Beszédjel átalakítása mozgó száj képévé siketek kommunikációjának segítésére

TAKÁCS GYÖRGY, TIHANYI ATTILA, BÁRDI TAMÁS, FELDHOFFER GERGELY, SRANCSIK BÁLINT

Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar
{takacs.gyorgy, tihanyia, bardi, flugi, sraba}@itk.ppke.hu

Kulcsszavak: audiovizuális beszédfeldolgozás, fej animáció, multimodális kommunikációs, szájrólolvasás

Siketek kommunikációs segédeszközeként egy beszédjelet közvetlenül szájmozgás-képpé átalakító rendszert fejlesztettünk. Az előzetes vizsgálati eredmények alapján, képzett jeltolmácsokkal készítettünk kép- és hangfelvételeket. Az MPEG-4 szabványnak megfelelő egységet használtunk fejmodellnek a beszédszervek mozgásának megjelenítésére. Egy neurális hálózat számolja ki a jellegzetes pontok főkomponens súlytényező értékeit a beszédjelből. A fejmodell vezérlő paramétereit a rendszer a főkomponens súlyértékekből származtatja. A rendszer terveink szerint egy alkalmas mobiltelefonon is futtatható. A tesztvizsgálat során siket személyek a szavak közel 50%-át értették meg helyesen a fejmodell által megjelenített mozgókép alapján.

1. Bevezetés

Siket emberekben hosszú gyakorlás után fantasztikus szintre fejlődik ki a beszéd megértése pusztán a szájmozgást nézve. Az volt a tervünk, hogy erre alapozott kommunikációs segédeszközt készítünk siket felhasználók számára, amely pusztán a szájrólolvasáson alapul és egy alkalmas mobiltelefon készüléken megvalósítható. Az általunk kifejlesztett rendszerben egy beszélő fej fontos részeit jelenítjük meg a színes grafikus kijelzőn. A mozgó fej vezérlő paramétereit közvetlenül a beszédjelből származtatott jellemzők alapján számoljuk ki. Tisztában vagyunk azzal, hogy az emberi beszédfolyamatnak ez csak egy részleges megjelenítése és azzal is, hogy elvéből fakadóan is hordoz hibákat. Arra számítunk, hogy korlátai ellenére a siketek hasznos kommunikációs segédeszközhöz juthatnak rendszerünkkel és természetes módon akár telefonon keresztül is szót érthetnek a hallók többségével. Reményeink szerint ezzel is lebontható egy akadály, ráadásul mindössze olyan hétköznapi eszközzel, mint egy megfelelő kategóriájú mobiltelefon. Természetesen rendszerünk nagyban épít a siketek kifinomult 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 nyelvi elemek (fonémák, vizémák) halmazára. Egy második lépésben pedig a diszkrét elemek halmazát alakították át mozgó fejjé. Nagy előnye a mi közvetlen rendszerünknek, hogy eredendően megőrzi a beszédfolyamat eredeti időbeli és energia-szerkezetét. Ezáltal a természetes beszédritmus eleve megőrződik. További előnye, hogy egy mobiltelefon korlátozott processzorteljesítménye, memóriakapacitása mellett is megvalósítható, és még ígéretesebb jellemzője, hogy elvileg nyelvfüggetlen.

Új ötlet a rendszerben, hogy a folyamatot nem átlagos beszélők jeleivel tanítottuk, hanem olyan hang és kép adatbázissal, amelyet képzett jeltolmácsok felvételeiből állítottunk össze. Az ő artikulációs stílusuk és dinamikájuk alkalmazkodott a siketek szájrólolvasási igényeihez.

Az irodalomban ismert szájrólolvasáshoz kapcsolódó mozgó fej alkalmazások érdekes csoportja foglalkozik azzal, hogy többletinformációt adjon a hallott beszédhez például zajos környezetben vagy nagyothallók számára [1,2,3]. A hallott és egyben látott beszédfolyamatban a szuperadditív megértés nagyobb, mint a külön modalitásban megértett elemek összege. Fontos kérdés, hogy hol és hogyan összegződnek az egyes modalitásból származó információ elemek. A mi alkalmazásunkban azonban csak a látás alapú beszédérzékelésre összpontosítottunk, mivel a célközösségben a hallás gyakorlatilag teljesen hiányzik.

