Vizuális beszéd előállítása beszédjelből

Gergely Feldhoffer

Tézisfüzet

Témavezető:
Takács György

Információs Technológiai Kar
Interdiszciplináris Műszaki Tudományok Doktori Iskolája
Pázmány Péter Katolikus Egyetem

Budapest, 2010

Tartalom

1 Bevezető **1**
1.1 Alrendszerek 1
1.1.1 AV képzés 1
1.1.2 Minőségi szempontok 2

2 Kutatási módszerek **2**
2.1 Adatbázis építés 2
2.1.1 Az alaprendszer 3

3 Új tudományos eredmények **4**
3.1 A közvetlen AV képzés természetessége 4
3.1.1 Módszerek 4
3.1.2 Eredmények 5
3.1.3 Következtetések 6
3.2 Időbeli aszimmetria 6
3.2.1 Kölcsönös információ 6
3.2.2 Többcsatornás kölcsönös információ becslés 7
3.2.3 Eredmények 8
3.2.4 Következtetések 8
3.3 Beszélőfüggetlenség 9
3.3.1 Szubjektív mérés 11
3.3.2 Objektív mérés 12
3.3.3 Következtetések 12
3.4 Vizuális beszédre kiterjesztett audio átvitellel működő távjelenlét alkalmazások 13
3.4.1 Vizéma alapú dekompozíció 13
3.4.2 Eredmények 14
3.4.3 Következmények 16

4 Publikációs lista **16**

Köszönetnyilvánítás

Szeretném megköszönni a témavezetést Takács Györgynek a sok iránymutatást, és mind jelenlegi mind régebbi munkatársaimnak, Tihanyi Attilának, Bárdi Tamásnak, Harczos Tamásnak, Srancsik Bálintnak és Oroszi Balázsnak. Hálás vagyok a doktori iskolának azért a környezetért, ami eredményes munkát tett lehetővé, személyesen különösen Nyékyné Gaizler Juditnak és Roska Tamásnak.

Hálás vagyok Hegedűs Ivánnak, Jung Gergely, Víg János, Tóth Máté, Szabó Gábor Dániel, Bányai Balázs, Mészáros László, Kovács Szilvia, Bucsi Szabó Solt, Krebsz Attila és Selmeci Márton jelenlegi és volt hallgatóknak, akik résztvettek a kutatócsoport munkájában.

Eredményeimet nem érhettem volna el azok nélkül a tapasztalatok nélkül, amiket Czap László, Takaashi Kuratate, Mihajlik Péter és Sasha Fagel kutatókkal találkozva volt szerencsém összegyűjteni.

Szeretném megköszönni továbbá doktorandusztársaimnak és barátaimnak, Weiss Bélának, Soós Gergelynek, Rák Ádámnak, Fodrózci Zoltánnak, Gaurav Gandhinak, Cserey Györgynek, Wágner Róbertnek, Benedek Csabának, Hegyi Barnabásnak, Bankó Évának, Iván Kristófnak, Pohl Gábornak, Sass Bálintnak, Miháltz Mártonnal, Lombai Ferencnak, Bérci Norbertnek, Tar Ákosnak, Veres Józsefnek, Kiss Andrásnak, Tisza Dávidnak, Vizi Péternek, Varga Balázsnak, Füredi Lászlónak, Bálint Bencének, Laki Lászlónak, Lővei Lászlónak, Mihalicza Józsefnek, és még sokaknak a sok segítséget, érdeklődést, hasznos és érdekes vitákat és beszélgetéseket.

Köszönöm a végtelen türelmet a Tanulmányi Osztálynak, és a technikai segítséget Tholt Péternek, Csillag Tamásnak és Rec Tamásnak.

Végül, de nem utolsósorban köszönöm a türelmes és szerető támogatást feleségemnek és családomnak.

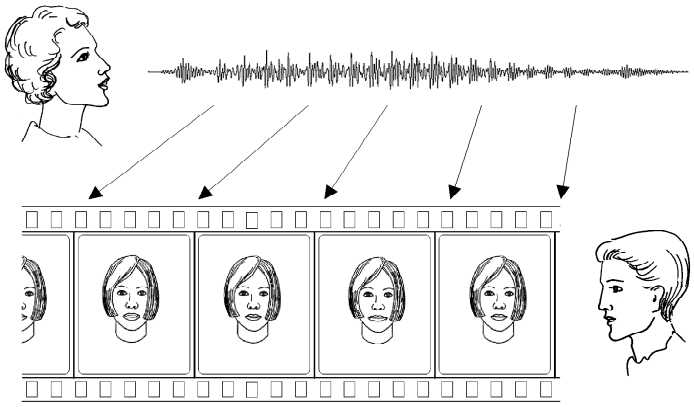

Fig. 1: A beszédhang alapú vizuális beszéd szintézis (Audio To Visual Speech, ATVS) feladata.

