Audiovizuális beszédfelismerés és beszédszintézis

PhD értekezés tézisei

Czap László

Tudományos vezetők:
Dr. Gordos Géza
Dr. Vicsi Klára

Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék

2004.

I. Bevezetés

A gépi beszédfelismerés az utóbbi években jelentősen fejlődött. Az alkalmazás és a környezet megfelelő körülhatárolásával a mindennapokban is használható rendszerek születtek. Angol nyelvre a diktált szöveg lejegyzése, vagy más nyelvekre több ezer szavas szókészletű szófelismerő mára elérhetővé vált. A legújabb kutatások eredményeképpen ezek a rendszerek robusztusabbá váltak. Zajos környezetben, változó csatornaparaméterek és beszédstílus mellett azonban megbízhatóságuk jelentősen romlik, meg sem közelítik az emberi beszédértés alkalmazkodó képességét. A beszéd vizuális modalitása az egyik ígéretes kiegészítő információforrás, amely mentes az akusztikus környezet és a zaj zavaró hatásaitól.

A multimodális emberi kommunikációban az akusztikus és vizuális jelet zseniálisan kombináljuk a maximális érthetőség érdekében. A vizuális modalitás előnyei az emberi beszédfelismerésben elsősorban három területen mutatkoznak meg: segíti a hangforrás, a beszélő helyének meghatározását, megkönnyíti az akusztikus jel szegmentálását, kiegészítő információval szolgál az artikuláció helyének meghatározásához.

Ha a hang gyenge minőségű, vagy hallássérült a megfigyelő, jobban hagyatkozik a szájról olvasásra. *Jobban hallom a TV-t, ha felteszem a szemüvegem.* Az emberi beszédértéstől jelentősen elmaradó gépi felismerőket hasonlíthatjuk a környezet vagy képességei által korlátozott emberi felfogóhoz abban a tekintetben, hogy a kiegészítő vizuális jel a gépi beszédfelismerők felismerési hatékonyságát is javíthatja, különösen zajos környezetben. Dolgozatomban a vizuális modalitás által hordozott információval egészítem ki a beszédhang elemzését, a szájról olvasást próbálom gépi úton megvalósítani.

Az ember-gép kapcsolatban új távlatokat nyithat az audiovizuális beszédfeldolgozás másik ága, a 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.

II. A kutatás célja

A gépi beszédfelismerési feladatot a folyamatos beszéd felismerésének célkitűzését szem előtt tartva közelítem meg. Célom olyan akusztikus, illetve audiovizuális motor kifejlesztése, amely egy folyamatos beszédfelismerő bemeneti modulja lehet. Egyik feladatomnak a kétmódusú (bimodal) felismerés vizuális lényegkiemelésének megoldását tekintem. A választott jellemzők kinyerését a videó jelből az ajakkontúrok kijelölésével végzik, de máig nincs rá megbízható eljárás. Célul tűztem ki egy olyan geometriai lényegkiemelés kifejlesztését, amely az ajakkontúrok követését nem igényli.

A vizuális lényegkiemelés másik iskolája a pixel bázisú megközelítés. Meg kívántam vizsgálni, hogy a geometriai vagy a pixel bázisú lényegkiemelés biztosít jobb audiovizuális beszédfelismerési eredményeket.

Az artikuláció statikus és dinamikus viselkedésének vizsgálata alapadatokat szolgáltatott a vizuális beszédszintézishez. vállalkoztam az artikulációt utánzó háromdimenziós fejmodell dinamikus működtetésének kidolgozására.

A kétmódusú beszédfelismerési feladat – magyar nyelvű audiovizuális adatbázis hiányában – beszédadatbázis felvételét igényelte. A válogatás célja a szövegek részleges lefedésére alkalmas leggyakoribb félszótagok tanítására és tesztelésére alkalmas anyag összeállítása volt. A félszótagok vizsgálatára egy lényegesen bővebb szókészletű akusztikus adatbázist is összeállítottam, hiszen a másik lényeges kérdés, amelyre a választ keresem dolgozatomban, hogy mit célszerű a gépi beszédfelismerés során a beszéd felismerendő nyelvi egységének tekinteni.

