Virtuális bemondó

CZAP LÁSZLÓ

Miskolci Egyetem, Villamosmérnöki Intézet, Automatizálási Tanszék
czap@mazsola.iit.uni-miskolc.hu

Reviewed

Kulcsszavak: beszédérthetőség, vizuális beszédszintézis, beszéd- és hallássérültek távközlése

Magyar nyelvű, vizuális szövegfelolvasó fejlesztéséről számol be a cikk. Az animáció háromdimenziós fejmodell mozgatásán alapul. Az artikuláció kialakításához felhasználtuk a fellelhető hangalbumok anyagát, a dinamikus vizsgálatnál saját vizuális beszédfelismerési kutatási eredményekre támaszkodtunk. A koartikulációs hatások figyelembe vételéhez a jellemzőket domináns, rugalmas és határozatlan osztályokba soroltuk, ezek alapján határoztuk meg a mozgásfázisok közötti interpolációt. A természetesség javítása érdekében többek között álvéletlen fejmozgásokat és pislogást programozunk. A fejmodell működtetése során megvalósítjuk alapérzelmek kifejezését is.

1. Bevezetés

Mindenki előtt ismert, hogy a beszéd érthetőségét javítja, ha látjuk a beszélő személy arcát, ezzel együtt az artikulációját. Ez a vizuális információ különösen sokat segít zajos környezetben és hallássérültek esetében. A gépi beszédkeltés jól kidolgozott rendszereinek természetes kiegészítője a mesterséges beszélő fej. Az arcanimáció megvalósítása a beszédartikuláció modellezésére mindössze két évtizeddel ezelőtt kezdődött. A mai szemmel kezdetleges eszközökkel végzett első próbálkozások a vizuális beszédszintézis úttörőmunkáját jelentették. A 3D modellezés fejlődése, a számítástechnikai eszközök kapacitásának robbanásszerű bővülése és a természetes artikuláció analízise életszerű, fotorealisztikus finomságú modellek kidolgozását tette lehetővé.

Az elmúlt évtizedben a terület dinamikusan fejlődött, egyre több alkalmazás jelenik meg. Az ember-gép kapcsolatban új távlatokat nyithat az audio-vizuális beszédszintézis és beszédfelismerés. Dialógus és oktató rendszerekben az érthetőséget és az attraktivitást nagyban javítja a beszédanimáció. Multimédiás alkalmazásokban a virtuális bemondó vagy szereplő tágítja a művészi szabadság határait. Hallássérültek beszélni tanítását segítheti a helyesen artikuláló virtuális bemondó, amely átlátszó arcával a természetes beszélőnél jobban megmutatja a hangképzés részleteit.

Hangvezérelt beszélő fejek fejlesztésén dolgoznak hallássérültek segítésére távközlési alkalmazásokban. A fejlett magyar nyelvű akusztikus beszédszintézis mellett hiánypótló célzattal kezdtünk vizuális beszédszintetizátor fejlesztéséhez.

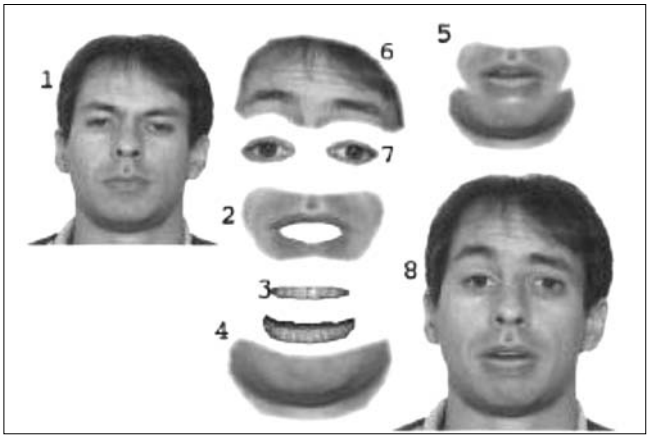
1. ábra Fotorealisztikus és transzparens megjelenítés

2. A beszédanimáció

Az első működőképes vizuális beszédszintetizátorok kétdimenziós modell mozgásfázisainak előállítására épültek, kezdetben előre tárolt képek előhívásával. A kulcskeretek közötti fázisokat gyakran képmorfológiai módszerekkel állították elő. A kétdimenziós modell nem teszi lehetővé a természetes fejmozgások, a beszédet kísérő gesztusok és érzelmek kifejezését.

