FIATAL MŰSZAKIAK TUDOMÁNYOS ÜLÉSSZAKA XVII.

Kolozsvár, 2012. március 22–23.

A SZAVAKON TÚLI KOMMUNIKÁCIÓ AZ AUDIOVIZUALIS BESZÉDSZINTÉZISBEN

CZAP László, PINTÉR Judit Mária

Abstract

The intelligibility of speech can be improved by showing the articulation of the speaker. This visual support is essential in noisy environment and for people with hearing impairment. An artificial talking head can be a natural supplement to sophisticated acoustic speech synthesis. A visual feature database for 3D modelling has been developed and a Hungarian talking head has already been created [1]. In this research, the general approach was to use both static and dynamic observations of natural speech to guide facial animation. A three-level dominance model has been introduced that takes co-articulation into consideration. Each articulatory feature has been grouped into one of three classes: dominant, flexible or uncertain. The acoustic features of speech and the articulation are then linked to each other by a synchronising process. A few features – semi-random head movement, eye blink, eyebrow rising – has been introduced to improve the naturalness of Hungarian talking head. Animation of basic emotions also has been carried out.

Key words:

audio-visual speech processing, speech synthesis, facial animation, talking head

Összefoglalás

A beszéd érthetősége jelentősen javítható a beszélő artikulációjának megjelenítésével. Ez a vizuális kiegészítés különösen fontos zajos környezetben és a hallássérültek esetében. A mesterséges beszélő fej természetes kiegészítője lehet a fejlett akusztikus beszédszintetizátoroknak. A vizuális jellemzőket tartalmazó adatbázist hoztunk létre a 3D modellezéshez, és egy magyar beszélő fejet alakítottunk ki. Ennek a kutatásnak az volt a célja, hogy gyakorlott beszélők megfigyelése alapján vezéreljük az arcanimációt. A koartikulációs hatások figyelembe vételéhez a jellemzőket domináns, rugalmas és határozatlan osztályokba soroltuk, ezek alapján alakult ki a mozgásfázisok közötti interpoláció. A természetesség javítása érdekében többek között álvéletlen fejmozgásokat és pislogást programozunk. A szemöldök mozgatása fontos szerepet játszik a gesztus kialakításában. A fejmodell működtetése során megvalósítjuk alapérzelmek kifejezését is.

Kulcsszavak:

audiovizuális beszédfeldolgozás, beszédszintézis, arcanimáció, beszélő fej

1. Bevezetés

Az elmúlt évtizedekben a gépi beszédfeldolgozás dinamikusan fejlődött, egyre több alkalmazás jelenik meg. Az ember-gép kapcsolatban új távlatokat nyithat az audio-vizuális beszédszintézis. 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 megmutathatja a hangképzés részleteit (1. ábra). Hangvezérelt beszélő fejek fejlesztésén dolgoznak hallássérültek segítésére távközlési alkalmazásokban.

1. ábra. Fotorealisztikus és transzparens megjelenítés

2. A beszédanimáció

Az első élethű modellek alapja a felületi hatásokat utánzó, bőrszövettel borított drótváz alakítására alapozott háromdimenziós animáció. Ennek paraméterei megfigyeléssel, vagy képfeldolgozási módszerekkel természetes beszélők képeiről leolvashatók [2]. Minden modell mozgásánál külön figyelmet kell fordítani a jellemzők összehangolt változtatására, mert könnyen természetellenes hatás alakulhat ki.

A beszéd akusztikus alapegységének, a fonémának vizuális megfelelője, a *vizéma*. A vizémák készlete szűkebb a fonémákénál, hiszen néhány fonéma artikulációja vizuálisan megegyezik. Nem látható pl. 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 [3, 4]. A 2. ábrán példát mutatunk be arra, hogy mennyire hasonló egy fényképen látható és egy 3D-s beszélő fejen beállított, ugyanazon hangra jellemző artikuláció [5].

2. ábra. A beszélő fényképe és a 3D fejmodell

3. Dinamikus működés

A dinamikus analízis forrása a vizuális beszédfelismerési kutatások során nyert eredményekből összeállított adatbázis [1]. Ebből származnak az ajkak nyitásának és szélességének időbeli változására vonatkozó adatok, valamint a nyelv és a fogak láthatóságát reprezentáló mozgás. A dinamikus jellemzők a kulcskeretek közötti interpoláció megválasztásában nyújtanak segítséget.