A szájmozgás dinamikája és természetessége tűnik az alkalmazás kritikus elemének. Számos közlemény számol be arról, hogy milyen bonyolult eljárásokkal érik el a beszélő fej modell megfelelő dinamikáját és természetességét [4,5,6].

Mi ezt pusztán azzal kívántuk elérni, hogy különös figyelemmel választottuk ki az adatbázisba bevont beszélő személyeket. Ezek tehát nem az átlagos népességet, hanem a siketek számára legjobban érthető beszélőket reprezentálják. Mi ezzel a trükkel oldottuk meg a nagyobb szájmozgás dinamikát igénylő követelményeket.

2. Adatbázistervezés és összehasonlítás

2.1. Előzetes szájrólolvasási mérések

A kezdeti vizsgálatokban mértük a siketek szájrólolvasási képességeit, feltérképeztük mindennapi kommunikációs problémáik lényegét. A részleteket korábbi cikkünkben ismertettük [13], itt most csak a végkövetkeztetéseket foglaljuk össze.

Legfontosabb végkövetkeztetéseink egyike volt, hogy a szájról olvasott beszéd érthetősége nagyon függ az artikuláció minőségétől. A szájrólolvasás sokkal nagyobb figyelmet igényel, mint a beszéd megértése hallás útján. A teljes beszédfolyamatról csak részleges információt ad, ezért a tévesztések eleve gyakoribbak. Az olyan artikuláció, amely eleve kiemeli a megkülönböztető jegyeket, valamint a lassú beszédtempó nagyon sokat segít a helyes megértésben. A hallók között messze legjobban teljesítik ezeket a követelményeket a képzett jeltolmácsok. Ãk napi kapcsolatban állnak a siketekkel és ezért eleve alkalmazkodik artikulációjuk a szájrólolvasás igényeihez. Ezért határoztuk el, hogy tanító adatbázisunkat jeltolmácsok kép- és hangfelvételeiből állítjuk össze, még akkor is, ha a végső használatkor bárkinek a hangja szolgálhat jelbemenetként.

Megtanultuk az előzetes kísérletek során azt is, hogy a siketeknek komoly nehézségeik vannak a természetes nyelv komplikált nyelvtani szabályaival. Elektronikus leveleiket, SMS üzeneteiket is elemezve látszik, hogy ugyanez megnyilvánul írott kommunikációjukban is. Amikor az érthetőséget teljes mondatok, rövid közlendők formájában adott nyelvi egységekkel próbáltuk mérni, akkor tapasztaltuk, hogy nem képesek a teljes üzenet pontos, szó szerinti visszaadására, hanem csak a legfontosabb üzenetelemek maradnak meg emlékezetükben. Sokszor csak az a kulcselem, amelyre az előzetes információk alapján a figyelmük középpontjába kerül. A ragokkal, toldalékokkal sem nagyon foglalkoznak. Konkrét nevek, személyes névmások fontosabbak számukra.

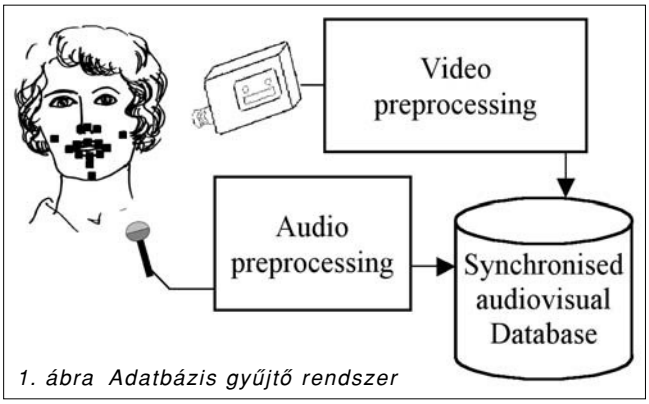
Egy hirtelen témaváltás is igen nehezen követhető számukra. Ennélfogva az érthetőségvizsgálatok szokásos szövegei és módszerei eleve nem használhatók esetükben. A tényleges érthetőség mérése érdekében ezek szövegösszefüggéstől lehetőleg mentes szöveget használnak, ami teljesen idegen a siketek kommunikációs stratégiájától. Emiatt speciális szövegű adatbázist alakítottunk ki mind a tanító, mind a tesztelő anyaghoz.

