Látható beszéd: beszédhang alapú fejmodell animáció siketeknek

Feldhoffer Gergely[1], Bárdi Tamás[1]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar, 1083 Budapest, Práter utca 50/a, Magyarország
{flugi, bardi}@itk.ppke.hu

Kivonat: Elkészült egy fej-animációs rendszer, ami siketek számára beszédjelből olyan szájmozgást állít elő, hogy a siket felhasználó azt megérthesse. Egy olyan audiovizuális adatbázis készült el hozzá, amihez professzionális jeltolmácsok közreműködését kértük. Az animációs részhez szabványos MPEG-4 fejmodellt használtunk. Sikerült olyan reprezentációt találni főkomponens analízis segítségével, aminél egy neuronháló képes volt az akusztikus jellegvektorokból kiszámítani az animációs paramétereket. A rendszer fontosabb elemei mobiltelefonra is elkészültek. Siket felhasználókkal végzett teszt 50% körüli felismerési pontosságot mutatott ki a képi adatokból és a hangból számolt animációra is.

1 Bevezetés

Ennek a cikknek a témája egy olyan rendszer, ami siketeknek próbál segítséget nyújtani, hogy a kommunikációs lehetőségek egy újabb irányát használni tudják. A legtöbb siket ember gyermekkorában a halló emberekkel való kommunikálásnak legalapvetőbb módjaként a szájról olvasást tanulja meg. Emellett jeltolmácsok, illetve az írásbeli lehetőségek azok, amiket használni tudnak, hogy a halló emberek üzeneteit megértsék. A másik irány, a siket ember üzenete a halló emberekhez, szintén tanulható. A siket fiatalok megtanulnak bizonyos szintig beszélni, ami ugyan árulkodik a képességbeli elmaradásról, de kis tanulással illetve megszokással érthető. Ez azt jelenti, hogy a telefonos kommunikációnak csak az egyik iránya hiányzik, a halló ember üzenetét nem tudja fogadni a siket, de a siket tud beszélni. Ha tehát a beszédjelből olyan szájmozgás animációt tudunk előállítani, amit a siket meg tud érteni, akkor nincs akadálya a kétirányú kommunikációnak.

A szakirodalom szerint a szájmozgás hangból történő teljes pontosságú visszaállítása lehetetlen feladat, mert többféle hanghoz tartozhat ugyanaz a mozgás, illetve többféle mozgáshoz tartozhat ugyanazon hang. Amit ebben a cikkben bemutatunk, az nem mond ellent ennek a tézisnek. Ennek az az oka, hogy nem az a cél, hogy az eredeti szájmozgást számoljuk ki a beszédjelből, elegendő, ha az egyik olyan animációt sikerül előállítani, ami az adott beszédhang-sorozatra olyan mértékben jellemző, hogy a siket ember képes megérteni.

A rendszerünk egyik legfontosabb tulajdonsága, hogy elkerültük a diszkrét osztályozást, többek között a fonémákra vagy vizémákra való bontást, ezzel elvi akadálya nincs a nyelvfüggetlenségnek, az kizárólag az adatbázis összeállításán múlik. Létezik vizéma alapú rendszer [5][2], ami egy beszédfelismerő és egy arc szintetizátor összekötése. Ez a rendszer siketek számára nem elegendően élethű. A mi megoldásunk egyik fontos előnye, hogy megtartja a természetes ritmust, és nem vét hibás fonéma meghatározásból eredő hibát.

Szintén fontos tulajdonsága a rendszernek, hogy programozható mobiltelefonon megvalósítható módszerekre szorítkoztunk. Ennek az oka az, hogy a siket felhasználó nem szívesen használ olyan eszközt, ami felhívja a figyelmet a fogyatékosságára, a mobiltelefon viszont társadalmilag nem megbélyegző.

2 Adatbázis

2.1 Előzetes felmérések

A munka előtt felmértük, hogy milyen típusú kommunikációs formát tudnának a siketek megérteni és elfogadni. Ennek a felmérésnek már az elején világossá vált, hogy nem szívesen használnának olyan vizualizációt, amit tanulniuk kell, így a beszélő száj animációjában találtuk meg azt a formát, ami megoldható és használható is. Természetesen kényelmesebb lenne a siket embernek a szöveggé alakítás, mert szájról olvasni kimerítő tevékenység, de folyamatos beszéd felismerése telefonon keresztül máig egy igen nehéz, nagy adatbázisokat és számítási kapacitást igénylő feladat, ami programozható telefonon a mai teljesítmény mellett nem reális cél.