III. A kutatás és a vizsgálatok módszere

Előzetes szakirodalmi tájékozódás, illetve a célkitűzések megfogalmazása után került sor a *szubjektív teszte*kre, amelyeket a vizuális modalitás emberi beszédértést segítő hatásának vizsgálatára végeztem.

Más célokra kifejlesztett eljárásokat adaptáltam az audiovizuális beszédfelismerés céljaira. A képmomentumok és a mozgásbecslés módszereit használtam a lényegkiemelésre, a medián szűrést a rugalmas artikulációs jellemzők interpolálására.

Más kutatók által alkalmazott módszereket is *adaptáltam* pl.: a diszkrét koszinusz transzformációt pixel bázisú lényegkiemelésre, a Bayes döntést az utólagos integrálásra. Erre példa a HTK rejtett Markov modell szoftvercsomag alkalmazása is.

Munkám során a kutatási eredmények vezettek gyakorlati alkalmazáshoz, máskor az *elméleti* kutatási *eredmények* elérését a *gyakorlati alkalmazás* igénye motiválta, erre példa az artikuláció dinamikus viselkedésének vizsgálata és a vizuális beszédfelismerési eredmények alkalmazása a vizuális beszédszintézis fejlesztésére.

Kutatási módszerem volt még a rendelkezésre álló szakirodalom alapján és saját eredményeim felhasználásával a *meglévő modellek integrációja és új modellek alkotása*. Ezt a módszert követtem pl.: a magyar nyelv vizéma készletének kialakításakor és az artikuláció háromszintű dominancia modelljének megalkotása során.

Az antropológia területéről származó *résztvevő megfigyelés* módszerét az artikuláció dinamikus viselkedésének vizsgálatakor alkalmaztam bemondók, beszélgető partnerek és önmagam megfigyelésével. A *kvalitatív* megfigyeléseket a jellemzők *kvantitatív* meghatározása követte, az *analízis szintézis által* módszerrel, a vizuális beszédszintézis paramétereinek finomításakor.

IV. Új tudományos eredmények

1. A vizuális beszéd jellemzői

A természetes emberi beszédérzékelés szubjektív vizsgálatával megmutattam, hogy elsősorban a hangképzés helyének azonosítását és az elöl képzett hangok felismerését segíti az artikuláció megjelenítése.

A gépi audiovizuális beszédfelismerő tervezéséhez elengedhetetlennek tartottam a kétmódusú emberi beszédértés vizsgálatát. A második szubjektív teszttel azt akartam megvizsgálni, hogy az arc mely részei hordozzák a beszédfelismerés szempontjából lényeges információt. A teljes arc, a száj (ajkak, nyelv, fogak), az ajkak, illetve az ajkak szélességével és nyitásával megegyező méretű ellipszis kísérte a zajos beszédhangot. A száj megmutatása az ellipszissel végzett kísérlethez képest harmadával növelte a helyes felismerés arányát. Később a gépi vizuális felismeré

si eredmények ugyanennyivel javultak, ha az *ajakszélesség* és *ajaknyílás* mellet a *szájnyílás intenzitását* is figyelembe vettük. Az ajakméretek és az ellipszis egyenrangú jellemzők, a növekedés is azonos mértékű, tehát az előbbi három jellemző reprezentálja a száj artikulációs szerepét. Nincs szükség további (egyes kutatóknál tíznél több) geometriai paraméterre. Az intenzitási tényező – amely a nyelv és a fogak láthatóságát hivatott képviselni – hátul képzett hangoknál a legkisebb (pl.: k, u), közepes értékű, ha elöl képzett hangoknál a nyelv látható (pl.: e, i), legnagyobb az értéke, ha a fogakat látjuk a szájnyílásban (pl.: s, cs).

1. 1. tézis: Szubjektív tesztekkel és a gépi vizuális beszédfelismerési eredmények elemzésével megmutattam, hogy az ajakszélesség, az ajaknyílás és a szájnyílás intenzitási tényezője megfelelően írja le a száj vizuális jellemzőit.

