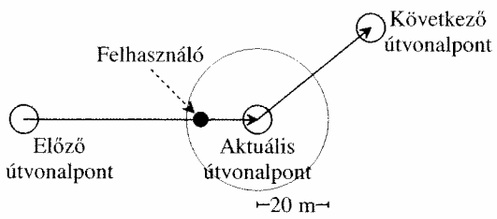


Navispeech

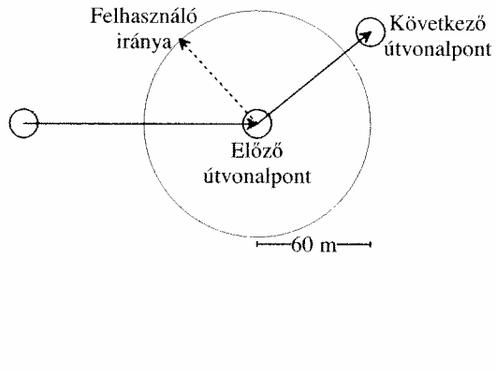
Elértük a(z) „Magyar tudósok körútja - Warga László utca sarok" útvonalpontot.

Következő pont: 3. Warga László utca - Bogdánffy utca sarok

Az előző útvonalpont felől nézve balra kell fordulni.

OK Mégsem

12.11. ábra. Útvonalpont elérése szimulációs ábrán (bal) és a képernyőre kiírt üzenet az útvonalpont elérésekor (jobb)



Navispeech

Letért az útvonalról. Forduljon meg, visszavezetem az előző útvonalponthoz.

Beállítások Kilépés

12.12. ábra. Az útvonalról való letérés szimulációs helyzetképe (bal), figyelmeztetés letéréskor (jobb)

A betűnagyság és a színek beállítására a *Beállítások – Kijelző beállítások* menüpontban van mód. Amennyiben a megjelenített szöveg nem fér el a képernyőn, a „Le" és „Fel" kurzorbillentyűkkel görgethető a megfelelő irányba.

Az új generációs *Symbian* operációs rendszert futtató okostelefonok lehetővé teszik a képernyő 90 fokos elforgatását (ennek neve „*landscape wiew*", míg az eredeti a „ *portrait view*"). A két nézet között lehet az alkalmazás futtatása közben is váltani. Ekkor a kijelzőn lévő szöveg tördelése és görgethetősége a fordított képarányhoz igazodik (12.13. ábra).

Navispeech

Aktuális menetirány:

Kelet

Beállítások Kilépés

Navispeech Kilépés

Aktuális menetirány:

Kelet

Beállítások

12.13. ábra. Függőleges nézet (bal) és vízszintes (jobb)

12.8. Beszédjel átalakítása mozgó száj képévé siketek kommunikációjának segítésére

Takács György

Ebben a fejezetben olyan eljárást mutatunk be, amelyik a beszédjelből mozgó száj képet készít, ezzel segítve a siketek kommunikációját (Feldhoffer et al. 2007). Siket emberekben hosszú gyakorlás után fantasztikus szintre fejlődik ki a beszéd megértése pusztán a szájmozgást nézve. Erre alapozva kommunikációs segédeszköz készíthető siket felhasználók számára, amely pusztán a szájról olvasáson alapul, és egy alkalmas mobiltelefon készülékben vagy egy IPTV „set-top-box" egységében megvalósítható. A Pázmány Péter Katolikus Egyetem Információs Technológiai Karán (PPKE-ITK) a Siketek és Nagyothallók Országos Szövetségének közreműködésével kifejlesztett rendszerben egy beszélő ember szájmozgásának képe jelenik meg grafikus kijelzőn. A rendszer a beszédszervek mozgását utánzó mozgó fej vezérlő paramétereit közvetlenül a beszédjelből származtatja.