A feladat rögzítése után felmértük, hogy milyen paraméterek befolyásolják a beszélő emberről készült felvételek érthetőségét. A felmérések azt igazolták, hogy egy mobiltelefon méretű kijelzőn, aminek a felbontása 320x208 pixel, érthető egy olyan felvétel, amin csak a száj környéke látható. Ugyanakkora felismerési arányt tapasztaltunk nagyobb felbontású, vagy nagyobb kijelzőkön, és időbeli felbontásban elegendőnek bizonyult 12 kép/másodperc is. Ez azt jelenti, hogy a mobiltelefonon megvalósított rendszer érthetőségét a készülék technikai paraméterei nem korlátozzák.

Ennek a felmérésnek a legfontosabb tanulsága az volt, hogy az érthetőség minden technikai paraméternél jobban függ attól, hogy a beszélő mennyire képes igazodni a siketek igényeihez. Azt a felvételt, amin a beszélő már sokat foglalkozott siket emberekkel, a legrosszabb technikai körülmények között is majdnem pontosan értették, míg a gyakorlatlan beszélőket a felszereléstől függetlenül kevéssé.

Egy másik felmérésben arra voltunk kíváncsiak, hogy a felismerésben számít-e az, hogy a felvételen látható arc árnyalataiból az eredeti háromdimenziós fej rekonstruálható. Az erre a kérdésre adott válasz dönti azt el, hogy szükség van-e a száj mozgásának háromdimenziós rögzítésére, vagy elegendő kétdimenziós, pontkövető vagy kontúrkereső algoritmusok használata. A felvételeket torzítottuk, hogy eltűnjenek az árnyalatok, bizonyos felvételeken csak a szájkontúr volt kivehető. A felismerendő szavak véletlen kétjegyű számok voltak. A siketek a torzított videókat ugyanolyan pontosan megértették, mint az eredetieket. Ezért használtunk a későbbiekben olyan

rögzítési eljárást, ami egyetlen kamerának a képéből dolgozva nyeri ki a száj állapotát.

A teszteléshez fontos a siketek kontextus követésének elemzése. Írásbeli és írásos-mozgóképes vegyes teszteket végeztünk, ahol meggyőződtünk arról, hogy annak ellenére, hogy írásban sok nyelvtani hibával kommunikálnak, a kontextust jól megértik, történetek szereplőinek cselekményeit akkor is követni tudják, ha az nyelvtanilag csak a toldalékokból derül ki, holott a toldalékokat írásban nem használják. Kiderült az is, hogy a siketek szókincse szűk, például, amikor egy állatról szólt a szövegkörnyezet, akkor a „gebe” szót azért nem ismerték fel, mert nem tartozik a szókincsükhöz. A feladat szempontjából ez nem okoz problémát.

2.2 MPEG-4

A világban sokféle arc-animációs rendszer terjedt el, többségükben az MPEG-4 szabvány szerint paraméterezhetőek. Ez a szabvány rögzíti az arc főbb pontjait (feature point, továbbiakban tartópont), azokat képes normálni, és így arc állapotokat lehet különböző alakú fejre illeszteni. Léteznek MPEG-4 szabvány szerint működő szövegfelolvasó rendszerek, amik egy beszédszintetizátor és egy fejmodell párhuzamos meghajtásával működnek. Munkánkhoz a Cosi és társai [3] által létrehozott Lucia fejmodellt használtuk.

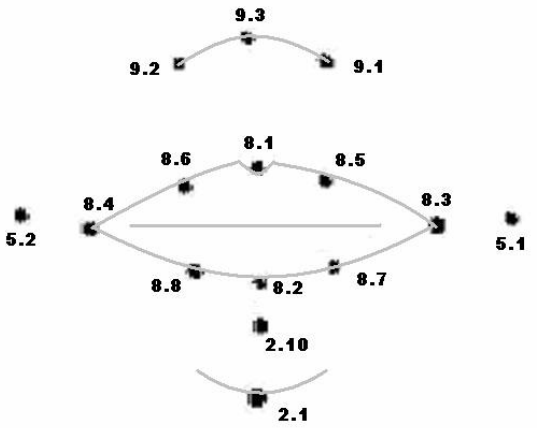

Fig. 1. Az MPEG-4 általunk is használt tartópontjai.