Az előzetes vizsgálatok fontos kérdése volt, hogy két vagy három dimenziós fejmodellt kell-e megvalósítani, és hogy mennyire fontos a harmadik (mélység) dimenzió a szájrólolvasás során. Ennek eldöntésére természetes beszélő személyek videofelvételeit mutattuk siketeknek olyan torzítások után, amelyek a mélységinformációt csökkentették. Az egyik esetben csak a kék színösszetevőt tartottuk meg és a piros és zöld színösszetevőket kivettük a képből. További felvételeknél pedig csak fehér vagy fekete képpontok maradtak az eredeti képből egy alkalmas küszöbszintet választva. Meglepő módon ezek a torzítások alig csökkentették a szájrólolvasás pontosságát, pedig a mélységinformációt kiölték a felvételekből.

További kísérleteinkben arra kerestünk választ, hogy a jobb mobiltelefonoknál szokásos képernyő méret és felbontás elegendő-e a szájrólolvasáshoz. Amennyiben a kijelzőn megjelenő kép a szájat és környékét mutatja (a beszédinformációt hordozó legfontosabb részeket), akkor ez a méret és felbontás eredeti videofelvételek esetén elegendő a gyakorlatilag teljes megértéshez.

2.2. Felvételek

Az adatbázis nem más, mint különböző bemondók összerendezett hangfelvételeinek és képfelvételeinek rendszere. A felvételek jelét azonos időkeretekben összeszinkronizálva dolgoztuk fel *(1. ábra)*. A bemondók fejét puha korlátokkal rögzítettük, hogy a fej ingatását megakadályozzuk. Az egyes pontokat abszolút koordinátáikkal jellemezhetők.



1. ábra Adatbázis gyűjtő rendszer

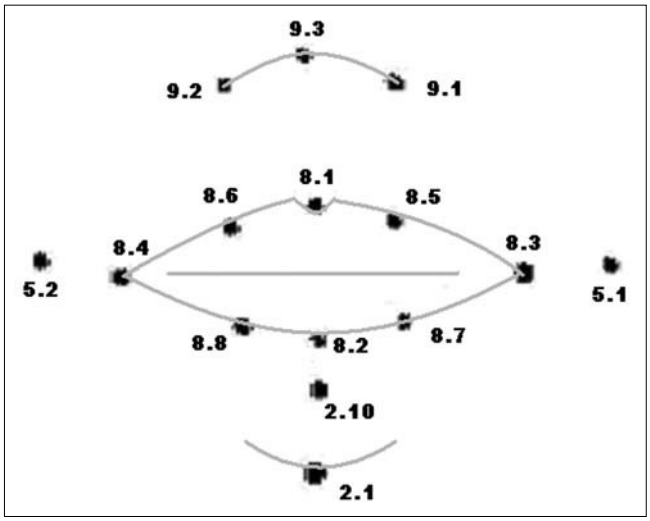
Jelen állapotában a rendszer bemondófüggő, de már dolgozunk a személytől független megoldáson is.

Az MPEG-4 szabvány az emberi arcot 86 jellemző ponttal (Feature Point, FP) írja le. Előzetes kísérleteink alapján ezekből 15-öt választottunk ki a szájnak és környezetének leírására. A felvételek során ezeket a pontokat könnyen lemosható és egészségre nem ártalmas festékkel jelöltük meg a bemondók arcán. A beszédfolyamat képének leírása az MPEG-4 szabvány szerinti jellemző pontokkal több szempontból is előnyös. Egyrészt a száj és arc mozgásának tömör és elég pontos leírására alkalmasak az FP koordináták, másrészt a bevált szabványos fejmodellek alkalmazhatók ezekkel a pontokkal vezérelve, így az igen összetett modellek alapvető fejlesztésére nem kellett erőinket pazarolni. Amint az előző pontban kifejtettük csak képzett jeltolmácsokkal készítettünk felvételeket.

