Audiovizuális beszédfelismerés és beszédszintézis

PhD értekezés

Czap László

Tudományos vezetők:

Dr. Gordos Géza
Dr. Vicsi Klára

Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék

2004.

Családomnak az elrabolt időért.

Az értekezés bírálatai és a védésről készült jegyzőkönyv a Dékáni Hivatalban elérhető.

Tartalomjegyzék

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 (Hermansky, Morgan, 1994). A beszéd vizuális modalitása az egyik ígéretes kiegészítő információforrás, amely mentes az akusztikus környezet és a zaj zavaró hatásaitól.

A multimodális kommunikációban több érzékszervünk vesz részt. A hallás mellett a látás a legfontosabb információforrásunk. A tapintás ezeknél kisebb szerepet játszik. A szaglásunk jelentősége a kommunikációban a törzsfejlődés során minimálisra zsugorodott. Az ízlelés szinte kizárólag a táplálkozással kapcsolatosan jut szerephez.

Bernsen (2002) összekapcsolja a modalitást a médiummal, mint az információ valamely formájának fizikai hordozójával. A média rokonítható az érzékszervekkel, amelyekre hat. A grafikus médium pl.: a látással, az akusztikus közeg a hallással társítható.

A multimodális emberi kommunikációban az akusztikus és vizuális jelet zseniálisan kombináljuk a maximális érthetőség érdekében. Még a gépi bimodális beszédfelismerés megszületése előtt, az ötvenes években demonstrálták, hogy zajos környezetben a vizuális jel segíti a beszéd jobb megértését (Sumby, Pollack, 1954). 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 (Summerfield, 1987).

A természetes, közvetlen emberi kommunikációban résztvevő modalitások együttesen fejtik ki hatásukat (Pease, 1987). A folyamat bonyolultságára jellemző Griffin (2003) megállapítása: *„A képszerű gondolkodásra való hajlam arra is nagy hatással van, hogy milyen módon közelítjük meg az interperszonális üzeneteket. Az interperszonális kommunikációt meghatározhatjuk olyan folyamatként, amely kölcsönösen elfogadott jelentést hoz létre egyedi szituációkban, ám a definíció érvényessége egyedül attól függ, hogy kinél milyen képzeteket vált ki."*

Általában csak a távközlésben kapcsolódik ki a csatornák egy része (Buda, 2000). A mai beszédfelismerő rendszerek a verbális csatorna dekódolására szorítkoznak. A vokális csatorna nem verbális része, a mimika, a tekintet, a testtartás, a gesztus és más mozgásos kommunikációs elemek a technika fejlődésével és a jelenségek mélyebb megismerésével az ember-gép kommunikációban is részt vehetnek, egyre több modalitás juthat szerephez.

Massaro (1998) kísérletekkel igazolta, hogy a modalitásokat egymás kiegészítésére használjuk. 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 beszédértést meg sem közelítő 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 angol szakzsargonban meghonosodott a *speechreading* kifejezés, amely a *speech recognition* és *lipreading* egyesítésével fejezi ki az akusztikus és a vizuális modalitás integrálását a beszédfelismerésben. (Tágabb értelemben beleértik a gesztusok figyelembe vételét is.)

Az audiovizuális beszédfelismerést beszélőfüggő feladatban közelítem meg. A vizuális lényegkiemelés személyfüggetlen megoldása ma még nem áll rendelkezésünkre. Külön kutatási témát jelentene az akusztikus és vizuális jellemzők kapcsolatának vizsgálata. Elég, ha csak a hangokat jellemzően elül vagy hátul képzők beszédének akusztikus és vizuális különbségeire gondolunk.

A vizuális lényegkiemelés gépi megvalósításához célszerűnek láttam elemezni az emberi audiovizuális kommunikációt. Mássalhangzó felismerési kísérletben tanulmányoztam, hogy a vizuális jel a beszéd mely jellemzőinek felismerését segíti jobban. A 2. fejezetben ismertetett szubjektív tesz-

ten alapuló vizsgálataim eredményeképpen a hangképzés helyének meghatározásában és elsősorban – nem meglepő módon – az elül képzett hangok felismerési arányaiban várható javulás.

A 3. fejezetben leírt újabb szubjektív teszttel arra kerestem a választ, hogy az arc mely részei hordozzák a beszédfelismerés szempontjából lényeges információt. Az ajakformához képest a nyelv és a fogak láthatósága jelentős mértékben javította mind a magánhangzók, mind a mássalhangzók felismerési arányát. Ez a körülmény ösztönzést adott a nyelv és a fogak láthatóságát kifejező paraméter keresésére. Az ajkak belső szélességét és nyílását választottam az ajakforma jellemzésére, a szájnyílás intenzitási tényezőjét a nyelv és a fogak láthatóságának kifejezésére.

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 egy megoldását tekintem. Kétmódusú beszédfelismerésen a hang és az artikulációs jellemzők alapján végzett beszédfelismerést értem. Egymódusú a beszédfelismerés, ha csak az akusztikus, vagy pusztán a vizuális jelet dolgozza fel.

A választott jellemzők kinyerését a videó jelből először színes képen végeztem. A lényegkiemelésnél el akartam kerülni a marker jelek alkalmazását, a beszélő ajkának könnyebb azonosítására piros rúzst alkalmaztam. Az arc különböző részeinek maszkolására kifejlesztett eljárás – amelyet a

második szubjektív teszt előkészítésére használtam – alkalmasnak bizonyult a választott jellemzők meghatározására. A feldolgozáshoz a képmomentumokat használtam. A vizuális lényegkiemelés továbbfejlesztése során az általánosabb alkalmazhatóság érdekében a szín információt nem kívántam felhasználni. A magas szintű, vagy geometriai alapú feldolgozás ismert eljárásainak általános jellemzője, hogy az ajkak külső és belső kontúrjának követésével az artikulációs szervek látható részeinek méretét vagy helyzetét olvassák le. A geometriai lényegkiemelésre a 4. fejezetben javasolt módszerem nem igényli az ajakkontúrok követését. A lényegkiemelést képi hasonlóság vizsgálatra vezettem vissza.

A vizuális lényegkiemelés másik iskolája a pixel bázisú megközelítés. Az alacsony szintű, vagy pixel alapú feldolgozás a száj környezetének képpontjait veti valamilyen transzformáció alá, és a transzformált jellemzők egy redukált készletét használja a felismeréshez. Az 4. 2. alfejezetben a diszkrét koszinusz transzformációs lényegkiemelés adaptálását és a jellemzők válogatását ismertetem.

A vizuális lényegkiemelés eredményeképpen két paraméter rendszert kapunk, egyet az akusztikus, egyet a vizuális jel leírására. A legnagyobb kihívás a két modalitás integrálása a legjobb felismerés érdekében, az ember ugyanis a legkülönbözőbb minőségű modalitásokat mindig úgy integrálja, hogy a kétmódusú beszédfelismerés felülmúlja mindkét modalitás külön mért eredményeit. A két modalitás integrálása a vizuális lényegkiemelésnél is kevésbé kidolgozott terület. A korai és kései egyesítés összehasonlítását egy egyszerűsített feladaton a 6. fejezetben írtam le. A korai integrá-

lás esetében az akusztikus és vizuális jellemzők egyesítése megelőzi a felismerés alapegységeire osztályozást. A kései integrálásnál a két modalitás alapján külön-külön elvégzett osztályozás után történik az egyesítés (Benoît, et al., 1998).

A kétmódusú beszédfelismerési feladat – magyar nyelvű audiovizuális adatbázis hiányában – beszédadatbázis felvételét igényelte. Az adatbázisban saját bemondással szótagok és szavak szerepelnek a tanításhoz, valamint szavak illetve szófűzérek a teszteléshez. 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. A félszótagok a szótagoknak a magánhangzó közepén kezdő és záró félszótagra bontásával alakulnak ki.

Agglutináló nyelvek esetében a szóalapú feldolgozás folyamatos beszéd felismerésére nem alkalmas. Nincs általánosan elfogadott becslés a magyar nyelvben előforduló szóalakok számára vonatkozóan, annyi azonban bizonyos, hogy kezelhetetlen mennyiségről van szó. Nyilvánvaló, hogy a fonéma szintű felismerés a hangok egymásra hatása miatt nem lehetséges. A szónál rövidebb, a fonémánál hosszabb alakzatokat célszerű választani felismerendő nyelvi egységként. Kísérleteket végeztem a hangpár és a félszótag alapú gépi beszédfelismerés összehasonlítására. Gépi beszédfelismerési kísérletekkel összehasonlítottam a hangpár és a Vicsi

Klára (Vicsi, Vigh, 1995) által javasolt félszótag alapú felismerési alapegységekkel elérhető eredményeket, mind az audiovizuális, mind az akusztikus adatbázison. Szignifikáns különbség tapasztalható a hangpár alapú feldolgozás javára. (Az idegen szavak kerülése végett használom a hangpár kifejezést. Egy fonéma közepétől a következő fonéma közepéig terjedő beszédszakaszt értem rajta. A szakirodalomban a diphone és a diad kifejezések fedik le a fogalom tartalmát.)

A beszédfelismerési vizsgálatoknál összehasonlító elemzést végeztem a pusztán akusztikus illetve a vizuális jellel kiegészített jellemzőkkel. A beszéd különböző alapegységeinek helyes felismerési aránya önmagában ugyan nem mérvadó, de a tanításnál és a tesztelésnél minden összeállítás esetében ugyanazokat a lépéseket követtem, az eredmények így egymással összevethetők. Az audiovizuális beszédadatbázison a geometriai és a pixel alapú lényegkiemelés összehasonlítását is elvégeztem. Megvizsgáltam a modalitások súlyozásával és a jel/zaj viszony becslésével elérhető javulást változó minőségű beszéd felismerése esetén. A felvétel körülményeinek és eszközeinek javításával, a rejtett Markov modell fejlesztésével, de leginkább a tanító alakzatok számának többszörözésével a felismerési eredmények jelentősen javíthatók.

Kutatásaim másik területe a vizuális beszédszintézis, természetes vagy szintetizált beszéd kiegészítése egy virtuális bemondó képével. Célom élethű fejmodell vezérlési algoritmusainak kidolgozása, amelynek artikulációja elfogadható egy természetes kiejtési módnak. A vizuális beszédelemzés eredményeinek felhasználásával háromdimenziós fejmodell mű-

ködtetéséhez szolgáltattam alapadatokat. A 8. fejezetben ismertetett audiovizuális beszédszintézis megvalósításához meghatároztam a magyar beszédhangok vizémáit - a fonémák vizuális megfelelőit. A vizuális beszédelemzés eredményeire elsősorban az artikuláció dinamikus jellemzőinek meghatározásához volt szükség, ennek leírása magyar nyelvre még várat magára. Az artikuláció dinamikus működtetésére dominancia osztályokat vezettem be. Módszert dolgoztam ki a rugalmas jellemzők szűrésére és a beszédtempóhoz igazodó artikuláció előállítására.

A doktori eljárás megindítása óta születtek a vizuális beszédszintézissel kapcsolatos eredmények, az értekezésnek a Doktori Tanács által elfogadott címe ezt még nem tartalmazta. A tanszéki vita során született a javaslat a cím kiegészítésére.

Köszönetnyilvánítás

Ezúton szeretném megköszönni Gordos Géza több évtizedes szakmai útmutatását és biztatását, Vicsi Klára tudományos vezetői segítségét és Lajtha György értékes lektori megjegyzéseit.

1. A kutatás elhelyezése a tudományág eredményei között

A szájról olvasás régről ismert technikájának a gépi beszédfelismerés szolgálatába állítása mintegy két évtizeddel ezelőtt kezdődött. A számítógépek sebessége és tárkapacitása, a képfelvevő és képmegjelenítő eszközök fejlődése lehetővé tette a képfeldolgozás eredményeinek szélesebb körű alkalmazását.

A gépi audiovizuális beszédfelismerő tervezéséhez elengedhetetlennek tartottam az emberi beszédértés vizsgálatát. Az akusztikus beszédfelismerés tapasztalataiból tudjuk, hogy a gépi beszédfelismerők eredményei messze elmaradnak az ember teljesítményétől, a humán kétmódusú beszédfelismerés vizsgálata a távlati célok kitűzésében segíthet. Az első szubjektív teszttel arra kerestem a választ, hogy a hangképzés mely jellemzőjét segít pontosítani az artikuláció megfigyelése, mit várhatunk a vizuális jel elemzésétől. Az eredmények számszerűen kifejezték az irodalomban általánosságban megfogalmazott állítást a képzés helyének pontosabb azonosítására.

A kommunikáció során a teljes arc, sőt az egész test megfigyelésére lehetőségünk van. Nehezen különíthető el, hogy az arc mely részlete segítette a beszéd felismerését. Az arc részeinek elfedésével elkülönítettem az ajkakat, hogy az ajakforma hatását szubjektív teszttel ellenőrizhessem. Az ajkak leírására a külső-belső méreteket illetve területeket szokás használni. Az ajkak belső méretei alapján egy ellipszist rajzoltam annak vizsgálatára,

hogy az ajakszélesség és ajaknyílás megfelelően reprezentálja-e az ajakformát. A maszkolás változtatásával a száj – ajkak, nyelv, fogak – megmutatásával a nyelv és a fogak kiegészítő támogatását mértem. Ennek legfontosabb célja, hogy érdemes-e erőfeszítéseket tenni a nyelv és a fogak láthatóságának leírására. A teljes arc megfigyelésével végzett kísérlet a pixel bázisú feldolgozástól várható eredményeket próbálja megbecsülni.

A szubjektív tesztek eredményei alapján jelöltem ki a vizuális jellemzőket, amelyeket a lényegkiemelés során meg kell határozni. Az ajkakat az ellipszis teszt alapján az ajakszélességgel és ajaknyílással írom le. A nyelv és a fogak láthatóságát a geometriai jellemzőkkel dolgozó kutatók többsége nem veszi figyelembe (Bregler, Konig, 1994; Dalton et al., 1996; Benoît, et al., 1996; Bernstein, Auer, 1996). Luettin az active shape model profiljai közötti választásra (ajkak zárva / száj nyitva, fogak láthatók / száj nyitva, fogak nem láthatók), valamint a nyelv és a fogak láthatóságának leírására használja az intenzitást. Petajan egy-egy bináris változóval jelzi, hogy a nyelv, illetve a fogak láthatók-e (Petajan, Graf, 1996). A nyelv és a fogak láthatósága a szájról olvasás tapasztalatai és a szubjektív tesztek szerint fontos vizuális paraméterek, ezért kerestem a leírásukra alkalmas jellemzőt.

A vizuális lényegkiemelés feladata a választott jellemzők kinyerése a képből. A geometriai alapú lényegkiemelés ismert eljárásainak közös vonása, hogy az ajkak külső és belső kontúrjának követését kívánják meg (Yuille et al., 1992; Silsbee, 1994; Cootes et al., 1995, 1998; Kass et al., 1988). Az ajakkontúrok meghatározására azonban mind a mai napig nem sikerült

megbízható eljárást kifejleszteni (Pérez et al., 2003). A vizuális lényegkiemeléshez olyan eljárás kidolgozását tűztem ki célul, amely nem igényli az ajakkontúrok követését. Az első próbálkozások során még felhasználtam a szín információt. Piros rúzs alkalmazásával az ajkakat színméréssel sikerült kijelölni, a feldolgozáshoz a képfeldolgozásban alakzatfelismerésre ismert geometriai momentumok módszerét adaptáltam (Hu, 1962; Mukundan et al. 1998). Az ajakméretekre a képellipszis modell alapján adtam becslést. A geometriai momentumokból kölcsönöztem a nyelv és a fogak láthatóságának leírására az intenzitási tényezőt is. Ez a bináris változóknál lényegesen finomabban írja le a szájnyílás világosság változását. Az intenzitási tényező hozzáadása az ajakméretekhez jelentősen javította a gépi audiovizuális felismerési eredményeket. A geometriai momentumokat vizuális jellemzőknek tekintő lényegkiemelési módszeremről az MIT Pressnél megjelenés alatt álló összefoglaló kiadvány említést tesz (Bailly et al., 2004).

Az általánosabb alkalmazhatóság érdekében fekete-fehér képekre is kidolgoztam az ajakkontúrok követését nem igénylő eljárást. A jellegzetes ajakfomákat és a nyelv és a fogak eltérő láthatóságát mutató képek iteratív válogatásával prototípus könyvtárat hoztam létre. A lényegkiemelést képi hasonlóság vizsgálatra vezettem vissza.

Az irodalomban a lényegkiemelés másik jelentős iskolája a geometriai alapú megközelítés mellett a pixel bázisú feldolgozás. Ennek előnye, hogy a teljes artikulációs terület képének transzformációjával információ veszteség nélkül, nemcsak a száj, hanem környezete is részt vesz a feldolgo-

zásban. Ezzel a teljes arc megfigyelésének esetéhez közelíthetünk. A pixel bázisú lényegkiemelés vizsgálatához a diszkrét koszinusz transzformációs eljárást adaptáltam (Potamianos et al., 1998; Nakamura et al., 2000; Neti et al., 2003). Célom a geometriai és a pixel bázisú lényegkiemelés felismerési eredményeinek összehasonlítása. Ezt eddig csak szó vagy fonéma alapú felismerőkkel, az akusztikus és vizuális modalitás kései integrációjával végezték el (Scanlon, Reilly, 2001; Matthews et al. 2001). A folyamatos beszédfelismerésnél általánosabban használható hangpár alapú felismerési feladaton, korai integrálással vetettem össze a két eljárást.

Az akusztikus és vizuális jel egyesítése a legjobb felismerés érdekében az audiovizuális beszédfelismerés egyik legnagyobb kihívása (Duchnowski et al., 1994) A Bayes döntésen alapuló integrálási modell (Duda, Hart, 1973) adaptálása a harmadik fejezet magánhangzó felismerési adatbázisára lehetővé tette a korai és kései integrálási modell összehasonlítását. Ebben a kísérletben -12 dB jel/zaj viszony esetén csak a kései integrálás múlta felül mind az akusztikus, mind a vizuális egymódusú felismerési eredményeket. Sajnos a félszótag vagy hangpár alapú felismerés esetében a potenciális változatok száma annyira megnő, hogy a kései integrálás megvalósítására ma még nincs lehetőség.

A mai sikeres beszédfelismerő rendszerek szóalapú felismerést végeznek. Agglutináló nyelvekre ezek az eljárások nem adaptálhatók. A fonémák a koartikulációs hatások miatt nem ismerhetők fel biztonsággal. A szó és hang közötti időtartamú alapegységekre több jelölt is kínálkozik. A szótagok és hanghármasok nagy száma megnehezíti a kellő számú tanító min-

tát tartalmazó adatbázis létrehozását. A hangpár általánosan elfogadott alternatíva. Vicsi Klára (1995) kimutatta, hogy egy szöveg részleges lefedéséhez a félszótagokból kell a legkevesebb. Mind az audiovizuális, mind a bővebb akusztikus adatbázison elvégeztem a hangpár és félszótag alapú gépi felismerés összehasonlítását.

A gépi beszédfelismerési eredmények azt mutatták, hogy a pixel bázisú lényegkiemelés a teljes artikulációs terület feldolgozásával többet javít az akusztikus felismerési eredményeken mint a geometriai jellemzők, de nem kínálja az artikuláció analízisének lehetőségét. A beszédfelismerési kísérletek és a vizuális lényegkiemelés eredményei alkalmasnak bizonyultak egy vizuális beszédszintetizátor (beszélő fej) alapadatainak szolgáltatására. A vizuális beszédszintézis mintegy egy évtizedes múltra tekint vissza. A fonémák vizuális megfelelőit az angol kifejezés (viseme) mintájára vizémának neveztem el. Az irodalomban fellelhető eredmények, elsősorban Massaro (1998) munkássága inspirálta a vizémák meghatározását magyar nyelvre. A szakirodalom szerint a vizémák kulcskereteket (keyframe) alkotnak, és ezek között interpolációval alakulnak ki a közbenső alakzatok. Ez a megközelítés túlságosan intenzív száj- és nyelvmozgáshoz vezetett. A koartikulációs hatások figyelembe vételére dominancia osztályokat vezettem be, és minden vizéma minden jellemzőjét a domináns, rugalmas, vagy határozatlan osztályba soroltam. 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. Az általam kidolgozott algoritmusok programozását két diplomaterv keretében egyetemi hallgatók végezték (Mátyás, 2003; Ferenczi, 2004).

2. Az emberi audiovizuális beszédérzékelés vizsgálata

A beszédfelismerés néhány kérdése csak az emberi kommunikáció vizsgálatával válaszolható meg. Ebben a fejezetben a kiegészítő vizuális jel hatását vizsgálom a beszédértésre. A zajmentes beszéd megértése szinte hibátlan vizuális támogatás nélkül is, ezért a méréseket additív zajjal terhelt beszéddel végeztem. A vizsgálatok során az audiovizuális beszédfeldolgozás néhány alapkérdésére kerestem a választ: Mennyivel javul a mássalhangzó felismerés, ha az akusztikus jelet vizuális jellel egészítjük ki? Segíti-e a vizuális jel a gerjesztés módjának meghatározását? Van-e különbség a hangképzés helye szerint a hangok vizuális támogatottsága között (Czap, 1998)?

Ezeknek a kérdéseknek a megválaszolása érdekében mássalhangzó érthetőség vizsgálatot végeztem. A V_1CV_1 típusú hangsorok ugyanazon magánhangzók között tartalmazták a felismerendő mássalhangzót. Különböző jel-zaj viszony mellett végeztem a vizsgálatot, a zaj normál eloszlású volt, majd több (4) beszélő hangjának összeadásából keletkezett zavaró jel (beszédkórus).

Ismeretes, hogy a szájról olvasás segíti a beszéd megértését, különösen zajos környezetben és hallássérültek esetében. Ehhez hasonlóan javíthatja a korlátozott képességű gépi beszédfelismerő felismerési arányát a videó jelből nyerhető információ.

A bimodális beszédfelismerés megértéséhez először célszerű megvizsgálni, hogy az emberi beszédfelismerésben mennyire járul hozzá az artikuláció megjelenítése a beszéd felismeréséhez.

2. 1. A mérés körülményei

A kísérletben résztvevő személyek fonetikai előképzettség nélküli, műszaki szakokon tanuló egyetemi hallgatók voltak. A V_1CV_1 típusú hangkapcsolatok (pl.: ete, ama) kétszeri meghallgatása után a mássalhangzót jegyezték le. Minden magánhangzó- mássalhangzó pár szerepelt a minták között. Korlátozott időt (kb. 2 másodperc) kaptak a válaszra. Egyik esetben csak a hangot hallották, a másik esetben a hang mellett a beszélő képét is látták. Előbb az akusztikus sorozatot, majd a képpel kiegészített minták sorát teszteltük. A szavak sorrendje a két mérési sorozatban egymástól függetlenül, véletlenszerűen változott. A képet egy közös TV készüléken figyelték, a hangot is közös hangszóróból hallották. A pillanatnyi jel-zaj viszony -6, 0, 6 és 12 dB volt. Az eltérő minőségű minták egy mérési sorozaton belül véletlenszerűen váltakoztak. A minták sorrendjének kialakításánál ügyeltem arra, hogy a korábbi mintákból ne tanulhassanak a kísérlet során.

A mérésben alkalmanként résztvevő 10-15 hallgató minden mintát közvetlenül egymás után kétszer hallgatott meg. Az ezt követő kb. 2 másodperces szünet alatt adták meg a választ. A tesztlapon a sorszám mellett kihagyott helyre írták be a mássalhangzó betűjelét. Egy tesztlapon 20-25 szó szerepelt, rendszerint két sorozatot teszteltünk egy alkalommal, a mérési gyakorlatok elején. Összesen tízféle tesztlapot használtunk.

A pillanatnyi jel-zaj viszonyon azt értem, hogy rövid beszédszakaszokra (5 ms) határoztam meg a zajgenerátor kimenő jelszintjét. Ezt úgy értem el, hogy a beszédszakasz mintáinak számával megegyező hosszúságú zaj mintáinak kisorsolása után meghatároztam a beszédszakasz és a zaj aktuális összenergiáját. A beállítani kívánt és az így kapott pillanatnyi jel/zaj viszony összehasonlítása után megkaptam, hogy hányszorosára kell változtatni a zaj amplitúdóját.

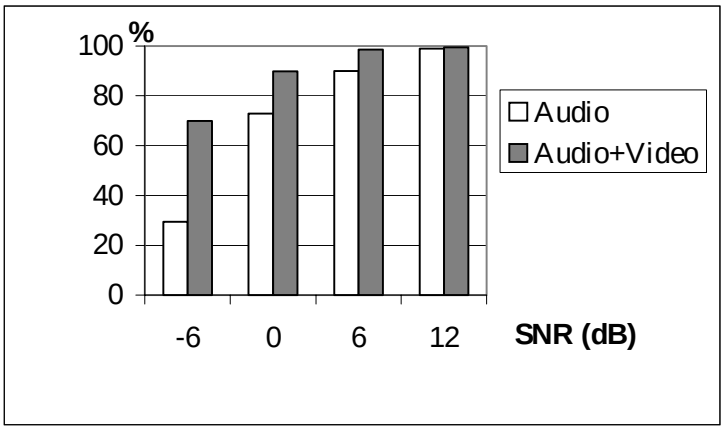
A szó teljes időtartama alatt állandó zaj amplitúdó mellett a kisebb energiájú mássalhangzókat jobban sújtotta volna az átlagos zaj, mint a nagyobb energiájú mássalhangzókat és a magánhangzókat. Azért választottam ezt a megoldást, mert nem a zajforrás modellezése volt a célom, hanem a zaj zavaró hatását vizsgáltam. A V_1CV_1 hangsor esetében az összenergiát elsősorban a magánhangzó határozza meg, így a vizsgált mássalhangzó idejében alkalmazott zaj nemcsak a vizsgált mássalhangzó, hanem a környező magánhangzó energiájától is függne a szó ideje alatt állandó zajteljesítmény esetén. Pl.: az *ata* és *iti* hangsoroknál egy átlagolt teljesítményű zaj az *a* nagyobb energiája miatt az *ata t*-jénél sokkal nagyobb lenne, mint az *iti t*-jénél. Kétféle zajforrást használtam:

- a beszéd sávjában (0-11025 Hz) normál eloszlású zaj
- 4 beszélő hangjának összeadásával keletkezett zavaró jel

A MATLAB *randn* függvényének teljesítmény sűrűség spektruma egyenletes a jel frekvenciatartományában. Az eredmények 10 166 válasz kiértékelésén alapulnak. Egy válasznak egy hallgató egy hangsor meghallgatása után beírt mássalhangzóját tekintem. A válaszok száma a tesztlapon szereplő minták száma és a kísérletben résztvevő hallgatók száma szorzatának összegzésével adódik.