A tartópontokon kívűl létezik az MPEG-4-ben magas szintű animációs leírás is, ami vizémákra alapul, ami a fonémákhoz tartozó száj-állapotokat jelenti. Ezeket nem használtuk, mert a rendszerünk nem dolgozik fonémaszinten.

2.3 Az adatbázis rögzítése

Az előzetes felmérések tanulsága szerint az adatbázist hivatásos jeltolmácsok szereplésével készítettük. A száj körüli tartópontokat kellett tehát adatbázisban rögzíteni. Ezeket olyan videó felvételekkel készítettük el, ahol a modell arcán megjelöltük az FP pontokat.

A videó felvételt automatikus módszerekkel dolgoztuk fel, ami a következő lépéseket jelenti: színkiemelés, binarizálás, dilatáció, erózió. A kapott eredményeket manuálisan javítottuk, ahol az automatikus módszer hibázott. Ezek a helyzetek jellemzően a felpattanók gyors mozgása, vagy a csücsörítés megváltozott fényviszonyai miatt léptek fel.

A felvétel közben jó minőségű hangfelvétel készült (48kHz, 16 bit), amit szinkronizáltunk a videó felvétellel. (Fig. 2)

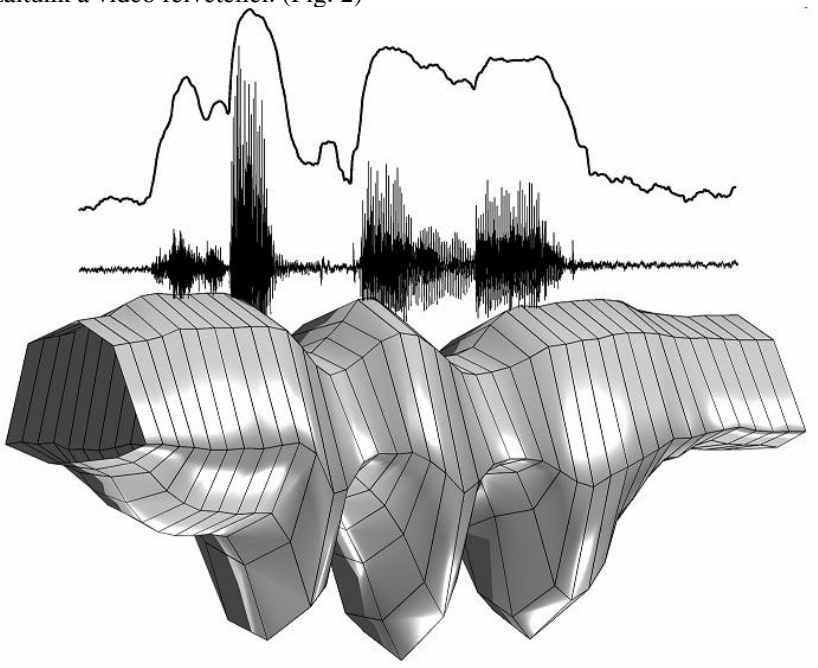

Fig. 2. A „szeptember” szó az adatbázisban. Az ábrán az energia, a beszédjel, és a száj külső kontúrja látható az időben.

A beszédjelen a videó sebességéhez igazított ablakozást használtunk, a PAL rendszer 25 kép/másodperc sebességénél ez 40ms. Így minden képhez tartozott egy hangszakasz. Ez a hangszakasz 48kHz mintavételezés mellett 1920 mintát jelent. Ezen szakaszhosszból kiválasztottuk a maximális kettőhatványt, ez 1024 jel, amin Radix-2 FFT algoritmust futtattunk Hamming ablakkal. Az eredményt mel skála szerint 16 sávban összegeztük, majd cosinus transzformációval kiszámítottuk a cepstrumot (MFCC).

A telefonon futó alkalmazásnál 8kHz mintavételezés mellett az ablakméretek 320-nak illetve 256-nak adódnak. A magasabb mintavételezéssel tanított hálózatot használni lehet az alacsonyabb mintavételezésű készülékeken a mel skála igazításával.