A vizuális lényegkiemelés feladata a választott vizuális jellemzők kinyerése a képből. A geometriai alapú lényegkiemelés ismert eljárásainak közös vonása, hogy az ajkak külső és belső kontúrjának követését kívánják meg, erre azonban máig nincs megbízható eljárás. Az ajakkontúrok követése helyett a lényegkiemelést képi hasonlóság vizsgálatra vezettem vissza. Artikulációs könyvtárat hoztam létre, amelyben az orrhegy alapján kijelölt artikulációs terület képei szerepelnek. A képeket úgy válogattam, hogy a minták tartalmazzák a jellegzetes ajakformákat. Külön prototípus képek ábrázolják a hasonló ajakformájú, de különböző nyelvállású formákat. Amennyiben a fogak láthatósága is eltérő, ezt újabb képek jelenítik meg. Ezeken a képeken elvégeztem – a jellegzetes pontok esetleg manuális kijelölésével – a lényegkiemelést. Az adatbázis összes képének feldolgozásával meghatározva képkockánként a hasonlóság mértékét, a legkevésbé hasonló alakzatokat felvettem az artikulációs könyvtárba. A műveletet addig ismételtem, amíg a legkevésbé hasonló alakzatok jellemzői bele nem simultak a környezetükbe. Statisztikát készítettem a prototípus képekről. Ez azt fejezte ki, hogy melyik prototípus alakzat hány képkockához volt a legközelebbi az összes tanító képkocka közül. Azokat a képeket, amelyek csak néhány képhez bizonyultak a legközelebbinek kivettem a prototípus könyvtárból. A prototípus alakzatok iteratív válogatása után 88 alakzat képviselte a jellegzetes képeket. A képi hasonlóság vizsgálat során az artikulációs könyvtár minden elemére megkeresve a legmegfelelőbb pozíciót, minden referencia alakzatra kapunk egy hasonlósági mértéket. Ez alapján kiválasztható a leginkább hasonló referencia elem, ennek vizuális jellemzőit örökli a vizsgált képkocka.

1. 2. tézis: A választott vizuális jellemzők meghatározását képi hasonlóság vizsgálatra vezettem vissza. A prototípus alakzatokból artikulációs könyvtárat hoztam létre. Az eljárás újdonsága, hogy nem igényli az ajakkontúrok követését, ami az ismert módszerek közös jellemzője.

A módszer számos előnnyel jár az ismert eljárásokkal összehasonlítva:

- nem igényli az ajkak sem külső, sem belső kontúrjának meghatározását,
- tetszőleges jellemzőket választhatunk a lényegkiemelésre, ezeket csak a kiválasztott képekre kell meghatározni, manuális támogatás is nyújtható,
- a száj környezetét is figyelembe veszi, feldolgozási területe a geometriai alapú és a pixel bázisú módszerrel feldolgozott terület között helyezkedik el,
- mérsékelt számításigényű, valós időben is elvégezhető a mai PC-ken,
- fekete-fehér képeken elvégezhető,
- vektorkvantáláson alapuló feldolgozás esetén közvetlen bemenetként szolgálhat.

Hátránya, hogy beszélőfüggetlen feladathoz az artikulációs könyvtár bővítésére van szükség, ami a feldolgozási idő növekedéséhez vezet.

2. A geometriai és pixel bázisú lényegkiemelés összehasonlítása

Az audiovizuális beszédfelismerésben használt geometriai és pixel bázisú lényegkiemelés összehasonlítását csak szavakra, illetve fonémákra végezték el, a kedvezőbb eredményeket mutató kései integrálással. A szónál kisebb alapegységek vizsgálatára, ahol csak a korai integrálási modell használható, nem került sor.

A kései integrálás folyamatos beszédfelismerésnél a jelöltek igen nagy száma miatt ma megoldhatatlan. Korai integrálásnál azonban nem biztos, hogy az önmagukban legjobb vizuális jellemzőket használhatjuk, hiszen az akusztikus jellemzőkkel kölcsönhatásban nem feltétlenül bizonyulnak optimálisnak. Példaként megemlítem, hogy a pixel bázisú lényegkiemelés együtthatóinak gondos válogatásával sikerült az egymódusú vizuális hangpár felismerési arányt 64,5 %-ra növelni. Ezeket a jellemzőket az önmagukban 88,4 %-os hangpár felismerési arányt mutató akusztikus jellemzőkkel egyesítve, a kétmódusú hangpár felismerési arány 81,8 %-ra csökkent.