2. 2. Az eredmények

Az eredményeket a jel-zaj viszony függvényében láthatjuk, az első ábra a helyes mássalhangzó felismerés arányát mutatja különböző jel-zaj viszony mellett.



2. 1. ábra. Mássalhangzó felismerési arányok a jel-zaj viszony függvényében

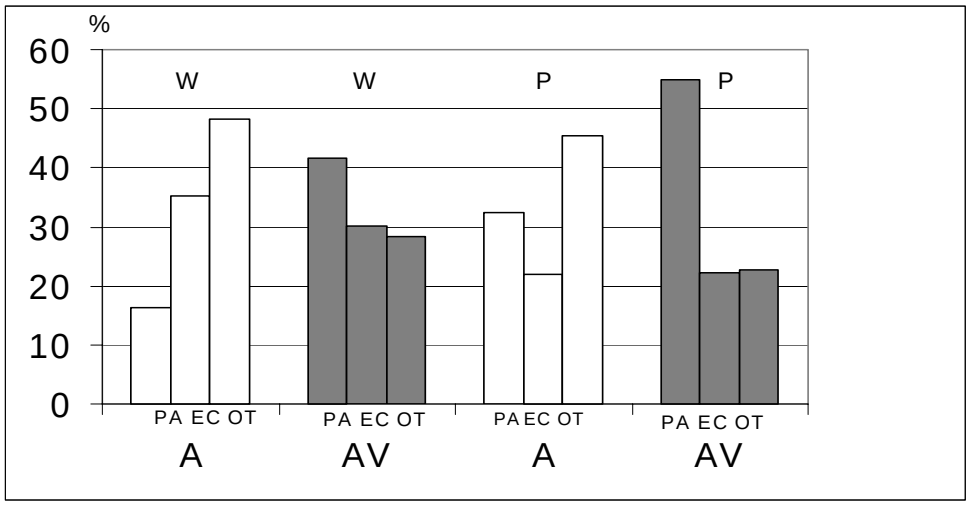
A helyes mássalhangzó felismerés aránya ebben a kísérletben az audiovizuális felismerésnél közelítőleg megegyezik a 6 dB-lel jobb jel-zaj viszony esetében tapasztalt akusztikus felismerési aránnyal. A mérési eredmények igazolják az általánosságban megfogalmazott állítást, hogy a vizuális jel különösen zajos beszédnél segíti a beszédfelismerést. Minél kisebb a jel/zaj viszony, annál nagyobb a mássalhangzók helyes felismerési arányainak különbsége az artikuláció vizuális megjelenítésével segített mássalhangzó felismerés javára.

2. 3. A tévesztések elemzése

A tévesztések elemzésével arra keresem a választ, hogy a beszélő képe a hangképzés mely jellemzőit segít pontosabban felismerni. A hangképzés helyének azonosítását és a gerjesztés módjának meghatározását hasonlítottam össze a vizuális támogatással és anélkül tapasztalt tévesztések elemzésével. A valódi, és a hibás válaszban feltételezett hangok képzési helyét, illetve gerjesztési módját hasonlítottam össze. A téves válaszokat három csoportba soroltam:

- a helyes és a hibás mássalhangzó képzési helye azonos, a valódi mássalhangzó helyett ugyanazon a helyen képzett hangot jelöltek meg (bilabiális, labio-dentális, alveoláris, prepalatális, palatális, veláris) (*PA*)
- a helyes és a hibás mássalhangzó gerjesztési módja azonos {zárhang (*p, t, k, ty, b, d, g, gy*), réshang (*f, sz, s, h, v, z, zs*), félmagánhangzó (*m, n, ny, r, l, j*), affrikáta (*c, cs*)} (*EC*)
- egyéb: a valódi és a feltételezett mássalhangzónak sem a képzési helye, sem a gerjesztés módja nem egyezik meg (*OT*)

Az 2. 2. ábrán látható eredmények -6 és 0 dB jel-zaj viszonyhoz tartoznak, mivel +6 és 12 dB-nél kevés tévesztés mutatkozott. (A *dz* és *dzs* hangok nem szerepeltek a kísérletben.)



2. 2. ábra. Tévesztési osztályok. A helyes és hibás hang képzési helye (PA), gerjesztési módja (EC) vagy egyik sem (OT) azonos, akusztikus (A) és audiovizuális (AV) jelekkel vizsgálva. A zavaró jel normál eloszlású zaj (W), illetve beszédkórus (P).

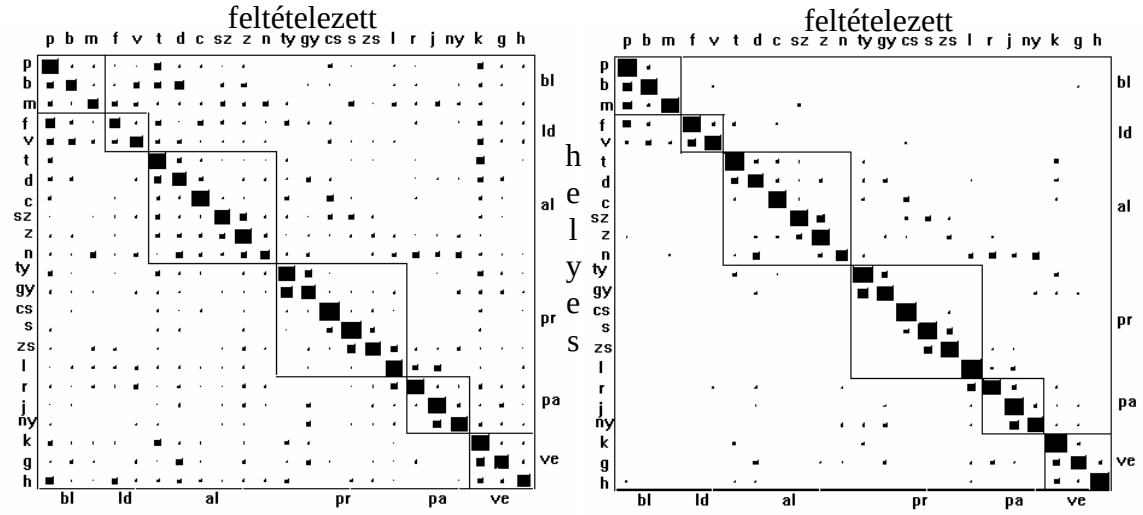
Audiovizuális tesztek esetén a hangképzés helyét megfelelően azonosító válaszok teszik ki a tévesztések legnagyobb részét. Csak hang esetén a hibák legnagyobb részénél sem a képzés helyét, sem a gerjesztés módját nem találták el. Akusztikus jelnél a képzés helye a tévesztések 16,3 %-ában volt helyes additív, normál eloszlású zajjal terhelten. Ez az arány megfelel a véletlen válaszok egy képzési helyre eső arányának. (A hatféle képzési hely egyike.) A hangképzés helyét helyesen kijelölő téves válaszok aránya a kiegészítő vizuális jel hatására 41,5 %-ra nőtt.

A gerjesztési módok azonosítása valamivel sikeresebb, ha a beszélő látható, a vizuális jel a gerjesztés módjának elkülönítését mérsékelten segíti. A gerjesztés módja a téves válaszok 45,4 %-ában volt helyes akusztikus, míg 58,6 %-a audiovizuális teszt esetén. (Ezek az arányok tartalmazzák azokat

a gerjesztési módot helyesen megadó téves válaszokat is, amelyek a képzés helyét is jól határozták meg.)

A téves válaszok elemzése alapján megállapítható, hogy a vizuális támogatás a képzés helyének azonosításában nyújt jelentős segítséget.

Az 2. 3. ábrán a teljes tévesztési mátrix látható. Az audiovizuális mássalhangzó felismeréshez tartozó ábra láthatóan rendezettebb a képzés helye szerint. Az akusztikus vizsgálat esetében a hibásan értékelt mássalhangzók 16,3 %-a egyezett meg a hangképzés helye szerint a helyes megoldás képzési helyével, a vizuális jellel kiegészítve ez az arány 41,5 %-ra nőtt.

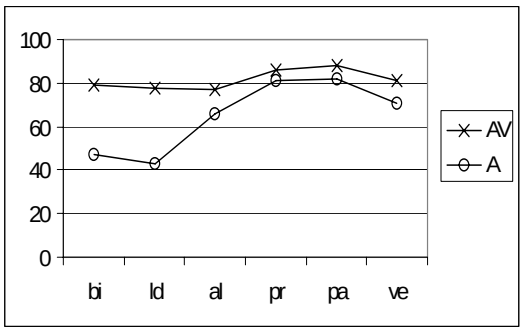


1. 2.

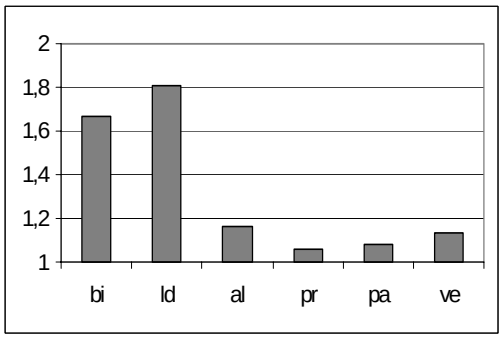
2. 3. ábra. A tévesztési mátrix Hinton diagramja akusztikus (1.) és audiovizuális (2.) vizsgálat alapján. (A folt területe arányos a függvényértékkel.)

A 2. 4. ábra a képzés helye szerinti bontásban tartalmazza a helyes mássalhangzó felismerési arányt (1.), valamint az audiovizuális és akusztikus

felismerési arányok hányadosát (2.), ez a hangképzés helye szerint kifejezi az audiovizuális eredmények javulását az akusztikusokhoz képest.



1.



2.

2. 4. ábra. A helyes mássalhangzó felismerés aránya a képzés helye szerint (1.), a felismerési arány növekedése a csak akusztikus jellel végzett vizsgálathoz képest (2.).

A humán mássalhangzó felismerési kísérletek alapján a vizuális jellel kiegészített gépi felismerőtől azt várhatjuk, hogy a hangképzés helyét jobban tudjuk azonosítani, a felismerési arány javulására elsősorban az elül képzett hangoknál számíthatunk.

3. Az arc mely részei hordozzák a releváns vizuális információt?

A 2. fejezet kísérletei is igazolták, hogy az emberi beszédfelismerés a vizuális információ felhasználásával jobb eredményeket mutat, mint pusztán az akusztikus jelre hagyatkozva. Nem tudjuk azonban azonosítani, hogy a kép mely részlete segítette a beszéd jobb felismerését, az akusztikus és vizuális jel egyesítése automatikus művelet (Tiippana, 1999). A gépi szájról olvasás tervezéséhez célszerű ismerni, hogy az emberi kommunikációban az arc mely részei mennyire segítik a beszéd felismerését. Érthetőség vizsgálatot végeztem zajos beszéddel úgy, hogy az arc egyes részeinek elfedésével a beszélő arcának csak részletei voltak láthatók (Czap, 2000). Ezeket a vizsgálatokat zajos beszéddel végezhetjük el, hiszen a jó minőségű beszéd tökéletesen érthető, a vizuális támogatás hatása nem mérhető.

Várakozásaink szerint az ajakforma lényeges vizuális jellemző. Fontos kérdés, hogy a nyelv és a fogak láthatósága számottevő javulást okoz-e, érdemes-e a lényegkiemelésnél erőfeszítéseket tenni a leírásukra. Vajon az arc egyéb részei is hozzájárulnak a beszédfelismerési eredmények további javulásához, ha a beszélő egész arca látható? Ezekre a kérdésekre kerestem a választ szubjektív teszt segítségével.

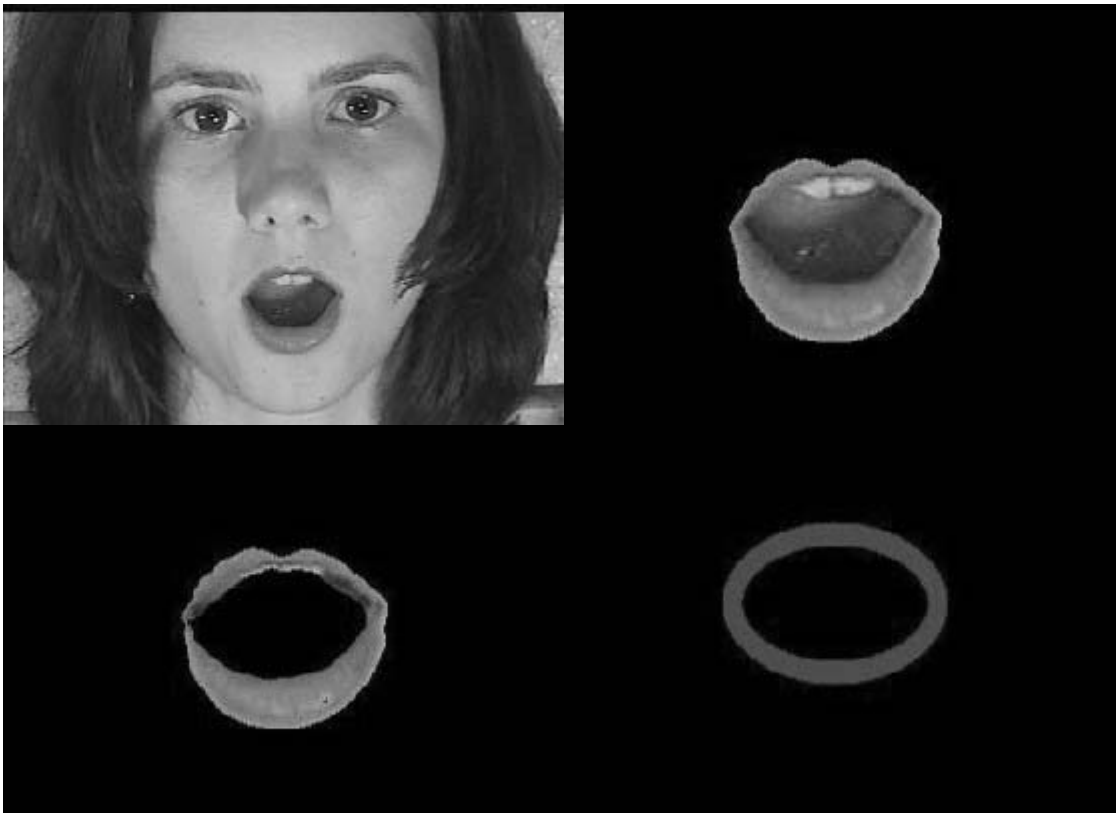
3. 1. Érthetőség vizsgálat

Az arc különböző részletei által hordozott vizuális támogatás mérésére érthetőség vizsgálatot végeztem. Mássalhangzó felismeréshez V_1CV_1 hangsor (pl.: eke, ata) középső mássalhangzóját kellett a tesztlapra a sorszámmal jelzett rubrikába beírni. Magánhangzó felismerésekor C_1VC_1 szótagok (pl.: lol, tet) középső magánhangzóját kellett megadni.

A mérésben 78, fonetikai ismeretekkel nem rendelkező egyetemi hallgató vett részt. A 10-15 fős csoport egy közös TV készüléken nézte a videó jelet és egy közös hangszóróból hallotta a hangot, minden hangsort kétszer egymás után. A válaszra korlátozott idő – kb. 3 másodperc – állt rendelkezésre. A tesztlapra a sorszám melletti sorba írták be a megfelelő betűt. Egy sorozat 20-25 mintát tartalmazott, egy alkalommal általában két sorozatot hallgattak végig a tesztlapok beszedése és kiosztása alatti rövid szünettel. Egy mérés kb. 10 percet vett igénybe.

A vizuális jel az alábbiak valamelyike lehetett (3. 1. ábra):

- a beszélő arca
- a beszélő szája (ajkak, fogak, nyelv)
- a beszélő ajkai
- az ajkak méreteit utánzó ellipszis



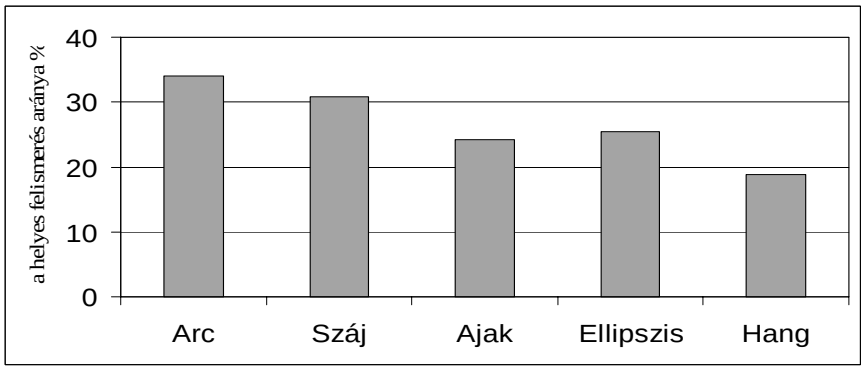
3. 1. ábra. Vizuális jel az arc különböző részeinek maszkolásával és ellipszissel

Az akusztikus jel zajos beszéd volt. A magánhangzó felismerési kísérleteket -18 dB pillanatnyi jel-zaj viszony mellett végeztük. A mássalhangzó felismerési vizsgálatoknál a jel-zaj viszony -6 dB volt. A pillanatnyi jel-zaj viszonyhoz szükséges zaj amplitúdót 5 ms-onként állítottam be.

Egyes kísérleteknél csak az akusztikus jel volt jelen (sötét képernyő), máskor hang nélkül, csak a kép alapján próbáltuk a magánhangzót vagy a mássalhangzót felismerni. Az eredmények 11 623 válasz kiértékelése alapján születtek, ezek közül 9 625 a mássalhangzók, 1 998 a magánhang-

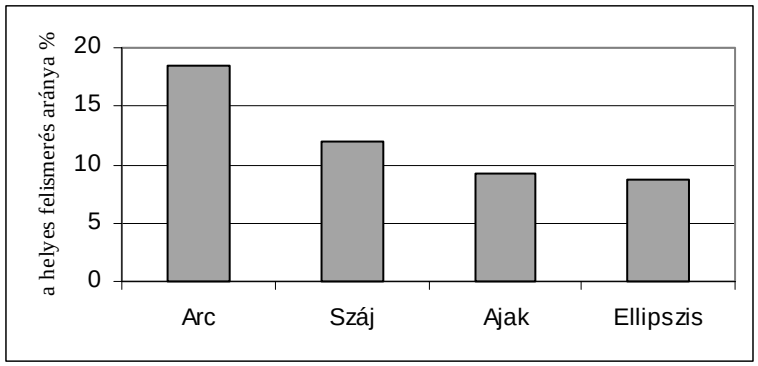
zók felismerését szolgálta. Egy hallgató egy hang betűjelének leírásával adott egy választ.

A 3. 2. ábra a mássalhangzó érthetőség vizsgálat eredményét mutatja audiovizuális jel esetén.



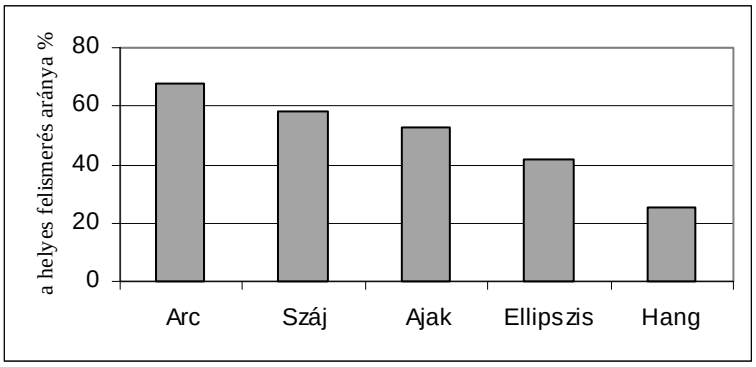
3. 2. ábra. A mássalhangzók felismerési aránya. A hang pillanatnyi jel-zaj viszonya: -6 dB, a képen az arc részletei, vagy a száj vízszintes és függőleges méretével megegyező kis- és nagy tengelyű ellipszis volt látható. A *Hang* megnevezésű oszlop esetében a képen csak a minta sorszáma volt látható, vizuális információt nem kapott a hallgató.

A 3. 3. ábrán a mássalhangzókra csak a kép alapján, hang nélkül kapott felismerési arányok láthatók.



3. 3. ábra. A mássalhangzók felismerése csak a kép alapján

A 3. 4. ábrán a magánhangzók felismerési eredményeit láthatjuk.



3. 4. ábra. A magánhangzók felismerési arányai. A pillanatnyi jel-zaj viszony -18 dB.

Várakozásunkkal egyezően, minél többet látunk a beszélő arcából, annál jobban segíti a kép a beszéd felismerését. Mivel jelentéssel nem bíró szavakról van szó, az arckifejezés nem növelhette az érthetőséget az egész arc megmutatása esetén sem. A javulás a száj (ajkak, fogak, nyelv) figyeléséhez képest inkább annak tulajdonítható, hogy az arc redői kiemelik a szájmozgást, segítik az artikuláció pontosabb követését.

A nyelv és a fogak láthatósága észrevehetően javítja a felismerési arányokat, a vizuális lényegkiemelésnél érdemes ezek leírására törekedni.

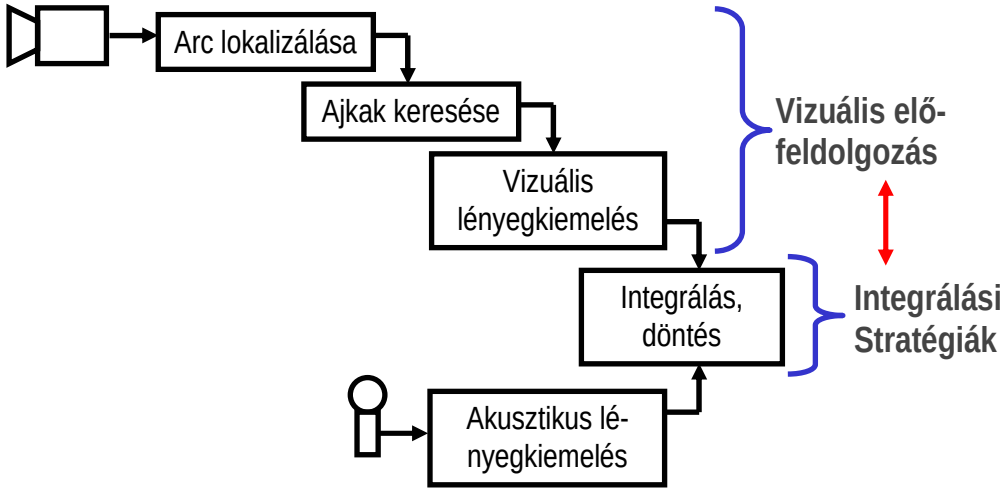
Magyarázatra szorul, hogy a 3. 2. ábrán – a másik kettőtől eltérően – az ellipszis modell eredményei felülmúlták az ajkakkal mért felismerési arányokat. A választ a hangképzés helye szerinti elemzéssel kaptam meg. A bilabiális hangok mutattak kiugróan jó eredményeket. Ezt azzal magyarázom, hogy zárt ajkaknál a belső ellipszis eltűnt, a homogén elliptikus folt

jobban megmutatta az ajkak záródását, mint az ajkak tónusos képe. A 23 mássalhangzó 13 %-át kitevő *p, b* és *m* esetében a 20% körüli átlagnál sokkal jobb felismerési arány magyarázza a vártnál néhány százalékkal jobb eredményt.

Az ellipszis modell annak ellenőrzését szolgálta, hogy a száj szélessége és nyitottsága (az ajkak belső kontúrjának vízszintes és függőleges mérete) alkalmas-e az ajakforma leírására. Annak ellenére, hogy szokatlan ábrázolással szembesültek a kísérletben résztvevők, az ellipszis modell eredményei megközelítették az ajkak megmutatásával mért eredményeket.

4. A vizuális lényegkiemelés

A bimodális beszédfelismerés egyik kulcskérdése a vizuális lényegkiemelés. A kutatások egyik jelentős területe a vizuális információt hordozó jellemzők meghatározása. Az akusztikus jel leírására kipróbált módszerek állnak rendelkezésre, a vizuális jellemzők kiválasztása és ezek leolvasása a képről azonban még kevésbé kidolgozott, tehát a másik lényeges kérdés, hogy a választott jellemzők hogyan nyerhetők ki a képből. Ehhez szükség van az arc lokalizálására, a száj körüli terület kijelölésére és a jellemzők meghatározására. (4. 1. ábra)



4. 1. ábra. Az audiovizuális beszédfelismerés sémája

A lényegkiemelés célja a több tíz megabit per szekundumos sebességű videojel adatmennyiségének radikális csökkentése és az artikuláció szempontjából releváns, a vizuális beszédet leíró jellemzők kinyerése. Nem elhanyagolható szempont a számítások komplexitása, a végrehajtási idő. A

kétmódusú beszédfelismerés kezdetén – mintegy két évtizeddel ezelőtt – a videojel feldolgozását a beszélő arcán színes „szépségflastrommal” megjelölt nevezetes pontok (pl.: állcsúcs, szájsarkak) azonosítása és néhány geometriai jellemző mérése jelentette. A vizuális beszédfelismerés egyik atyja, Petajan (1984) küszöbdetekcióval határozta meg az ajkak szélességét, nyitását és területét. Úttörő munkájának rövid idő alatt számos követője akadt és kialakultak az audiovizuális beszédfeldolgozás kutató műhelyei. A lényegkiemelés és az integrálás különböző módszereit dolgozták ki, amelyek azonban nehezen hasonlíthatók össze, mivel más-más adatbázison dolgoztak. Nemzetközi projektek keretében létrehoztak audiovizuális adatbázisokat, de a vizuális lényegkiemelésre még nem alakult ki követendő módszer.

Kedvelt egyszerűsítés a jól azonosítható marker jelek (Benoît, 1996a) arcra festése. Pl.: a kék rúzs (Dalton, 1996) színméréssel egyszerűen elkülöníthető, hiszen ez a színösszetevő egyébként jellemzően nem fordul elő a bőr színében. Ma is elsősorban kutatási célú elemzés folyik, a beszélő által tolerálható előkészítés elfogadható. Régebben nem volt ritka a beszélő fejének fix pozícióba rögzítése sem, ma enyhe természetes mozgást kezelni lehet. A technikai háttér és az eljárások fejlődése ma már lehetővé teszi a kötöttségek enyhítését.