A testmodellezés fejlődése a háromdimenziós modellezésre terelte a kutatók figyelmét.

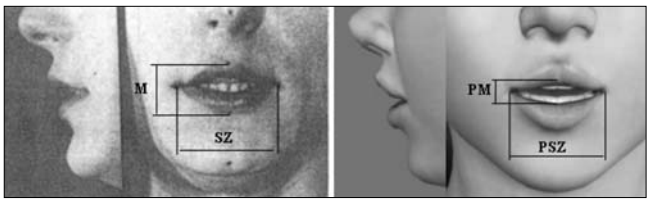
2. ábra Kétdimenziós fejmodell elemei [1]



A 3D modellek egyik típusa az arcizmok megfeszítésével szimulálja az arckifejezéseket. Az ilyen modellek valósághű eredményt nyújtanak, de a kívánt arckifejezés előállítása rendkívül számításigényes és a valóságos izomtónusok nem mérhetők. Ma még ígéretesebb a pusztán felületi hatásokat utánzó, a bőrszövettel borított drótváz alakítására alapozott animáció. Ennek paraméterei megfigyeléssel, vagy képfeldolgozási módszerekkel természetes beszélők képeiről leolvashatók [2]. Minden modell mozgatásánál külön figyelmet kell fordítani a jellemzők összehangolt változtatására, mert könnyen természetellenes hatás alakulhat ki.

2.1. A beszéd vizuális alapegysége

A beszéd legkisebb akusztikus egységének, a fonémának (hangzó) vizuális megfelelője, a vizéma. A vizémák készlete szűkebb a fonémákénál, hiszen néhány fonéma artikulációja vizuálisan megegyezik. Nem látható például a zöngésség, de a képzés helyében megegyező, időtartamban vagy intenzitásban eltérő hangok is azonos artikulációs mozgásokkal jelennek meg. A hangképző szervek jellemző helyzete magyar beszédhangokra megtalálható alapvető munkákban [4,5,6]. A *3. ábrán* példát mutatunk be arra, hogy mennyire hasonló egy fényképen látható [5] és egy 3D-s beszélő fejen beállított ugyanazon hangra jellemző artikuláció [6].



3. ábra A beszélő fényképe és a 3D fejmodell

A magyar beszédhangok vizéma készletét a [4]-ben megadott mintaszavak artikulációs jellemzőiből alakítottuk ki. Az eredményt az *1. táblázat* mutatja, a hangokat a magyar helyesírási betűképükkel jelöljük.

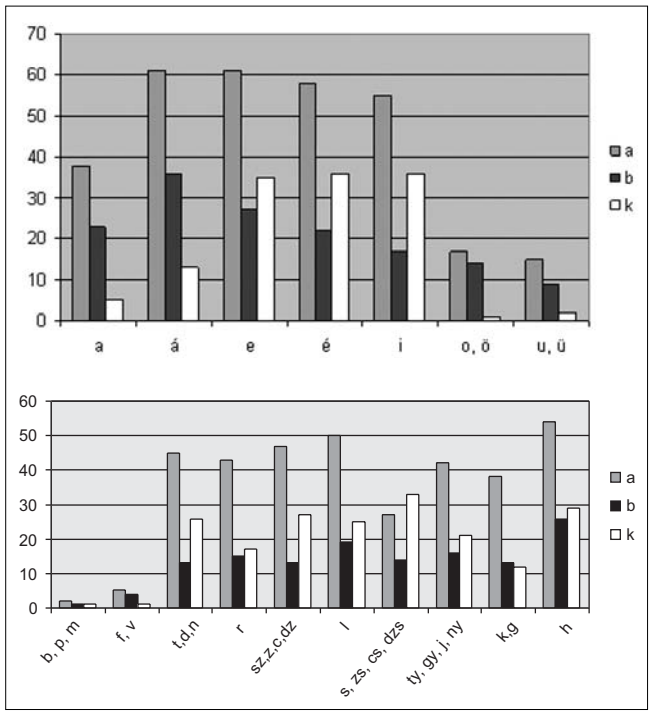
1. táblázat A magyar nyelv vizéma készlete