A felvételekhez egyszerű kamerákat használtunk: 720x576 pontos felbontással, másodpercenként 25 képpel, PAL szabvány szerint. Ez azt jelenti, hogy 40 ms hosszú időablakokban készülhettek az összeszinkronizált kép- és hangelemzések. A felvételek a szájat és környékét rögzítették annak érdekében, hogy a kiválasztott jellemző pontok helyzete minél kisebb hibával meghatározható legyen. A fej többi részét (bár a szem környéke, vagy akár a hozzáfűzött tekintet is hordoz tartalmi információt) nem vontuk bele vizsgálatainkba. A képfelvételeket ezután emberi beavatkozás nélkül dolgoztuk fel. A képjelet a kontraszt, a fényesség és telítettség tekintetében úgy torzítottuk, hogy a jellemző sárga pontok minél jobban kiemelődjenek. A sárga pontokat végül az RGB komponensek kompará-

lásával detektáltuk. A binarizált képen először dilatációs műveleteket végeztünk, hogy biztosan összefüggő képpont halmazt nyerjünk, majd lépésenként kívülről eróziós folyamattal szedtünk le képpontokat, míg 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) pontok között középen legyen, mivel ezek a pontok mozognak a 15 közül legkevésbé *(2. ábra)*.

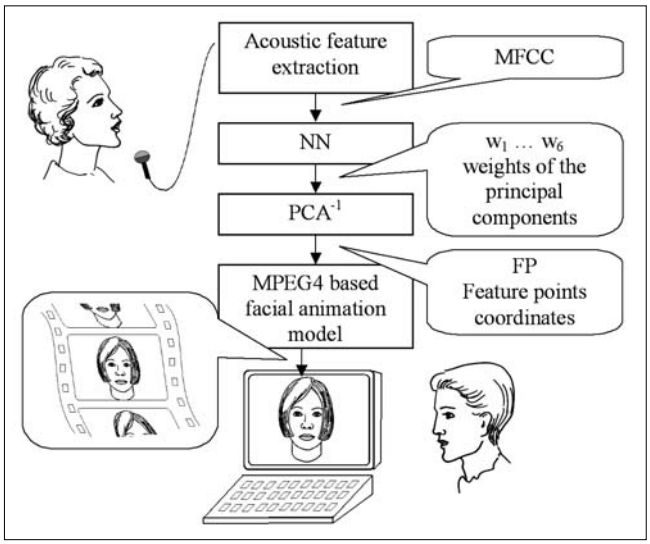


2. ábra Az MPEG-4 jellemző pontok kiválasztott részhalmaza a száj körül

A beszédjelet egy hangcsatornában rögzítettük 48 kHz mintavételezéssel, 16 bites mintákkal.

A tanító és tesztelő adatbázis szövegét a 2.1 pontban leírt követelmények szerint választottuk ki. Eszerint a felvételek kétjegyű számokat, hónapok neveit, a hét napjait tartalmazták.

3. ábra A beszéd-szájmozgás átalakító rendszer elemei



3. A beszédjel átalakítása szájmozgás-képpé

A fejlesztés állapotában a rendszer lényegében egy személyi számítógépen futó programrendszer. A *3. ábra* szerint itt az alapelemek feladatait és kapcsolódását tekintjük át. Az egyes elemek részleteit a 3.1-3.4 pontok fejtik ki

A mintavételezett beszédjelen minden 40 ms keretben meghatároztuk a mel skála szerinti kepsztrum együttható 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 $[w_1,...w_6]$. A főkomponens elemzés (PCA) inverz műveletével nyerjük a fejmodell vezérléséhez ténylegesen szükséges FP koordináta értékeket. Ez egy lineáris kombinációs műveletet jelent csupán. Az FP koordinátákat meghatározzuk minden időkeretre. Erre láthatunk egy példát az *5. ábrán*.

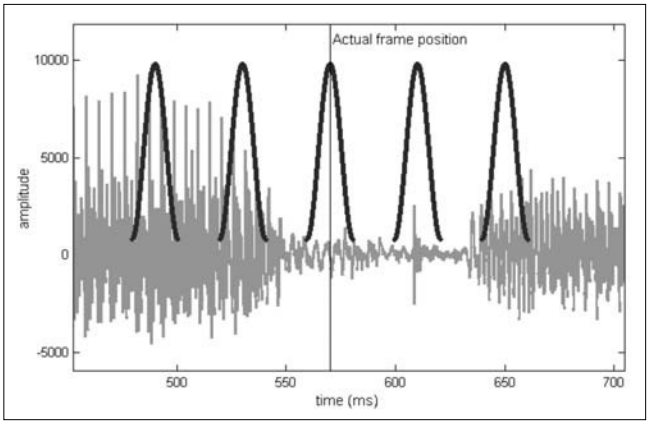
Rendszerünk utolsó eleme a nyílt forráskódú LUCIA beszélő fej rendszernek egy enyhén módosított változata. A modellt az FP koordinátákkal vezéreljük és a mozgó kép megjelenik a kijelzőn. A részletek a 3.4 fejezetben találhatók.

