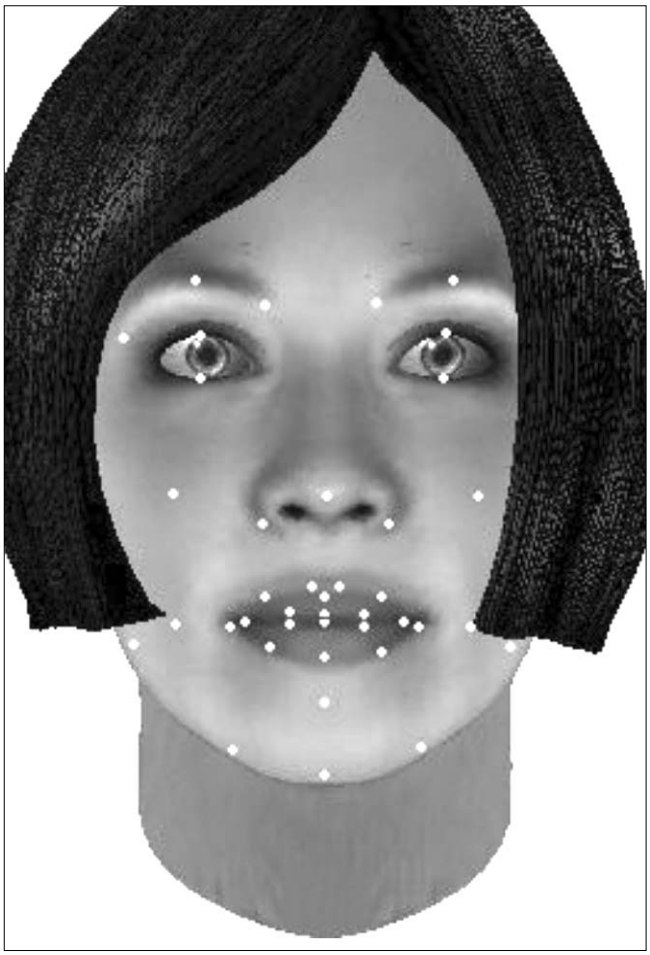
Azt a módszert választottuk, hogy minden mozgatott FP-hez meghatároztuk a modellünk egy-egy háló csúcspontokkal leírt egybefüggő részét. Ez jelentősen gyorsítja a mozgató algoritmusokat, mivel nem kell a teljes fej összes rácspontjának távolságát meghatározni minden egyes FP helyzetétől, hanem elegendő a kijelölt részhalmaz pontjainak a figyelembe vétele a számítások folyamán. Az alsó és felső ajakrész szétválasztását szemlélteti a *8. ábra*.

Az ábrán sötétebb pontok jelölik a száj alsó szélét. Ezekre a pontokra hatnak, ezeket mozgatják a 8.2; 8.7; 8.8 tartópontok (lásd 3. ábra).

Minden FP-hez tartozik egy mozgatási hatókör – egy gömb alakú térrész – és azon a hatókörön belül levő rácspontok elmozdulását határozza meg az adott FP elmozdulása az MPEG-4 rendszerben maghatározott skálázás szerint.

8. ábra
Az alsó szájszélet meghatározó hálópontok





9. ábra LUCIA modellen alkalmazott FP-k

A 3D grafikus modellt az MPEG-4 rendszernek megfelelően ki kell egészíteni a 3 dimenzióban értelmezett FP-kel, és azok hatókörének meghatározásával, valamint az egyes FP-k által mozgatható rácspontok halmazával *(9. ábra)*.