Magánhangzók	Mássalhangzók
e	b, p, m
é	f, v
i	t, d, n
ö, o	r
ü, u	sz, z, c, dz
á	l
a	s, zs, cs, dzs
	ty, gy, j, ny
	k, g
	h

Néhány megjegyzés a vizémák osztályozásához:

- a csoportosítás elsősorban ajakforma alapján történt, a nem látható nyelvállás eltérő lehet (pl.: o-ö, u-ü)
- a nem jelzett hosszú magánhangzók a rövid párjuknál szűkebb szájnyílással vannak jelen
- az artikuláció előállításához ennél bővebb készlettel dolgozunk

A *4. ábra* a vizémák ajakméreteit és intenzitási tényezőit ábrázolja.

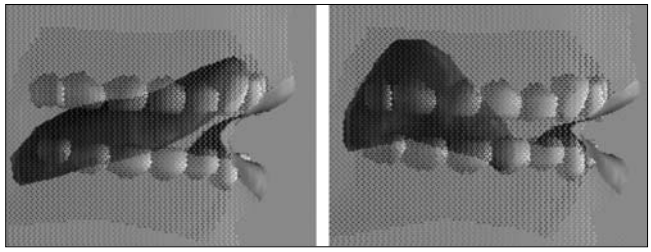


4. ábra A vizémák ajakszélessége (a), ajaknyílása (b) és a szájnyílás átlagos világossága (intenzitás, k). A méretek pixelben, az intenzitás a fehér (255) világosságának arányában látható

Az eddig megjelent beszédhangok atlasza [4], illetve magyar hangalbumok [5,6] alapján meghatározhatók a vizémák legfontosabb paraméterei, ezekből alakul ki az a kulcskeret (keyframe) készlet, amely az artikuláció kiindulási alapja [7].

A legfontosabb jellemzők az ajkak és a nyelv működtetéséhez tartoznak. Az alapvető ajakjellemzők: nyitás (tág-szűk), szélesség (széles-keskeny), Az ajkak nyitása szoros összefüggésben van az állkapocs mozgásával (nyitott - zárt). A száj szélessége tehát az ajaknyitással és az ajakkerekítéssel, illetve az ajakréssel, áll összefüggésben. Az állkapocs helyzete a nyitás mellett a fogak láthatóságával is összefügg. A nyelvállást *(5. ábra)* a nyelv függőleges helyzete (fent-lent),

5. ábra
Jellemző nyelvállások: balra az n, jobbra a k-g hangokra



vízszintes mozgása (elül-hátul), hajlítása (domború-homorú), és a nyelvhegy formája (széles-keskeny, vékony-vastag) befolyásolják.

A statikus jellemzők alapján beállíthatók a beszédhangok állandósult szakaszára jellemző artikulációs paraméterek, kulcskeretek.