Általános feltételek esetén, a legkisebb korlátozás alkalmazásával, a beszélő szabad mozgása mellett végeznénk el a vizuális lényegkiemelést. A televíziós képfelbontás nagy távolságból vagy széles látószöggel nem teszi lehetővé az artikulációs terület elég finom ábrázolását. Használható mére-

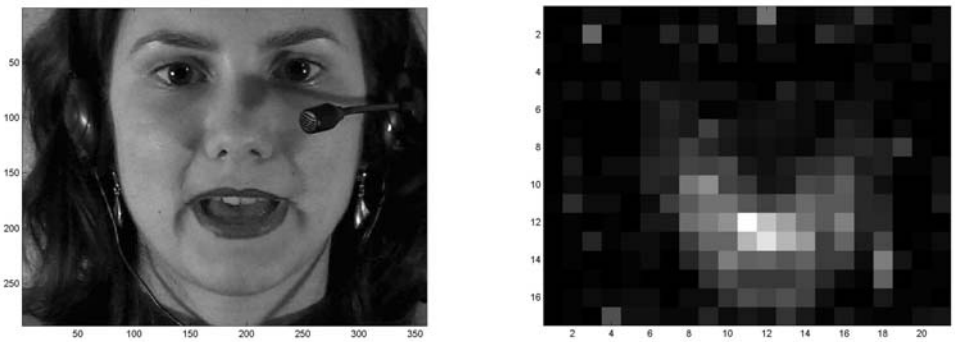
tű és felbontású képet akkor kapunk, ha a kamera mozgatásával követjük a beszélő mozgását, a zoom és a fókusz folyamatos állításával biztosítjuk a megfelelő nagyítást és képélességet. Diplomamunkák (Berecz, 2000; Drótos, 2000) keretében eljárást dolgoztunk ki a beszélő követésére. A bőrszín alapján az arc megkereshető, majd a kamera mozgató zsámoly elektronikus vezérlésével a kép centrumába állítható. Az elektronikusan vezérelhető objektív zoom és fókusz értékének megfelelő beállításával portré jellegű kép állítható elő.

A beszélő szabad mozgásának követését azonban ezzel még nem tekinthetjük megoldottnak. Mozgás közben a beszélő rendszerint nem szemben áll a kamerával. Kis szögelfordulás vetemítéssel kezelhető. Amennyiben a szemek és az orrhegy-orrlyuk által alkotott háromszöget sikerül kijelölni, a vetemítés paraméterei meghatározhatók. Nagyobb szögelfordulás és a megvilágítás helyfüggése, valamint a fény beesési szögének változása a fej elfordulásával túlságosan változatos képeket eredményez. A szabad mozgású beszélő követése a jövő feladata, a mai kutatási irányok alapján valószínűsíthető, hogy az első alkalmazások web-kamerás, vagy más rögzített pozíciójú (pl. autós) környezetben fognak születni (Neti, 2003; Bailly, 2004). A kísérletben szereplő felvételek fix kamera beállítással, ülő helyzetben természetes fejmozgású beszélővel készültek.

A következő feladat az artikuláció szempontjából fontos terület (region of interest) kijelölése. Beszéd közben az arc legintenzívebb mozgást végző területe a száj és az áll környezete. Képsorozat vizsgálata alapján mozgásbecsléssel megállapíthatjuk, hogy egy képrészlet mennyivel tolódott el az

előző képhez képest. Az így kapott mozgásvektorok összegzéséből megállapítható a bemondó képein a legintenzívebb mozgást mutató terület. A mozgásbecslés a mozgóképek tömörítésére szolgáló eljárások egyik szokásos lépése. Az előző képkockán szereplő képrészletek eltolásával próbálja összerakni a képet. A mozgásvektorok azt fejezik ki, hogy a pl.: 16x16 pixeles képrészlet az előző képből milyen irányú és nagyságú eltolással állítható elő.

A 4. 2. ábrán 8 képkockából (160 ms) határoztam meg mozgásvektorok abszolút értékének összegét, ennek intenzitás képét látjuk a beszélő képe mellett. A legintenzívebb mozgás a száj területén tapasztalható.



4. 2. ábra. A beszélő képe és a mozgásvektorok abszolút értékének összege.

Másik lehetőség a vizsgálandó terület kijelölésére az arc jellegzetes helyeinek azonosítása alapján megtalálni a feldolgozandó képrészletet. Olyan alakzatot kell keresni, amely nagy megbízhatósággal azonosítható a képen. Erre a szemek és az orrhegy-orrlyuk területe a legalkalmasabb (Massaro, Stork, 1998; Nankaku et al., 1999). A szem szemüvegeseknél és pislogás idején nehezen található meg, ezért ebben a dolgozatban a feldol-

gozandó terület az orrhegyhez képest kijelölt, a száj környezetét nagy biztonsággal tartalmazó ablak kivágásával alakult ki. A terület kijelölésekor feltételezzük, hogy a kamera távolsága és látószöge állandó.

A képkockánkénti feldolgozás, a választott jellemzők leírása, diszkrét idejű változókat eredményez. Rendkívül fontos a hang és kép közötti szinkron, valamint több kamerás (pl.: elöl- és oldalnézet) felvételek esetében a vizuális jelek egymás közötti szinkronizálása. McGrath (1985) kutatásai szerint az audiovizuális beszédfelismeréshez a hang és a kép 40 ms-nál kisebb elcsúszása még nem okoz veszteséget, ez PAL rendszerű felvételnél egy képidőnek felel meg.

4. 1. A geometriai alapú lényegkiemelés

A lényegkiemelés egyik fő kérdése, hogy milyen jellemzők hordozzák a beszédfelismerés szempontjából lényeges vizuális információt. A másik fontos kérdés, hogy a kiválasztott vizuális jellemzők hogyan nyerhetők ki a képből. A geometriai alapú felismerésnél az artikuláció különböző paramétereivel próbáljuk leírni a vizuális beszédjel változásait. A leggyakoribb vizuális jellemzők:

- az ajkak szélessége, vagyis az ajkak belső kontúrjának vízszintes mérete
- az ajaknyílás, az ajkak belső függőleges mérete
- az előbbi jellemzők az ajkak külső kontúrjára vonatkozóan
- az ajkak belső kontúrja által bezárt terület
- az ajkak külső vonalával határolt terület
- az áll mozgása pl.: az orrhegyhez képest
- oldalnézeti képen az ajkak előre mozgása
- a nyelv és a fogak láthatóságának leírására általánosan alkalmazott jellemző még nem alakult ki

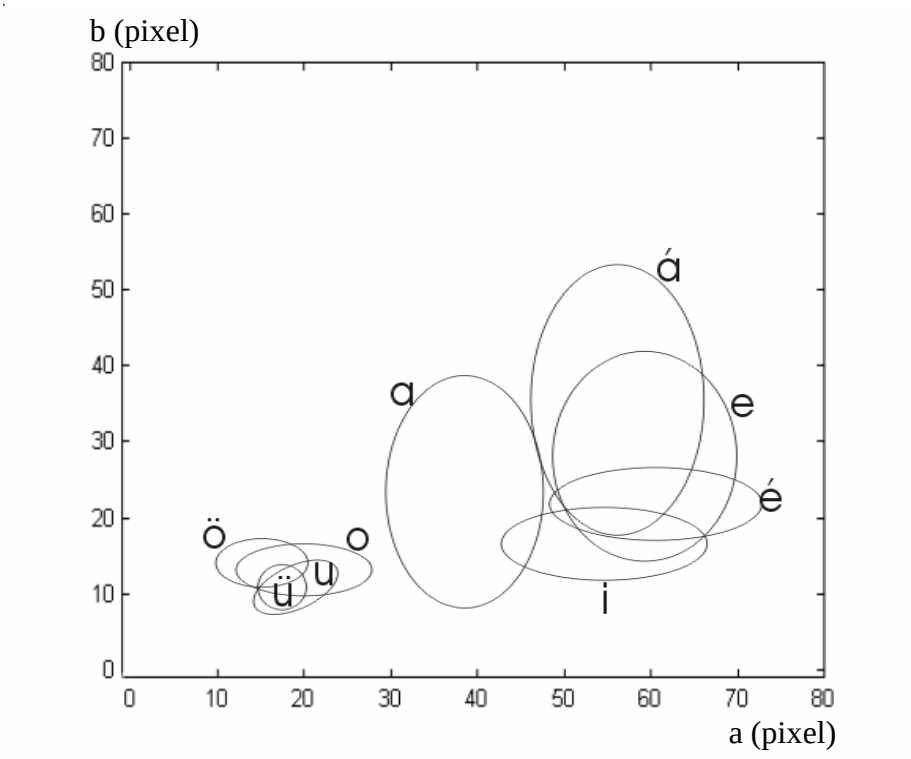
A felsorolt paraméterek erősen redundánsak, szerepüket több kutató vizsgálta (Benoît, 1995; Hennecke et al., 1996; Lavagetto, 1996). Az egymással összefüggő jellemzők közül pl.: a szájnyílás belső méretei (4. 8. ábra *a* és *b*) meghatározzák a területet, így az nem jelent új információt. Az *a* és *b* aránya eltérő ajakkerekítéses és ajakréses hangoknál. Ajakréses hangoknál a külső kontúr méretei, területe csak kevéssel nagyobb a belső mére-

teknél, ajakkerekítéses hangoknál a külső méretek számottevően nagyobbak. Az ajkak előre mozgása az ajakkerekítéses hangokra jellemző. Ezek a változók kevés új információt hordoznak az *a* és *b* arányához képest. Az áll mozgása összefüggésben áll a száj nyitottságával és a fogak láthatóságával, hiszen az alsó fogsor az állal együtt mozog. Redundáns jellemzők alkalmazása robusztusabb felismeréshez, vagy ellentmondó jellemzők esetén az eredmények romlásához vezet. A vizuális jellemzőknek az akusztikus jelet kell erősíteniük, a két paraméter egymásra hatása módosíthatja egyes jellemzők hasznosságát.

4. 1. 1. A geometriai jellemzők kiválasztása

Az audiovizuális beszédfelismerés szubjektív tesztekkel végzett elemzése alapján kezdtem a vizuális lényegkiemeléshez. A szájról olvasás gépi megvalósításának természetes megközelítése, hogy az artikuláció látható jellemzőit használjuk a vizuális jel leírására. Az ajaknyílás és ajakszélesség mellett a nyelv és a fogak láthatóságát leíró változót kerestem.

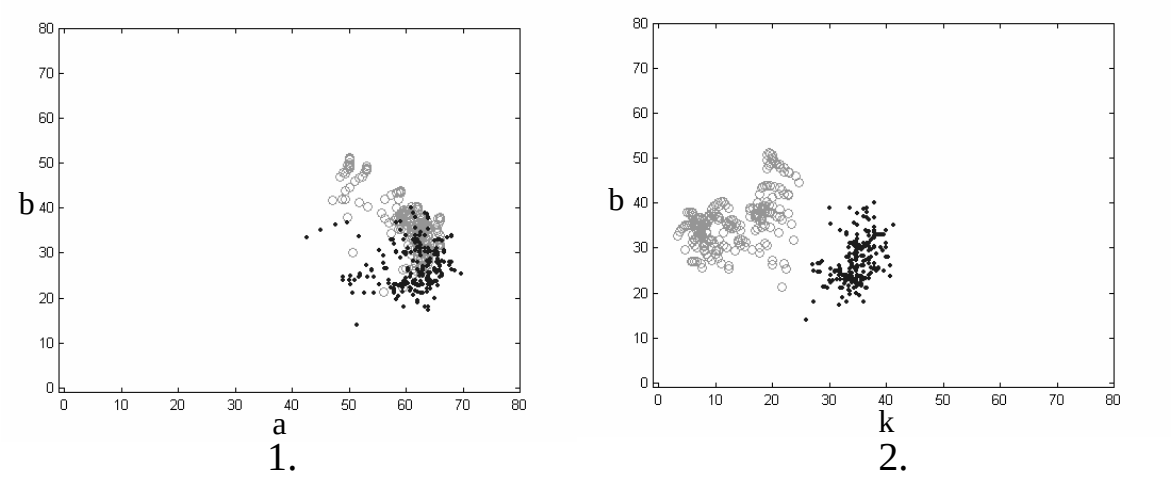
A mérési eredmények alapján a magyar magánhangzók állandósult szakaszának a geometriai lényegkiemelés eredményeképpen kapott ajakszélesség és ajaknyílás tartományait a 4. 3. ábrán ábrázoltam.



4. 3. ábra. A magánhangzók állandósult szakaszának ajakszélesség (*a*) és ajaknyílás (*b*) tartományai. A vízszintes tengely az ajakszélességet, a függőleges az ajaknyílást mutatja pixelben kifejezve.

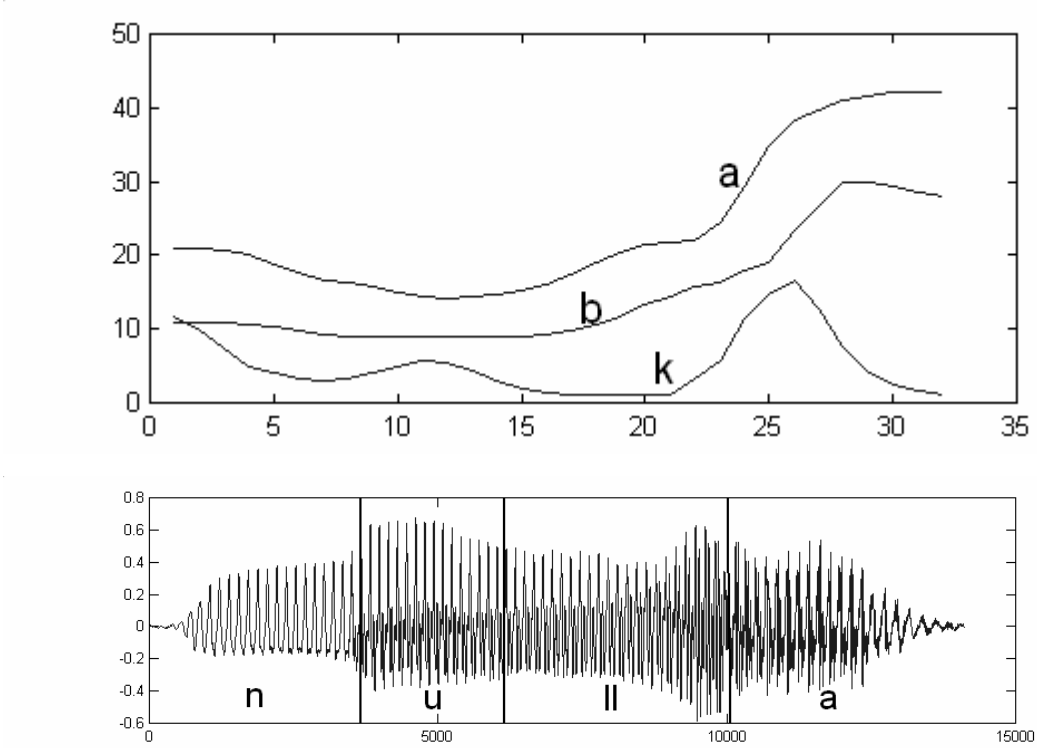
Az ajakszélesség és az ajaknyílás az ajakforma leírására hivatott. A képmomentumokból származtatható intenzitás faktor (*k*) - amely valójában a szájnyílás átlagos világossága - a nyelv és a fogak láthatóságának jellemzésére szolgál. A *k* intenzitás faktor a hátul képzett hangoknál a legkisebb (pl.: k, u). Közepes értékű, ha elül képzett hangoknál a nyelv látható (pl.: e, i). Legnagyobb a *k* értéke, ha az állkapocs zárt állása miatt a fogakat látjuk a szájnyílásban (pl.: s, cs).

A 4. 3. ábra tanúsága szerint a magánhangzók ajakméretei átfedéseket tartalmaznak. Egyes hangok nem különíthetők el az ajakméretek alapján, eltérő azonban az intenzitási tényezőjük. Pl.: Az *á* és az *e* hangok ajakméreteiben jelentős az átlapolódás. Az *e* hangnál a nyelv elöl helyezkedik el, az intenzitás faktor lényegesen magasabb, mint az *á* esetében. Az *e* és az *á* hang középső szakaszának értékeit láthatjuk a 4. 4. ábrán az ajaknyílásszélesség *(a-b)*, illetve az intenzitás-aiaknvílás *(k-b)* síkban.



4. 4. ábra. Az *á* (o) és *e* (.) hangok 1: ajakszélesség (*a*) – ajaknyílás (*b*) illetve 2: intenzitás (k) - ajaknyílás (*b*) párjai.

Az artikuláció dinamikus vizsgálatánál is értékes jellemző lehet a *k* intenzitási tényező. A 4. 5. ábrán a *nulla* szó kimondása közben láthatjuk az ajakszélesség, az ajaknyílás és az intenzitási tényező változását. Az ajakméretek folyamatosan változnak az *u* és *a* jellegzetes értékei között. Az *l* ajakmérete nem domináns, csak az intenzitási tényező változása mutatja meg, hogy a nyelv az *l* után a fogmedertől az alsó nyelvállásba megy át.



4. 5. ábra. Az ajakszélesség (*a*), az ajaknyílás (*b*) és az intenzitási tényező (*k*) változása a *nulla* szó kimondása közben.

A gépi audiovizuális felismerési kísérleteknél – az ajkak előre mozgásának kivételével – megpróbáltam az *a*, *b* és *k* paramétereket a 4. 1. pontban felsorolt jellemzőkkel kiegészíteni. Összhangban a szubjektív teszt eredményeivel és az irodalmi adatokkal, a felismerési eredmények az új paraméterek bevezetésével nem javultak. Az előbbi indokok alapján a gépi

beszédfelismerés vizuális jellemzőinek az ajakszélességet (*a*), az ajaknyílást (*b*) és az intenzitási tényezőt (*k*) választottam.

A választást a későbbi gépi beszédfelismerési eredmények is alátámasztják. Vizsgáljuk meg a vizuális jel hatására bekövetkezett felismerési arány változását a humán és gépi kísérlet eredményei alapján egyaránt!

A szubjektív tesztek eredményei a száj (ajkak, nyelv, fogak) és az ellipszis modell megfigyelése alapján:

4. 1. táblázat. A szubjektív tesztek mássalhangzó és magánhangzó felismerési eredményei a száj és az ajkakat reprezentáló ellipsis megfigyelése alapján

Ábra	3. 2.	3. 3.	3. 4.
Száj	30,8%	12,0%	57,8%
Ellipszis	25,4%	8,7%	41,5%
Hányados	**1,21**	**1,38**	**1,39**

A 4. 2. táblázat az automatikus hangpár felismerési eredményeket mutatja az ajakszélesség (*a*) és ajaknyílás (*b*), valamint az előbbi jellemzők intenzitási tényezővel (*k*) kiegészített paraméterrendszerével.

4. 1. táblázat. Gépi hangpár felismerési eredmények az ajkakat (*a*, *b*) illetve az ajkakat és a nyelv és a fogak láthatóságát leíró jellemzők (*a*, *b*, *k*) alapján

Ábra	7. 1.	7. 2.
a,b,k	22,7%	36,9%
a,b	16,9%	26,6%
Hányados	**1,34**	**1,39**

Az ellipszis, illetve az ajakszélesség és ajaknyílás egyenértékű jellemzők, hiszen az *a* és *b* alapján az ellipszis megrajzolható. Ha elfogadjuk, hogy a száj megfigyelése az ellipszishez képest ugyanannyival növelte a felismerési eredményeket, mint a *k* intenzitási tényező hozzáadása az (*a, b*) ajakjellemzőkhöz képest, az *a, b* és *k* jellemzők az artikulációt a száj megfigyelésével egyenrangúan írják le.

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.

4. 1. 2. Geometriai lényegkiemelés színes képekre

Az előző részben választott jellemzők meghatározása a kép alapján egyszerűbbnek tűnt a színinformáció felhasználásával, ezért az első próbálkozásaim színes képek feldolgozásán alapultak. A 3. fejezetben ismertetett szubjektív tesztek videó jelének előkészítése során a piros szájrúzs színmérésen alapuló maszkolásával sikerült az ajkakat és a szájat (ajkak, nyelv és fogak) kiemelni a képből. Képfeldolgozási ismereteim alapján az alakzat felismerési célokra használt geometriai momentumokra terelődött a figyelmem. A képmomentumokat – számítási egyszerűségük miatt – az elsők között alkalmazták objektumok formájának leírására (Hu, 1962; Mukundan, Ramakrishnan 1998; Dougherty, Giardina, 1987). Az elmúlt négy évtized során sok képfeldolgozási feladatban bizonyították alkalmazhatóságukat. A digitális képekből származtatható jellemzők a képen szereplő objektumok alakjáról nyújtanak hasznos információt. Előnyeik leginkább az alakzat felismerési feladatokban aknázhatók ki, hiszen olyan mennyiségek származtathatók belőlük, amelyek invariánsak az eltolásra, nagyításra és forgatásra.

A *(p+q)*-ad rendű geometriai momentumok definíciója:

$$m_{pq} = \iint_{\zeta} x^p y^q f(x, y) dxdy, \qquad \text{p, q=0,1,2,3….} \tag{4.1.}$$

ahol ζ az objektum helye az *x-y* síkban, amelyet a kép *f(x,y)* világosságfüggvénye jellemez.

A különböző rendű geometriai momentumok a kép világosság eloszlását jellemzik a síkban. A képmomentumok rendszere ilyenformán alkalmas az objektumok alakjának leírására.

Néhány geometriai momentumnak fizikai jelentés is tulajdonítható:
Definíció szerint a nulladrendű momentum *(m_{00})* a kép összintenzitását reprezentálja. Az elsőrendű momentumok *(m_{01} , m_{10})* az intenzitás momentumokat jelentik az *x*, illetve az *y* tengely körül. A kép világosságának súlypontja *(x_0,y_0)* az elsőrendű momentumok normálásával kapható meg.

$$x_0 = m_{10}/m_{00}\ ; \qquad y_0 = m_{01}/m_{00} \tag{4.2.}$$

Az origónak a súlypontba tolásával egyszerűen számíthatók olyan momentumok, amelyek nem függnek az objektum helyétől. Ennek folytán jutunk az eltolásra érzéketlen centrális momentumok meghatározásához:

$$\mu_{pq} = \iint_{\zeta} (x - x_0)^p (y - y_0)^q f(x, y) dxdy, \qquad \text{p,q=0,1,2,3….} \tag{4.3.}$$

A másodrendű centrális momentumok különösen fontosak. A kép világosság függvényének átlag körüli szórásnégyzetét μ_{20} és μ_{02} képviseli, μ_{11} a kovarianciát adja.

A másodrendű centrális momentumok úgy is tekinthetők, mint a kép tehetetlenségi nyomatékai a koordinátarendszer tengelyeivel párhuzamos, a súlyponton átmenő referencia tengelyekre nézve. A kép fő tehetetlenségi tengelyei úgy definiálhatók, mint a súlyponton átmenő két merőleges egyenes, amelyet referencia rendszerként használva μ_{11} eltűnik. A kép így

kapott I_1, I_2 fő tehetetlenségi nyomatékai a referencia tengelyek körül az alábbi összefüggésekkel határozhatók meg:

$$I_1 = \frac{(\mu_{20} + \mu_{02}) + [(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2]^{1/2}}{2}, \tag{4.4.}$$

$$I_2 = \frac{(\mu_{20} + \mu_{02}) - [(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2]^{1/2}}{2}. \tag{4.5.}$$

Ezek a kifejezések felhasználhatók egy olyan ellipszis meghatározására, amelynek a tehetetlenségi nyomatékai megegyeznek az eredeti kép momentumaival. Az *a* és *b* mennyiségek meghatározzák az ellipszis fél nagy- és kistengelyét:

$$a = 2\,(I_1/\mu_{00})^{1/2}\,; \qquad b = 2\,(I_2/\mu_{00})^{1/2} \tag{4.6.}$$

Az egyik fő tehetetlenségi nyomaték tengely *x* tengellyel bezárt Θ szöge:

$$\Theta = \frac{1}{2}\tan^{-1}\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right) \tag{4.7.}$$

A kép ellipszis rendelkezik még egy további tulajdonsággal is: egyenletes a világossága, értéke *k* az ellipszisen belül, nulla azon kívül. Ezzel, és a nulladrendű momentummal definiálható a *k* intenzitás faktor

$$k = \mu_{00}/(\pi ab)\,. \tag{4.8.}$$

Az

$$s = (I_1+I_2)/m_{00} \tag{4.9.}$$

értéket gyakran az objektum szétterültségének, az

$$e=(I_2 - I_1)/(I_2 + I_1) \tag{4.10.}$$

mennyiséget az alakzat elnyúltságának nevezik.

Az emberi audiovizuális felismerési kísérletben szereplő felvételeken az ajkak színmérésen alapuló maszkolásával kapott bináris kép alapján az ajkak külső méreteit reprezentáló ellipszis kis- és nagytengelye meghatá-

rozható. Az ezen belüli terület invertálásával számítható az ajkak belső méreteire jellemző ellipszis kis- és nagytengelye. Az ajkak belső méreteire jellemző ellipszis megvastagításával kapott kép szolgált a 3. fejezetben végzett érthetőség vizsgálat egyik képjeleként. (3. 1. ábra)

Az *a* és *u* hangok ejtésekor kapott ellipsziseket a 4. 6. ábra mutatja.

4. 6. ábra. Az ajkak színmérés alapján, küszöbdetekcióval kapott bináris képe és a számított képellipszis az *a* (balra) és az *u* (jobbra) hang ejtése közben.

1. (*a=55, b=35*) 2. (*a=58, b=35*)

4. 7. ábra. Példa az ajakforma színmérés alapján, küszöböléssel binárisra vágott képére túl magas (1.) és túl alacsony (2.) küszöbérték mellett. Az ábrán feltüntettem a képellipszis modellből számított ellipszis méreteit is.

A geometriai momentumokból számított képellipszis méretek robusztusnak bizonyultak a maszkoláskor használt színhasonlósági küszöbszintekre. A 4. 7. ábra a túl magas és túl alacsony küszöbérték képét és a hozzájuk rendelt ellipszisek tengelyeinek adatait tünteti fel. A lényegesen különböző alakzatok ellenére az ellipszisek tengelyei csak kevéssé térnek el.