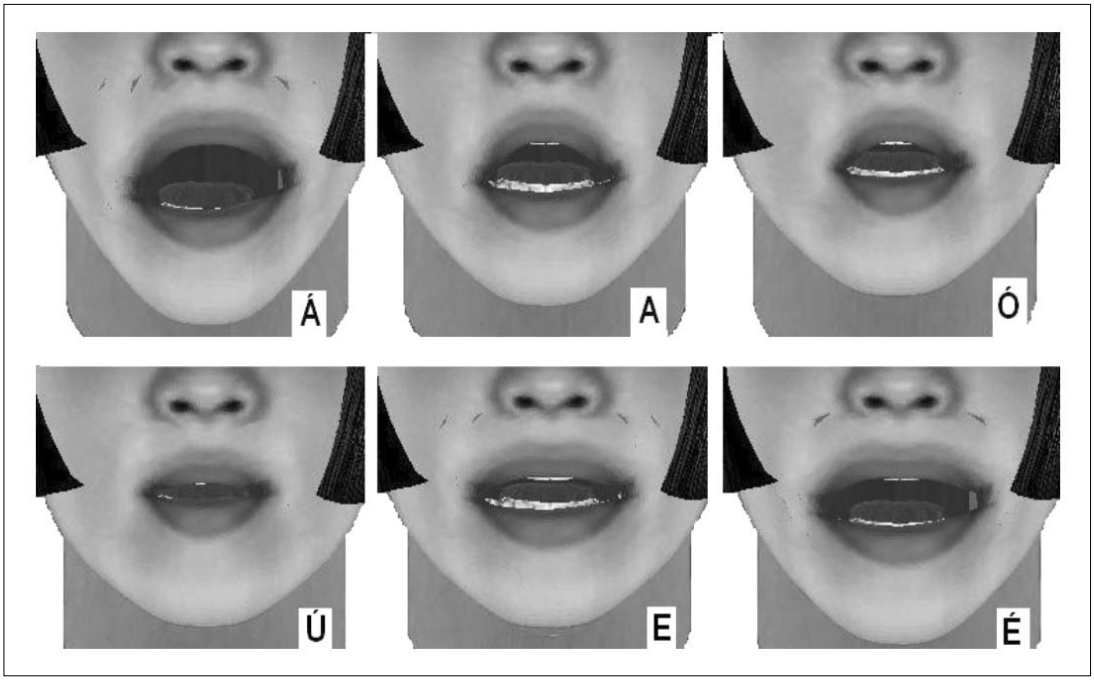
A beszédjelből szájmozgást előállító projekt során az előzőekben részletezett módon kialakított LUCIA modellt alkalmaztuk. A projekt eredeti elképzelései szerint a megvalósításkor a beszédjelből közvetlenül az FP mozgatási paramétereket állítottunk elő, tehát nem volt szükség arra, hogy az egyes vizémákat külön-külön meghatározzuk és előállítsuk, de a hosszan kitartott magánhangzók tiszta fázisainál jól megkülönböztethető szájállásokat hozott létre a fejmodell *(10. ábra)*.

5. Mérési eredmények, következtetések

Az animációs rendszerünk komponenseinek ellenőrzésére szájról olvasási kísérleteket végeztünk siket tesztalanyokkal. A szájról olvasási feladatok nehézségét úgy állítottuk be, hogy körülbelül 95% és 100% közötti felismerési arányt kapjunk a vetített eredeti videó felvételekre, hogy referenciaként szolgálhasson.

Ilyen jó arányt az előzetes kísérletek leírásánál [3] már ismertetett módon a felismerendő szövegben használt szókincs és nyelvtan erős szűkítésével, valamint egy jól artikuláló jeltolmács szerepeltetésével értünk el. Ezután mértük a felismerési arányt, úgy, hogy a videófelvétel helyett az animált beszélőfej-modell volt látható, ugyanakkor minden más kísérleti körülményt változatlanul hagytunk.

A fejmodellre való áttérés két lépcsőben történt. Az elsőben a felvételeken festékpöttyel megjelölt MPEG-4 pontok koordinátáit igyekeztünk átvinni a modellre: vagyis a fejmodell vázát képező háló megfelelő csomópontjait minden képkockán a felvételen mért koordinátájú pozíciókba mozgattuk. Ezzel azt kívántuk elérni, hogy a modell közvetlenül utánozza a jeltolmács artikulációját, ebben a lépésben a hang még nem játszott szerepet.



10. ábra Példák magyar nyelvű jellegzetes magánhangzó szájállásokra (viziémákra)

A második lépcsőben a beszédhang alapján számított koordináták szerint vezéreltük a fejmodellt. Itt már csak a hangbemenetre volt szükség az animáció előállításához [3].

A kísérlet során a felismerési arányok a következők szerint alakultak:

– eredeti felvételek (referencia): 97,1%;
– animáció a jeltolmácsra festett tartópontok alapján vezérelt modellnél (1. lépcső): 54,9%;
– animáció a hang alapján (2. lépcső): 47,9%.