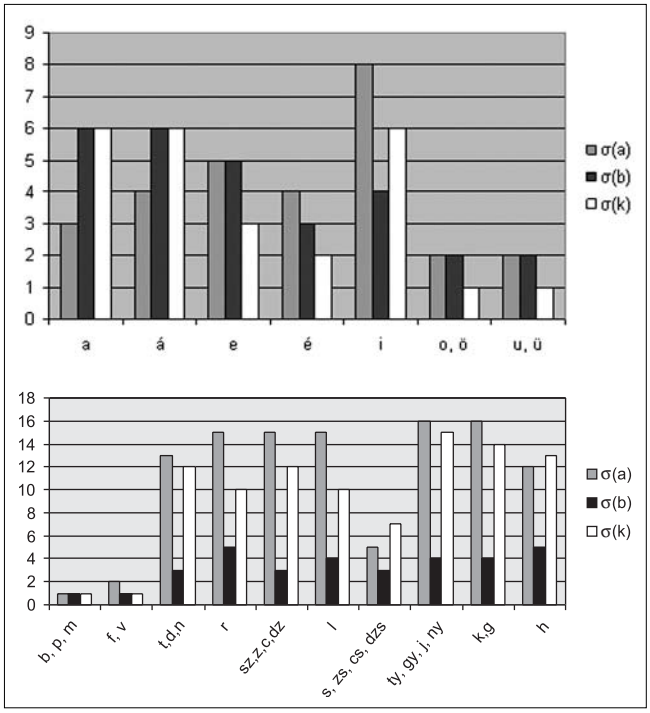
2.2. Dinamikus működés

A folyamatos magyar beszéd dinamikus jellemzőinek átfogó leírása még várat magára. Az analízis során a hangalbumokban található pillanatképek korlátozottan használhatók, és csak a mintaszavakra vonatkoztathatók. A dinamikus analízis másik forrása a saját, vizuális beszédfelismerési kutatások során nyert eredményekből összeállított adatbázis [8]. Ebből származnak az ajkak nyitásának és szélességének időbeli változására vonatkozó adatok, valamint a nyelv és a fogak láthatóságát reprezentáló intenzitás faktor, a szájüregre vonatkozóan. Ezek a kulcskeretek közötti interpoláció megválasztásában nyújtanak segítséget.

A koartikulációs hatások figyelembe vételéhez túl kellett lépnünk az úgynevezett „keyframe” modellen. A vizémák minden jellemzőjét (például ajak- és nyelvállások) osztályoztuk domináns jellegük alapján. Egyes paraméterek a környezettől függetlenül felveszik jellegzetes értékeiket, mások a környezetükbe simulnak. A vizuális beszédfelismerés adatainak szórása alapján a vizémák jellemzőit három kategóriába soroltuk:

- *domináns* –
  nem enged koartikulációs hatásoknak
- *rugalmas* –
  a környezete befolyásolja az adott jellemzőt
- *határozatlan* –
  a környezete alakítja ki az adott jellemzőt

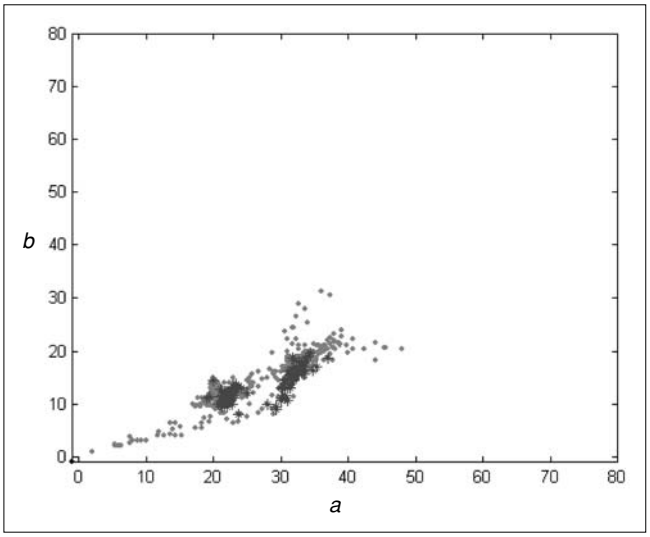
6. ábra A vizémák jellemzőinek szórása



A dominancia meghatározásához elsősorban a jellemzők szórását használtuk fel, de segítséget nyújt a látható jellemzők grafikus ábrázolása, az átmeneti és az állandósult szakaszok eloszlása is.

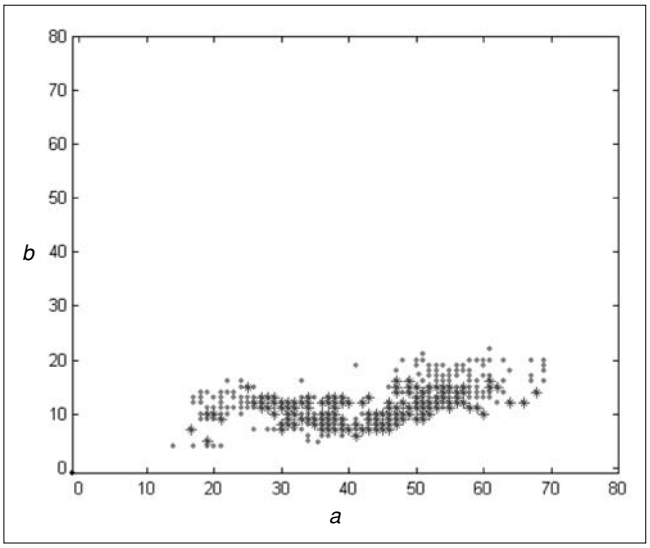
A *7. ábrán* eltérő árnyalattal láthatók az s hang átmeneti és kvázistacionárius szakaszának ajakméretei. A szomszédos hangok által meghatározott kezdeti- és végállapotok között az ajakméretek egy szűkebb területet foglalnak el.