Az ajakforma jellemzésére a képellipszis modellből származtatott *a* és *b* tengelyeket használtam fel, a *k* intenzitás faktor reprezentálta a nyelv és a fogak láthatóságát egy gépi magánhangzó felismerési kísérletben. A vizuális beszédfelismerésben az arc különböző részeinek szerepét vizsgáló szubjektív tesztnél használt felvétel képeiből kivágott képkockák szolgáltatták a bemeneti adatokat. A manuálisan szegmentált magánhangzók középső három keretének *a, b* és *k* értéke szolgált a tanítás és tesztelés alapjául. A C_1VC_1 szótagok mássalhangzóit a hangok különböző képzési helyéről választottam (*b, v, t, l, j* és *k*), a magánhangzók rövid alakja az *á*-val és *é*-vel kiegészítve alkották a kilenc magánhangzót (*a, á, e, é, i, o, ö, u, ü*). Az összes párosítást (6*9=54) ötször mondta be a női beszélő, ebből három a tanítást, kettő a tesztelést szolgálta. A tanító alakzatok a tesztelésben nem vettek részt.

Egy feed-forward neurális hálózatot tanítottam be, konjugált gradiens back propagation algoritmussal. A magánhangzók felismerési aránya kizárólag a vizuális jel alapján 81%-ra adódott. Ez a kutatás kezdeti szakaszában elért eredmény az útkeresés egyik fázisának tekinthető.

4. 1. 3. Geometriai lényegkiemelés fekete-fehér képekre

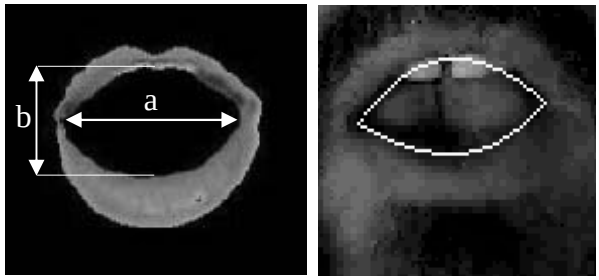
A vizuális lényegkiemelés továbbfejlesztésekor általánosan használható megoldást kerestem, amely a felvétel körülményeitől kevéssé függ. Ezért eltekintettem a színinformáció felhasználásától és a beszélő arcának markerezésétől. Az ajaknyílás, ajakszélesség és intenzitás értékeket fekete-fehér képekből kívántam meghatározni.

A vizuális jellemzők kinyerésére ismert eljárások közös vonása, hogy a száj belső és külső kontúrjának követését követelik meg (Yuille et al., 1992; Silsbee, 1994; Cootes et al., 1995, 1998; Kass et al., 1988; Shdaifat et al. 2003). Ezek a módszerek rendkívül számításigényesek. A számítási kapacitás növekedésével ez a probléma enyhül, de a legújabb módszerek sem elég megbízhatók. Egy friss közlés szerint (Pérez et al., 2003) képsorozat feldolgozásával a képkeretek 5,8 %-ánál nem sikerült követni az ajakkontúrokat. Amennyiben egyetlen kép alapján próbálták megoldani a feladatot, az ajakkontúrok meghatározása a képkeretek 27,4 %-ánál volt sikertelen.

Mivel az ajakkontúrok követése igen nehézkes, olyan módszer kifejlesztésére törekedtem, amely ezt nem igényli. Az általam javasolt eljárás a feldolgozásra kijelölt terület képi hasonlóságán alapul.

A következő bekezdésben tárgyalt prototípus alakzatok jellemzőinek kialakításánál megengedhetőnek tartom a kézi beavatkozást, ezért a jellegzetes pontok kijelöléséhez adható gépi támogatást nem részletezem. A proto-

típus alakzatok képein a szájsarkak elmozdulás vektorok vagy manuális segítség alapján kijelölhetők. Az ajkak méretét reprezentáló ajakszélesség (*a*) és ajaknyílás (*b*) ezek alapján adódik. A szájnyílás belső területét a belső szájsarkak és a szájnyílás felső illetve alsó széle közé rajzolt parabolák által határolt területtel közelítem, amelyre a *k* intenzitás faktor meghatározható.



4. 8. ábra. Az ajkak belső méretét reprezentáló jellemzők és az intenzitási tényező meghatározásához parabolákkal kijelölt terület.

Artikulációs könyvtárt 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 láthatóság 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 prototípus alakzatok képei megtalálhatók a CD mellékleten.

1. 2. tézis

A választott vizuális jellemzők meghatározását a 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 bázisú és a pixel bázisú módszerrel feldolgozott terület között helyezkedik el
- mérsékelt számításigényű, valós időben is elvégezhető a mai PC-ken
- fekete-fehér képeken elvégezhető
- a vektorkvantáláson alapuló feldolgozás esetén közvetlen bemenetként szolgálhat

Hátránya, hogy beszélőfüggetlen feladathoz az artikulációs könyvtár bővítésére van szükség, ami a feldolgozási idő növekedéséhez vezet. A kutatás jelenlegi fázisában a beszélőfüggő audiovizuális beszédfelismerés tűzhető ki reális célként. Külön kutatási terület lehet az artikuláció személyfüggése, a vizuális és az akusztikus jellemzők kapcsolata.

A hasonlóság vizsgálatára a képtömörítésre kifejlesztett, mozgásbecslésre szolgáló algoritmusok alkalmasnak bizonyultak. A feldolgozandó terület az orrhegyhez képest kivágott ablakkal rendelkezésre áll. A vizsgált kép feldolgozásra kijelölt területe minden irányban 15 pixellel nagyobb a prototípus képek méreténél. Erre azért van szükség, mert a kijelölés pontossága nem éri el a pixeles finomságot, ezért a vizsgált képet elmozdulás vektorokkal el kell tolni, és ez alapján meg kell keresni artikulációs könyvtár összehasonlítás alatt álló referencia képéhez leginkább hasonló pozíciót. Az így kapott hasonlóság mérték fogja jellemezni az adott prototípus kép és a vizsgált kép hasonlóságát. 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 a prototípus alakzatnak a vizuális jellemzőit rendeltem az adott képkockához.

A képtömörítésnél alkalmazott mozgásbecslés az elmozdulás vektor meghatározását szolgálja, kijelölve azt a képrészletet, amely az előző képkockán a legjobban hasonlít a vizsgált képrészlethez. Esetünkben az elmozdulás vektor közömbös, a hasonlóság mértéke az eljárás kimeneti adata.

A hasonlósági mértékek rendszerint távolság függvényeken, illetve a keresztkorrelációs függvényen alapulnak (Rao, Hwang, 1996):

Az abszolútérték hiba:

$$M_k(i,j) = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left|R_k(m,n) - X(m+i,n+j)\right| \qquad (4.11.)$$

A négyzetes hiba:

$$M_k(i,j) = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left[R_k(m,n) - X(m+i,n+j)\right]^2 \qquad (4.12.)$$

Ahol $M_k(i, j)$ a k-adik referencia elemhez tartozó hasonlósági mérték (i, j) eltolás esetén. Ennek minimuma lesz a k-adik referencia elemhez tartozó hasonlóság $-15 < i < 15$, $-15 < j < 15$ pixel. M, N a vizsgált ablak mérete: $M=81$, $N=71$ pixel. R_k a k-adik referencia kép, X a vizsgált kép világosság mátrixa.

A keresztkorrelációs függvényen alapuló hasonlóság:

$$M_k(i,j) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} R_k(m,n)X(m+i,n+j)}{\left[\sum_{m=1}^{M}\sum_{n=1}^{N} R_k^2(m,n)\right]^{1/2}\left[\sum_{m=1}^{M}\sum_{n=1}^{N} X_k^2(m+i,n+j)\right]^{1/2}} \qquad (4.14.)$$

A jelölések megfelelnek az előbbi hasonlóság függvényeknél alkalmazottaknak, az eltérés csupán annyi, hogy a keresztkorrelációs függvény esetében a maximumot keressük, ez jelenti a leginkább hasonló alakzatot. A geometriai bázisú lényegkiemelésnél a keresztkorrelációs függvényt használtam a képhez legközelebb álló referencia alakzat kiválasztására.

4. 2. A pixel bázisú lényegkiemelés

A számítógépek sebességének és tárkapacitásának további növekedésével a geometriai bázisú rendszerek mellett előtérbe került a pixel alapú feldolgozás. Ebben az esetben a száj és környezete, de akár az egész kép minden pontja részt vehet az elemzésben. A pixel bázisú feldolgozás előnye, hogy az artikulációs terület feldolgozásával a képi információt veszteség nélkül a felismerés szolgálatába állíthatja. A teljes kép feldolgozása esetén a gesztusok figyelembe vételére is lehetőség nyílik. A geometriai alapú feldolgozás lehetővé teszi az artikuláció elemzését, a hangképzés statikus és dinamikus jellemzőinek mérését. A pixel alapú feldolgozás ezeket a lehetőségeket nem kínálja. További hátránya a pixel bázisú feldolgozásnak, hogy érzékenyebb a megvilágítás változásaira és személyfüggő felismeréshez használható.

A pixel bázisú feldolgozás hátrányaival kapcsolatban meg kell említeni, hogy a megvilágítás változását a geometriai bázisú rendszerek is csak kismértékben tolerálják. Pl.: ha a beszélőt szemből nem éri fény, a szájnyílása mindig sötét. Ha azonban éppen a kamera irányából világítjuk meg, a nyelv a hátul képzett hangoknál is látható lesz, a kétdimenziós képen a mélység nem érzékelhető.

A pixel bázisú lényegkiemelés az ajkak környezetének kijelölésével, rendszerint a képpontok számának decimációval történő redukálása után elvégzett transzformációt jelenti. A transzformációk a képsíkból a síkfrekvencia tartományba konvertálják a képeket. Erre a célra a diszkrét koszi-

nusz transzformációt (DCT) választottam, amelynek előnye a Fourier transzformációval szemben, hogy valós függvényeket valós függvényekbe konvertál (Potamianos et al., 1998; Nakamura et al., 2000; Neti et al., 2003). A száj környezetének kijelölése a geometriai bázisú feldolgozáshoz megtörtént, a sorok és oszlopok számát decimálással harmadolva 27x23 pontos képet kapunk. A transzformált jel vízszintes és függőleges irányú síkfüggvényeiből a 8-8 legkisebb vízszintes és függőleges síkfrekvenciájú 64 bázisfüggvény együtthatóiból válogattam a vizuális jellemzőket. A nagy síkfrekvenciájú komponensek a textúrát képviselik.

4. 2. 1. A jellemzők válogatása

Az akusztikus és vizuális jel integrálásának természetszerűleg a csak hang alapján kapott felismerési arányok növekedését kell eredményeznie, egyébként nincs értelme a vizuális jel erőforrásigényes feldolgozásának. Különösen a vizuális jelből nyerhető adatok mennyiségét szükséges redukálni, hiszen egy kép több százezer képpontból áll, leírása milliós nagyságrendű adatot jelent. A beszédfelismerés szempontjából releváns jellemzők meghatározása a lényegkiemelés célja.

A geometriai alapú lényegkiemelés esetében a jellemzők hasznosságát humán kísérletekkel is lehet ellenőrizni, felhasználhatjuk a szájról olvasás tapasztalatait. Pixel bázisú jellemzőknél nehezen kapcsolhatók az együtthatók látható jellemzőkhöz, ezért humán kísérleti eredményekre nem építhetünk. Különböző matematikai eljárásokat alkalmaznak a vizuális jellemzők válogatására (Potamianos et al., 2001) illetve az adatok utófeldolgozására, amelyek a jellemzők transzformációjával csökkentik az adatok mennyiségét és az osztályozási hibákat.

A jellemzőválogatást a vizuális adatokra elvégezve maximalizálható a vizuális modalitás teljesítménye, majd az optimalizált vizuális adatok kombinálhatók az akusztikus jellemzőkkel (Duchnowski et al. 1994; Scanlon et al. 2003). A két modalitás külön-külön optimális jellemzőinek csak a kései integrálásnál – ahol külön-külön döntést hozunk a két modalitás jele alapján, és utólag kombináljuk a két eredményt – nyilvánvaló az előnye. Korántsem biztos azonban, hogy korai integrálás esetén a külön optimali-

zált jellemzők egymással kölcsönhatásban is optimálisnak bizonyulnak. Folyamatos beszédfelismerésnél a potenciális jelöltek nagy száma miatt csak a korai integrálás oldható meg, ezért különösen fontos az egyesített jellemzők vizsgálata (ld. 5. fejezet). Sajnos a beszéd és az akusztikus, illetve a vizuális modalitás integrálása bonyolultabb folyamat annál, hogy okos megfontolások alapján a várakozásainknak megfelelő eredményeket kapjuk.

A várakozásoktól eltérő eredmények a vizuális jellemzők válogatása és az akusztikus és vizuális jelek egyesítése során is adódtak. A diszkrét koszinusz transzformált 64 legkisebb síkfrekvenciájú együtthatóinak gondos válogatásával sikerült 16 jellemző alapján a vizuális hangpár felismerési arányt 64,5 %-ra növelni. Ezeknek a jellemzőknek az önmagukban 88,4 %-os hangpár felismerési arányt mutató akusztikus jellemzőkhöz csatolása után az egyesített hangpár felismerési arány 81,8 %-ra csökkent. Korai integrálás esetén tehát korántsem biztos, hogy a legjobb vizuális beszédfelismerési eredményt produkáló jellemzőket használhatjuk. Nem triviális, hogy az akusztikus jellemzőket erősítő – a vizuális beszédfelismerés szempontjából nem feltétlenül a legjobb – paraméterek is felülmúlják a geometriai jellemzők hasznosságát. Mivel az akusztikus jellemzőkkel kölcsönhatásban nem feltétlenül a legjobb vizuális felismerési eredményeket mutató jellemzők bizonyulnak optimálisnak, kevésbé hatékony vizuális paraméterekkel kell beérnünk. Emiatt nem ültethetők át automatikusan a hangpár alapú, csak korai integrálással egyesíthető audiovizuális felismerésre a geometriai és pixel alapú lényegkiemelés összehasonlításakor szavakra és fonémákra kapott eredmények.

2. 1. tézis

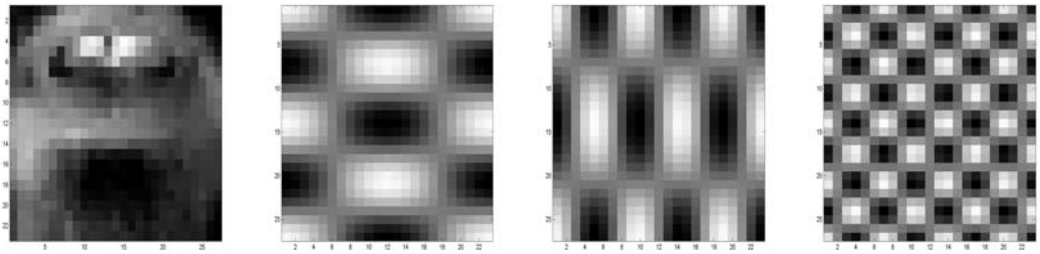
Ellenpéldával igazoltam, hogy a kései integrálás esetében felhasználható, a legjobb vizuális beszédfelismerési eredményeket adó pixel bázisú jellemzők nem vihetők át automatikusan a korai integrálási modellbe. Az akusztikus és vizuális jellemzők kölcsönhatása miatt a pixel alapú vizuális jellemzőket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

A kudarcélmény után, amit az akusztikusnál gyengébb audiovizuális hangpár felismerési eredmények jelentettek, a tanítás-tesztelés folyamatában végeztem a jellemzők válogatását. Ezzel akartam biztosítani, hogy a beszédfelismerési eredmények legalább ne romoljanak a vizuális jel felhasználása által.

A válogatást a következő lépésekben végeztem: Az akusztikus jellemzőket először egyetlen DCT együtthatóval kiegészítve elvégeztem a tanítást és tesztelést. A hangpár felismerési arányokat meghatároztam egyenként mind a 64 együtthatóra. A legnagyobb növekedést a (4, 2) indexű DCT együttható – jelöljük *p*-vel – eredményezte. Ezt választottam az első vizuális jellemzőnek. Az akusztikus jellemzőkhöz és az előbbiekben kiválasztott *p*-hez hozzáadva ismét végigpróbáltam az összes együtthatót, a második vizuális jellemző megtalálása érdekében. A legnagyobb növekedést a (2, 5) indexű, *q*–val jelölt együttható okozta. Az akusztikus jellemzőket, valamint a *p* és *q* együtthatókat harmadik vizuális paraméterként kiegészítve újra elvégeztem a tanítást és tesztelést az összes jellemzőre, így lett a harmadik vizuális jellemző az *r* , indexe: (8, 7). A *p*, *q* és *r* mel-

lett nem akadt olyan DCT együttható, amely negyedikként javította volna a hangpárok felismerési arányát.

Az első, egyedüli vizuális jellemzőhöz (*p*) az akusztikus jel mellett a válogatás módszere miatt a legjobb arány tartozott. A később legjobb másodiknak bizonyult vizuális jellemző (*q*) az akusztikus jel mellett egyedüli vizuálisként a 4. legjobb, a legjobb harmadik (*r*) egyedüliként a 12. legjobb felismerési arányt mutatta. A később legjobb harmadiknak bizonyult együttható (*r*) a *p* mellett másodikként tesztelve csak a 7. volt a rangsorban. Nem biztos, hogy nincs három – vagy több – másik együttható, amely a kiválasztottaknál jobb eredményt mutatna, de legalább az összesen több száz tanítási és tesztelési forduló során minden körben nőtt az audiovizuális felismerési arány az akusztikushoz képest. A *p*, *q* és *r* együtthatókhoz tartozó DCT bázisfüggvényeket egy mintaképpel a 4. 9. ábra mutatja.



I p q r

4. 9. ábra. Egy mintakép (I) és a legnagyobb növekedést eredményező együtthatókhoz (p, q, r) tartozó DCT bázisfüggvények.

A geometriai alapú elemzés kidolgozását akkor is elsőrendű fontosságúnak tartom, ha a pixel bázisú jellemzők hatékonyabbnak bizonyulnak, hiszen az artikuláció dinamikus vizsgálatát csak a geometriai paraméterek segítik.

5. Az akusztikus és vizuális modalitás integrálása

Az audiovizuális beszédfelismerés másik kulcskérdése a vizuális lényegkiemelés mellett, hogy az akusztikus és vizuális jel hogyan integrálható a legjobb felismerés érdekében. A humán beszédfelismerési kísérletek azt mutatják, hogy az integrált eredmények minden körülmények között felülmúlják mind az akusztikus, mind a vizuális egymódusú arányokat (Potamianos et al. 2003).

Az audiovizuális beszédfelismerés az akusztikus beszédfelismerésből fejlődött ki, annak eredményeire épít. Kézenfekvő megoldásnak tűnt, hogy az akusztikus és vizuális jellemzőket konkatenálva a bevált rejtett Markov modell, vagy neurális hálózat alapú felismerőt a megnövelt dimenziójú paraméterekkel tanítjuk és teszteljük. Ezt korai integrálási modellnek nevezték el (Hennecke et al., 1996). A másik szélsőséges lehetőség, hogy külön-külön döntést hozunk az akusztikus és vizuális modalitás alapján, és a két eredményt utólag egyesítjük (kései integrálás). Próbálkoztak közbenső megoldásokkal is, amikor a feldolgozás valamely fázisában figyelembe veszik a másik modalitást (Massaro, 1998; Benoît et al., 1996). Egyes eljárások lehetővé teszik a két modalitás megbízhatóságának figyelembe vételét, és pl. a jel/zaj viszony becslése alapján súlyozhatjuk az akusztikus és vizuális jeleket (Adjoudani, and Benoît, 1996; Teissier et al., 1997).

A 3. fejezetben szereplő C_1VC_1 szavak felismerési kísérlete lehetőséget kínált a korai és kései integrálás összehasonlítására magánhangzó felisme-

résí feladat keretében. A Bayes döntési eljárás alkalmas a két modalitás kései integrálására. Az akusztikus és vizuális csatorna a posteriori valószínűségeit a két modalitásra külön-külön betanított neurális hálózatok kimeneti aktivitásai adják. Az *i*-edik osztály $P(\omega_i)$ a priori valószínűségének becslésére a vizuális jel neurális hálózatának kimeneti aktivitását használtam, míg az akusztikus jel neurális hálózata adta a becslést a $p(x \mid \omega_i)$ feltételes valószínűségre. A Bayes osztályozó úgy dönt az *x* megfigyelés alapján ω_i-re, hogy $max\{ P(\omega_i \mid x) \}$ $i=1,2\ldots n$ -re teljesüljön (Duda, and Hart, 1973). A $P(\omega_i \mid x)$ a posteriori valószínűségek a Bayes féle döntéselmélet felhasználásával számíthatók a $P(\omega_i)$ a priori valószínűségekből és a $p(x \mid \omega_i)$ feltételes valószínűségekből:

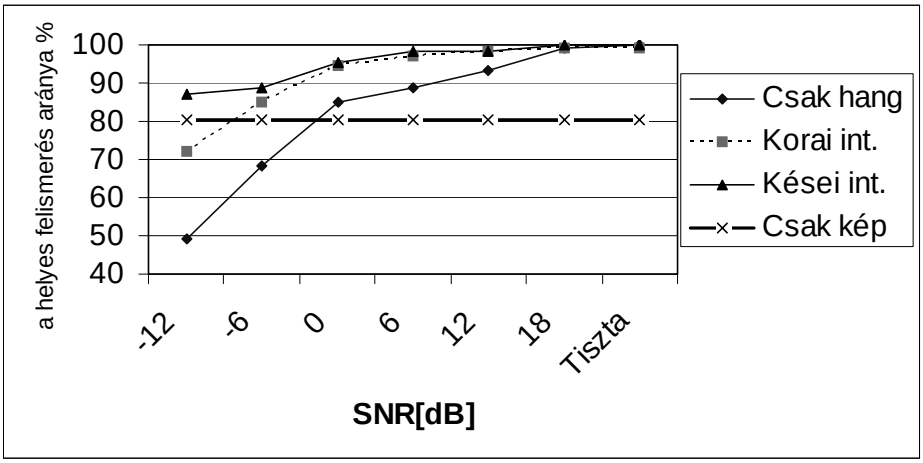
$$P(\omega_i \mid x) = \frac{p(x \mid \omega_i)P(\omega_i)}{p(x)} \quad \text{, ahol} \tag{5.1.}$$

$$p(x) = \sum_{i=1}^{n} p(x \mid \omega_i)P(\omega_i) \tag{5.2.}$$

Massaro és Stork (1998) kimutatta, hogy a vizuális és az akusztikus jel között erős korreláció áll fenn, egy kategórián (szó, szótag, hangpár, hang) belül azonban a jellemzők véletlen eloszlásúak, feltételesen függetlenek. Ebben az esetben a két modalitás valószínűségei szorzatának maximuma szolgáltatja az optimális döntést.

Az 5. 1. ábra különböző jel-zaj viszony mellett az akusztikus, a korai, valamint kései integrálású audiovizuális és a csak vizuális jellel mért automatikus felismerési eredményeimet mutatja. Gyenge minőségű beszédnél csak a kései integrálási modell eredményei múlták felül mind az akuszti-

kus, mind a vizuális egymódusú felismerési arányokat. A vizuális jel neurális hálózatának gyenge kimeneti aktivitása több esetben sikeresen rontotta le az akusztikusan jó eredményt mutató hamis jelöltek esélyeit. A korai integrálási modell is jelentősen növelte a felismerés biztonságát a pusztán akusztikus felismerés eredményeihez képest, ebben a kísérletben a hibák száma feleződött a vizuális jel hatására.

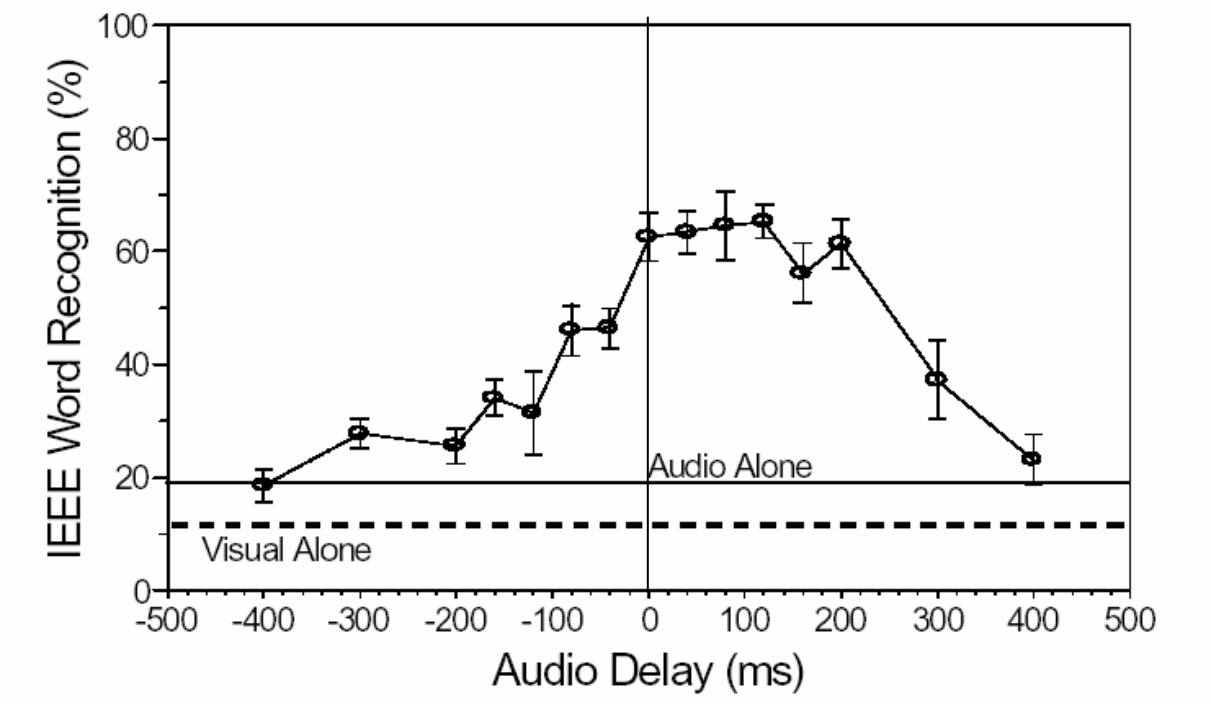


5. 1. ábra. A korai és kései integrálású kétmódusú felismerés eredményei az akusztikus és vizuális egymódusú eredményekkel összehasonlítva.