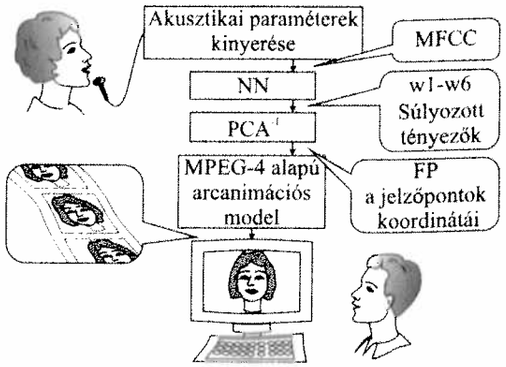
Kiinduló megfontolások. Tisztában voltak a fejlesztők azzal, hogy a teljes emberi beszédfolyamatnak ez csak egy részleges megjelenítése, de számoltak azzal, hogy korlátai ellenére a siketek hasznos kommunikációs segédeszközhöz juthatnak ezzel a megoldással. A rendszer nagyban épít a siketek kifinomult szájról olvasási képességeire és a közvetlen kommunikációban kialakult folyamatos kiegészítő és hibajavító mechanizmusaira. Jelfeldolgozási szempontból a rendszer sarkalatos eleme, hogy időkeretenként meghatározott folyamatos jellegű beszédjellemzőkből folyamatos képjellemzőket számol. Az eddig ismert megoldások leképezték a folyamatos beszédfolyamatot diszkrét elemek (vizémák) halmazára, egy második lépésben pedig a diszkrét elemek halmazát alakították át mozgó fej képévé. Nagy előnye az ismertetésre kerülő közvetlen átalakításnak, hogy megőrzi a beszédfolyamat eredeti időbeli

torzítottuk, hogy a jellemző sárga pontok minél jobban kiemelődjenek. A sárga pontokat végül az RGB komponensek komparálásával detektáltuk. A binarizált képen először dilatációs műveleteket végeztünk, hogy biztosan összefüggő képponthalmazt nyerjünk, majd lépésenként kívülről eróziós folyamattal szedtünk le képpontokat, amíg egyetlen pixel maradt, amit a jellemző pont közepének tekintettünk. Ez az automatikus eljárás legfeljebb 1–2 pixel eltérést eredményez a manuálisan kiválasztott középponthoz képest.

Tekintettel arra, hogy az egyes FP jellemző pontok vízszintesen 40–60, függőlegesen 80–140 pixel tartományban mozognak, az FP meghatározás fenti hibája elfogadható. A koordinátarendszert úgy választottuk meg, hogy középpontja az orr két oldalára helyezett (9.1 és 9.2 a képen) pontok között középen legyen, mivel ezek a pontok mozognak a 15 közül legkevésbé (12.15. ábra).

A beszédjelet hangcsatornában rögzítettük 48 kHz mintavételezéssel, 16 bites mintákkal. A tanító és tesztelő adatbázis szövegét a korábban leírt követelmények szerint választottuk ki. Eszerint a felvételek kétjegyű számokat, hónapok neveit, a hét napjait tartalmazták.

A beszédjel átalakítása szájmozgás képévé. A fejlesztés állapotában a rendszer lényegében egy személyi számítógépen futó programrendszer. A 12.16. ábrán az alapelemek feladata és kapcsolódása szerepel. A mintavételezett beszédjelen minden 40



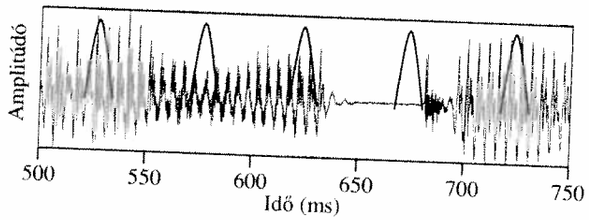
12.16. ábra. A beszéd-szájmozgás átalakító rendszer elemei