A videó felvételből nyert pontokon lényegkiemelést végeztünk főkomponens analízissel (1). Az első 6 főkomponens használatával az eredeti adatok 1-2%-os hibával történő rekonstruálása lehetséges, ami átlagosan 1 pixel elmozdulást jelent a PAL szabványú képen. Ez a hiba elhanyagolható, hiszen 720x576 pixeles felbontás mellett van tartópont, ami 120 pixelnyi tartományban mozog.

$$w_{1\ldots6} = P^{-1}B\Big|_{p_1^{-1}\times..\times p_6^{-1}} \tag{3}$$

A 16 dimenziós hangi és a 6 dimenziós képi adatokat neuronháló tanításával kapcsoltuk össze. A neuronháló bemenete 5 egymást követő MFCC ablak, a kimenete pedig a középső ablakhoz tartozó képkocka főkomponens-térben adódó koordinátái. A neuronháló 40 rejtett neuront tartalmaz.

Felmerül a kérdés, hogy az itt használt 40ms hosszú ablak alkalmas-e a beszédből szájmozgássá alakításhoz. A beszédfeldolgozó rendszerek mindig rövidebb ablakot használnak. A beszédre jellemző szájmozgásnál a külső szemlélő azt a hatást látja, amit a szájmozgató izmok okoznak. Ezek az izmok jóval lassabban képesek csak mozogni, mint a nyelv, vagy más, a hangot befolyásoló izmok. Még ennél is fontosabb, hogy megkülönböztethetünk fonémákat dominancia szempontból. [4] Azt nevezzük domináns fonémának, ami meghatározza a száj állását, ilyenek a magánhangzók és az ajakhangok. A domináns fonémák a szomszédos nem domináns fonémákra eső szájállást is befolyásolják. A rendszer 5 egymást követő szakasszal dolgozik, melyek összesen 200ms időtartamot fognak át. Ez az átlagos fonémahosszak alapján valószínűsíti, hogy az 5 ablak közül legalább egy domináns fonéma tiszta fázisába essen. (Fig. 3) Ez a neuronhálózat tanulása szempontjából már elegendő. A 200ms késés nem jelent problémát, mert az információ csak ebben a modalitásban érkezik, és a 200ms belül van az emberi dialógusok toleranciahatárán.

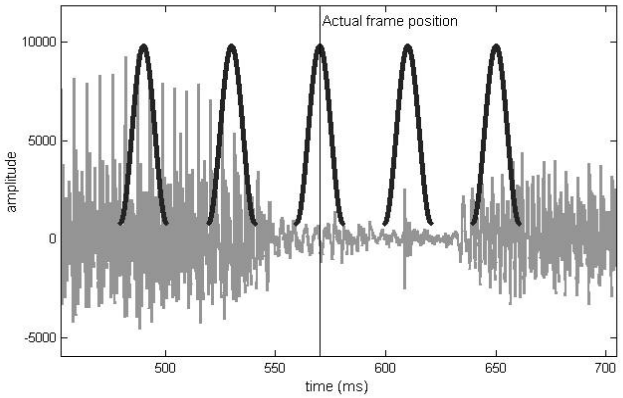

Fig. 3. Az öt egymást követő ablak. Két ablak kezdőpontja között 40ms a távolság.

2.4 Az adatbázis tartalma

Az adatbázisban szerepelt egy rövid általános tanító rész, ami a magánhangzókat egyenlő számban tartalmazta. Az adatbázis nagyobbik részét olyan anyag teszi ki, ami lehetőséget ad közvetlen tesztelésre siket felhasználókkal. A rendszer minőségének a mérésére ugyanis úgy kerülhet sor, ha a felhasználók látnak olyan szintetizált képet is, amit az adatbázis képi anyagából nyerünk, és ennek a lehető legjobb minőségű módja az, ha az adatbázisban rögzítünk olyan frázisokat, amik a tesztelés során egyben felhasználhatóak. A teszt során olyan szituációt modelleztünk, amiben figyelembe vettük a siketek jó kontextus követését, ezért megszorítottuk az anyagot olyan tartalomra, mint egy- és kétjegyű számok, hónapok nevei, hét napjai. Ezek folyó szövegben mindig úgy helyezkednek el, hogy a kontextusból tudható a halmaz, de az, hogy melyik elem a több közül, az bizonytalan.