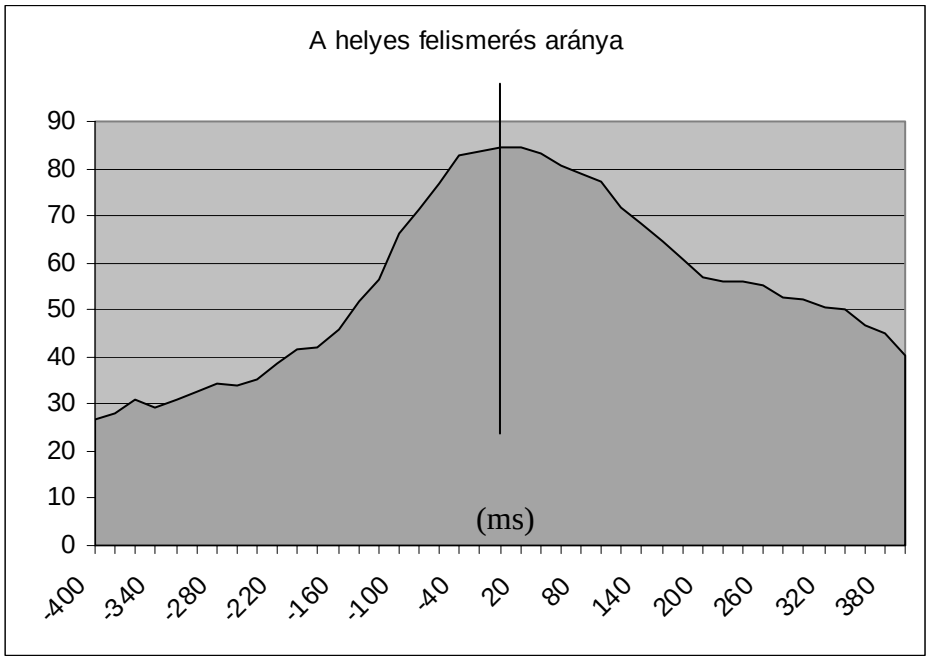
A kései integrálás könnyen megvalósítható izolált szavas beszédfelismerőknél, vagy az előbbihez hasonló fonéma felismerési kísérletnél, ahol az összes jelöltre meghatározható az együttes valószínűség. Kapcsolt szavas, de még inkább a folyamatos beszédfelismerés esetében a potenciális jelöltek száma annyira megnő, hogy képtelenség az összes lehetséges esetre az akusztikus és vizuális jelek valószínűségeit meghatározni és a végén egyesíteni a két modalitás eredményeit. Folyamatos beszédfelismerési kísérleteknél tehát a korai integrálási modellt van módunk alkalmazni.

Amikor a beszéd megértéséhez az akusztikus modalitás mellett a vizuális modalitást is felhasználjuk – humán és gépi felismerésnél egyaránt – az integrálást megnehezítheti a két jel időbeli elcsúszása, a szinkron hibája. Egyes távközlési rendszerek esetében az átviteli út késleltetése a videó- és az audió jelre eltérő lehet. Real-time rendszerekben a feldolgozási idő eltérése is okozhat szinkronhibát. A híradásokban a tudósító hangjának és képének az eltolódása megnehezíti a beszédértést, kép nélkül könnyebben felfognánk.

A humán beszédfelismerési kísérletek az időben eltolt akusztikus és vizuális jelre aszimmetrikus eredményt mutatnak: a kísérleti alanyok jobban tolerálják a hang késleltetését. Ha a hang siet a képhez képest, csak kis eltérést viselünk el, már 40 ms fölött számottevő romlás lép fel. A helyesen felismert beszédrészletek arányának minimális csökkenését tapasztalták, amikor a hang legfeljebb 200 ms-ot késett a képhez képest (McGrath, Summerfield, 1985; Grant et al. 2003). A 5. 2. /1. ábrán az emberi szófelismerési eredmények változását látjuk a hang késésének függvényében a képhez képest. Az 5. 2. /2. ábrán az értekezés tárgyát képező gépi felismerő hangpár felismerési arányai hasonló aszimmetriát mutatnak a hang késleltetésének függvényében.



1.



2.

5. 2. ábra. A felismerési eredmények változása humán (1.) (Grant et al. 2003) és gépi (2.) beszédfelismerés esetében a hang késleltetésének függvényében.

A humán és gépi beszédfelismerési eredmények hasonló aszimmetriája a hang és a kép elcsúszásának függvényében reményt adhat arra, hogy a gépi felismerők valamit megragadhatnak az emberi felismerés sajátosságaiból.

6. Beszédatbázisok és nyelvi modellek

Az automatikus beszédfelismerés összetett, ma még csak részlegesen megoldható feladat. Nehezíti a magyar nyelv felismerését, hogy a szavanként esetleg több száz féle toldalék – képző, jel, rag – a szóalakokat megsokszorozza, az elsősorban angol nyelvre kifejlesztett szóalapú felismerők nem adaptálhatók. Felmerül a kérdés, hogy agglutináló nyelveknél mi lehet a felismerés alapegysége, amely általánosabban, minden nyelvre alkalmazható. Minél nagyobb egységeket választunk, annál könnyebb megkülönböztetni egymástól a szótárelemeket, viszont annál több betanítandó elemmel kell számolnunk. A skála egyik végén a több millió szóalak (esetleg több szóból álló egység), másik végén a minimális számú, de a koartikulációs és egyéb hatások miatt azonosíthatatlan fonéma áll. A kettő között különböző feltételek szerinti kompromisszumokkal választhatunk hangpárokat, félszótagokat, hanghármasokat, szótagokat vagy más alapegységeket. Vicsi Klára (1995) vizsgálta, hogy egy hosszú szöveg részleges lefedéséhez hány kettőshang, félszótag illetve szótag szükséges (6. 1. táblázat) az alábbi követelmények szerint:

- fedje le a nyelv egészét, vagy legalább annak túlnyomó részét
- számossága ne legyen túl nagy
- az elemek realizációja ne legyen túl erősen kontextusfüggő

6.1.táblázat. Hosszú szöveg részleges lefedéséhez szükséges elemszám (Vicsi, 1996)

Kumulatív gyakoriság	hangpár	félszótag		szótag
		kezdő	záró	
50%	71	31	12	116
75%	183	75	41	402
90%	399	134	95	1011
95%	586	176	150	1604
99%	956	259	317	3380

A szótag és a táblázatban nem szereplő hanghármas alkalmazása a nagyszámú betanítandó elem miatt nehézkes. A félszótag és a hangpár versengésében a szöveg részleges lefedéséhez a félszótagból kell kevesebb, a teljes lefedéshez hangpárból. A félszótag mellett szól, hogy általában hosszabb az időtartama, de illeszteni csak a kezdő és záró félszótag határán tudjuk, az így képzett szótagok határán a hangok egymásra hatását nem tudja figyelembe venni. A hangpárok mindkét végükön illeszkednek a szomszédos hangokhoz. Felmerül a kérdés, hogy a félszótagok és hangpárok felismerését az előbbi tulajdonságok mennyire segítik, vagy gátolják, melyik hatás érvényesül erősebben.

Az adatbázisok összeállításánál a felismerés alapegységének vizsgálatát tartottam szem előtt.

6. 1. Az audiovizuális beszédadatbázis

Magyar nyelven nem áll rendelkezésre audiovizuális beszéd adatbázis, ezért a szerző bemondásával, házi videó berendezés és közönséges PC mikrofon felhasználásával felvett hang- és videó anyagon folyt az elemzés. Az audiovizuális adatbázis a leggyakoribb félszótagok betanítását szolgálta. A tanító alakzatok egyik részét az azonos magánhangzóhoz tartozó kezdő és záró félszótagok kombinációiból képzett szótagok alkották. A tanító alakzatok másik része és a tesztelő alakzatok számok és dátumokban előforduló szavak, illetve szófűzérek. A tanításra 486 szótag és 79 szó, a tesztelésre 35 szófűzér szolgált.

A hangpár alapú felismeréshez ugyanezt az adatbázist használtam, kihagyva a csupán egyszer előforduló hangpárok szavait (29 db.). A tanító és tesztelő alakzatokra a félszótagoktól eltérően, a hangpárok gyakoriságát fegyelembe véve osztottam fel a szavakat. Az audiovizuális adatbázis 163 hangpárt tartalmaz. Ez az összes lehetséges hangpár mintegy 20 %-a.

6. 1. 1. Az akusztikus lényegkiemelés

Az audiovizuális vizsgálatoknál a videó szabvány behatárolja a lehetőségeket. A PAL szabvány szerinti, másodpercenkénti 25 kép félképekre bontásával 50 képet tudunk nyerni, így a vizuális időkeret 20 ms-ra adódik. A szinkronitás végett a hangnál is ezt a lépésközt kell választani. A 22 050 Hz-es mintavételi frekvenciájú hang 512 mintája alkot egy keretet, a 20 ms-os lépésköz 441 mintának felel meg, ami 16%-os átlapolódást jelent.

A hangfelvétel körülményei normál irodai környezetnek felelnek meg. A hang a szobában működő, ventilátor zajt termelő PC-n került rögzítésre. Az utcáról beszűrődő zaj mellett a számítógép tápegysége által okozott zaj jelentős mértékű.

Az akusztikus lényegkiemelés megvalósítására hozzáférhető, elfogadott megoldást kerestem. Az MFCC paraméterek meghatározásának MATLAB implementációját választottam (Brookes, 1996). Keretenként 12 együttható mellett a differenciális jellemzőket (Δ) és ezek differenciáját (ΔΔ) használtam fel.

6. 1. 2. A vizuális előfeldolgozás

Az audiovizuális adatbázis videó felvételei egyféle beállítással, az ülő helyzetben természetes fejmozgás mellett, speciális világítási előírások nélkül készültek.

Az általánosabb alkalmazhatóság érdekében a színinformációt nem akartam felhasználni, ezért a képek feldolgozása a színes képek intenzitás képpé alakításával kezdődött. A vizuális előfeldolgozás keretében a videó felvételből ki kell vágni a szótárelemhez tartozó szakaszt. A hangot hangfájlba másolva, a videó jelet képkockánként elmentve szétválasztjuk a két modalitást. A váltottsoros letapogatású képeket félképekre bontva – a páros és páratlan sorokat külön elmentve – a képfeldolgozás eszközeinek igénybe vételével félképenként elvégezhető a vizuális lényegkiemelés. A képkeretek félképekre bontásának következtében a vizuális jel mintavételi időköze 40 ms-ról 20 ms-ra csökken.

A vizuális előfeldolgozás szavanként több órás időtartama miatt az audiovizuális adatbázis mindössze 600 szóból, illetve szófűzérből áll. A feldolgozott félképek száma 21 740, a minták teljes időtartama mintegy 7 és 1/4 perc.

6. 2. Az akusztikus beszédadatbázis

Mivel az audiovizuális adatbázis mindössze 600 elemből állt, egy lényegesen több szóból álló adatbázison is meg akartam vizsgálni, hogy a felismerés alapegységeként a félszótag vagy a hangpár az alkalmasabb választás. Az akusztikus adatbázis saját bemondással 8000 szóból áll a tanítás céljaira, 1400 szó a tesztelést szolgálja. A szavak bemondásakor törekedtem a gondos artikulációra, kerülve a modoros kiejtést.

Az 1 996 589 szóból álló klasszikus és modern próza alapján – amely 4 238 066 szótagot tartalmazott – készült statisztikai elemzés szerint az adatbázis szavaiban szereplő 121 kezdő félszótag kumulatív gyakorisága 70,2%. A 83 záró félszótag kumulatív gyakorisága 80,7%, a kiválasztott félszótagok a szótagok 60,1%-át, a szavak 32,6%-át fedték le.

Ugyanezen az adatbázison hangpár alapú felismerést is végeztem, az akusztikus adatbázis 440 hangpárt tartalmaz. Ez a teljes hangpár készletnek mintegy fele.

A félszótagok, a hangpárok és a HTK szoftvercsomag szintaktikája szerint leírt kapcsolódási szabályok megtalálhatók a CD mellékleten.

6. 3. A fonotipikus fonetikai átírás

Beszéd közben a kiejtett hangok nem mindig felelnek meg a helyesírás szabályai szerint lejegyzett betűkhöz tartozó hangoknak. A szomszédos hangok egymásra hatása az írásképnek megfelelő fonéma egy másik fonémával helyettesítését eredményezi. Az adatbázisban szereplő szavak fonotipikus fonetikus átírása előzte meg a tanítást. A karakterek átírásakor az alábbi szabályokat alkalmaztam:

- az *a-á* és *e-é* párok kivételével a magánhangzók rövid alakja szerepel
- figyelembe vettem a képzés helye szerinti és a zöngés-zöngétlen hasonulást (Grétsy, Kovalovszky, 1980)
- szóösszetételeknél nem vettem figyelembe a fentiektől eltérő kiejtést

A folyamatos beszédfelismerésre dolgozatomban igen korlátozott feltételek mellett teszek kísérletet:

- személyfüggő a tanító és a tesztelő kép- és hanganyag
- a felvétel körülményei nem változtak
- az akusztikus, illetve audiovizuális előfeldolgozásra és a mintaillesztésre korlátozódott a tevékenység

A félszótag vagy hangpár alapú mintaillesztést nyelvi elemzésnek kellene követnie, amelynek során az akusztikai illesztésnél legjobbnak bizonyult elemek sorozatából a legvalószínűbb szavakat vagy mondatokat kellene kiválasztani a szótárt és a nyelvtani ismereteket tároló tudásbázisból. A

nyelvi elemzés nélkülözhetetlenségére jellemző, hogy a fonetikus átirat visszaalakítása a megfelelő szóalakra sem egyértelmű. Pl.: láptól – lábtól

A beszélő személye, lelkiállapota, az akusztikus környezet és az átviteli csatorna változásai, a zaj, a hangképzés közben keletkezett zörejek stb. sokkal jobban megzavarják a gépi felismerőt, mint az emberi felfogót. Ez természetes, hiszen az intelligencia egyik fokmérője az alkalmazkodóképesség.

Az előző két bekezdésben említett feladatok és nehézségek indokolják, hogy a feladatra korlátozó feltételekkel vállalkoztam.

6. 4. A nyelvi modellek és a HTK szoftvercsomag

A HTK rejtett Markov modell (HMM) építésére szolgáló szabad felhasználású szoftvercsomag, amelyet elsősorban beszédfeldolgozásra, ezen belül is beszédfelismerésre fejlesztettek. A rejtett Markov modellt és a szoftvercsomagot felhasználói szinten ismertem meg, kutatást, fejlesztést ezzel kapcsolatban nem végeztem, ezért csak az alkalmazás körülményeire térek ki. A HMM részletei alapvető munkákban megtalálhatók (Rabiner, Juang, 1993).

A HMM felépítése a nyelvi modell megadásával kezdődik. Azt kell eldönteni, hogy milyen elemekből épül fel a beszéd, és ezek az elemek hogyan kapcsolódnak egymáshoz. Félszótag alapú felismerésnél a kezdő félszótag magánhangzóval végződik, ehhez kapcsolódhatnak az ugyanolyan magánhangzóval kezdődő záró félszótagok. Az így kialakuló szótagok határain a kontextust nem tudjuk figyelembe venni, hiszen bármilyen szótag csatlakozhat hozzá.

A hangpár alapú felismerésnél rekurzív módon generálódik a hangpárok láncolata, miután minden hangpárra megadtuk, hogy melyek követhetik, illetve előzhetik meg. A hangpár alapú felismerésnél a felismerés alapegységei mindkét végükön illeszthetők a szomszédos hangpárokhoz, a kontextusfüggés jól meghatározható. Tovább finomítható a modell hármas hangok megadásával, ehhez azonban jóval több tanító alakzatra van szükség, ezt a kiterjesztést nem végeztem el.

A félszótag alapú felismerés finomításakor – amit a felismerési eredményeknél részletezek – a szótagok hangpár közbeiktatásával csatlakoznak egymáshoz, a kontextusfüggés beilleszthető a nyelvi modellbe. A szabályok megegyeznek a hangpár alapú felismerés megadásánál alkalmazottakkal.

A *label* fájlokban az előbbiekben meghatározott nyelvi modell szerint megadjuk a tanító és tesztelő szövegek fonetikusan átírt változatának felbontását. Lehetőség van az egységek határának megadására kézi szegmentálás után. A könnyebbség kedvéért a szegmentálást a tanítás folyamatára bíztam.

A beszéd megadása a lényegkiemelés eredményeképpen kapott vektorokkal történik. Lehetőség van saját lényegkiemelési eredmények bevitelére, a *user* típusú jellemzőn keresztül. A lényegkiemelést az ismert és széleskörben használt MFCC jellemzőkkel végeztem, az algoritmus MATLAB implementációja (Brookes, 1996) segítségével. Az együtthatókon kívül a differenciákat (Δ) és ezek differenciáit (ΔΔ) használtam a felismeréshez.

A HMM definíciója a legegyszerűbb esetben lényegében az állapotok számának megadását jelenti. A félszótag alapú felismerésnél a félszótag hosszától függően 4-6 állapotot választottam. A hangpár alapú felismerésnél egységesen négyállapotú HMM-ek szerepeltek. Ez alól a szünet kivétel, ott a *silence model* miatt öt állapotra van szükség. Az audiovizuális felismerésnél a modalitások súlyozhatósága végett az akusztikus és vizuá-

lis vektorokat két független folyamként (*stream*) kezeltem a normál eloszlású additív zajjal végzett kísérletekben. Próbaképpen két alternatív eloszlást (*Gauss mixture*) tartalmazó HMM-mel is végeztem tanítást és tesztelést. A tanítás négy fordulóban történt, amit a *silence model* beiktatása után még három újabb forduló követett. Ha a tesztelés során a felismerés sikeres, akkor az adott kiejtésben a felismerés alapegységeit a saját modellje állítja elő a legnagyobb valószínűséggel.

A rejtett Markov modell definíciós fájljai megtalálhatók a CD mellékleten.

7. Beszédfelismerési eredmények

Az audiovizuális adatbázison a geometriai és a pixel bázisú vizuális lényegkiemelést, valamint a félszótag és hangpár alapú felismerést hasonlítom össze. A vizuális lényegkiemelés geometriai és pixel alapú jellemzőit értékelem pusztán a vizuális jel alapján végzett felismerési kísérletben és az audiovizuális, integrált felmérésben egyaránt. Az akusztikus és a vizuális modalitás súlyozott figyelembevétele lehetőséget kínál különböző minőségű beszéd estén az együttes felismerési arány javítására.

Az audiovizuális adatbázisnál lényegesen több szót tartalmazó akusztikus adatbázis megbízhatóbb összehasonlításra ad módot a felismerés alapegységét illetően. Az eredmények értékelésénél a háromféle hiba, a helyettesítés (*S*), a beszúrás (*I*) és a törlés (*D*) együttes figyelembe vételére van szükség. A HTK az eredmények kiértékelésére két arányt kínál: A helyesen felismert egységek (*H*) viszonyát az összeshez (*N*): *H/N*, illetve a beszúrásokat is tekintetbe vevő *(H-I)/N* arányt. Az eredmények megadásakor a reálisabb második értéket tüntetem fel.

7. 1. Audiovizuális beszédfelismerési eredmények

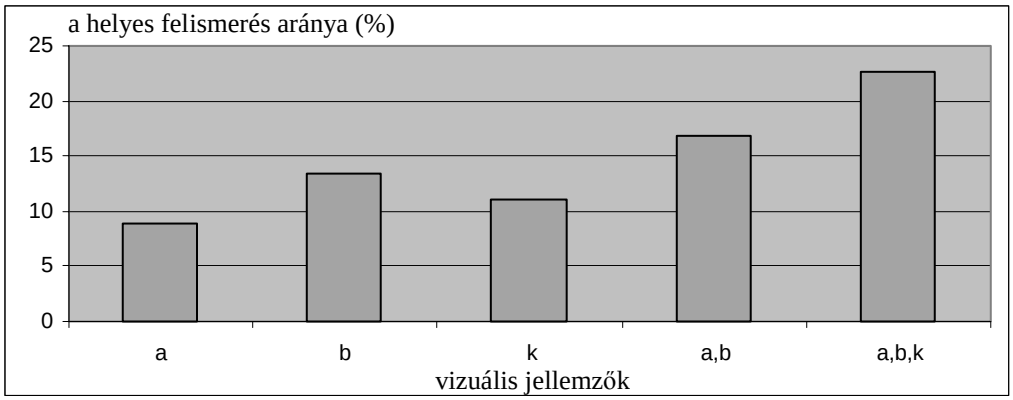
Az audiovizuális adatbázison végzett kísérletek az akusztikus és a vizuális modalitás külön-külön mért felismerési eredményeinek vizsgálatára adnak módot. A legnagyobb kihívás, hogy a kétmódusú felismerés felülmúlja az egymódusú eredményeket. A felismerés alapegysége a félszótag, illetve a hangpár.

A geometriai és pixel alapú vizuális lényegkiemelés összehasonlítására is lehetőséget kínál az audiovizuális adatbázisok között közepes méretűnek számító 600 szót illetve szófűzért tartalmazó személyfüggő beszédadatbázis.

7. 1. 1. A félszótag alapú felismerés eredményei

A geometriai alapú vizuális lényegkiemelésnél az ajakszélesség (*a*), az ajaknyílás (*b*) valamint a nyelv és a fogak láthatóságát reprezentáló intenzitási tényező (*k*) paraméterekkel jellemeztem az artikulációt.

A pusztán vizuális jel alapján végzett felismerési kísérletben egyes jellemzők külön-külön mért eredményeit, illetve az ajakméretek és az intenzitási tényezővel kiegészített ajakjellemzők által elért növekményt figyelhetjük meg a 7. 1. ábrán. Önmagában az intenzitási tényező az ajakszélességnél jobb, az ajaknyílásnál gyengébb eredményt mutat. Az intenzitás faktor hozzáadása az ajakméretekhez észrevehető javulást okoz. Az eredmények közvetlenül nem hasonlíthatók össze a 2. fejezetben kapott humán felismerési eredményekkel, hiszen azok fonémákra (vizémákra) vonatkoztak, de az intenzitási tényező bevezetése az ajakméretek kiegészítéseként ugyanolyan javulást okozott, mint a humán kísérletben a száj (ajkak, nyelv, fogak) láthatósága az ellipszishez képest.

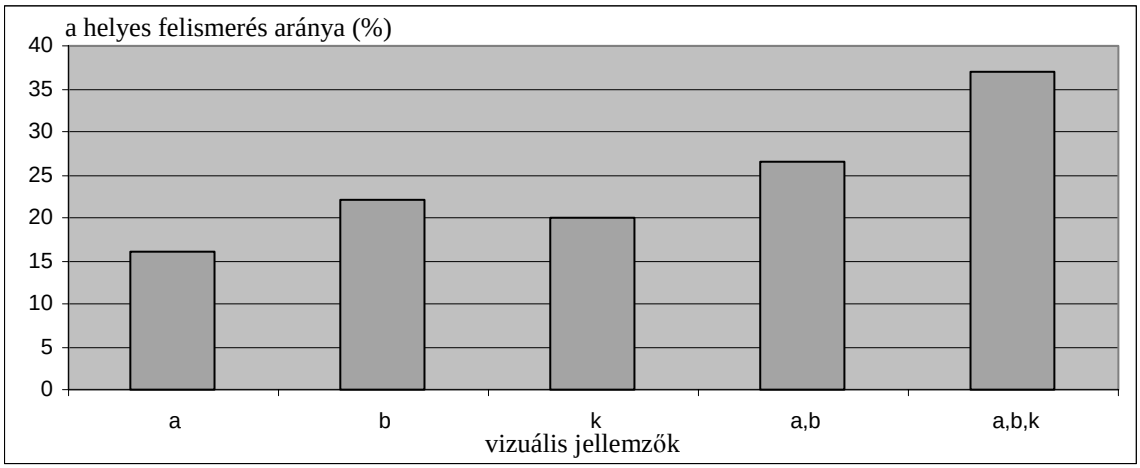


7. 1. ábra. A félszótagok felismerési arányai különböző geometriai jellemzők esetén: ajakszélesség (a), ajaknyílás (b), intenzitási tényező (k) ajakjellemzők (a,b) és mindhárom (a, b, k).

Az MFCC és differenciáira alapozott akusztikus félszótag felismerési arány 69,1%-ra adódott. A vizuális jellemzőkkel kiegészített kétmódusú félszótag felismerési arány 75,9%. A pusztán akusztikus felismerő hibáinak negyede-ötöde eliminálható a vizuális jel hozzáadásával.

7. 1. 2. A hangpár alapú felismerés eredményei

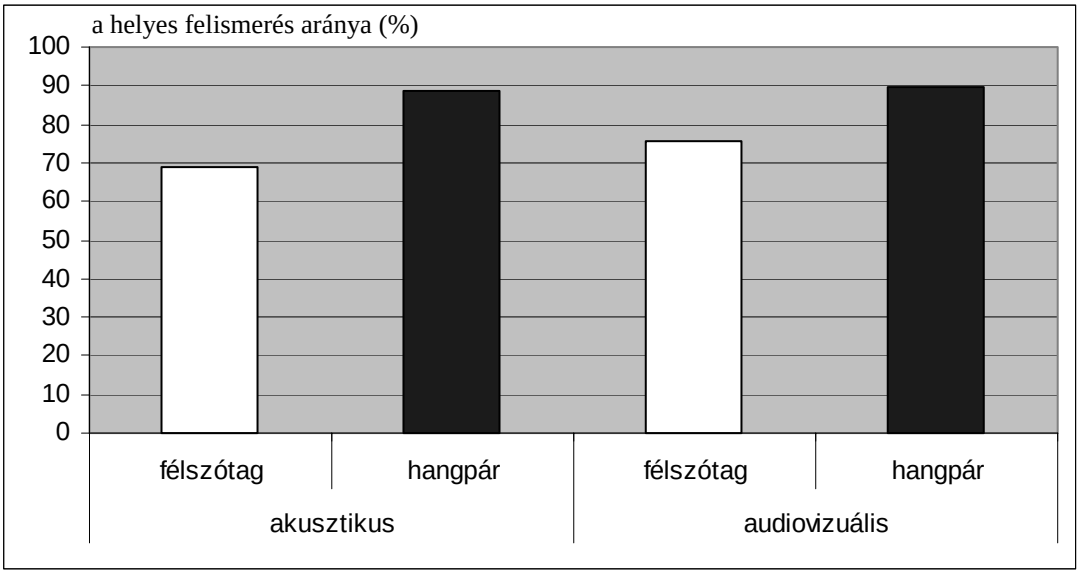
A hangpár alapú kísérletben a helyes felismerés aránya – kizárólag a vizuális jel vizsgálatával – az egyes jellemzőkre, illetve együttes alkalmazásuk esetén a 7.2. ábrán látható eredményeket mutatta.



7. 2. ábra. A geometriai jellemzők hangpár felismerési arányai.

A hangpár alapú felismerési eredmények az akusztikus felismerés esetében is jelentős mértékben meghaladták a félszótag alapú felismerési arányokat. A hangpárok helyes felismerési aránya 88,4%-ot ért el. A geometriai alapú vizuális jellemzőkkel kiegészítve az akusztikus változókat, az audiovizuális helyes felismerési arány 89,8%-ra növekedett. A vizuális kiegészítés a hibák nyolcadát küszöbölte ki.