3.1. Akusztikai lényegkiemelés

A bejövő beszédjelen először egy magasemelő szűrési műveletet hajtunk végre $H(z) = 1\text{-}0.983z^{-1}$ karakterisztikával. Ezután 21.33 ms hosszúságú Hamming-ablakkal súlyozzuk a jelet. Az ablakban lévő jelből 16 elemű mel-frekvenciás kepsztrum együttható vektort számolunk.

A koartikuláció jelenségének a beszédfolyamat képi ábrázolásánál legalább akkora jelentősége van, mint a hangjelek feldolgozásakor. A beszédszervek képe 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.

4. ábra Egy fonémaátmenet jellemzése öt egymás utáni keret alapján



A siket partnerek számára a lassabb beszédtempó vezet eredményre. Gyakorlott jeltolmácsok a beszédhangok tisztafázisú 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 *(4. ábra)*.

A neurális hálózat bemenetére tehát mindig 5 egymás utáni elemzési ablak kepsztrum vektora kerül.

3.2. A neurális hálózat

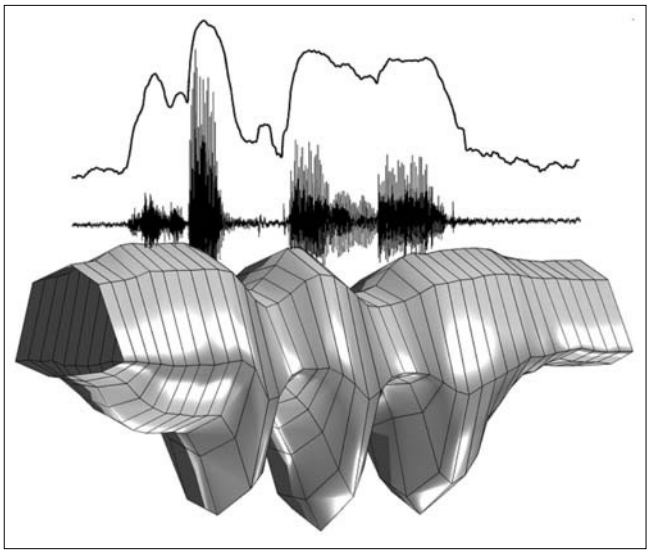
A visszacsatolt neurális hálózatot a hagyományos hibajel visszaterjedéses módszerrel tanítottuk, azzal a programmal, amelyet David Anguita fejlesztett ki és tett közzé [8].

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 koordináta értéke a középső időkeretben.

A tanító adatbázis 5450 időkeretet tartalmazott. A hálózat tanítását 100.000 ciklusban végeztük. A neurális hálózat modell a bemeneti és kimeneti változók értékeit a -1, 1 értéktartományba normálta. Az MFCC és PCA változókat mind ebbe a tartományba transzformáltuk lineárisan az MFCC vektor energia 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.

5. 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 hullámformát mutatja. Az alsó ábrán látható felület az ajakkontúrokat mutatja.



3.3. Főkomponens analízis (PCA)

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 dimenzió redukció végrehajtására a főkomponens elemzés módszerét (Principal Component Analysis, PCA) alkalmaztuk. Ez felfogható mozgáskomponensek szerinti felbontásra, amint ezt a *6. ábra* mutatja. Az első 6 PCA vektort választottuk a száj és környékének leírására az alábbi egyenlet szerint

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

ahol *P* jelöli a PCA vektorok (30x30) méretű sajátérték vektorát, *B* a 30 dimenziós vektor készlet, *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.