3 A rendszer

A kész rendszer főbb alkotóelemei a hangkezelő rész, ami a telefon mikrofonját olvasva elvégzi az adott hosszúságú ablakokra vágást, a jellegvektorok kiszámítása, neuronháló, főkomponens analízis, és fejmodell szintetizálás. (Fig. 4)

Az MFCC adatok előállításához Radix-2 FFT algoritmust használunk, ami egy programozható mobiltelefonon is alkalmas valósidejű elemzésre.

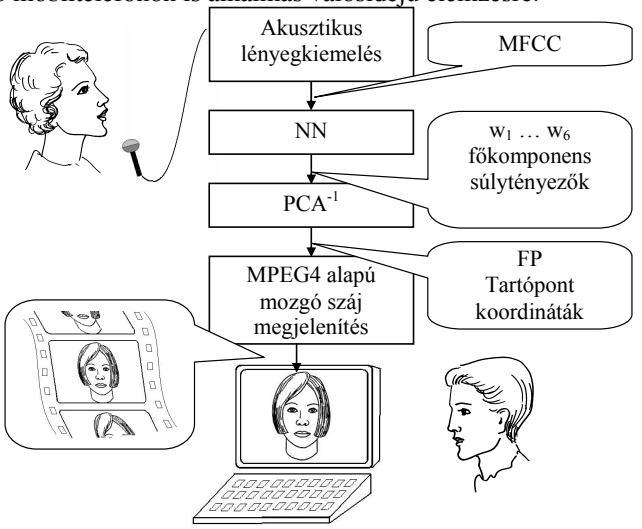

Fig. 4. A rendszer áttekintése.

3.1 Neuronháló

A tesztelt rendszer neuronhálózata 80 bemenettel rendelkezik, ami 5 keret, egyesével 16 MFCC együttható. A kimeneten 6 érték jelenik meg, ami az adatbázis adataiból nyert főkomponensek első 6 összetevő által kifeszített térben 6 koordináta. A neuronhálózat egy hagyományos hiba visszaterjesztéses neuronháló (backpropagation), aminek a számítási kapacitását egy mátrixszorzásra alapuló technikával gyorsított fel Davide Anguita. [1] Ez a neuronháló 1 000 000 epoch tanítás után került tesztelésre siket felhasználókkal.

3.2 PCA

A főkomponens analízisnek több előnye van. Elsősorban a rendszer működőképességének szempontjából fontos, hogy komolyabb adatveszteség nélkül dimenziót lehet csökkenteni. Ez esetünkben a 30 dimenziós pixeltér és a 6 dimenziós főkomponensek által kifeszített tér közötti különbség. A PCA használata előtt futtattunk neuronháló tanításokat pixeltérben is, és több mint 300 millió epoch után sem volt használható a hálózat kimenete. Az ilyenformán tanított neuronhálózat is egyetlen szabadsági fok mentén mozdult el: az állkapocs nyílása. A probléma az volt, hogy neuronhálózatnak kellett megtanulnia azt is, hogy a száj körüli pixelek általában hogyan helyezkednek el, és az együtt mozgásuk általában milyen. Ezt a terhet veszi le a neuronhálóról a PCA térben felírt száj állapot leírás. A PCA koordinátarendszerében ugyanis az általános mozdulat-összetevők már megvannak, egyszerűen ezek együtthatóival reprezentáljuk az állapotot.

A másik előnye a PCA-nak, hogy igen egyszerű használni. A pixeltérbe való visszaszámolás egy konstans mátrixszorzással megoldható. (2)

$$\overline{B}_k = \left(\underline{w}_k + \underline{c}\right) \cdot P \tag{4}$$

Matlab kódot készítettünk, ami a gyűjtött adatbázisból automatikusan kiszámítja a főkomponenseket, átírja az adatbázis pixeltérben rögzített értékeit PCA térbe, és elkészíti azt a C kódrészletet, amivel a visszakódolás elvégezhető. Azért használunk generált kódot kimentett értékek fájlkezeléssel való feldolgozása helyett, mert a mobil platformon így jóval egyszerűbb dolgozni.