2. 1. tézis: Ellenpéldával igazoltam, hogy a kései integrálás esetében felhasználható, a legjobb vizuális beszédfelismerési eredményeket adó pixel bázisú jellemzők nem vihetők át automatikusan a korai integrálási modellbe. Az akusztikus és vizuális jellemzők kölcsönhatása miatt a pixel alapú vizuális jellemzőket a tanítás-tesztelés folyamatában válogattam.

A folyamatos audiovizuális beszédfelismerés szempontjait szem előtt tartó összehasonlító vizsgálat eltér a szóalapútól, mert:

- a potenciális jelöltek nagy száma miatt a más kutatók által alkalmazott kései integrálás nem alkalmazható,
- az előbbi ok miatt a felhasznált képi jellemzők nem feltétlenül a legjobb vizuális paraméterek.

A kései integrálásnál használt jellemzők és a kapott eredmények ugyanúgy nem ültethetők át a korai integrálási modellre, mint a szóalapú felismerés eredményei a szónál kisebb alapegységgel dolgozó felismerőkre.

2. 2. tézis: Megvizsgáltam a korai integrálási modellt, eredményei azzal a nem triviális tanulsággal jártak, hogy a pixel bázisú jellemzők felismerési eredményei a szónál kisebb alapegység esetén, és a korai integrálás kedvezőtlenebb feltételei mellett is felülmúlják a geometriai jellemzőkéit.

A kísérletekhez használt adatbázisok összeállításánál a részleges lefedést biztosító leggyakoribb félszótagok tanítására és tesztelésére válogattam a szavakat. Az audiovizuális adatbázis 600 szót illetve szófűzért, az akusztikus adatbázis 9400 szót tartalmaz.

Gépi beszédfelismerési kísérletekkel összehasonlítottam a hangpár és a félszótag alapú felismerési alapegységekkel elérhető eredményeket, mind az audiovizuális, mind az akusztikus adatbázison. Ezekben a kísérletekben a hangpár alapú beszédfelismerés jobb eredményeket mutatott, mint a félszótag alapú. Az eredmények erősítik azt a meggyőződést, hogy a felismerés alapegységének ki kell fejeznie az elemek kontextusfüggését.

Audiovizuális felismerőnél lehetőségünk van a hangminőség romlása esetén a vizuális modalitás súlyának növelésére. Kísérleteim tanúsága szerint különösen gyengébb minőségű beszédnél múlja felül a változó súlyozású felismerő az akusztikus és vizuális modalitást azonos súlyozással kezelő társáét.

3. A beszéd vizuális alapegységének osztályozása

A vizuális beszéd elemzésekor gyűjtött tapasztalatok egyik alkalmazási lehetősége a vizuális beszédszintézis. A mesterséges vagy természetes beszédet az artikulációt utánzó háromdimenziós fejmodell képével egészítjük ki. Tetszőleges szöveget kísérhet az animáció, magyar nyelvű, vizuális szövegfelolvasó kifejlesztése a kutatás célja.

A beszéd legkisebb akusztikus egységének, 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.

3. tézis: A geometriai lényegkiemelés eredményeként alapadatokat szolgáltattam vizuális beszédszintetizátor (beszélő fej) működtetéséhez, meghatároztam a vizéma (a fonéma vizuális megfelelője) osztályokat a magyar beszédhangokra.

A magyar beszédhangok vizéma készletét a hangalbumokban található mintaszavak artikulációs jellemzőiből kiindulva, saját audiovizuális mérési eredményekkel kiegészítve alakítottam ki. Egyes hangok vizuális megjelenése nyilvánvalóan azonos, mások fonetikai ismeretek alapján sorolhatók egy osztályba. A geometriai méretek és az intenzitási tényező alapján további összevonások lehetségesek.

8. 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

4. Az artikuláció dinamikus viselkedésének modellezése

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 eredményekből összeállított adatbázis. 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.

4. tézis: A háromdimenziós fejmodell dinamikus működtetéséhez háromszintű dominancia modellt vezettem be. Definiáltam, és a magyar vizémákra meghatároztam a domináns, rugalmas és határozatlan paramétereket. Kidolgoztam a jellemzők dominancia osztályok szerinti approximációját, valamint a beszédtempó figyelembe vételének módját.