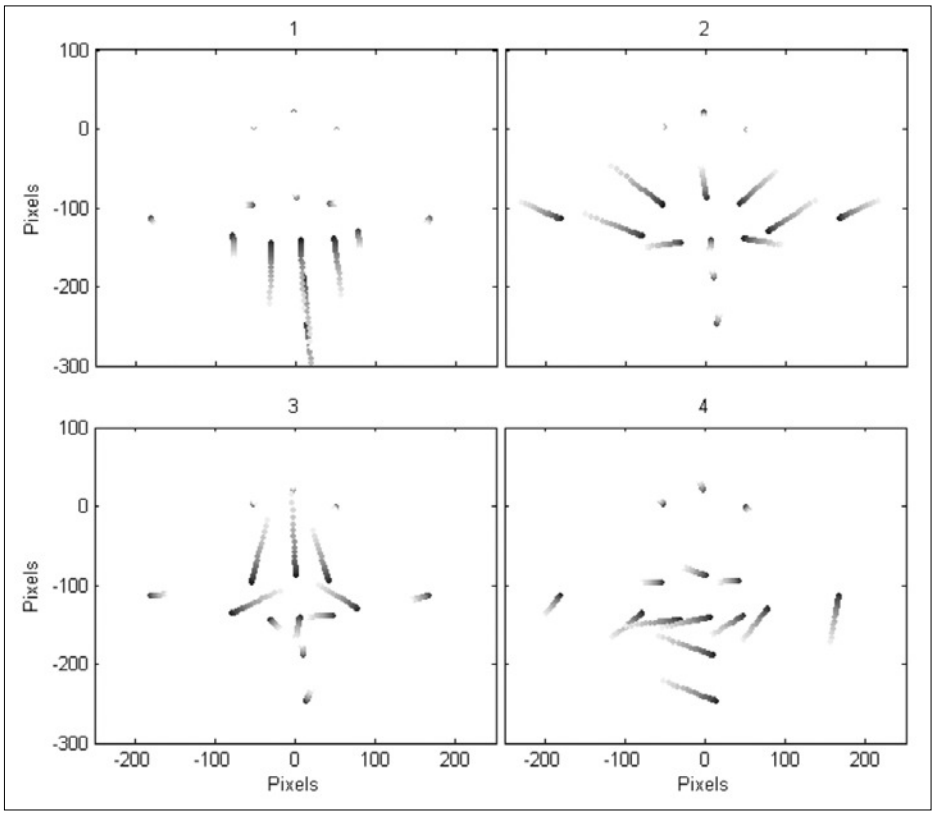
A betanított hálózat kimenő értéke egy 6-dimenziós térben jelenik meg. Ebből a jellemző pontok koordinátái a következő egyenlet segítségével határozhatók meg.

$$\overline{B}_k = \left(\underline{w}_k + \underline{c}\right)\cdot P \tag{2}$$

Mivel *P* értékét a tanítás során határozzuk meg, ezért ez a művelet mindössze 180 szorzást igényel keretenként.

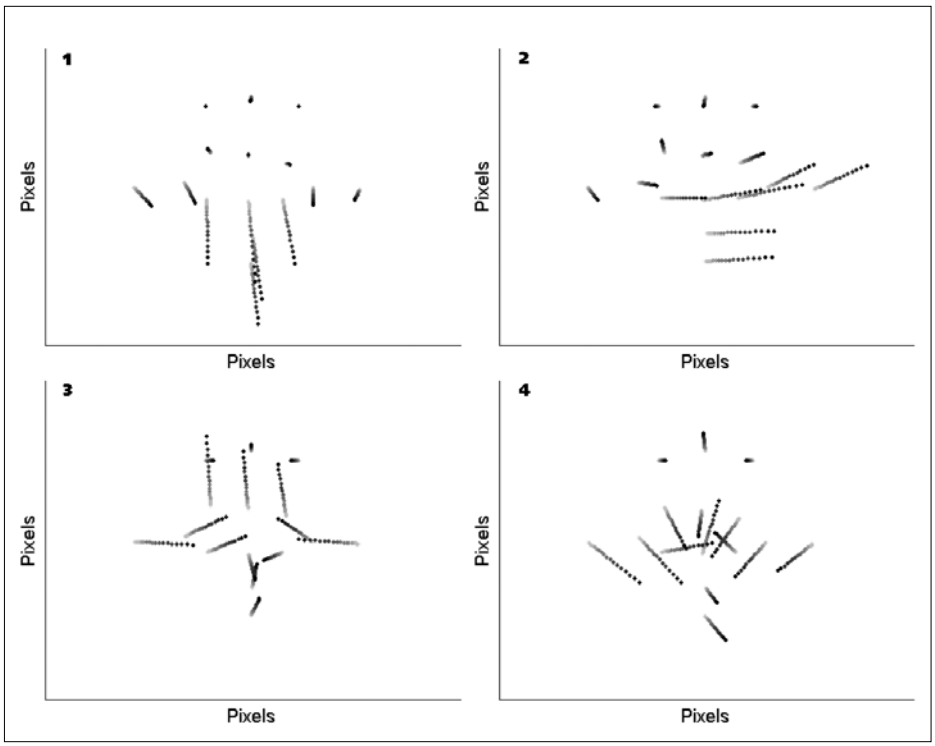
A főkomponens analízis ebben az esetben több, mint egy egyszerű mechanikus tömörítő eljárás. A PCA vektorok értékes információt hordoznak a bemondó beszédstílusáról is és a felvétel minőségéről is. A PCA vektorok – bár automatikus eljárás eredményeként adódnak – az egyes vizémák jól azonosítható megkülönböztető jegyeihez kapcsolódnak. Szépen kiolvasható ez a 6. ábrán is. 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”). Jól megfigyelhető, hogy 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 dimenzió sorrendje 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 megkü-