A 7. 3. ábrán összefoglalva látjuk a félszótag és hangpár alapú akusztikus és audiovizuális felismerési arányokat.

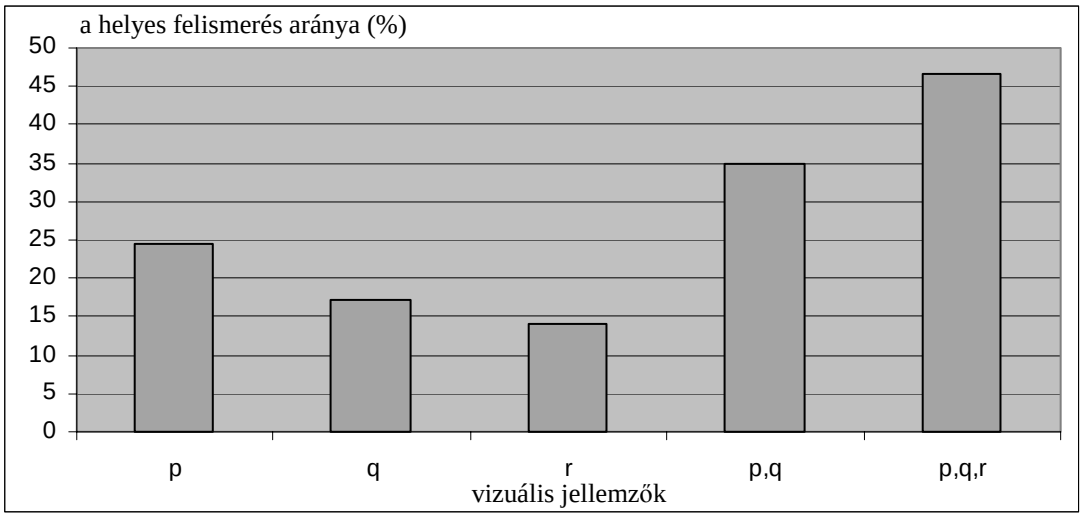


7.3. ábra. A félszótag és hangpár alapú felismerés összehasonlítása.

Összefoglalva a félszótag és hangpár alapú felismerési eredményeket, megállapíthatjuk, hogy a félszótagokra válogatott adatbázison is a hangpár alapú felismerési eredmények kedvezőbbek. A félszótag alapú megközelítés mellett szól, hogy általában hosszabb idejű mint a hangpár, és a hosszabb beszédszakaszok általában könnyebben megkülönböztethetők. A hangpár mellett szól, hogy mindkét végén illeszthető a szomszédos fonémához. Az eredmények azt mutatják, hogy a felismerés alapegysége tekintetében többet nyom a latba a kontextusfüggés figyelembe vétele mint a beszédszakasz időtartama.

7. 1. 3. A pixel bázisú lényegkiemelés eredményei

Az artikuláció látható részének leírására a szájról olvasás tapasztalatai alapján természetes megközelítésnek tekinthetjük a geometriai jellemzők megadását. A könnyen értelmezhető és természetes paraméterek jelentősen elősegítették az artikuláció jobb megismerését. Ha nem ragaszkodunk a látható jellemzőkhöz, a kép tetszőlegesen transzformált paraméterei szolgálhatják az audiovizuális beszédfelismerést. A DCT együtthatók –a vizuális lényegkiemelés fejezetben tárgyalt – válogatásával a pixel bázisú vizuális kiegészítés a geometriai lényegkiemelésnél jelentősebb javulást okozott a hangpár alapú felismerésben.

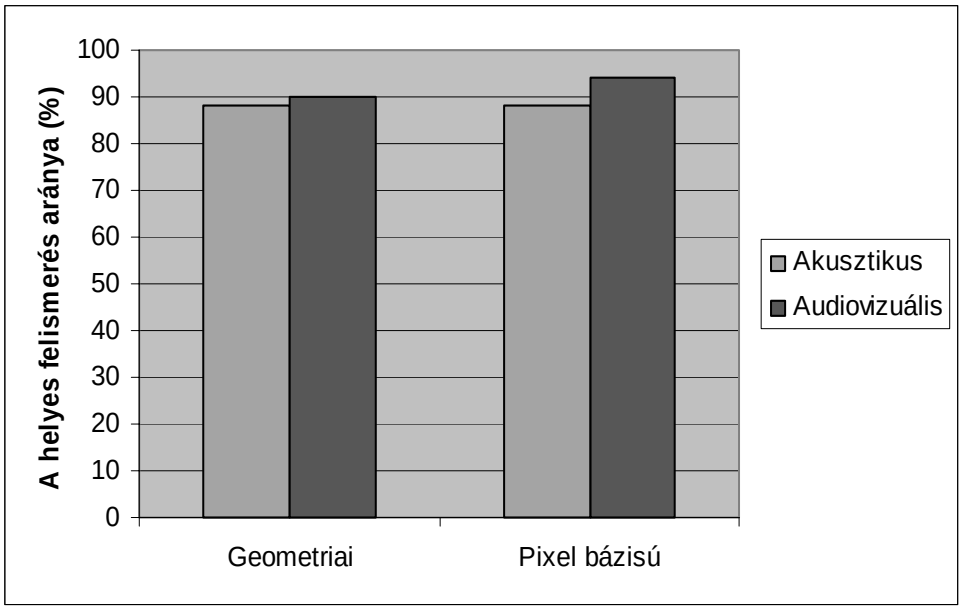


7.4. ábra. A pixel bázisú jellemzőkhöz tartozó vizuális felismerési arányok.

A geometriai alapú felismerésnél a rejtett Markov modell egy *Gauss mixture*-t és egy *stream*-et tartalmazott. A pixel bázisú felismerés esetében további finomításokat végeztem. Az egyik fimomítási lehetőség a változók több folyamra (*stream*) bontása. Az akusztikus és vizuális jellemzők egyenlő súlyozása esetén a felismerési eredmények megegyeznek az egy

folyamot tartalmazó eredményekkel. A két *stream* bevezetése lehetőséget kínál a két modalitás eltérő súlyozására, amelynek kiaknázását a következő alfejezetben mutatom be (Young et al. 2003).

A HMM finomítására a másik lehetőség több *Gauss mixture* alkalmazása. Két *Gauss mixture* tanítása és tesztelése valamivel javította mind az akusztikus, mind az audiovizuális felismerési eredményeket. A helyes akusztikus hangpárfelismerés aránya 88,4%-ról 88,7%-ra, az audiovizuális felismerési arány 93,6%-ról 94,2%-ra emelkedett. A pixel bázisú lényegkiemelés eredményeképpen a felismerési hibák száma mintegy felére csökken.



7. 5. ábra. A geometriai és a pixel bázisú vizuális jellemzőkkel mért audiovizuális hangpár alapú felismerés eredményei.

A 4. 2. 1. pontban a pixel bázisú jellemzők válogatásánál láttuk, hogy az optimalizálás nélkül kiválasztott, az egyedüliként legjobb 16 együttható 64,5 %-os vizuális beszédfelismerési arányt mutatott. A 7. 4. ábrán feltüntetett eredmény ennél csaknem 20 százalékponttal gyengébb. Ez utóbbi az akusztikus jellemzőkkel kölcsönhatásban legjobb hangpár felismerési eredményeket mutató együtthatók vizuális felismerési aránya. A korai integrálással tehát lényegesen gyengébb egymódusú vizuális felismerési eredményeket mutató együtthatókkal tudjuk kiegészíteni az akusztikus jellemzőket. Ennek ellenére a pixel bázisú jellemzőkkel végzett audiovizuális felismerés eredményei felülmúlták a geometriai jellemzőkkel kapott hangpár felismerési arányokat.

A geometriai és pixel alapú lényegkiemelés összehasonlítását csak szó alapú elemzéssel végezték el (Scanlon, Reilly, 2001; Matthews et al. 2001). Pl.: a *Tulips* adatbázis mindössze az 1, 2, 3, 4 angol számjegyeket tartalmazza, 12 bemondóval, beszélőnként kétszer. Az *AVletters* adatbázis az angol abc betűzését tartalmazza 10 beszélővel, egy bemondó három sorozatával (Matthews et al. 2001).

A folyamatos audiovizuális beszédfelismerés szempontjait szem előtt tartó összehasonlító vizsgálat azért tér el a szóalapútól, mert

- a felismerendő alakzatok száma nagyobb (163 hangpár)
- 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

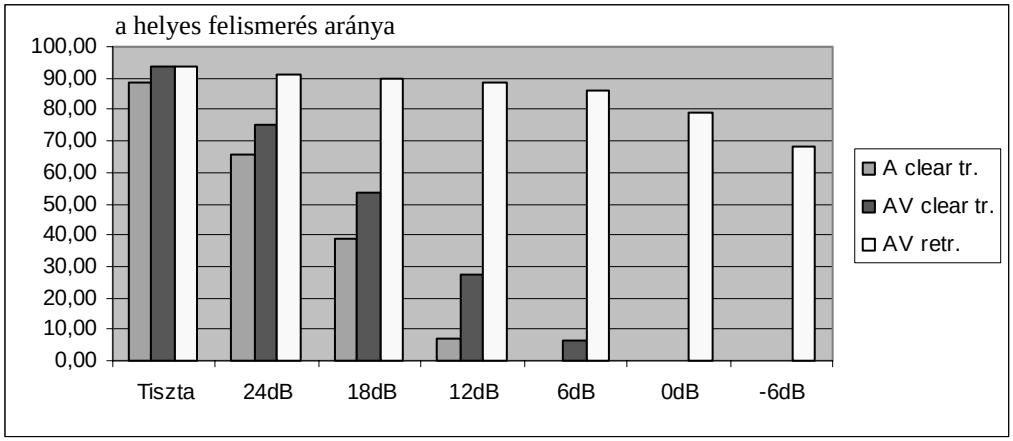
A kései integrálásnál használt jellemzők és a kapott eredmények ugyanúgy nem ültethetők át a korai integrálási modellre, mint a szóalapú felismerés eredményei a szónál kisebb alapegységgel dolgozó felismerőkre.

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.

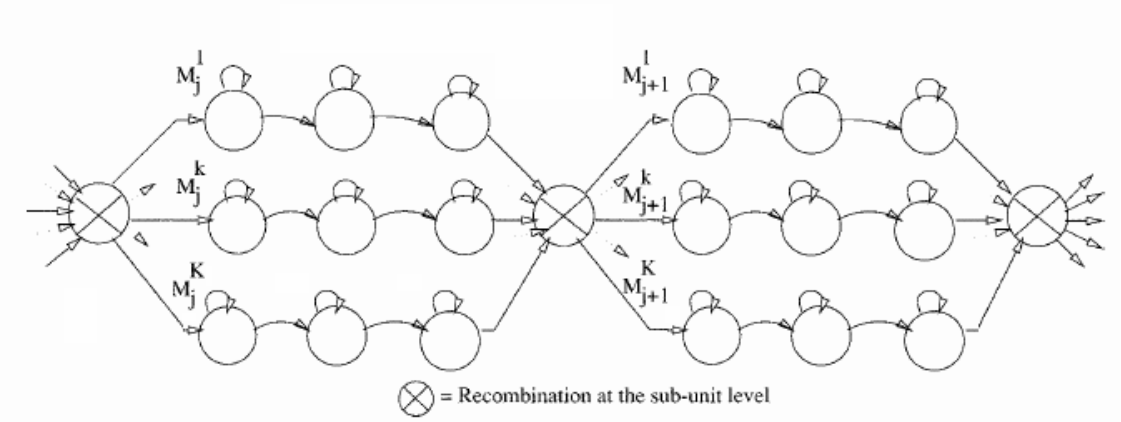
7. 1. 4. A modalitások súlyozása

Az egészséges emberi felfogó elsősorban zajos környezetben veszi hasznát a vizuális modalitásnak. A gépi beszédfelismerők teljesítménye sokat romlik, ha az akusztikus környezet változik. Amennyiben az akusztikus környezet állandó, módunkban áll a zajos beszéddel tanítani a felismerőt. Ezt az esetet modellezi az a kísérlet, amikor a beszédhez különböző energiájú Gauss zajt keverünk, és a tanítást minden jel/zaj viszonyra külön elvégezzük, ezzel lényegében minden jel/zaj viszony mellé egy-egy felismerőt rendelünk. (7. 6. ábra *AV retr.*) Amennyiben nem tudunk a környezethez optimalizált felismerőt választani, bele kell törődnünk, hogy a zajmentes beszéddel tanított felismerő teljesítménye csökken. A zajmentes beszéddel tanított akusztikus felismerő teljesítménycsökkenését a zaj növekedésével a 7. 6. ábrán az *A clear tr.*, az audiovizuális felismerőét az *AV clear tr.* oszlopok mutatják.



7. 6. ábra. Minden jel/zaj viszonyra külön betanított audiovizuális (AV retr.), illetve tiszta beszéddel tanított akusztikus (A clear tr.) és audiovizuális (AV clear tr.) felismerő hangpár felismerési arányai.

Audiovizuális felismerőnél lehetőségünk van a hangminőség romlása esetén a vizuális modalitás súlyának növelésére. Az akusztikus és vizuális jellemzőket külön folyamatként értelmezve, mindkét modalitásnak megfelel egy-egy Markov modell. A párhuzamos folyamatok a felismerési alapegységek határain szinkronizáltak, az elemeken belül azonban az átmenetek egymáshoz képest időben eltolódhatnak. A 7. 7. ábra a több folyamatra bontott rejtett Markov modell sémáját mutatja (Dupont, Luettin, 2000).



7. 7. ábra. Több *stream*et tartalmazó rejtett Markov modell felépítése.

A maximum a posteriori döntés a felismerési alapegységeknek azt a sorozatát keresi, amelyre *K=2 stream* esetén a

$$P(\Lambda \mid O^a, O^v) = \frac{P(O^a, O^v \mid \Lambda)P(\Lambda)}{P(O^a, O^v)} \qquad (7.\ 1.)$$

maximális, ahol Λ az egyes megoldásokat, O^a, O^v a vizsgált akusztikus, illetve vizuális megfigyeléseket (vektorokat) jelenti. A két modalitás feltételes függetlensége (Allan, 1994; Massaro, Stork, 1998) esetén:

$$P(O^a, O^v \mid \Lambda) = P(O^a \mid \Lambda)P(O^v \mid \Lambda) \qquad (7.\ 2.)$$

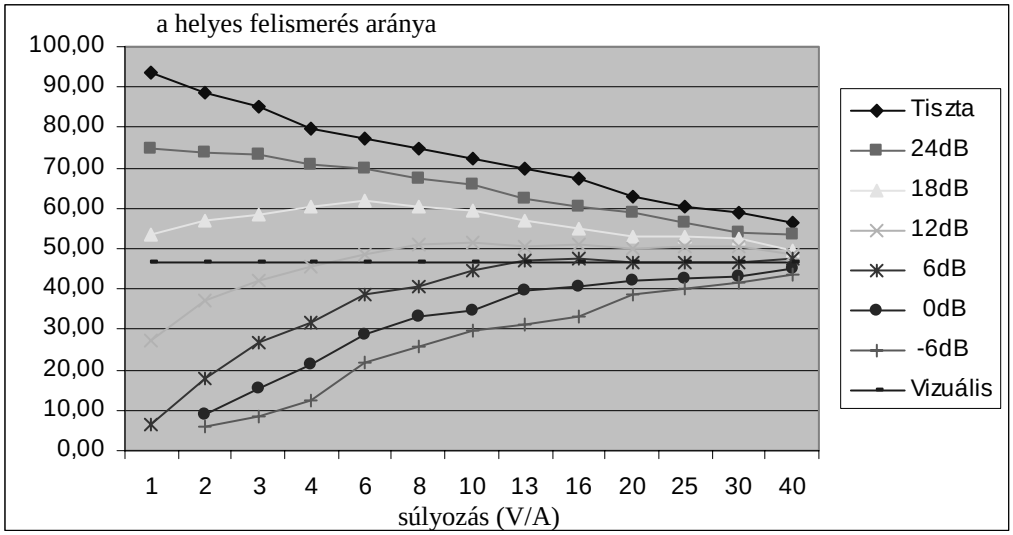
Ha a generált alapegységek valószínűségeit el akarjuk tolni az egyik modalitás irányába, lehetőségünk van a modalitások eltérő súlyozására. Legyen:

$$P(O^a, O^v \mid \Lambda) = P(O^a \mid \Lambda)^{wa} P(O^v \mid \Lambda)^{wv} \qquad (7.3.)$$

, ahol *wa* és *wv* az akusztikus illetve a vizuális modalitás súlya.

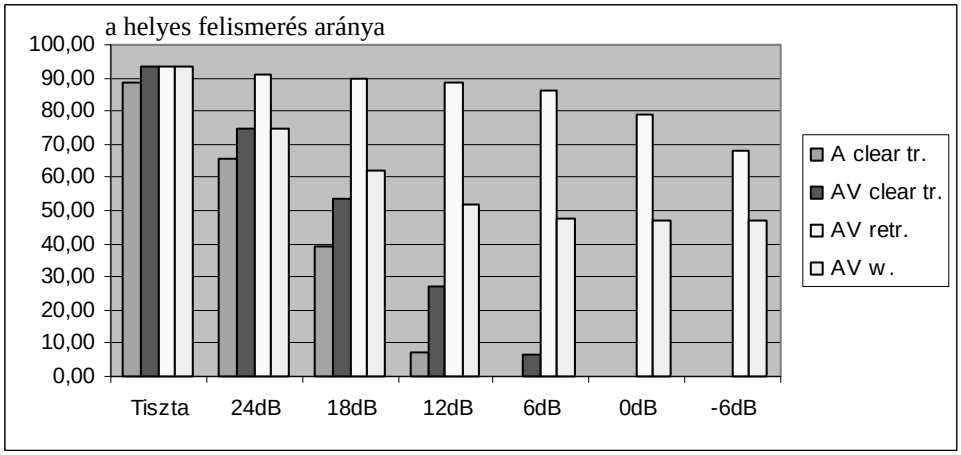
A hozzáadott zaj nélküli beszéddel végzett tanítás során *wa=wv=1*. A két *stream*re osztott akusztikus és vizuális jellemzőkkel végzett tanítás után a zajmentes beszédre kapott eredmények pontosan megegyeztek az egy *stream*be egyesített jellemzőkkel elért felismerési arányokkal. Ez azt sugallja, hogy nem származik előny a felismerési alapegységen belüli átmenetek időbeli elcsúsztathatóságából. A szinkron az állapotok szintjén is teljes, az 5. 2. /2. ábrával összhangban, amely szerint kismértékű csökkenés már kis elcsúszásnál fellép. A zajos beszéddel folytatott tesztelés során a vizuális jel súlyának növelésével lehetőségünk van az artikulációs jellemzők erősítésére.

A különböző minőségű beszéd modellezéséhez különböző jel/zaj viszonyú akusztikus jelet generáltam normál eloszlású zaj hozzáadásával. A zajmentes beszéddel tanított felismerő akusztikus és vizuális modalitásának eltérő súlyozásával megvizsgáltam a különböző jel/zaj viszonyú beszédhez tartozó felismerési arányokat, az eredményeket a 7. 8. ábra mutatja. A vizuális modalitás súlyának növelésével a helyes felismerés aránya az egymódusú vizuális felismerési arányhoz tart.



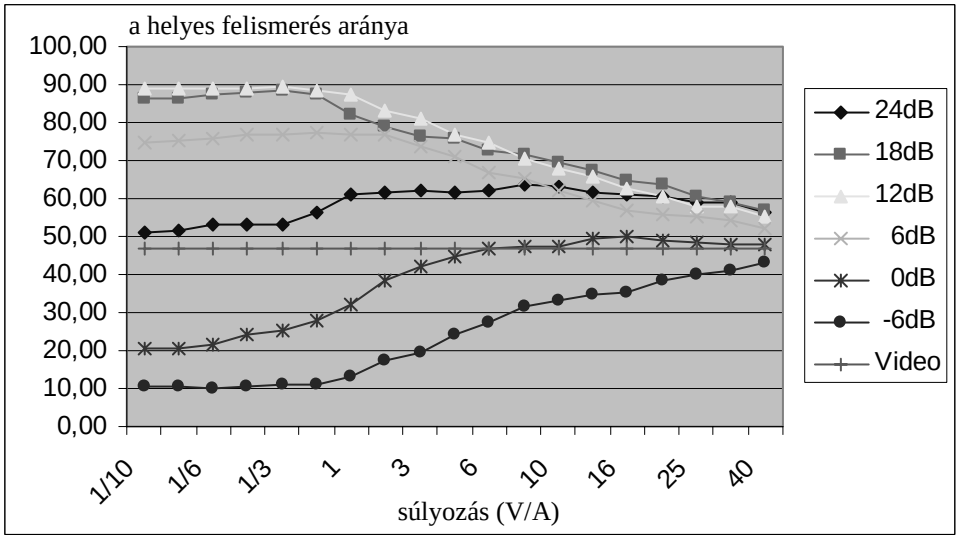
7. 8. ábra. A helyes felismerés aránya a vizuális modalitás különböző súlyozásával (V/A) a jel/zaj viszony függvényében.

A 7. 9. ábrán a hozzáadott zajt nem tartalmazó beszéddel tanított fix súlyozású audiovizuális felismerő (*AV clear tr.*) teljesítményét hasonlíthatjuk össze a maximális felismerési arányt biztosító súlyozással kapható (*AV w.*) helyes felismerési arányokkal. Az ábra tanúsága szerint különösen gyengébb minőségű beszédnél múlja felül a változó súlyozású felismerő az akusztikus és vizuális modalitást egyaránt egységnyi súlyozással kezelő társáét.



7. 9. ábra. A zajmentes beszéddel tanított audiovizuális felismerő 7. 6. ábra szerinti oszlopai, kiegészítve a jel/zaj viszony függő (AV w.) súlyozás eredményeivel.

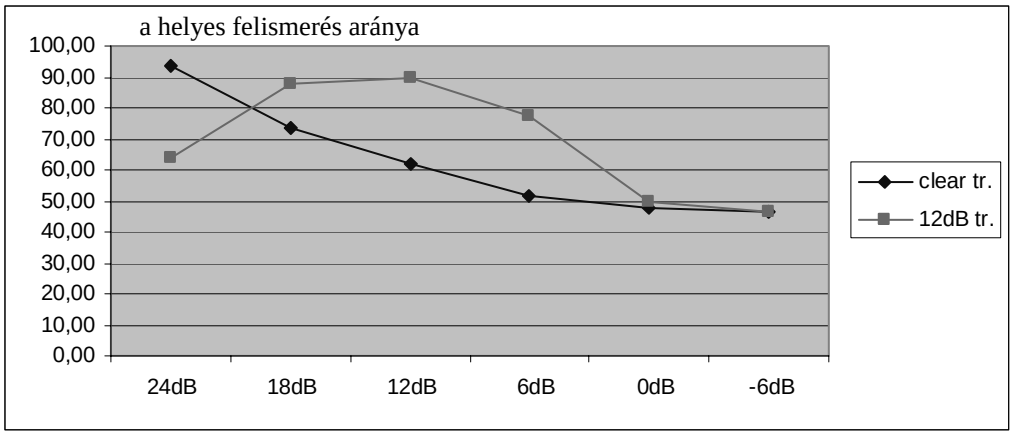
Mivel a zajmentes beszéddel tanított felismerő hibája a modalitások rögzített súlyozása mellett a jel/zaj viszony csökkenésével meredeken nő, érdemes megvizsgálni, hogy változó környezetben a tanítást zajos beszéddel elvégezve milyen felismerési arányokat kapunk. A 7. 10. ábrán látható eredményeknél a tanító beszéd jel/zaj viszonya 12 dB volt.



7. 10. ábra. 12 dB jel/zaj viszonyú beszéddel tanított felismerő felismerési arányai a vizuális és akusztikus modalitás (V/A) különböző súlyozásával.

A beszédfeldolgozásban kevésbé járatos megfigyelő számára meglepő lehet, hogy a jel/zaj viszony javulása zajos jellel tanítás esetén csaknem olyan mértékű csökkenést okoz a felismerési arányokban, mint a jel/zaj viszony romlása.

A könnyebb összehasonlíthatóság érdekében a 7. 11. ábrán a zajmentes (*clear tr.*) és a 12 dB jel/zaj (*12 dB tr.*) viszonyú beszéddel tanított felismerő eredményeit látjuk az akusztikus és a vizuális modalitás optimális súlyozása esetén.

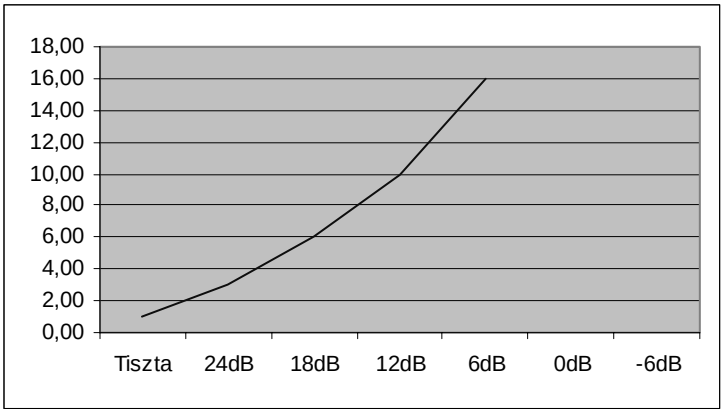


7. 11. ábra. Zajmentes (clear tr.) és a 12 dB jel/zaj viszonyú (12 dB tr.) beszéddel tanított audiovizuális felismerő helyes felismerési arányai a jel/zaj viszony függvényében.

Amennyiben a vizuális és akusztikus modalitás súlyozását változtatni akarjuk, szükség van a jel/zaj viszony becslésére. Mivel a 7. 8. és 7. 10. ábra tanúsága szerint a maximumok laposak, a jel/zaj viszony közelítő becslése is sokat javíthat a felismerő teljesítményén. A jel/zaj viszonyt megbecsülhetjük a beszédszünet és az aktív beszédszakasz teljesítményé-

nek arányával, az *energiabillegés* vizsgálatával (Gordos, 2004). Benoît (1996b) az akusztikus modalitás megbízhatóságát azzal minősíti, hogy a gépi beszédfelismerő győztes alakzatának valószínűsége mennyivel emelkedik ki az első néhány jelölt átlagából.