ms keretben meghatároztuk a mel-skála szerinti kepsztrumegyüttható-vektort (Mel-Frequency Cepstrum Coefficients, MFCC). Ezeket a jellemző vektorokat vezettük a neurális hálózat (NN) bemenetére, amely a kimenetein kiadja a szájmozgás pillanatnyi állapotát tömörítetten leíró súlytényező vektort [w1-től w6-ig]. A főkomponenselemzés (Principal Component Analysis PCA) inverz műveletével nyerjük a fejmodell vezérléséhez ténylegesen szükséges FP koordinátaértékeket. Ez egy lineáris kombinációs műveletet jelent csupán. Az FP koordinátákat meghatározzuk

minden időkeretre. A rendszer utolsó eleme a nyílt forráskódú LUCIA beszélőfej-rendszernek egy enyhén módosított változata (Cosi et al. 2003). A modellt az FP koordinátákkal vezéreljük és a mozgó kép megjelenik a kijelzőn (lásd később).

Akusztikai lényegkiemelés. A bejövő beszédjelen először egy magasemelő szűrési műveletet hajtunk végre $H(z) = 1\text{-}0{,}983z\text{-}1$ karakterisztikával. Ezután 21,33 ms időtartamú Hamming-ablakkal súlyozzuk a jelet. Az ablakban lévő jelből 16-elemű mel-frekvenciás kepsztrumegyüttható-vektort számolunk.

A koartikuláció jelenségének a beszédfolyamat képi ábrázolásánál legalább akkora jelentősége van, mint a hangjelek feldolgozásakor. A beszédszervek pillanatnyi állása szempontjából vannak domináns és változó fonémák. A domináns fonémák kifejezetten megszabják a száj és környezete képét, viszont a változó típusok képét a környező domináns fonémák nagyban befolyásolják. Ebből fakadóan a beszédjelből a beszédszervek képét becslő algoritmusnak a szomszédos kereteket is felölelő környezetre is tekintettel kell lennie. A siket partnerek számára a lassabb beszédtempó



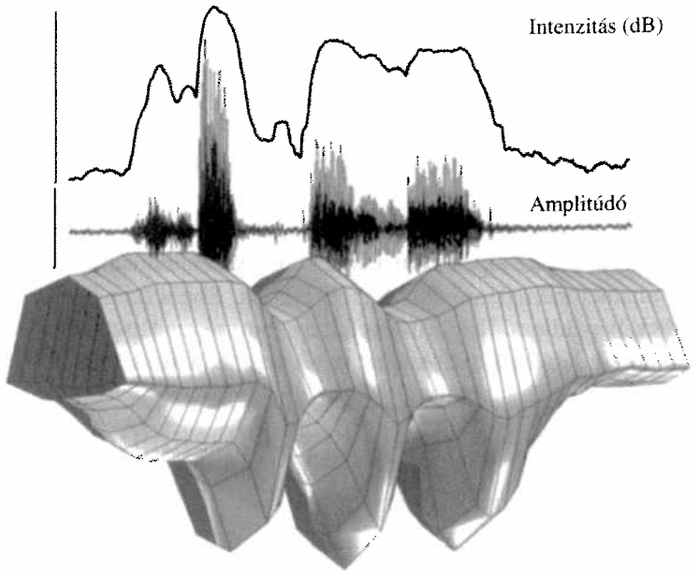
12.17. ábra. Egy hangátmenet jellemzése öt egymás utáni keret alapján

a kedvező. Gyakorlott jeltolmácsok a beszédhangok tiszta fázisú részét világosan és kiemelve képzik. Másodpercenként 5–10 beszédhangot ejtve és 40 ms hosszú elemzési időkereteket tekintve 5 elemzési ablak egyike bizonyosan ráesik a beszédfolyamat legalább egy domináns fonémájára (12.17. ábra). A neurális hálózat bemenetére tehát mindig 5 egymás utáni elemzési ablak kepsztrumvektora kerül.