6. ábra
A jellemző pontok helyzete az első, második, harmadik és negyedik főkomponens szerint kifejezve képzett jeltolmács beszédje alapján.

7. ábra
A jellemző pontok x-y koordinátái az első, második, harmadik és negyedik főkomponens értékével kifejezve gyakorlatlan bemondó felvételei alapján.



lönböztető komponenseket (korrektív komponens például az érzelmet kifejező összetevő). Ezt mutatja be a *7. ábra*, ahol a második főkomponens fejezi ki, hogy a bemondó nagyon jellemzően, ferdén mozgatja a száját. Ezért nem is érdemes felhasználni felvételét a hálózat tanítására.

3.4. 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 és munkatársai [10].

A LUCIA modell az MPEG-4 szabványra épült. Az eredeti fejmozgató (FAP) paraméterek vizéma alapú rendszert figyelembe véve lettek kialakítva, a szájrólolvasás igényrendszerét nem vették tekintetbe a fejlesztésnél.

Ezért volt szükség némi módosításra, hogy a modell képes legyen a jellemző pont koordináták közvetlen fogadására. A közvetlen vezérlés bőrön látható pontok mozgási lehetőségeinek anatómiai alapú megkötöttségeinek finomabb figyelembe vételét követelte meg. Ennek részleteiről [8] ad tájékoztatást.

4. Kísérletek és eredmények

4.1. Előzetes vizsgálatok

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.

4.2. Mérési módszerek és eredmények

Pusztán szájrólolvasás alapján nem lehet azonos képzési helyű és módú fonéma párokat megkülönböztetni (például baba-papa). Természetes módon az észlelő személy a szövegösszefüggésre alapozva automatikusan korrigálja vagy kiegészíti a szájról leolvasott információt. Párbeszéd esetén egy visszakérdezés tisztázni képes a többértelmű üzenetet. Vizsgálatainkban kirekesztettük 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 mérése során is. Mindig megadtuk, hogy milyen zárt halmazból kell a lehetséges választ várniuk.

A vizsgálószövegben ezért kétjegyű számok, hónapok nevei, a hét napjainak nevei szerepeltek előre megadott kategória szerint.

A mérések során a modell teljes fejét, szájmozgását mutatta a kivetített mozgókép nagyméretű 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).

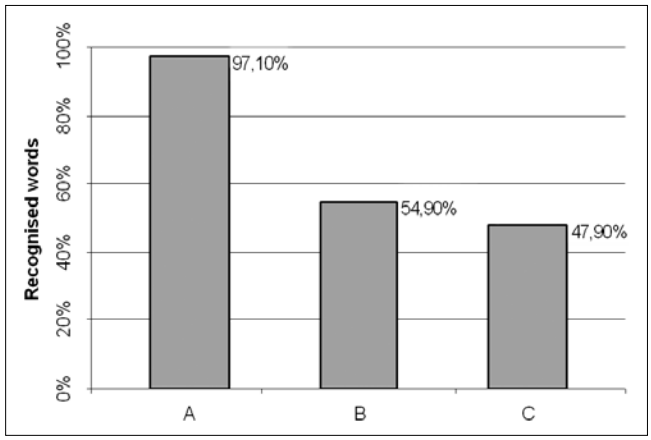
A siket vizsgálószemélyek válaszaikat írásos formában adták meg egy előkészített űrlapon. A végső eredményeket adó vizsgálat részvevői már több alkalommal rész vettek az előzetes vizsgálatokban, így mindegyikük gyakorlott mérőszemélynek volt tekinthető. A végső vizsgálat 70 szó megértését regisztrálta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

Az eredmények a *8. ábrán* láthatók.