Egyes paraméterek a környezettől többé-kevésbé függetlenül felveszik jellegzetes értékeiket, mások a környezetükbe simulnak. A vizémák minden jellemzője (például ajak- és nyelvállások) osztályokba sorolandó domináns jellegük alapján. A besorolás alapja az lehet, hogy különböző hangkörnyezetben mennyire változnak az adott hang jellemzői. A vizuális beszédfelismerés adatainak szórása alapján a vizémák jellemzőit három kategóriába soroltuk:

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

4. A természetesség javítása

A beszélő természetes fejmozgása, mimikája hírolvasó bemondók felvételeinek tanulmányozása alapján alakult ki. Ennek nyomán álvéletlen mozgások, például visszafogott bólogatás, a fej enyhe oldalra billentése és átlag körül szóródó pislogási periódus alkalmazható. A prozódia tükröződése a fejmozgásban, illetve az arcmimikában nehezen algoritmizálható. Így pl. a mondathangsúly kifejezése nehézségekbe ütközik. Az intonáció azonban felhasználható a szemöldök mozgásának vezérlésére. Kijelentő mondatnál a szemöldök is enyhén ereszkedő, a mondathangsúlynál emelhető a szemöldök. A gépies, periodikusan ismétlődő mozgások elkerülésére a szabályalapú tendenciák mellett a valószínűségi alapon megvalósított mozgás dominál. Az akusztikus jel energiájával arányos a fej lefelé mozgásának, a szemöldök emelkedésének és a pislogásnak a valószínűsége. A szemmozgás a fejmozgás korrigálását szolgálja, hogy a tekintet mindig a nézőre szegeződjön, egyéb szemmozgatás kézi beavatkozást igényel. Dialógus rendszerekben a szerepváltást segíthetik a gesztusok, az értő figyelmet a szemöldök emelésével jelezhetjük, bólogatással is visszaigazolhatjuk figyelmes hallgatásunkat. Ezek a műveletek manuálisan állíthatók be.

Előartikuláció és szűrés: A kimondás megkezdését kb. 300 ms időtartamú csend előzi meg. Ez alatt az idő alatt a levegővételt az ajkak megnyitása imitálja. Ezután az ajkak alaphelyzetéből elkezdődik az első domináns vizéma kialakítása. Ezzel a kiegészítéssel – amit előartikulációnak nevezhetünk – már az első hang megszólalása előtt kialakul az ajakforma, hasonlóan a természetes kiejtéshez.

Érzelmek kifejezése: A beszéd multimodális jellegéhez hozzátartoznak a gesztusok is. A testbeszéddel árnyaljuk mondandónkat, megerősítjük, vagy éppen cáfoljuk verbális üzenetünket. Arcanimációs rendszerekben az arckifejezések érzelmi töltését programozni lehet. Az Ekman [6] által meghatározott hét alapérzelem jelenti a választékot, skálázható és keverhető formában: semleges, haragos, ellenszenves, szorongó, boldog, szomorú, meglepett. Erre láthatunk példát a 3. ábrán.

3. ábra. Szomorúság, harag, öröm és meglepetés kifejezése

5. Összefoglalás

Audiovizuális beszédszintetizátor rendszerünk számos nem verbális jellemzőt állít be az animáció természetességének elősegítésére. Az előartikuláció, a moderált álvéletlen fejmozgás, pislogás, a beszéd tempójához alkalmazkodó szűrés javítja az attraktivitást. Az érzelmek kifejezésével tovább fokozhatjuk a beszélő fej vonzerejét. A vizuális beszédszintetizátor működésére példák találhatók az alábbi címen: <http://mazsola.iit.uni-miskolc.hu/~czap/mintak>

6. Köszönetnyilvánítás

A bemutatott kutató munka a TÁMOP-4.2.1.B-10/2/KONV-2010-0001 jelű projekt részeként az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg.

Irodalomjegyzék

- [1] Czap L.: '*Audio-Visual speech recognition and synthesis*'. PhD Thesis, Budapest University of Technology and Economics. (2004)
- [2] Massaro, D.W.: '*Perceiving Talking Faces*'. The MIT Press Cambridge, Massachusetts London, England, pp. 359-390. (1998)
- [3] Bolla K.: '*A Phonetic Conspectus of Hungarian*'. Tankönyvkiadó., Budapest. (1995)
- [4] Molnár J.: '*The Map of Hungarian Sounds*'. Tankönyvkiadó, Budapest. (1986)
- [5] Mátyás J.: '*Visual speech synthesis*'. MSc Thesis, University of Miskolc. (2003)
- [6] Ekman, P.: '*Facial expressions*' In: C. Blakemore & S. Jennett (Eds) Oxford Companion to the Body. London: Oxford University Press. (2001)

Pintér Judit Mária, egyetemi hallgató
 Munkahely: Miskolci Egyetem
 Cím: 3515 Miskolc, Egyetemváros
 Telefon: +36 (46) 565-111
 E-mail: pinter.judit.maria@gmail.com

Dr. Czap László, egyetemi docens, dékánhelyettes
 Munkahely: Miskolci Egyetem
 Cím: 3515 Miskolc, Egyetemváros
 Telefon: +36 (46) 565-111
 E-mail: czap@uni-miskolc.hu