A PCA további lehetőségeket is rejt. Ezek közül a legérdekesebb az, hogy a PCA alkalmas arra, hogy objektív véleményt formálhassunk az adatbázishoz készített felvétel minőségéről az olvashatóság tekintetében. Azt figyeltük meg, hogy a siketekkel dolgozó, a szájmozgására tudatosan figyelő és rutinos személyek főkomponensei jól megkülönböztethetőek a képzetlen szereplőkétől. A különbség a főkomponensek sorrendjében van. A főkomponens analízis ugyanis fontossági sorrendbe állítja a főkomponenseket, az első főkomponens az az iránya a sokdimenziós pontfelhőnek, amerre a szórás a legnagyobb. Esetünkben az állkapocs nyitásához kapcsolódó főkomponens az, aminek a súlya lényegesen megelőzi a többit, a szórás körülbelül 70%-áért felelős. Ebben minden szereplő közös: az első főkomponens mindig ez. A másodiktól a negyedikig azonban eltér a sorrend. A képzett, professzionális beszélő főkomponensei mind vizéma megkülönböztető szerepet töltenek be, olyan mozdulatokat, mint a száj széthúzása („csííz”), vagy a csücsörítés. (Fig. 5) A képzetlen beszélőknél igen komoly helyezést érnek el, a vizémákat nem megkülönböztető, beszéd-

szokásokhoz, emóciókhoz kapcsolható mozdulatok, mint például a száj tekerése, oldalra elhúzásával nyitása.(Fig. 6)

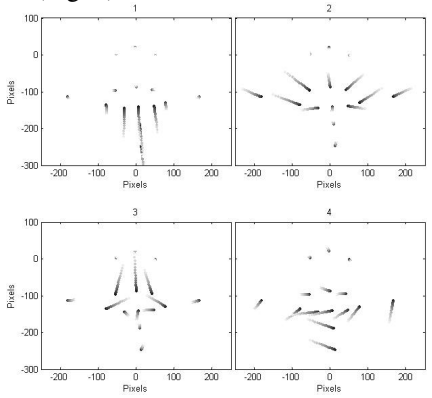

Fig. 5. Profi beszélő főkomponensei. Látható az első három mozdulatkomponens vizéma megkülönböztető szerepe.

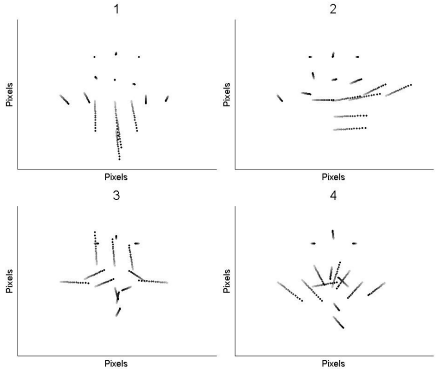

Fig. 6. Gyakorlatlan beszélő főkomponensei. Már a második főkomponens

3.3 Fejmodell

A folyamat végén a fejmodell áll, ami megkapja azon pixelek koordinátáját, ahol a tartópontoknak lenniük kell a 2 dimenziós felvételen. Ebből ki lehet számolni a 3 dimenziós MPEG-4 tartópontokat.[6] Ez úgy lehetséges, hogy a legtöbb mélységi komponenst elhanyagoljuk, csak az állkapocs hátramozdulását számítjuk bele a koordinátákba, illetve, hogy a száj normál állapotától vett távolságokból fejezzük ki az egyes tartópontok helyét. Az így kapott tartópontok alapján az arc szintézise azon alapul, hogy értelmezünk minden tartópont köré egy hatókört, ami azt a bőrfelületet reprezentálja, amit a tartópont magával húz, amikor mozog. Ez a technika a száj vonalánál igényel körültekintést, a hatóköröket úgy kell megadni, hogy az alsó és a felső ajak között ne legyen átfedés, különben a száj belső kontúrja nem nyílna ki. Az állkapocsnál is van egy kis probléma: az állkapcson elhelyezett tartópontnak az egész állkapcsot mozgatnia kell, de ez nyers formában olyan rajzfilmszerű mozgást eredményez, mintha az állkapocscsont függőlegesen eltolódva nyitná a szájat, nem pedig

elfordulva. Ezt a problémát úgy oldottuk meg, hogy tettünk egy el nem mozduló tartópontot az állkapocs tengelyének két oldalára. Ezzel elértük, hogy a súlyozott elmozdulások eredője jól imitálja az állkapocs lefelé elfordulását.