A *domináns* paraméterek a szomszédos hangoktól függetlenül felveszik jellegzetes értéküket. Ugyancsak a dinamikus viselkedés tekintetében, a *rugalmas* változók engednek a koartikulációs hatásoknak. Azokat a jellemzőket, amelyeknek értékét a környezetük határozza meg, *határozatlan*nak definiáltam.

A dominancia meghatározásához elsősorban a jellemzők szórását használtam fel, de segítséget nyújt a látható jellemzők grafikus ábrázolása, az átmeneti és az állandósult szakaszok eloszlása. Az ajakméretek változásának trajektóriája is támpontot ad a dominancia osztály meghatározásához.

A domináns és határozatlan jellemzők megvalósítása nem okoz nehézséget, a rugalmas jellemzők kialakítására a medián szűrést vezettem be. További szűrés segíti a mozgás simítását és a különböző sebességű bemondáshoz alkalmazkodást.

Néhány példa az artikulációs jellemzők beillesztésére a háromszintű dominancia modellbe:

8. 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)

8. 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 természetesség javítása érdekében álvéletlen mozgásokat, például visszafogott bólogatást, a fej enyhe oldalra billentését és átlag körül szóródó pislogási periódust alkalmaztunk. Megvalósítottuk az alapérzelmek kifejezését is.

V. Az értekezés témájában megjelent publikációk

Czap L, Gordos G, Németh G, Olaszy G, Tihanyi A: *An Integrated Approach to Text-to-Speech and Fixed Vocabulary Formant Synthesis*, ITG-Fachtagung. Bad Nauheim, Okt. 1988. ITG-Fachbericht 105, VDE-Verlag pp. 213-216. 1988.

Czap L.: *Voiced-Unvoiced Classification of Speech Signals by a Neural network Classifier*. International Conference of PhD Students. Miskolc. Section Proceeding, Engineering Science I. pp. 111-115. 11-17 August 1997.

Ajtonyi I, Czap L.: *Image processing in Measurement of Caustics*. IMEKO International Symposium on Development in Digital Measuring Instrumentation. Naples, Italy, 17-18. Proc. pp. 772-774. September 1998

Czap, L.: *Audio and Audio-visual Perception of Consonants Disturbed by White Noise and 'Cocktail Party'*. 5th International Conference on Spoken Language Processing,

Sydney, Australia, Proceedings Vol. 2, pp. 253-256. 30th November - 4th December 1998

Czap L.: *Vizuális és akusztikus emberi beszédfelismerés.* Automatika, Mérés- és Műszertechnika Konferencia. Siófok. Proc. pp. 26-31. 1998.

Ajtonyi I, Czap L.: *Evaluation of Crack Caustics by Neural Networks* European Control Cobference'99, Karlsruhe, Germany, Proc. Vol. F. pp. 686-688. 1999

Ajtonyi I, Czap L.: *Evaluation of Different Mode Crack Caustics by Neural Networks* 3rd IEEE International Conference on Intelligent Engineering Systems High Tatras, Stará Lesná, Slovakia, Proc. pp. 373-375. 1-3. November 1999

Czap L.: *Lip Representation by Image Ellipse. 6th International* Conference on Spoken Language Processing, Bejging, China, Proc. Vol. IV. pp. 93-96. 2000

Czap L.: *Feature Extraction for Speechreading.* International Carpathian Control Conference '04, Zakopane, Poland, Proc. Vol. I. pp. 821-824 2004

Czap L., Mátyás J.: *Beszélő fej.* Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Proc. pp. 196-202. 2003.

Czap L., Mátyás J.: *Talking Head* microCAD 2004. Miskolc, Proceedings of Section G. pp. 19-24. 2004

Czap L., Mátyás J.: *Beszélő fej.* ENELKO 2004. Proceedings pp. 20-24. Kolozsvár.

Czap L.: *Audiovizuális beszédfelismerés.* Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Proc. pp. 293-300. 2004.

Czap L.: *Virtuális bemondó.* Híradástechnika (Közlésre elfogadva)