7. ábra
Az s hang átmeneti (.) és állandósult (*) szakaszának ajakszélessége (a) és ajaknyílása (b)



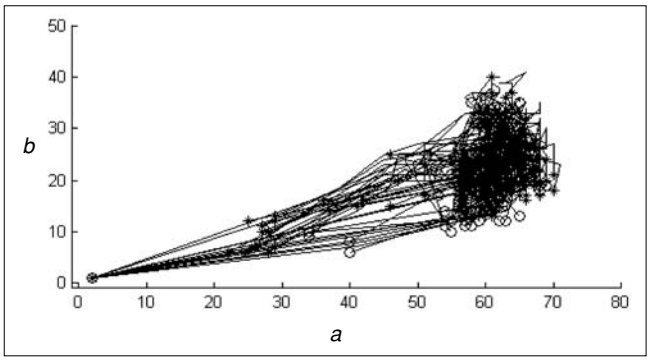
Az ajakméretek eloszlása a *j* hang átmeneti és állandósult tartományára a *8. ábrán* látható. Az ajakszélesség tartománya lényegében megegyezik az átmeneti és az állandósult időszakban, tehát széles tartományban a környezetéhez igazodik, a határozatlan osztályba sorolható. Az ajaknyílás az állandósult szakaszban szűkebb tartományt fed le, az ajaknyílás tekintetében a *j* vizéma domináns jelleget mutat.

8. ábra
A j vizéma ajakméreteinek eloszlása (átmeneti (.) és állandósult (*) szakasz)



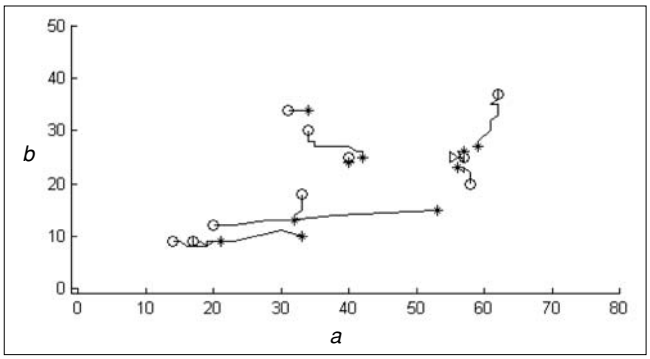
Az ajakméretek változásának trajektóriája is támpontot ad a dominancia osztály meghatározásához. A *9. ábra* az *e* hang ajakméreteinek változását mutatja. A görbék egyenként nem követhetők, de láthatóan tetszőleges kezdeti- és végállapot mellett áthaladnak egy sűrűn behálózott területen. Jól látható a magánhangzók ajakméreteire jellemző domináns jelleg.

9. ábra Az e vizéma ajakméreteinek változása



A domináns változókkal ellentétben, a határozatlan jellemzők nem tartanak jól meghatározható értékekhez. A *h* hanghoz tartozó trajektória példáit látjuk a *10. ábrán.* (A változások követhetősége végett csak néhány görbe szerepel.)

10. ábra
A h vizéma ajakméreteinek változása.
„" jelzi a kezdőpontot, „o" a végpontot*



A *2. táblázat* mutatja a vizémák ajakformára, a *3. táblázat* a nyelv vízszintes helyzetére vonatkozó csoportosítását.

2. táblázat
Dominancia jellemzők az ajakformára nézve

Domináns	magánhangzók, s, zs, cs, dzs
Határozatlan	k, g, r, h
Vegyes	p, b, m, l, j, n, ny, f, v, sz, z, c, dz,, d, t, ty, gy (ajaknyílás domináns, szélesség határozatlan)

3. táblázat
Dominancia jellemzők a nyelv vízszintes helyzetére nézve

Domináns	t, d, n, r, l, ty, gy, j, ny, s, zs, cs, dzs, sz, z, c, dz
Rugalmas	magánhangzók
Határozatlan	p, b, m, f, v, k, g, h

A dominancia beállításai a paraméterek interpolációját határozzák meg. A további módosítások – például hosszú magánhangzóknál állandósult szakasz beiktatása – finomítják az artikulációt.