A neurális hálózat. A visszacsatolt neurális hálózatot a hagyományos hibajel-visszaterjedéses módszerrel tanítottuk (Anguita–Back 1993). A hálózat három rétegben 80 csomópontot tartalmaz. A bemeneti réteg fogadja 80 ponton 5 egymás utáni időkeret 16–16 MFCC értékét. A rejtett réteg 40 csomópontot tartalmaz. A kimenő réteg 6 csomóponton szolgáltatja a 6 főkomponens súlyértékét, amelyekből előállítható a 15 jellemző pont (FP) x-y koordinátaértéke a középső időkeretben.

A tanító-adatbázis 5450 időkeretet tartalmazott. A hálózat tanítását 100 000 ciklusban végeztük. A neurális hálózatmodell a bemeneti és kimeneti változók értékeit a -1, 1 értéktartományba normálta. Az MFCC és PCA változókat mind ebbe a tartományba transzformáltuk lineárisan az MFCC vektorenergia-összetevőjének kivételével. A már betanított neurális hálózat programja igen gyorsan futtatható, mivel az egész adatbázist képviseli a hálózat súlytényező vektor, amely mindössze 3440

elemből áll. A hálózat kimeneti értékeinek valós idejű számolásához tehát egy alkalmas mobiltelefon erőforrásai elegendőek.



12.18. ábra. A 8.1–8.8 jelű jellemző pontok x-y koordinátái az idő függvényében a „szeptember" szó kiejtésekor. A felső folyamatos vonal a keretenkénti energiát ábrázolja dB-skálán, a középső görbe a hullámforma időfüggvénye. Az alsó ábrán látható felület az ajakkontúrokat mutatja.

Főkomponens-analízis. A képfelvétel minden időkeretében 15 jellemző pont írja le a száj és környékének pillanatnyi alakját. A kétdimenziós ábrázolás alapján ez egy 30 dimenziós térben egy ponttal jellemezhető. A rendszer tanítása sokkal hatékonyabbá vált azáltal, hogy a 30 dimenziót 6-dimenziós rendszerré tömörítettük. A dimenzióredukció végrehajtására a főkomponens-analízis módszerét (Principal Component Analysis, PCA) alkalmaztuk. Ez felfogható mozgáskomponensek szerinti felbontásra. Az első 6 PCA vektort választottuk a száj és környékének leírására az alábbi egyenlet szerint

$$w_{1\ldots6} = P^{-1}B\Big|_{p_1{}^1\times\ldots\times p_6{}^1}, \tag{12.6}$$

ahol P jelöli a PCA vektorok (30x30) méretű sajátértékvektorát, B a 30 dimenziós vektorkészlet, c pedig a választott origó, amely a zárt ajkakkal semleges arc súlytényezőinek 0 értékét jelenti. Ez az adattömörítés mindössze 1–3% hibát eredményezett, ami az adott megjelenítő eszközön a jellemző pontok 1–2 pixeles változását eredményezi akár x, akár y koordináta szerint nézve. Ez teljesen elfogadható közelítés. Mivel a hálózat tanításához használt w súlytényező 0 értéke a semleges archoz tartozik, ezért a súlytényező előjele is egy nagyon fontos információt hordoz:

megmutatja, hogy a pont merre mozdul el. A betanított hálózat kimenő értéke egy 6-dimenziós térben jelenik meg. Ebből a jellemző pontok koordinátái a következő egyenlet segítségével határozhatók meg:

$$\overline{B}_k = (\underline{w}_k + \underline{c}) \cdot P. \tag{12.7}$$

Mivel P értékét a tanítás során határozzuk meg, ezért ez a művelet mindössze 180 szorzást igényel keretenként. A főkomponens-analízis ebben az esetben több, mint egy egyszerű mechanikus tömörítő eljárás. A PCA vektorok értékes információt hordoznak a bemondó beszédstílusáról is és a felvétel minőségéről is. A PCA vektorok – bár automatikus eljárás eredményeként adódnak – az egyes vizémák jól azonosítható megkülönböztető jegyeihez kapcsolódnak. Az állkapocs függőlegesen látszó mozgása adja a legerősebb PCA komponenst. A száj vízszintes széthúzása adja a második főkomponens nagy részét (erre a mozgásra kéri fel a fényképész az érintetteket azzal, hogy mondják: „csíííz"). A harmadik főkomponens az ajakkerekítés mértékéhez kapcsolódik. Ezek miatt állítható, hogy a PCA vektorok eredendően kapcsolódnak a vizéma megkülönböztető jegyekhez. Ezen nézőpontból a PCA vektorok dimenziósorrendje rendelkezik kiemelt jelentőséggel. Képzett jeltolmácsoknál az első néhány főkomponens tartalmazza a vizéma megkülönböztető jegyeket. Gyakorlatlan bemondóknál azt tapasztaltuk, hogy a korrektív komponensek sorrendben megelőzik a vizémákat megkülönböztető komponenseket (korrektív komponens például az érzelmet kifejező összetevő).

Beszélő fejmodell. A szabad forráskódú programmal közzétett LUCIA fejmodell némileg módosított változatát használtuk a rendszerben. Ezt más célra, az érzelmeket is kifejező vizuális beszédmodell céljára fejlesztették (Cosi et al. 2003). A LUCIA modell az MPEG-4 szabványra épült. Az eredeti fejmozgató (FAP) paraméterek vizémaalapú rendszert figyelembe véve lettek kialakítva, a szájról olvasás igényrendszerét nem vették tekintetbe a fejlesztésnél. Ezért volt szükség némi módosításra, hogy a modell képes legyen a jellemző pontkoordináták közvetlen fogadására. A közvetlen vezérlés bőrön látható pontok mozgási lehetőségeinek anatómiai alapú megkötöttségeinek finomabb figyelembevételét követelte meg.

Kísérletek. Hasznosnak bizonyultak az előzetes méréseink a rendszer tökéletesítése és az adatbázis kialakítása szempontjából. Ennek során derült ki például, hogy képzett jeltolmácsokat célszerű alkalmazni a rendszer tanításánál. Az előzetes vizsgálatok mutattak rá arra is, hogy a szavak közötti szünetekre is különös figyelmet kell fordítani. Egy küszöbszint alatti háttérzaj nem okoz gondot. Nagyobb háttérzaj óhatatlanul elkezdi mozgatni szavak között is picit a szájat, és ez nagyon megzavarja a pusztán szájról olvasásra épülő beszédfelismerést. Az előzetes vizsgálatok során a siket kísérleti személyektől összegyűlt észrevételeket, javaslatokat gondosan figyelembe vettük a rendszer tökéletesítésénél és a vizsgálati módszerek finomításánál.

Mérési módszerek és eredmények. Pusztán szájról olvasás alapján nem lehet azo-

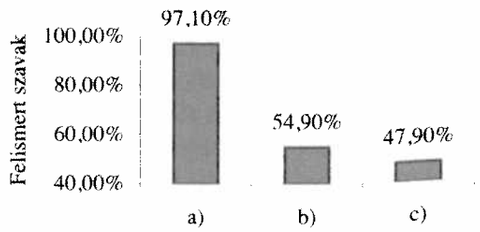
nos képzési helyű és módú fonémapárokat megkülönböztetni (például *baba-papa*). Természetes módon az észlelő személy a szövegösszefüggésre alapozva automatikusan korrigálja vagy kiegészíti a szájról leolvasott információt. Párbeszéd esetén a visszakérdezés tisztázni képes a többértelmű üzenetet. Vizsgálatainkban kizártuk a visszakérdezés lehetőségét, ezért olyan vizsgáló szöveget állítottunk össze, amely lehetőleg kizárja a kétértelműséget. A siketek az előzetes információk alapján mindig erősen leszűkített készletű lehetséges üzenetek közül egy kiválasztására összpontosítanak a szájról olvasott beszéd megértése során. Ezt a természetes mechanizmust célszerű volt követnünk a rendszer vizsgálata során is. Mindig megadtuk, hogy milyen zárt halmazból kell a lehetséges választ várniuk. A mérések során a modell teljes fejét, szájmozgását mutatta a kivetített mozgókép nagy méretű vetítővásznon. Természetesen hang nélkül. Így a töredékes hallással rendelkező vizsgálószemélyek sem hallhattak semmit a beszédjelből. A vizsgálati anyag véletlen rendben az alábbi eseteket tartalmazta:

a) a jeltolmács eredeti képfelvétele (hang nélkül),

b) a fejmodell mozgóképe, ahol a 15 vezérlő paraméter (FP) koordináták értékei jeltolmács képfelvételeiből származtak (hang nélkül),

c) a fejmodell mozgóképe, ahol a 15 vezérlő paraméter (FP) koordinátáit a rendszer a beszédjel paramétereiből számolta ki (a megjelenítés hang nélkül történt itt is).

A siket vizsgálószemélyek válaszaikat írásos formában adták meg egy előkészített űrlapon. A végső eredményeket adó vizsgálat részvevői már több alkalommal részt vettek az előzetes vizsgálatokban, így mindegyikük gyakorlott mérőszemélynek volt tekinthető. A végső vizsgálat 70 szó megértését regisztrálta és mintegy 30 percig tartott. Amikor jelezték, akkor a képfelvételt kérésükre megismételtük. A végső vizsgálatban 18 siket személy vett részt. Az eredmények a 12.19. ábrán láthatók.

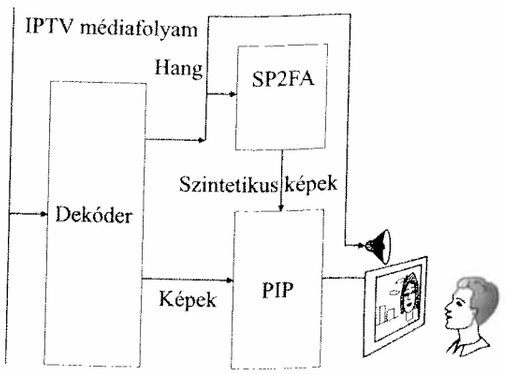


12.19. ábra. A helyesen megértett szavak aránya a) jeltolmács képfelvétele alapján, b) jeltolmács FP koordinátáival vezérelt fejmodell képe alapján, c) beszédjelből számolt FP koordinátákkal vezérelt fejmodell képe alapján

Értékelés. A jeltolmácsok eredeti képfelvételei alapján a szavak szájról olvasása körülbelül 3% felismerési hibát eredményezett. A 15 FP pont koordinátáival vezérelt fejmodell, ha a vezérlő paramétereket közvetlenül a jeltolmács képfelvételein megjelölt pontok koordinátáiból származtattuk, akkor 42% felismerési hibát adott. A méréseket követő megbeszéléseken a vizsgálószemélyek olyan szóbeli megjegyzéseket tettek, hogy hiányzott bizonyos helyzetekben a modellből a nyelv képe és néha a szájtól távolabbi részek mozgása is. Emiatt a fejmodell árnyaltabb vezérlése esetleg megfontolandó. A pusztán hangelemzésből számolt vezérlő paraméterekkel vezérelt fejmodell alapján mért szóérthetőség az előző esethez képest csak 7%-kal csökkent. Ez mutatja a rendszer alapvető eredményét, azaz annak igazolt tényét, hogy a hangjelből számolt vezérlő paraméterekkel jól megközelíthető a képjelből származtatott paraméterekkel vezérelt modell felismerési aránya. Mindez épít a siket személyek kifinomult felismerési képességeire, és kizárólag erre az esetre érvényes az előző megállapítás.