Jelen cikk szempontjából a felvételekről a LUCIA modellre való áttérés, vagyis az első lépcső érdekes. Itt elég jelentős romlás tapasztalható a felismerési arányban, ennek lehetséges (valószínű) okaira térünk ki röviden.

Megállapítható, hogy az általunk kiválasztott és a felvételeken megjelölt MPEG-4 pontok helyzete hiányosan (információ veszteséggel) reprezentálja azokat a látható beszédképzési jellemzőket, melyek a szájról olvasásban szerepet játszanak. A kísérletek után minden alkalommal kikértük a résztvevő siketek véleményét, hogy mely tényezők gátolták őket leginkább a szájról olvasásban.

A felvételek és az animációk között talán a legfontosabb különbség, hogy a fejmodellnek nincs nyelve. De ha a LUCIA modell lehetővé tenné a nyelv animálását, akkor is problémát jelentene, hogy nincsenek referencia adataink a nyelv pillanatnyi helyzetéről, nem tudjuk, hogyan is kéne mozgatni. A nyelvre a felvételeken nem festhettünk pontot. A nyelv hiányában például a *kilenc* vagy a *nulla* szavak felismerése gondot okozott az animáció esetében, míg a felvételeken jól látható volt a nyelv föntről lecsapódása az *l* hang után, így valamennyi tesztalanyunk könnyedén felismerte azokat.

A másik problémánk volt, hogy a felvételekhez csak az ajkak külső kontúrján tudtunk pontokat megjelölni, beljebb nem. Ezek viszont az ajakkerekítésről kevés információt tartalmaznak. Az animációkon elsősorban az ajakkerekítéses magánhangzók (pl. *u, ü*) voltak kifogásolhatók. Szintén a pontok elhelyezésére vezethető vissza, hogy nincs elegendő információnk a fogak láthatóságáról. Pedig elsősorban ettől függ az ajkakon belüli terület világossága, ami egy igen karakteres és könnyen észlelhető vizuális jellemző [4].

Az MPEG-4 szabvány eredeti célja egy olyan modell megalkotása, aminek segítségével tömöríteni, majd rekonstruálni lehet mozgó fej adatokat. Munkánk során megoldottuk, hogy a szabványra építve olyan minőségben mozgatható a száj és környezete, hogy ennek alapján a siketek a beszédet képesek szájról leolvasni.

Fontos eredménynek tartjuk azt is, hogy az animáció működik akkor is, ha nem képpontok mintavételezése alapján származtattuk a tartópont paramétereket, hanem beszédjelből számoltuk. Az eredményeink azt mutatják, hogy igen kis különbség van a mintavételezéssel vezérelt arc, és a beszédjel alapján vezérelt arcmodell felismerhetősége között.

További fejlesztést igényel a fejmodell finomítása. A száj külső körvonalán túl a belső kontúr, fogak vagy nyelv láthatósága tűnik a továbblépés első lehetőségének.

Köszönetnyilvánítás

A szerzők ezúton is kifejezik köszönetüket a Nemzeti Kutatási és Technológiai Hivatalnak a 472/04 szerződés keretében nyújtott támogatásáért.

Irodalom

[1] Cosi P., Fusaro A., Tisato G., "LUCIA: a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model", Proc. of Eurospeech 2003, Geneva, Switzerland, September 1, 2003, Vol. III, pp.2269–2272.

[2] Szirmai-Kalos László, Antal György, Csonka Ferenc, „Háromdimenziós grafika animáció és játékfejlesztés", ComputerBook Kiadó Kft., Budapest 2003.

[3] Takács György, Tihanyi Attila, Bárdi Tamás, Feldhoffer Gergely, Srancsik Bálint: „Beszédjel átalakítása mozgó száj képévé siketek kommunikációjának segítésére" Híradástechnika 2006/3, pp.31–37.

[4] László Czap, János Mátyás, „Virtual Speaker" Híradástechnika – Selected Papers 2005/6, pp.2–5.

[4] I. Pandzic, R. Forchheimer, „MPEG-4 Facial Animation: The Standard, Implementation and Applications", Wiley, 2002.