Ha a 7. 8. ábrán látható diagramon megvizsgáljuk, hogy egy adott jel/zaj viszonyhoz a vizuális jel milyen súlyozása esetén kapjuk a legjobb felismerési eredményeket, minden jel/zaj viszonyhoz rendelhetünk egy optimális súlyt. A 7. 12. ábrán a zajmentes beszéddel tanított felismerő optimális súlyozását láthatjuk a jel/zaj viszony függvényében. A 7. 8. ábra tanúsága szerint 0 és -6 dB jel/zaj viszony esetén az akusztikus jel rontja a felismerési arányokat, vagyis a vizuális modalitás önmagában jobb eredményre vezet.



7. 12. ábra. A zajmentes beszéddel tanított felismerő vizuális és akusztikus modalitásának (V/A) optimális arányai a jel/zaj viszony függvényében.

7. 2. Az akusztikus beszédfelismerési eredmények

Az akusztikus beszédadatbázis az audiovizuálisnál lényegesen több tanító és tesztelő szót tartalmazott. 8000 szó szolgálta a tanítást, 1400 szó a tesztelést. A szavakat részleges lefedettséget biztosító félszótagok alkotják.

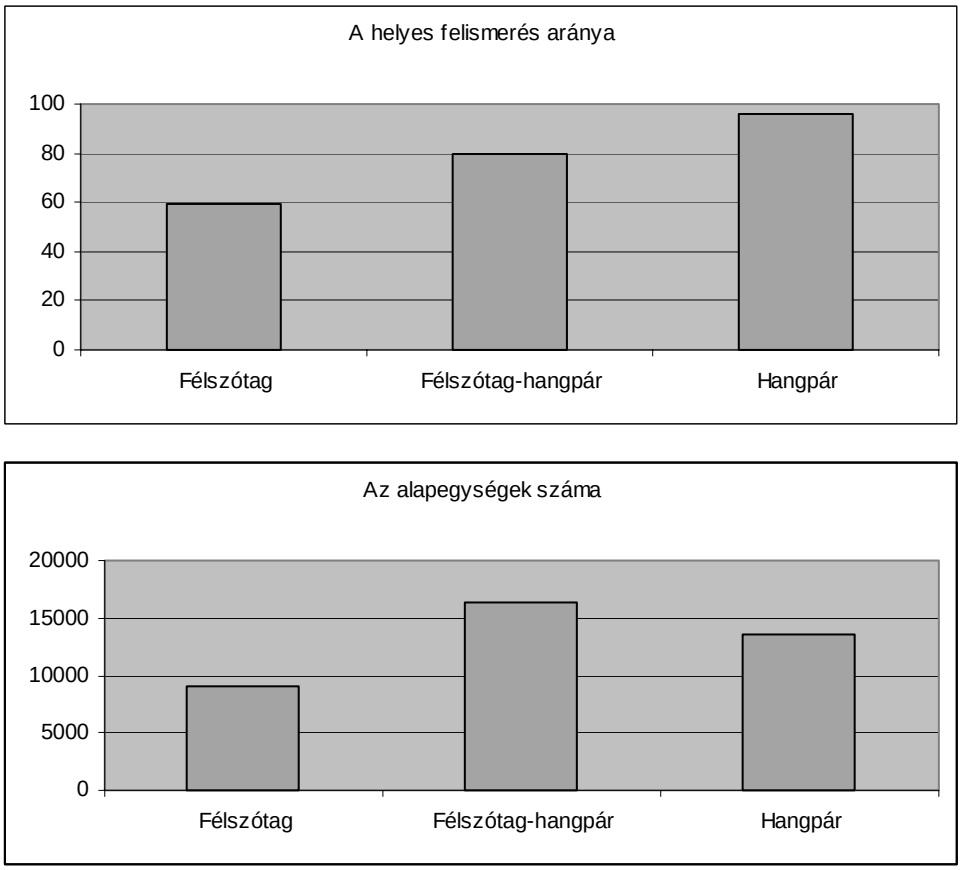
A beszédfelismerési kísérlet arra keresett az audiovizuális adatbázison végzett méréseknél megalapozottabb választ, hogy mit célszerű a folyamatos beszédfelismerésnél a felismerés alapegységének választani. A félszótag alapú felismerésnél a szótagokat a magánhangzó közepén kezdő és záró félszótagra bontjuk (Vicsi, 1995). A szótagoláskor nem az akadémiai elválasztási szabályokat alkalmazzuk, a következő szótagba egy mássalhangzót viszünk át. Pl.: ..V-V.., ..V-CV.., ..VC-CV.., ..VCC-CV…

A félszótagok helyes felismerési aránya 59,2%-ra adódott. A gyenge eredményt annak tulajdonítom, hogy a kezdő és záró félszótagokból képzett szótagok tetszőleges sorrendben követhetik egymást, a szótaghatáron a koartikulációs hatások figyelembe vételére nincs lehetőség. Kísérletet tettem a szótaghatárok illesztésére hangpárok beiktatásával. A záró félszótag utolsó hangjának felétől a következő kezdő félszótag első hangjának feléig terjed a közbeiktatott hangpár. Az elején és a végén szünettel kezdődő illetve végződő hangpár helyezkedik el. A láncolat láncszemeinek felépítése:

hangpár – kezdő félszótag – záró félszótag -

A folytatás újabb láncszem vagy az utolsó hangpár lehet. Az így megerősített kontextusfüggés eredményeképpen a félszótag-hangpár felismerési arány 79,8 % lett.

A hangpár alapú felismerési eredmény felülmúlta a félszótag és félszótag-hangpár alapú arányokat egyaránt. A hangpár alapú akusztikus felismerési arány 95,9 %-ra adódott. A felismerés különböző alapegységeire kapott eredményeket a 7. 13. ábrán látjuk. Az 1 400 szavas tesztelésre szolgáló szótár 9 010 félszótagból, vagy 16 389 félszótag-hangpár láncból, illetve 13 639 hangpárból áll. Ezek szerint a hangpárok átlagos időtartamánál a félszótagok ideje 50%-kal hosszabb, a félszótag-hangpár elemeké pedig kb. 20%-kal rövidebb. A leggyengébb eredményt a leghosszabb átlagméretű félszótag, a legjobb eredményt a közepes méretű hangpár alapú felismerés mutatta. Ez megerősíti az audiovizuális teszt eredményét, egy nagyobb hangadatbázison is fontosabbnak bizonyult a kontextus figyelembe vétele a felismerés alapegységének időtartamánál. A beszédfelismerés alapegysége mindenképp függ a környezetétől, az eredmények tanúsága szerint az alapegységnek ki kell fejeznie a kontextusfüggést. Mivel a fonémák határán átmeneti tulajdonságokat találunk, valószínűleg nem előnyös az alapegység határát fonéma határára tenni.



7. 13. ábra. A félszótag, a félszótag–hangpár és a hangpár alapú akusztikus felismerési eredmények és az alapegységek száma a szótárban.

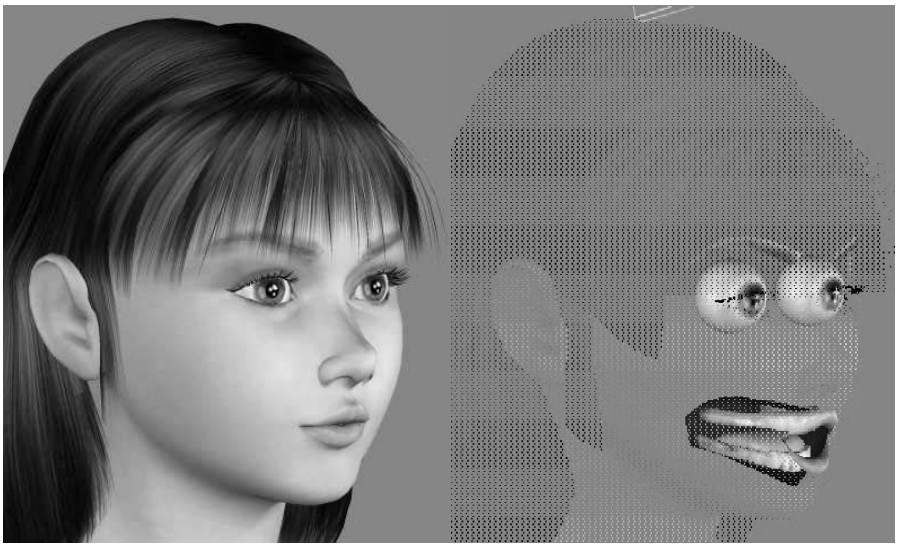
8. Audiovizuális beszédszintézis

A vizuális beszéd elemzésekor gyűjtött tapasztalatok egyik alkalmazási lehetősége a vizuális beszédszintézis. Ezen azt értjük, hogy a mesterséges vagy természetes beszédet az artikulációt utánzó háromdimenziós fejmodell képével egészítjük ki. Tetszőleges szöveget kísérhet az animáció, magyar nyelvű, vizuális szövegfelolvasót fejlesztünk (Czap, Mátyás, 2003, 2004). Az általam kidolgozott algoritmusok programozási feladatait témavezetésem mellett a Miskolci Egyetem két hallgatója diplomamunka keretében végezte (Mátyás, 2003; Ferenczi, 2004). A háromdimenziós fejmodell mozgatásán alapuló animáció kialakításához a saját mérési eredmények mellett felhasználtuk a fellelhető hangalbumok anyagát, ezek alapján osztályoztam a fonémák vizuális megfelelőit, a vizémákat. A saját vizuális beszédfelismerési kutatási eredményekre elsősorban a dinamikus vizsgálatnál támaszkodtam. A koartikulációs hatások figyelembe vételéhez a jellemzőket domináns, rugalmas és határozatlan osztályokba soroltam, 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.

8. 1. A vizuális beszédszintézis motivációja

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 lehet a mesterséges beszélő fej. Az arcanimáció megvalósítása a beszédartikuláció modellezésére mindössze két évtizeddel ezelőtt kezdődött. A mai szemmel nézve 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 audiovizuális 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 megmutathatja a hangképzés részleteit.

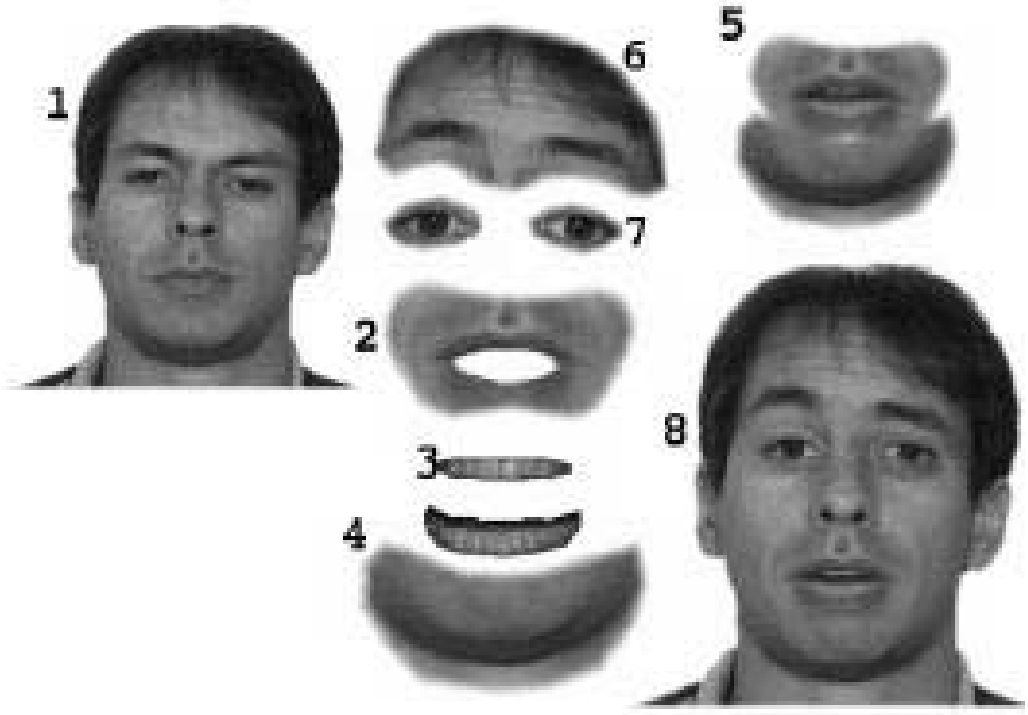


8. 1. ábra Fotorealisztikus és transzparens megjelenítés

Hangvezérelt beszélő fejek fejlesztésén dolgoznak hallássérültek segítésére távközlési alkalmazásokban (Karlsson et al. 2003). A fejlett magyar nyelvű akusztikus beszédszintézis mellett hiánypótló célzattal kezdtünk vizuális beszédszintetizátor fejlesztéséhez.

8. 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.



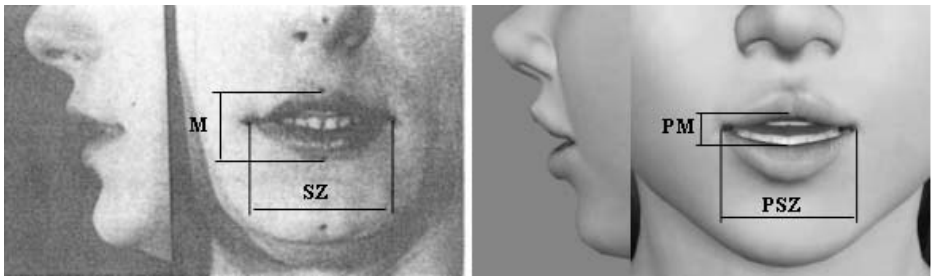
8. 2. ábra. Kétdimenziós fejmodell elemei (Cosatto, Grafat, 1998)

A testmodellezés fejlődése a háromdimenziós modellezésre terelte a kutatók figyelmét. 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 (Massaro, 1998). Minden modell mozgatásánál külön figyelmet kell fordítani a jellemzők összehangolt változtatására, mert könnyen természetellenes hatás alakulhat ki.

8. 3. A beszéd vizuális alapegysége

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. A hangképző szervek jellemző helyzete magyar beszédhangokra megtalálható alapvető munkákban (Molnár, 1986; Bolla, 1980, 1995) . A 8. 3. ábra példát mutat be arra, hogy mennyire hasonló egy hagyományos labiogram (Bolla, 1995; Mátyás, 2003) és egy 3D-s beszélő fejen beállított, ugyanazon hangra jellemző artikuláció.



8. 3. ábra Minta fotolabiogram és a renderelt 3D fejmodell

A magyar beszédhangok vizéma készletét mintaszavak (Bolla, 1995) 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.

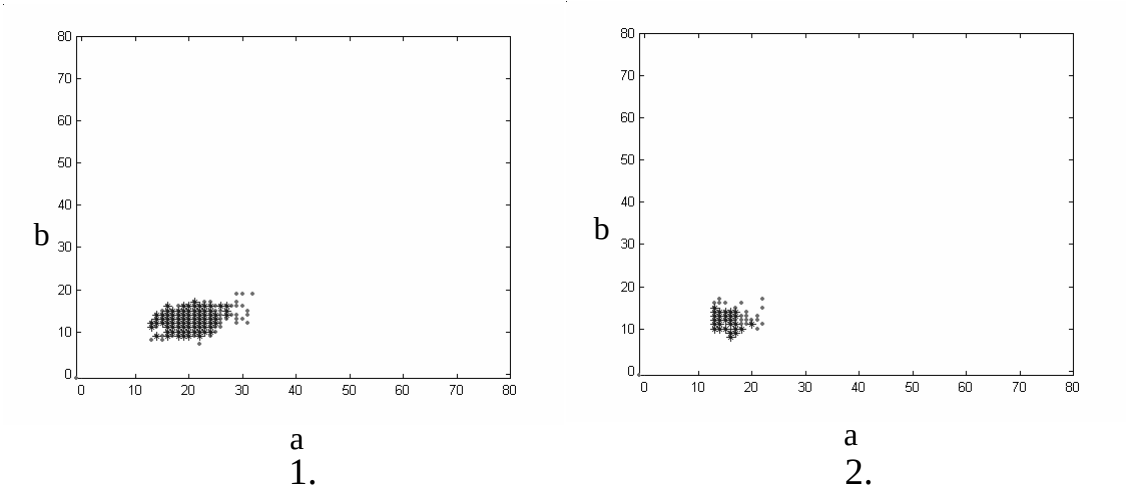
Az eredményt a 8. 1. táblázat mutatja, a hangokat a magyar helyesírási betűképükkel jelöltem.

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

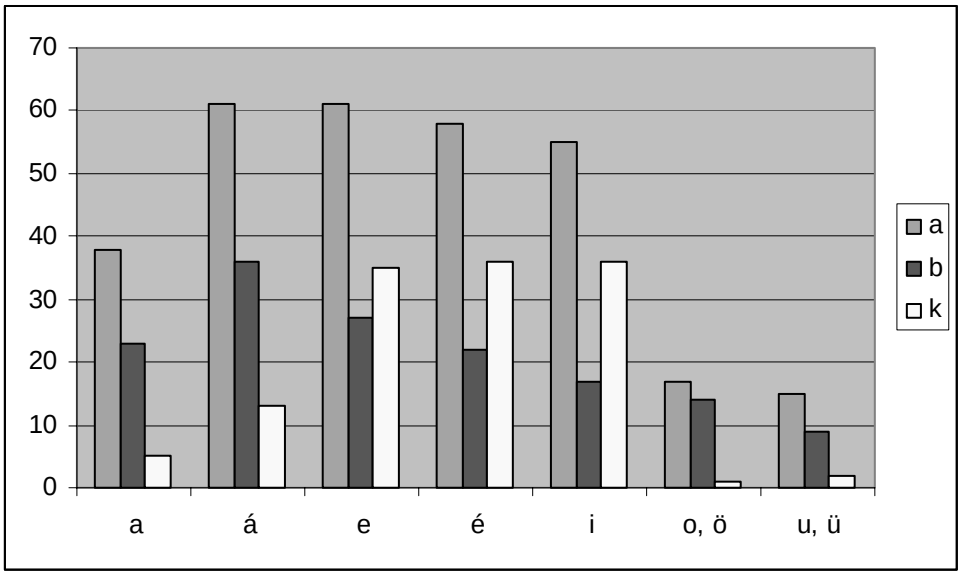
Néhány megjegyzés a vizémák osztályozásához:

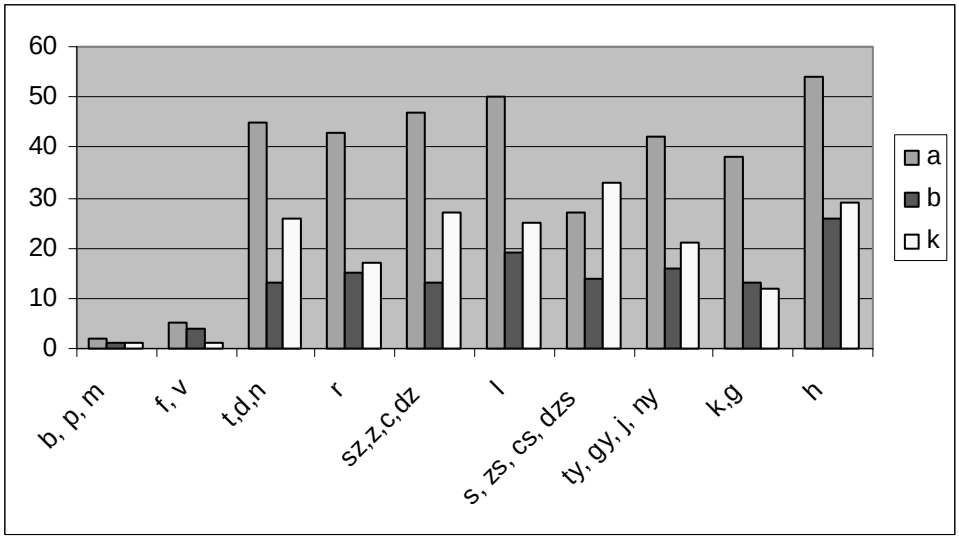
- a csoportosítás elsősorban ajakforma alapján történt, a nem látható nyelvállás eltérő lehet (pl.: o-ö, u-ü)
- a nem jelzett hosszú magánhangzók a rövid párjuknál szűkebb szájnyílással vannak jelen (8. 4. ábra)
- az artikuláció előállításához ennél bővebb készlettel dolgozunk



8. 4. ábra. Az *o* (1.) és a szűkebb ajaknyílású *ó* (2.) ajakszélességének (*a*) és ajaknyílásának (*b*) átmeneti (.) és állandósult (*) szakasza.

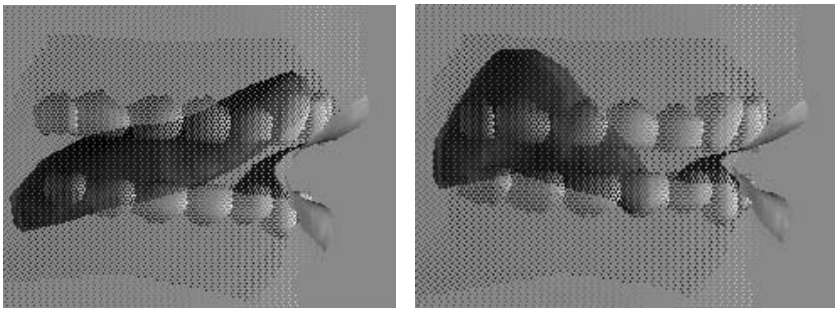
A 8. 5. ábra a vizémák ajakméreteit és intenzitási tényezőit ábrázolja.





8. 5. ábra. A vizémák ajakszélessége (*a*), ajaknyílása (*b*) és intenzitási tényezője (*k*). A méretek pixelben, az intenzitás a fehér (255) világosságának arányában látható.

Az eddig megjelent beszédhangok atlasza (Molnár, 1986), illetve magyar hangalbumok (Bolla, 1980, 1995) alapján meghatározhatók a vizémák legfontosabb paraméterei, ezekből alakul ki az a kulcskeret (keyframe) készlet, amely az artikuláció kiindulási alapja. 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 (8. 6. ábra) a nyelv függőleges helyzete (fent-lent), vízszintes mozgása (elöl-hátul), hajlítása (domború-homorú), és a nyelvhegy formája (széles-keskeny, vékony-vastag) befolyásolják.



8. 6. ábra. Jellemző nyelvállások: baloldalon az n-re, jobbra a k-g hangokra

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

5. tézis

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

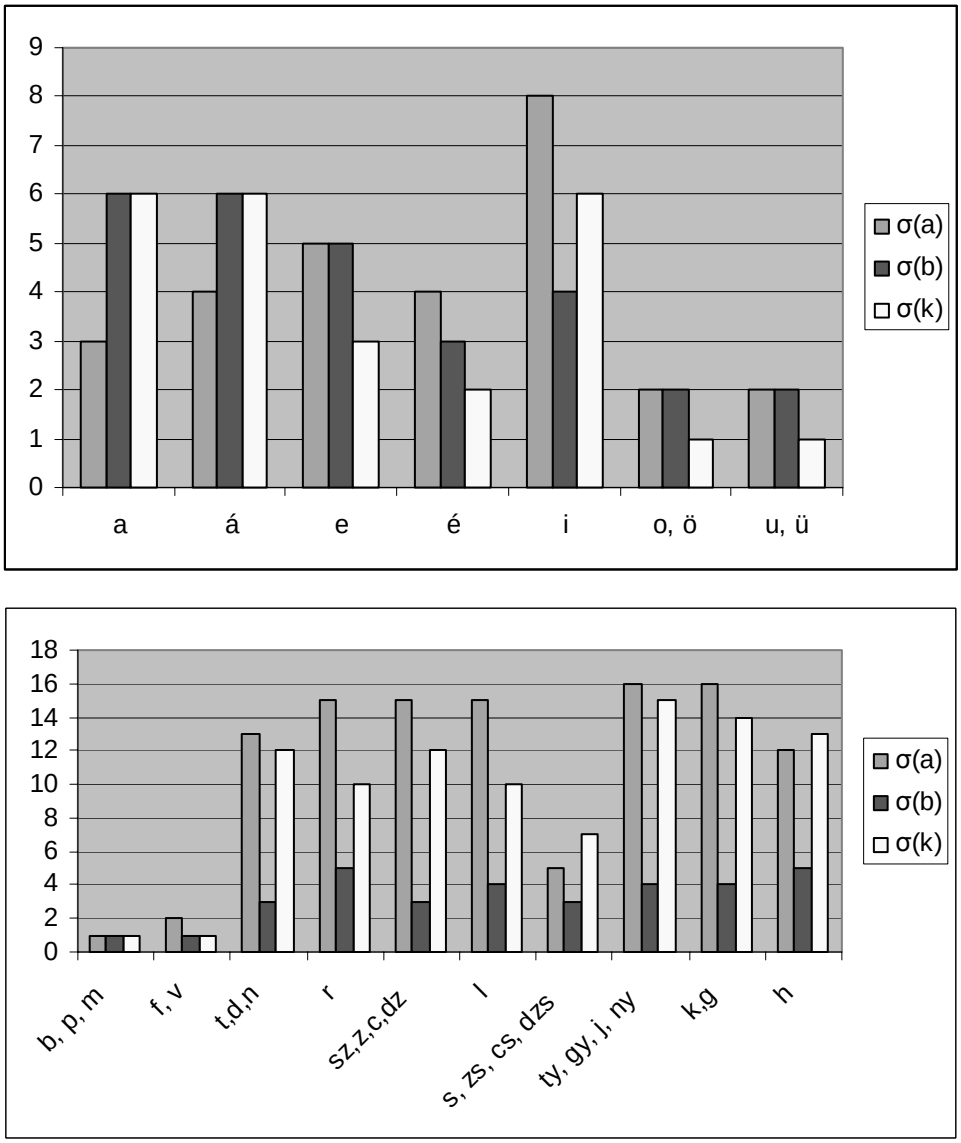
8. 4. Dinamikus működés

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 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.