4 Eredmények, fejlesztési lehetőségek

A rendszer által előállított animációkat bemutattuk siketeknek. Kontrollként valódi videó felvételt is mutattunk, illetve azt a fej-animációt is, aminek a paramétereit nem a hangból, hanem egyenesen az adatbázisban rögzített mozgásból állítottunk elő. Ezt az tette lehetővé, hogy az adatbázisban tárolt paraméterek és a fejmodell paramétereit igyekeztünk egyformára csinálni. A arcra felfestett pontok pontatlanságát utólag korrigáltuk, hogy a fejmodell mozgása a lehető legpontosabban kövesse az eredeti felvételt.

Összesen 70 rövid mozgóképet mutattunk, egy- és kétjegyű számokat, hónapneveket és a hét napjait. A mozgóképek között volt valódi felvétel, az adatbázishoz készült felvétel egy-egy részlete. Volt a képi adatok felhasználásával meghajtott szintetizált fej, és volt hangból számított fejmozgás. A valódi videó felismerési aránya 97% volt. Annak az animációnak, ami a képi adatok felhasználásával készült, 55%-os lett. A hangból számított animáció felismerése 48%-os volt.

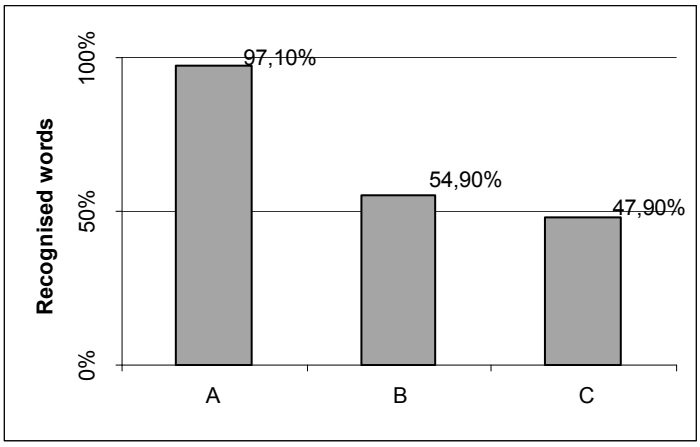

Fig. 7. Felismerési pontosságok a különböző mintákon: eredeti felvétel professzionális jeltolmáccsal (A), szintetikus arc, aminek a paraméterei az adatbázis képi adataiból származik (B), illetve hangból számolt szintetikus arc (C)

Ebből az a következtetés vonható le, hogy a fejmodellünk részletessége a fő probléma. Nyilvánvaló hiányosság, hogy nincs információ a száj belső kontúrjáról, a fogak és a nyelv láthatóságáról. Ezeket a paramétereket nem lehet pontok felfestésével megoldani, ezért a jelenlegi kutatás ebben az irányban is zajlik, új képfeldolgozó eljárásokat vizsgálunk meg. Ugyancsak kutatás folyik abban az irányban is, hogy hogyan lehet összekapcsolni több személy adatait, a jelenlegi rendszer ugyanis csak egy ember hangjának és szájmozgásának összekapcsolását tudja megtanulni.

Köszönetnyilvánítás

Ezúton köszönjük az együttműködést a Siketek és Nagyothallók Országos Szövetségének, és a siketeknek, akik segítettek tesztelni, a T-Mobile-nak, Dr Takács Györgynek, Tihanyi Attilának, Srancsik Bálintnak, és Harczos Tamásnak.

Bibliográfia

1. Anguita D.:Matrix Back Propagation – An efficient implementation of the BP algorithm. Technical report, DIBE – University of Genova (1993)
2. Beskow J.: Talking Heads, Model and Applications for Multimodal Speech Synthesis: Doctoral Dissertation, Stockholm (2003)
3. Cosi P., Fusaro A., Tisato G.: Lucia a New Italian Talking Head Based on a Modified Cohan-Massaro's Labial Coarticulation Model. Proceedings of Eurospeech 2003, Geneva, Switzerland (2003) 2269–2272
4. Czap L., Mátyás J.: Virtual Speaker, Híradástechnika, selected papers (2005) 2-5
5. Granström B., Karlsson I., Spens K-E.: SYNFACE – a project presentation, Proc of Fonetik, TMH-QPSR (2002) 93-96
6. Takács Gy., Tihanyi A., Bárdi T., Feldhoffer G., Srancsik B.: MPEG-4 modell alkalmazása szájmozgság megjelenítésére, Híradástechnika 2006 8.