További alkalmazási lehetőség: IPTV speciális hanginformáció siketek számára. A siketek számára létkérdés a televízióműsorokban is a szájmozgás követése az események megértése szempontjából. A leggyakoribb probléma Magyarországon a szinkronizált filmek tömege. A színész szája az eredeti (legtöbbször angol) nyelv rendszere szerint mozog, amiről a magyar siketek nem tudnak leolvasni semmit. A felirat sem tökéletes megoldás, mert nagyon leköti a néző figyelmét, s ha elég részletes, már alig marad figyelem a film élvezetére. További esetek is problémát okoznak a siketeknek: a politikai, a magazin- és a hírműsorokban gyakran betétrészletek láthatók alámondott hanginformációval, amelyek a háttérhang nélkül érthetetlenek. A természetfilmekben, városok, tájak ismertetését tartalmazó műsorokban narrátor mondja az alapvető információt, amelyet a képek, mozgóképek színesítenek, tesznek élvezetessé. Ezek üzenetének lényege nem érheti el a siket vagy erősen nagyothalló nézőket.

Az általunk kidolgozott, beszédjelet szájmozgássá átalakító rendszer alkalmas arra, hogy egy IPTV médiafolyamban (streamben) érkező televízióműsor műsorhangjának felhasználásával egy szintetikusan előállított emberifej-modell száját a beérkező (akár szinkronizált vagy narrátor) hangnak megfelelően mozgassa. A szintetikusan előállított képet hozzáadja az eredeti műsor képtartalmához, és azt együttesen jeleníti meg a felhasználó képernyőjén (kép a képben, PIP). A rendszer blokkjait a 12.20. ábra ismerteti, ahol az SP2FA egység azonos a 12.16a ábrán bemutatott rendszerrel (SPeech to FAcial Animation). A megvalósítás Direct Show keretben volt a legcélszerűbb, amelyet média kezeléséhez fejlesztett a Microsoft. A keretrendszer a médiafeldolgozásban már jól ismert, hálózatba szervezhető alapvető funkciókra épül. Kísérleti eredményeink igazolták, hogy lehetséges beszédjelből közvetlenül szájmozgást leíró jellemzők származtatása olyan pontossággal, ami lehetővé teszi siket személyek számára a beszéd gyakorlati hasznosságú megértését. Erre alapozva segédeszköz készíthető siketek számára, hogy megértsék csak telefonon vett beszédjelből a beszédüzenetet. A rendszer alapelemei olyan számítástechnikai erőforrással megvalósíthatók, amely rendelkezésre áll a mai legfejlettebb mobiltelefonokban. A fejmodell további finomításától reméljük a teljes rendszer olyan fejlődését, amely



12.20. ábra. IPTV speciális hanginformációt siketek számára olvashatóvá konvertáló rendszer alapelemei

révén elérhető a 20% alatti vizuális felismerési hiba, amely szint minden szempontból elfogadható értéket jelent. Emlékeztetünk arra, hogy a mobiltelefonok áldásaiból gyakorlatilag kirekesztett siketek közösségének ez forradalmi előrelépést jelentene jelenleg még fennálló akadályaik leküzdésében.

A fejlesztések során több érdekes új tudományos eredmény is született. A hangi és képi jeltartalom kölcsönös információja alapján mérhető, hogy egy fél fonéma időtartamával is előbbre járhat a száj mozgása, mint a beszédszerveinkkel keltett hang.

12.9. Beszédtanítás és beszédtechnológia

Vicsi Klára

A beszédterápiával, és a beszédkutatással foglalkozó szakembereket már sok évtizeddel ezelőtt foglalkoztatta, hogy miként lehet a technika vívmányait a beszédterápiában felhasználni. A megvalósított eszközök bonyolultak és nehezen használhatók voltak. Ahogy a számítástechnika és a beszédtechnológia fejlődik, a beszédoktató rendszerek megoldási lehetőségei nőnek. Az új kutatásokat a beszédfelismerés, a beszédszintézis, a beszédelemzés, a vizuális megjelenítés legújabb kutatási eredményei támogatják (Vicsi 2004). A beszédhibás vagy hallássérült emberek beszédoktatásán kívül egy egészen új irányzat annak a vizsgálata, hogy az idegennyelv-oktatásban is lehetne-e hasznosítani a számítógépes támogatás adta lehetőségeket (Computer Aided Language Learning, CALL).