4.3. Értékelés

A jeltolmácsok eredeti képfelvételei alapján a szavak szájrólolvasása 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.



8. 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*

A pusztán hangelemzésből számolt vezérlő paraméterekkel vezérelt fejmodell alapján mért szó érthetőség az előző esethez képest csak 7%-kal csökkent. Ez mutatja rendszerünk 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.

5. Összefoglalás

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

Köszönetnyilvánítás

A szerzők ezúton is kifejezik köszönetüket a Nemzeti Kutatási és Technológiai Hivatalnak a 472/04 szerződés keretében nyújtott támogatásáért.

A közös munka során igaz barátainkká vált siketek és jeltolmácsok lelkes közössége ösztönző példaként áll előttünk további kutatásaink során. Ezért nem csak áldozataikat köszönjük, hanem további segítségüket is kérjük.

Köszönjük Harczos Tamás kollégánk értékes ötleteit és munkáját is.

Irodalom

[1] D. W. Massaro,
Perceiving Talking Faces: From Speech Perception to a Behavioral Principle Cambridge,
Mass: MIT Press, 1998.

[2] D. W. Massaro, D. G. Stork,
"Speech Recognition and Sensory Integration,"
American Scientist, 86. 1998.

[3] M. Johansson, M. Blomberg, K. Elenius,
L. E. Hoffsten, A. Torberger,
"Phoneme recognition for the hearing impaired,"
TMH-QPSR, 2002., Vol. 44 – Fonetik, pp.109–112.

[4] K. H. Choi, J. N. Hwang,
"Constrained optim. for Audio-to Visual Conversion,"
IEEE Transactions on Signal Processing,
Vol. 52, No.6, June 2004, pp.1783–1790.

[5] R. Gutierrez-Osuna, P. K. Kakumanu, A, Esposito,
O. N. Garcia, A. Bojorquez, J. L Castillo, I. Rudomin,
„Speech-driven Facial Animation with Realistic Dynamics" IEEE Transactions on Multimedia,
Vol. 7, February 2005, pp.33–42.

[6] J. Beskow,
Talking Heads, Models and Applications for Multimodal Speech Synthesis (Doctoral Dissertation),
Stockholm, 2003.

[7] J. Ostermann,
"Animation of Synthetic Faces in MPEG-4",
Computer Animation, Philadelphia, Pennsylvania,
June 8-10, 1998., pp.49–51.

[8] Davide Anguita,
Matrix Back Propagation – An efcient implementation of the BP algorithm", Technical Report,
DIBE – University of Genova, November 1993.

[9] B. Granström, I. Karlsson, K-E Spens,
"SYNFACE – a project presentation",
Proc. of Fonetik 2002, TMH-QPSR 44: pp.93–96.

[10] Cosi P., Fusaro A., Tisato G.,
"LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model", Proceedings of Eurospeech 2003,
Geneva, Switzerland, September 1, 2003.,
Vol. III, pp.2269–2272.

[12] G. Salvi:
"Truncation error and dynamics in very low latency phonetic recognition", Proc. of ISCA workshop on Non-linear Speech Processing (2003).

[13] Bárdi Tamás, Feldhoffer Gergely, Harczos Tamás,
Srancsik Bálint, Szabó Gábor Dániel:
"Audiovizuális beszéd adatbázis és alkalmazásai",
Híradástechnika, Vol. LX, 2005/10, pp.24–28.

[14] Feldhoffer Gergely, Bárdi Tamás,
Jung Gergely, Hegedűs Iván Mihály:
"Mobiltelefon alkalmazások siket felhasználóknak",
Híradástechnika, Vol. LX, 2005/10, pp.29–32.