A koartikulációs hatások figyelembe vételéhez túl kellett lépni az úgynevezett „keyframe” modellen, mivel ez a megközelítés túlságosan intenzív száj- és nyelvmozgáshoz vezetett.. A vizémák minden jellemzőjét (például ajak- és nyelvállások) osztályoztam dominánciájuk alapján. Egyes paraméterek a környezettől függetlenül felveszik jellegzetes értékeiket, mások a környezetükbe simulnak. A vizuális beszédfelismerés adatai alapján a vizémák jellemzőit három kategóriába soroltam:

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

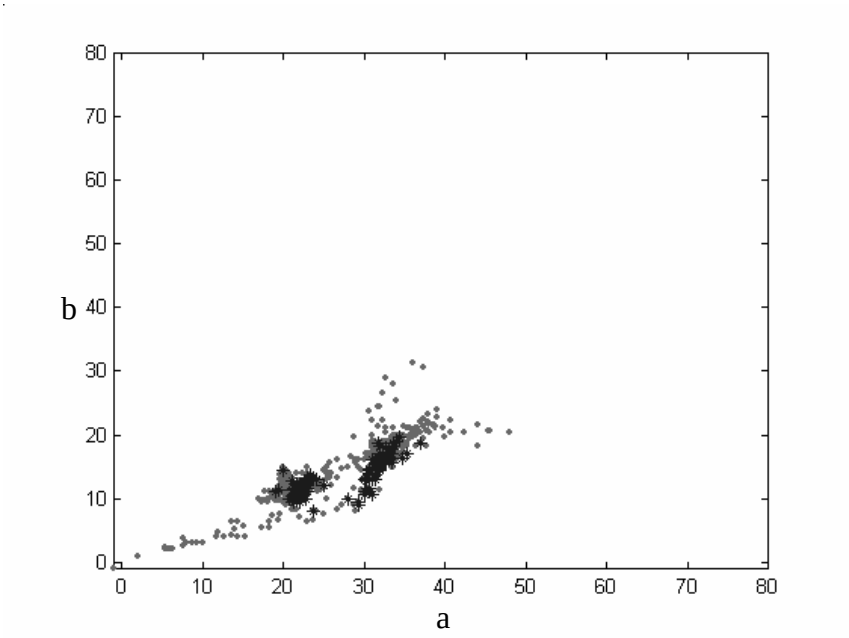
A 8. 7. ábra a vizémák paramétereinek szórását mutatja.



8. 7. ábra. A vizémák jellemzőinek szórása.

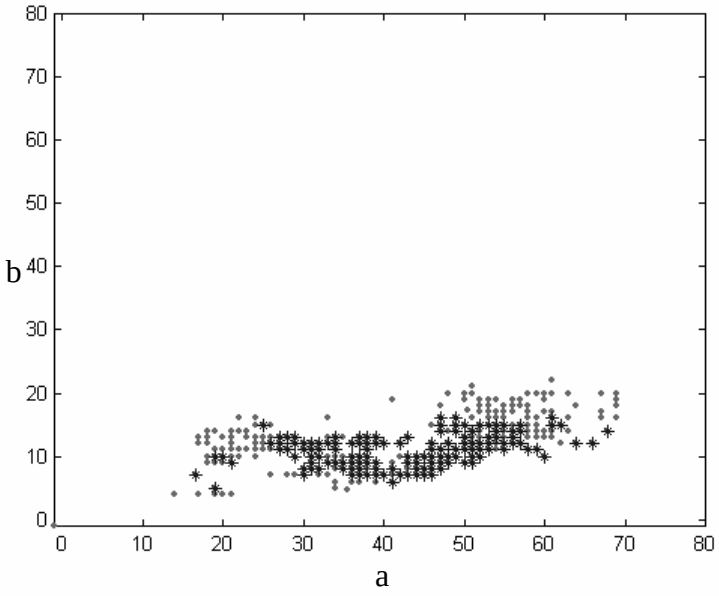
A dominancia meghatározásához elsősorban a jellemzők szórását használtam fel, de segítséget nyújt a látható jellemzők grafikus ábrázolása, az átmeneti és az állandósult szakaszok eloszlása is. A 8. 8. ábrán eltérő színnel láthatók az *s* hang átmeneti és kvázistacionárius szakaszának ajakméretei.

A szomszédos hangok által meghatározott kezdeti- és végállapotok között az ajakméretek egy szűkebb területet foglalnak el.



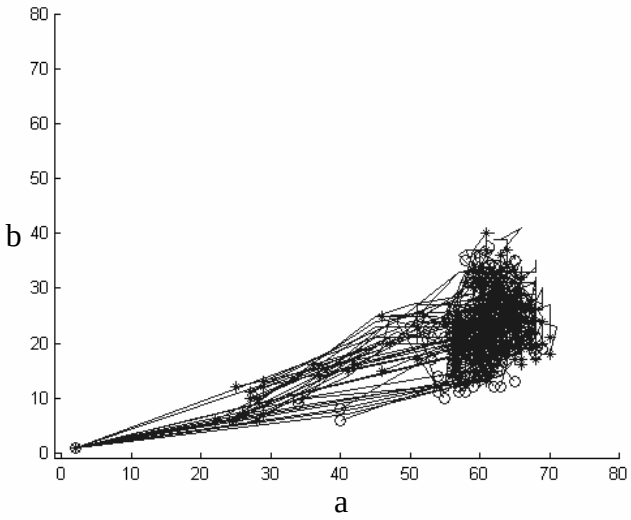
8. 8. ábra. Az *s* hang átmeneti (.) és állandósult (*) szakaszának ajakszélessége (*a*) és ajaknyílása (*b*).

Az ajakméretek eloszlása a *j* hang átmeneti és állandósult tartományára a 8. 9. 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.



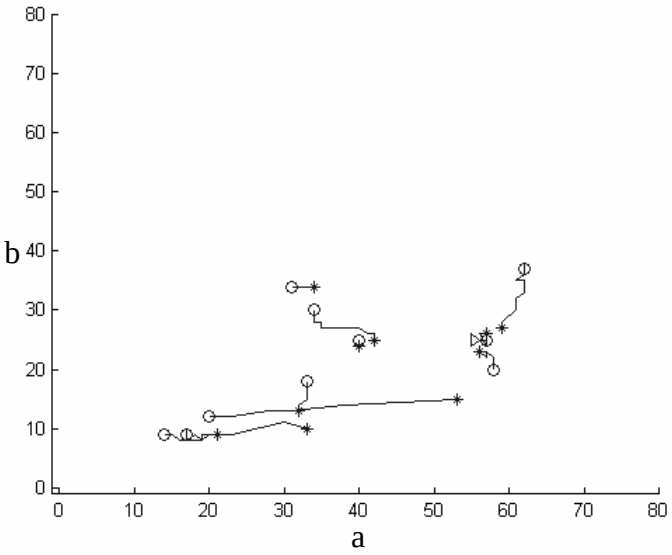
8. 9. ábra. A *j* vizéma ajakméreteinek eloszlása (átmeneti (.) és állandósult (*) szakasz)

Az ajakméretek változásának trajektóriája is támpontot ad a dominancia osztály meghatározásához. A 8. 10. 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.



8. 10. 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 8. 11. ábrán. (A változások követhetősége végett csak néhány görbe szerepel.)



8. 11. ábra. A *h* vizéma ajakméreteinek változása. „*” jelzi a kezdőpontot, „o” a végpontot.

A 8. 2. táblázat mutatja a vizémák ajakformára, a 8. 3. táblázat a nyelv vízszintes helyzetére vonatkozó csoportosítását.

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

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 háromszintű dominancia modell megalkotása után találtam rá Cohen és Massaro (1993) megoldására, akik a dominanciát folytonos változóként kezelik, az idővel felerősödő és elhaló hatást exponenciális függvényekkel közelítik.

8. 5. 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 pl. 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 ugyanarra a 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.

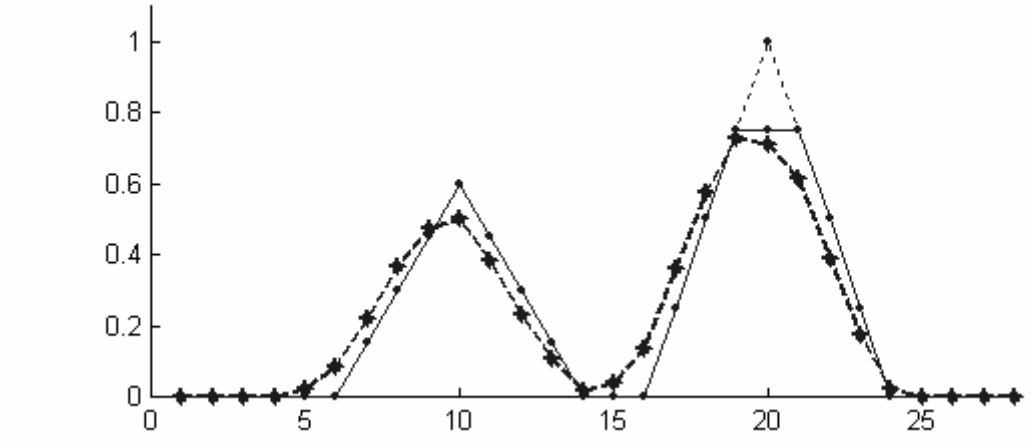
8. 5. 1. Előartikuláció és szűrés

Olaszy Gábor (2003) ötlete alapján a kimondás 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 neveztem el – már az első hang megszólalása előtt kialakul az ajakforma, hasonlóan a természetes kimondáshoz.

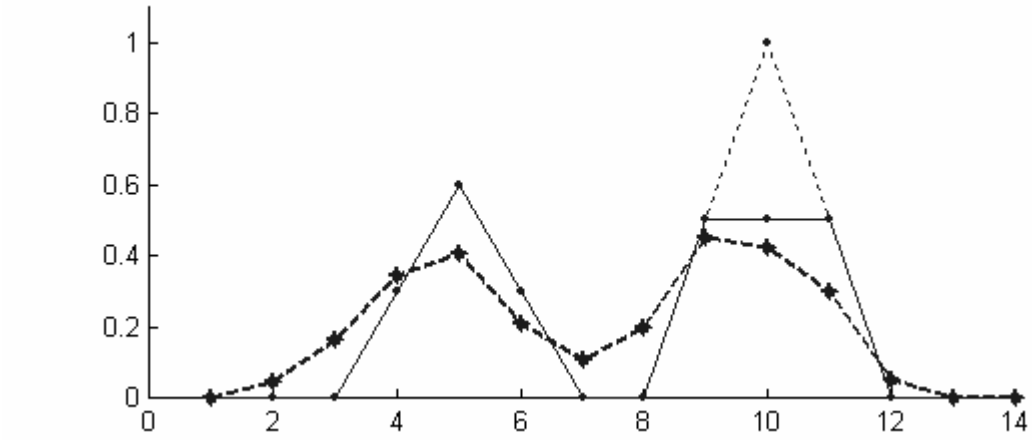
A természetes vagy szintetizált beszédhez szinkronizálás folyamán különböző sebességű beszéddel szembesültünk. Lassú beszédnél a vizémák jellemzői megközelítik névleges értéküket, gyors beszédnél az artikuláció elnagyoltabb. A rugalmas csoportba sorolt jellemzőkre is igaz, hogy gyors beszédnél a lekerekítés nagyobb. A rugalmas jellemzők kialakítására a medián szűrést alkalmaztam: 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 8. 12. 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.



1.



2.

8. 12. 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.

8. 6. 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 (1978) által meghatározott hét érzelem közül választhatunk: semleges, haragos, ellenszenves, szorongó, boldog, szomorú, meglepett. Erre láthatunk példát a 8. 11. ábrán.

8. 11. ábra Ellenszenves és boldog arckifejezés

A CD mellékleten néhány minta található a Beszélő fej különböző modelljeivel.

9. Tézisek

1. A beszéd vizuális jellemzői

A természetes emberi beszédérzékelés szubjektív vizsgálata megmutatta, hogy az ajakszélesség az ajaknyílás és a szájnyílás világossága a száj megfigyelésével egyenrangúan írja le az artikulációt.

1. 1. A választott vizuális jellemzők meghatározását a 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.

1. 2. A választott vizuális jellemzők meghatározását a 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.

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 alapegység vizsgálatára – ahol csak a korai integrálás jöhet szóba – nem került sor.

2. 1. 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.

2. 2. 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.

3. A beszéd vizuális alapegységének osztályozása

Az emberi, vagy gépi beszéd érthetőséget javító kísérője lehet az artikuláció vizuális megjelenítése.

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

4. Az artikuláció dinamikus viselkedésének modellezése

A hangképzés dinamikus viselkedésének vizsgálata elengedhetetlen a vizuális beszédszintetizátor működtetéséhez. Az artikuláció dinamikus tulajdonságait a magyar nyelvre még nem írták le.

4. 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 CD melléklet tartalma

A CD melléklet könyvtárai megegyeznek a kapcsolódó fejezetek számával.

3. /minták

Az arc különböző részeinek elfedésével kialakított vizsgáló jelekre találunk példákat, ilyen jellegű mintákkal zajlott a szubjektív mérés.

4. /artikulációs_könyvtár

A hasonlóságon alapuló lényegkiemelés prototípus képei.

6. 4. /diaddef, felszodef

A félszótag és hangpár alapú felismerés alapegységeit és kapcsolódási szabályait leíró szabályok a HTK szintaktikája szerint.

A fonémákra/vizémákra egykarakteres jelölést vezettem be:

á	é	cs	gy	ny	sz	ty	zs
A	E	C	G	N	S	T	Z

8. /átlátszó

A transzparens modell nyelvmozgása jobban látható mint a természetes beszélőé.

8. /modell3

Minta a legjobb ajakmozgású modellel.

8. 6. /érzelmek

Néhány példa az érzelem kifejezésére. Az érzelem skálázható, itt erős érzelem kifejezést állítottunk be.

Felhasznált irodalom

[1] Stork, D.G. and Hennecke, M.E. (Eds.) (1996) *Speechreading by Humans and Machines,* Springer-Verlag, Berlin.

Adjoudani, A., and Benoît, C. (1996). *On the integration of auditory and visual parameters in an HMM-based ASR.* In [1] pp. 461-471.

Allen J. B. (1994) *How do humans process and recognize speech*? IEEE Trans. Speech Audio Processing, vol. 2, no. 4, pp. 567–577. In Workshop on Robust Methods for Speech Recognition in Adverse Conditions

Bailly, G., Vatikiotis-Bateson, E., and Perrier, P. Eds., (2004) *Chapter to appear In: Issues in Visual and Audio-Visual Speech Processing.* MIT Press.

Benoît, C. (1995) *On the production and perception of audio-visual speech by man and machine.* In Y. Wang, et al., (Eds.),Multimedia & Video Coding, Plenum Press, NY.

Benoît, C., Adjoudani, A., Guiard-Marigny, T., Le Goff, B., and Reveret, L. (1998). *Multimodal integration for advanced multimedia interfaces (MIAMI).* Reports ESPRIT III, Basic Research Project 8579. Electronic download.

Benoît, C., Adjoudani, A. (1996b) *On the Integration of Auditory and Visual Parameters in an HMM-based ASR.* In [1] pp. 461-471.

Benoît, C., Guiard-Marigny, T., Le Goff, B., & Adjoudani, A. (1996a). *Which components of the face do humans and machines best speechread?.* In [1] pp. 315-328.

Berecz B. (2000) *Videokamera automatikus fókuszbeállítása.* Diplomaterv, Miskolci Egyetem.

Bernsen, N. O. (2002) *Multimodality in Language and Speech Systems – from Theory to Design Support Tool.* In Multimodality in

Language and Speech Systems. Kluwer Academic Publishers, Dordrecht/Boston/London.

Bernstein, L. E. & Auer, E. T., Jr. (1996). *Word Recognition in Speechreading.* In [1] pp. 17-26.

Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K. E., Öhman, T. (1997) *The Teleface Project Multi-Modal Speech-Communication for the Hearing Impaired.* Eurospeech'97 Proceedings

Bolla K. (1980) *Magyar hangalbum : A magyar beszédhangok artikulációs és akusztikai sajátságai* MTA Nyelvtudományi . Intézet., Budapest.

Bolla K. (1995) *Magyar fonetikai atlasz. A szegmentális hangszerkezet elemei* Nemzeti. Tankönyvkiadó., Budapest.

Bregler C., Omohundro, S.M. (1995) *Nonlinear image interpolation using manifold learning.* In G. Tesauro, D.S. Touretzky and T.K. Leen (Eds.), Advances in Neural Information Processing Systems, vol. 7, Cambridge, MA: MIT Press. pp. 973-980.

Bregler, C. and Konig, Y. (1994). "*Eigenlips" for robust speech recognition.* Proc. International Conference on Acoustics, Speech and Signal Processing, Adelaide, Australia, pp. 669–672.

Brookes M. (1996) *VOICEBOX*
www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

Buda Béla (2000) A *közvetlen emberi kommunikáció szabályszerűségei.* Animula Kiadó, Budapest, pp. 71-113.

Cohen, M. M., Massaro, D. W. (1993) *Modeling coarticulation in synthetic visual speech.* In N. M. Thalmann & D. Thalmann (Eds.) Models and Techniques in Computer Animation, Tokyo: Springer-Verlag.

Cohen, M. M. Walker, R. L. Massaro D. W. (1996) *Perception of Synthetic Visual Speech.* In [1]

Cootes, T.F., Edwards, G.J., and Taylor, C.J. (1998). *Active appearance models.* Proc. European Conference on Computer Vision, Freiburg, Germany, pp. 484–498.

Cootes, T.F., Taylor, C.J., Cooper, D.H., and Graham, J. (1995). *Active shape models - their training and application.* Computer Vision and Image Understanding, 61(1): 38–59.

Cosatto E., Grafat H. P. (1998) *2D Photo-realistic Talking Head* Computer Animation, Philadelphia, Pennsylvania, pp. 103-110.

Cosi, P. Caldognetto, E. M. (1996) *Lips and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications.* In [1]

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

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

Czap L. (1998a) Audio and Audio-visual Perception of Consonants Disturbed by White Noise and 'Cocktail Party'. 5th International Conference on Spoken Language Processing 30th November - 4th December, Sydney, Australia Proceedings Volume 2, p. 253-256.

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

Czap L. (2000) Lip Representation by Image Ellipse. 6th International Conference on Spoken Language Processing Proceedings Bejging, China, Proceedings Vol. IV. 93-96.

Czap L. (2004a) Audiovizuális beszédfelismerés. Magyar Számítógépes Nyelvészeti Konferencia, Szeged. (közlésre elfogadva)

Czap L. (2004b) Feature Extraction for Speechreading. International Carpathian Control Conference ’04, Zakopane, Poland, Proc. pp.

Dalton, B., Kaucic, R., Blake, A. (1996) *Automatic Speechreading Using Dinamic Contours.* In [1] pp. 373-382.

Dougherty, E.R. and Giardina, C.R. (1987). *Image Processing – Continuous to Discrete*. Vol. 1. Geometric, Transform, and Statistical Methods. Englewood Cliffs, NJ: Prentice Hall.

Drótos L. L. (2000) *Emberi arc követése kamera mozgatással és zoomolással.* Diplomaterv, Miskolci Egyetem.

Duchnowski, P., Meier, U., and Waibel, A. (1994). *See me, hear me: Integrating automatic speech recognition and lipreading.* Proc. International Conference on Spoken Language Processing, Yokohama, Japan, pp. 547–550.

Duda R., and Hart P. (1973). *Pattern Classification and Scene Analysis.* New York: Wiley & Sons.

Dupont S., Luettin J. (2000) *Audio-Visual Speech Modeling for Continuous Speech Recognition* IEEE Transactions on Multimedia, Vol. 2, No. 3, September

Ekman, P., Friesen, W. (1978) *Facial Action Coding System*. Consulting Psychologists Press. Inc.

Ferenczi P. (2004) *Vizuális beszédszintézis*. Diplomaterv, Miskolci Egyetem.

Gordos G. (2004) *Személyes konzultáció.* Budapest, 2004. 10. 25.

Grant K.W., van Wassenhove V., Poeppel D. (2003) *Discrimination of Auditory-Visual Synchrony.* AVSP 2003, Jorioz, France. Proc. pp. 31-35.

Grétsy L., Kovalovszky M. (Eds.) (1980) *Nyelvművelő kézikönyv.* Akadémiai Kiadó, Budapest.

Griffin, E. (2003) *Bevezetés a kommunikációelméletbe.* Harmat Kiadó, Budapest, pp. 50-53.

Hennecke, M.E., Stork, D.G., and Prasad, K.V. (1996). *Visionary speech: Looking ahead to practical speechreading systems.* In [1] pp. 331–349.

Hermansky, H. and Morgan, N. (1994). *RASTA processing of speech. IEEE* Transactions on Speech and Audio Processing, 2 (4): 578–589.

Hu, M. K. (1962). *Visual pattern recognition by moment invariants.* IRE Transactions on Information Theory, Vol. 8. (1) pp. 179-187.

Karlsson I., Faulkner A., Salvi G. (2003) *SYNFACE – A Talking Face Telephone*. Eurospeech 2003 Geneva, Switzerland: pp. 1297-1300

Kass, M., Witkin, A., and Terzopoulos, D. (1988). *Snakes: Active contour models.* International Journal of Computer Vision, 1(4):321–331.

Lavagetto, F., Lavagetto, P. (1996) *Time Delay Neural Networks for Articulatory Estimation from Speech.* In [1] pp. 437-444

Luettin, J., Thacker N.A., and Beet, S.W. (1996). *Active shape models for visual speech feature extraction.* In [1] pp. 383-390.

Massaro, D. W. (1998) *Perceiving Talking Faces.* The MIT Press Cambridge, Massachusetts Lon-don, England

Massaro, D.W., Stork, D.G. (1998) *Speech recognition and sensory integration.* American Scientist, May-June.

Massaro, D.W. (1998) *Perceiving Talking Faces.* The MIT Press Cambridge, Massachusetts London, England 359-390

Matthews, I., Potamianos, G., Neti, C., and Luettin, J. (2001). *A comparison of model and transform-based visual features for audio-visual LVCSR.* Proc. International Conference on Multimedia and Expo, Tokyo, Japan.

Mátyás J. (2003) *Vizuális beszédszintézis.* Diplomaterv, Miskolci Egyetem.

McGrath, M, Summerfield, Q. (1985) *Intermodal Timing Relations and Audio-visual Speech Recogniton.* Jurnal of Acoustical Society of America, 77(2): pp. 678-685,

Molnár J. (1986) *A magyar beszédhangok atlasza* Tankönyvkiadó, Budapest.

Mukundan, R., and Ramakrishnan K.R. (1998). *Moment functions in image analysis.* Singapore: Word Scientific Press. pp. 11-24.

Nakamura, S., Ito, H., and Shikano, K. (2000). *Stream weight optimization of speech and lip image sequence for audiovisual speech recognition.* Proc. International Conference on Spoken Language Processing, Beijing, China, vol. III, pp. 20–23.

Nankaku, Y., Tokuda, K., and Kitamura, T. (1999) *Intensity- and location normalised training for HMM-based visual speech recognition.* In Proceedings of the Eurospeech'99, Budapest, pp. 1287-1290.

Neti, Ch., Potamianos, G. (2003) *Audio-Visual Speech Recognition in Challenging Environment.* EUROSPEECH, Geneve

Olaszy G. (2003) *ITEM 345 sz. pályázat. Az animáció verifikálása.*

Pease (1989) *Testbeszéd*. Park Kiadó, Budapest.

Pérez, J. F. G. Lukas, K. Frangi, A. F. (2003) *Low Resource Lip Finding and Tracking Algorithm for Embedded Devices* EUROSPEECH, Geneve

Petajan E., Graf H. P. (1996) Robust Face Feature Analysis for Automatic Speechreading and Character Animation. In [1] pp. 425-436

Petajan, E.D. (1984). *Automatic lipreading to enhance speech recognition.* Proc. Global Telecommunications Conference, Atlanta, GA, IEEE Communication Society. pp. 265-272, pp. 265–272.

Potamianos G., Neti C., Deligne S. (2003) *Joint Audio-Visual Speech Processing for Recognition and Enhancement.* AVSP 2003, Jorioz, France. Proc. pp.95-104

Potamianos, G. and Graf, H.P. (1998). *Discriminative training of HMM stream exponents for audio-visual speech recognition.* Proc. International Conference on Acoustics, Speech and Signal Processing, Seattle, WA, pp. 3733–3736.

Potamianos, G., Luettin, J., and Neti, C. (2001a). *Hierarchical discriminant features for audio-visual LVCSR.* Proc. International Conference on Acoustics, Speech and Signal Processing, Salt Lake City, UT, pp. 165–168.

Rabiner, L., Juang, B. H. (1993) *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ,

Rao, K. R., Hwang, J. J. (1996) *Technics and Standards for Image, Video and Audio Coding.* Prentice Hall, Upper SaddleRiver, NJ,

Robert-Ribes, J., Piquemal, M., Schwartz, J-L., and Escudier, P. (1996). *Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition.* In [1] pp. 193-210.

Saldana, H. M. Pisoni, D. B. Fellowes, J. M. Remez, R. E. (1996) *Audio-Visual Speech Perception Without Speech Cues: A First Report.* In [1]

Scanlon P., Reilly R., de Chazal P. (2003) *Visual Feature Analysis for Automatic Speechreading.* AVSP 2003, Jorioz, France. Proc. pp. 127-132

Scanlon, P. and Reilly, R. (2001). *Feature analysis for automatic speechreading.* Proc. Workshop on Multimedia Signal Processing, Cannes, France, pp. 625–630.

Shdaifat I., Grigat R., Langmann D. (2003) *A System for Automatic Lip Reading.* AVSP 2003, Jorioz, France. Proc. pp. 121-126

Silsbee, P.L. (1994). *Motion in deformable templates.* Proc. International Conference on Image Processing, Austin, TX, vol. 1, pp. 323–327.

Sumby, W.H. and Pollack, I. (1954). *Visual contribution to speech intelligibility in noise.* Journal of the Acoustical Societyof America, *26*(2):212–215.

Summerfield, A.Q. (1987). *Some preliminaries to a comprehensive account of audio-visual speech perception.* In Dodd, B. and Campbell, R. (Eds.), Hearing by Eye: The Psychology of Lip-Reading. London, United Kingdom: Lawrence Erlbaum Associates, pp. 3–51.

Teissier, P., Schwartz, J-L., and Guérin-Dugué, A. (1997). *Models for audovisual fusion in a noisy-vowel recognition task.* Electronic Proceedings of the IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing. Princeton, NJ.

Tiippana, K., and Sams, M. (1999). *Attentional effects on audiovisual integration of speech.* Proceedings of the 7th European Summer School on Language and Speech Communication. Stockholm, Sweden.

Vicsi, K., Vigh, A. (1995) *Text independent neural network/rule based hybrid, continuous speech recognition.* EUROSPEECH'95. Madrid: pp. 2201-2204

Young S. et al. *The THK Book for HTK Version 3.2* Cambridge University Engineering Department 2003, http://htk.eng.cam.ac.uk/docs/docs.shtml

Yuille, A.L., Cohen D.S., and Hallinan, P.W. (1989) *Feature Extraction from Faces Using Deformable Templates.* In Proceedings of the Computer Vision and Pattern Recognition. Washington, DC. IEEE Computer Society Press pp. 104-109.

Yuille, A.L., Hallinan, P.W., and Cohen, D.S. (1992). *Feature extraction from faces using deformable templates.* International Journal of Computer Vision, 8(2): pp. 99–111.