Fontos az is, hogy a gyors technikai fejlődés mellett figyelembe vegyük a fonetikai, fonológiai, oktatási szempontokat, a beszédfejlesztés különböző lépéseit, a felhasználók károsodási mértékét vagy az életkor szerinti szellemi képességeket. Sajnos

az utóbbi években kialakított beszédoktató rendszerek legtöbbje csak átveszi változtatás nélkül a legújabb beszédtechnológiai eljárásokat és nem alakítja azokat a speciális alkalmazáshoz, feladathoz. Erre példa a gépi beszédfelismerők széles körű alkalmazása a kiejtés helyességének megítélésére. A helytelen kiejtést azonban ezek a rendszerek nem képesek detektálni, holott éppen az lenne a feladatuk (pongyola, renyhe ejtés, hadarás vagy esetleges hangkihagyás).

A beszédfelismerők használatán alapuló automatikus visszajelzés hatékonyságát illetően a tanárok véleménye sem igazán pozitív (Wallace 1998). Tapasztalataik szerint vagy nem megfelelőek ezek az automatikus ítéletek, vagy pedig nem elég érzékenyek a gépi megoldások az apróbb különbségek észrevételéhez, ami félrevezeti a tanulókat. Ezáltal a felhasználók rosszabb eredményeket érhetnek el, mint az automatikus visszajelzés használata nélkül. Ilyen automatikus visszajelzésen alapuló kiértékeléses eljárással dolgozik az ISTRA és ISLE (Interactive Spoken Language Education) nyelvoktató rendszer. E program is a fonémaalapú rejtett Markov-modelleket alkalmazó beszédfelismerési technológiát használja fel a kiejtés megítélésére. Hasonlóan működnek továbbá a TALK TO ME, TELL ME MORE angol nyelvoktató programok (`http://www.auralog.com/us/schools.html`) (Nouza 1999). E programok használhatóságát segítené, ha valamilyen más, például vizuális visszacsatolást is alkalmaznának. Néhány program hullámforma-megjelenítést ugyan használ, és a fonetikus, akusztikus szakember el is igazodik a hullámformán, de egy gyermek bizonyosan nem.

A kialakított rendszerek egyik csoportjánál magát az artikulációt mutatják be a tanulóknak a beszéd közbeni artikulációs mozgás grafikai megjelenítésével. Közvetlenül a beszédképző szervek pontos beállítására teszik a hangsúlyt. Ez az úgynevezett folyamatorientált megközelítés. Az előállított szintetikus arc artikulál a beszédhanggal szinkronban. Az artikuláció azonos idejű modellezése technológiailag nehéz feladat (Hardcastle et al. 1999, Gibbon–Hardcastle 1998). A paraméterekkel vezérelt vizuális beszédszintézis az arc 3D-s poligonális modelljén alapszik (Massaro 1998b, Cole et al. 1998). Ezek a rendszerek a beszéd vizuális képsorozatának jellemzéséhez, megjelenítéséhez nyomon követik és meghatározzák a beszélő szájmozgását (lásd a 9.12. fejezetet). Egy kedves beszélő fej van beépítve a Massaro és munkatársai (Massaro 1998b) által fejlesztett oktató rendszerbe, melynek neve CSLU Speech Toolkit. Ez egy kutatói segédeszköz, melynek honlapja a következő címen található: `http://cslu.cse.ogi.edu/toolkit`. Ezek az artikulációs mozgást modellező oktató rendszerek arra a vizuális visszacsatolásra építenek, ami a természetes beszédkommunikációban is jelen van. Hallássérült gyermekek esetén problémát jelent, hogy a belső hangképző szervek pozícióját nem lehet látni a képernyőn. A fejlesztők igyekeznek a modellekben a rejtett beszédszerveket is láthatóvá tenni, de egyelőre ezek a rajzok még elég riasztóak.

A beszédoktató rendszerek másik csoportjánál különböző akusztikai paramétereket jelenítenek meg. Tehát nem az artikulációs szervek beállítását hangsúlyozzák,