3. A természetesség javítása

A beszélő természetes fejmozgását, mimikáját hírolvasó bemondók felvételein tanulmányoztuk. Ennek nyomán álvéletlen mozgásokat, például visszafogott bólogatást, a fej enyhe oldalra billentését és átlag körül szóródó pislogási periódust alkalmaztunk. A prozódia tükröződése a fejmozgásban, illetve az arcmimikában nehezen algoritmizálható, így például a mondathangsúly kifejezése nehézségekbe ütközik. Az intonáció azonban felhasználható a szemöldök mozgatásának vezérlésére. A mondathangsúlynál is emelhető a szemöldök. A szemmozgást a fejmozgás korrigálására használjuk, hogy a tekintet egy pontra szegeződjön, egyéb szemmozgatás kézi beavatkozást igényel. Dialógus rendszerekben a szerepváltást segíthetik a gesztusok, az értő figyelést a szemöldök emelésével jelezhetjük, bólogatással is visszaigazolhatjuk figyelmes hallgatásunkat. Ezek a műveletek manuálisan állíthatók be.

3.1. Előartikuláció és szűrés

A kimondás megkezdése előtt kb. 300 ms időtartamú csendet iktatunk be. Ez alatt az idő alatt a levegővételt imitáljuk az ajkak megnyitásával. Ezután az ajkak alaphelyzetéből elkezdjük az első domináns vizéma kialakítását. Ezzel a kiegészítéssel – amit előartikulációnak neveztünk el – már az első hang megszólalása előtt kialakul az ajakforma, hasonlóan a természetes kimondáshoz.

A természetes vagy szintetizált beszédhez szinkronizálás folyamán különböző sebességű beszéddel szembesültünk. Lassú beszédnél a vizémák jellemzői megközelítik névleges értéküket, gyors beszédnél az artikuláció elnagyoltabb. A rugalmas csoportba sorolt jellemzőkre is igaz, hogy gyors beszédnél a lekerekítés nagyobb. A rugalmas jellemzők kialakítására a medián szűrést alkalmaztuk: A szűrésben résztvevő mintákat nagyság szerint sorba rendezzük, és a középső lesz a szűrt érték. A szűrést három mintára végezzük. Egy jellemző időfüggvényét három lépésben alakítjuk ki:

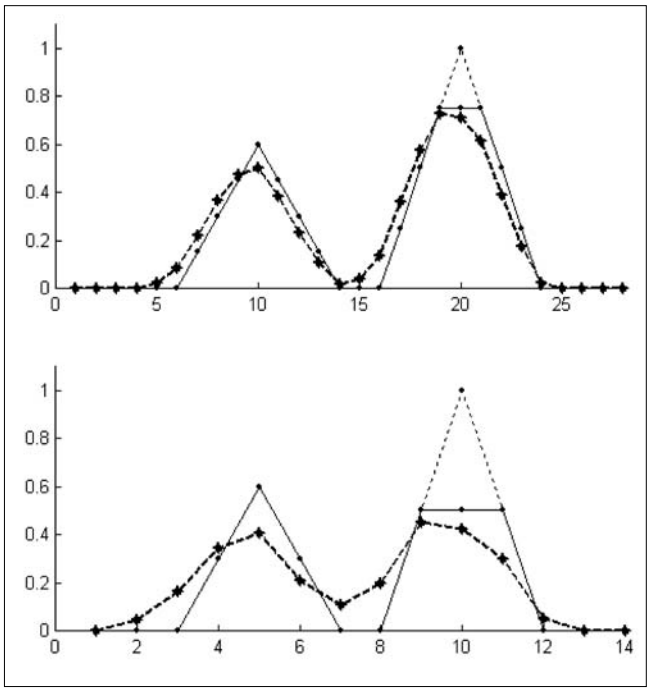
• A domináns és rugalmas vizémák értékei között – a határozatlanok nélkül – lineáris interpolációt végzünk.

• A rugalmas vizémák környezetében végrehajtjuk a medián szűrést. Ez kevesebb minta – gyors beszéd – esetén nagyobb csúcslevágást okoz.

• Az így kapott értékeken még egy simítást végzünk, amely az aktuális, a két megelőző és a követő mintákat érinti. A szűrt érték a négy minta súlyozott összege. A súlyozás állandó, nem függ a beszéd sebességétől. A simító szűrés egyrészt finomítja a mozgást, másrészt gyors beszédnél jobban lekerekíti a csúcsokat. A szintetizált beszéd analízise alapján a szűrés hatása előre erősebb (két keret) mint hátra (egy keret).

A *11. ábrán* gyors és lassú beszédnél követhetjük a medián szűrés és a simítás hatását pl.: a nyelv vízszintes helyzetére. A példában a lassú beszéd kétszer annyi keretből áll, mint a gyors kimondás. Az ábrán jól követhető a gyors beszédnél érvényesülő lekerekítés, a medián szűrés és a simítás hatására egyaránt.

11. ábra
Példa a domináns (1. csúcs) és rugalmas (2. csúcs) jellemző szűrésére és a lassú (1.) illetve gyors (2.) beszéd simítására. A lineáris interpoláció eredménye (···), a medián szűrés (—) és simítás (---) után.



3.2. Érzelmek kifejezése

A beszéd multimodális jellegéhez hozzátartoznak a gesztusok is. A testbeszéddel árnyaljuk mondandónkat, megerősítjük vagy éppen cáfoljuk verbális üzenetünket. Arcanimációs rendszerünkben az arckifejezések érzelmi töltését próbáltuk meg algoritmizálni és programozni. Az Ekman [9] által meghatározott hét érzelem közül választhatunk: semleges, haragos, ellenszenves, szorongó, boldog, szomorú, meglepett. Erre láthatunk példát a *12. ábrán*.

12. ábra Ellenszenves és boldog arckifejezés

4. Összefoglalás és kitekintés

A cikk célja vizuális szövegfelolvasó rendszer fejlesztésének bemutatása. A jelen fázisban az artikuláció dinamikus jellemzőinek további finomítását végezzük. A természetes vagy gépi beszédhez a szinkronizálás még nem teljesen automatikus, a következő feladatunk ennek megoldása. A fejlesztőrendszerünk a beszélő fej videó anyagát hosszadalmas számításokkal állítja elő, ami több órás feldolgozási időt is jelenthet.

Jelenleg – annak ellenére, hogy rendszerünk szövegfelolvasásra is alkalmas – csak olyan alkalmazásokra gondolhatunk, ahol előzetesen rögzített üzeneteket jelenítünk meg. Reményeink szerint a real-time animáció a közeli jövőben szuperszámítógépek nélkül is megvalósítható lesz és ezzel a tényleges virtuális bemondói, felolvasói alkalmazások is megvalósíthatók lesznek.

A vizuális beszédszintetizátor működésére példák találhatók az alábbi címen:

http://mazsola.iit.uni-miskolc.hu/~czap/mintak

Irodalom

[1] Cosatto E., Grafat H. P. (1998): 2D Photo-realistic Talking Head Computer Animation, Philadelphia, Pennsylvania, pp.103–110.
[2] Massaro, D.W. (1998): Perceiving Talking Faces, The MIT Press Cambridge, Massachusetts London, England, pp.359–390.
[3] Bernstein, L.E., Auer, E.T. (1996): Word Recognition in Speechreading. Speechreading by Humans and Machines. Springer-Verlag, Berlin Heidelberg, Germany, pp.17–26.
[4] Molnár József: A magyar beszédhangok atlasza, Tankönyvkiadó, Budapest, 1986.
[5] Bolla Kálmán: Magyar fonetikai atlasz, A szegmentális hangszerkezet elemei, Nemzeti Tankönyvkiadó, Budapest, 1995.
[6] Bolla Kálmán: Magyar hangalbum, A magyar beszédhangok artikulációs és akusztikai sajátságai, MTA Nyelvtudományi Intézet, Budapest, 1980.
[7] Mátyás János: Vizuális beszédszintézis, Diplomaterv, Miskolci Egyetem, 2003.
[8] Czap, L.: Lip Representation by Image Ellipse, ICSLP 2000 Bejging, China, Proceedings Vol. IV., pp.93–96.
[9] Ekman, P., Friesen, W. (1978): Facial Action Coding System Consulting, Psychologists Press. Inc.