1 Bevezető

Beszédhangból vizuális beszéd (Audio To Visual Speech, ATVS) előállítása napjainkban egyre népszerűbb és kutatottabb terület. A beszédfeldolgozás mérvadó konferenciáin új szekciókat nyitnak a multimodális kutatási irányoknak, az Interspeech 2008-ban önálló beszédhang alapú vizuális beszéd szintézis szekciót is tartott.

A terület felhasználási területei jellemzően két csoportba tömörülnek, egyfelől halláskárosultak számára vizuális kommunikációs segítség, másfelől számítógéppel előállított animációhoz hatékony segédeszköz lehet.

1.1 Alrendszerek

Minden ATVS három alapvető komponensből áll, a beszédhang előfeldolgozásából, a beszédhang modalitás és a vizuális modalitás megfeleltetéséből (AV képzés), és az arc képének előállításából.

A beszédhang előfeldolgozása jellegvektorok kiszámítását jelenti, amik hasznos és tömör formában reprezentálják a beszédet. A keletkező jellegvektorok dimenziószáma és reszintetizáláskor keletkező hibája a két legfontosabb minőségi szempont itt. Példaként a mel frekvencia együtthatók (MFC) ábrázolásában néhány mel skála szerinti csatornára osztjuk a spektrumot, amivel némi információt elveszítünk, azonban az időszeletenkénti több száz adat helyett egy tucat körüli együtthatóval írjuk le a jelet. Az olyan adatbázisoknál, amik neurális hálózatok tanítására készülnek, a dimenzionalitás elsődleges szempont.

1.1.1 AV képzés

A hang-kép leképezésnek több módját használják. Az egyik megközelítés szerint egy kész beszédfelismerő (Automatic Speech Recognition, ASR) rendszer után kell kötni

egy a megtalált fonémasorozaton értelmezett vizuális beszéd szintetizátort, ami a fonémákhoz vizémákat rendel, és a vizuális koartikulációs szabályokat alkalmazva egy arc mozgóképének paramétereit előállítja [1, 2]. Egy másik megközelítés, amit mi is használni fogunk, közvetlenül a hangból származtatott jellegvektorokból képez a vizuális beszéd paraméterterére[3, 4].

Napjainkban a kutatási programok lefedik a beszédhang előfeldolgozási módszereinek vizsgálatát vizuális beszéd szintézishez való alkalmasság szempontjából[5], az arc reprezentáció és kontroll módszereket[6], és az élethű természetességű arc megjelenítő rendszereket[7].

1.1.2 Minőségi szempontok

A hang-kép leképezéseket a következő szempontok szerint lehet értékelni.

- természetesség: mennyire hasonlít a keletkező mozgás egy valódi ember szájának mozgására.
- érthetőség: a kimenet mennyit segít a megértésben a szájról olvasóknak
- sebesség: a rendszer erőforrásigénye
- taníthatóság: lehet-e könnyen tanítani a rendszert példák megadásával, mekkora a javulás
- beszélőfüggőség: mennyire változik beszélőnként az érthetőség, és a természetesség
- nyelvfüggőség: mennyire bonyolult a teljes rendszert egy másik nyelvre átállítani. Elég-e az adatbázis cseréje, vagy szabályokat is kell módosítani
- robosztusság: hogyan változik a rendszer teljesítménye különböző akusztikai környezetben, például zajban

2 Kutatási módszerek

Közvetlen AV képző rendszereket építettem, amiket szubjektív és objektív mérésekkel értékeltem ki. A természetességet szubjektív véleménypontozással, az érthetőséget felismerési tesztekkel, a pontosságot és alkalmazhatóságot általában neurális hálózatok tanításával.

A kutatási program alapvetése az a feltételezés, hogy a képzett hang és a száj állása közötti fizikai kapcsolat elegendő lehet a jó minőségű AV képzéshez.

2.1 Adatbázis építés

A közvetlen AV képzéshez a modalitásokból mintapárokra van szükség, ahol a vizuális modalitás egy fej vagy száj állapotának leírása, a hang pedig valamelyik alkalmas

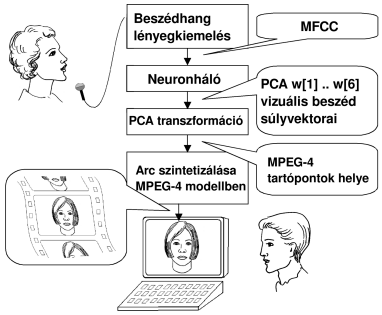

Fig. 2: A mi közvetlen ATVS rendszerünk felépítése.

beszéd előfeldolgozási eljárás eredménye. Mi az MPEG-4 szabvány szerinti fejleírást választottuk, pontosabban a szabvány szerinti tartópontok száj körüli részhalmazát. Felvételt készítettünk beszélő arcokról, ahol ezen pontokat megjelöltük, majd a felvételt feldolgozva az pozíciójukat megkerestük.

2.1.1 Az alaprendszer

A közvetlen AV képzést neurális hálózattal oldottuk meg. Az eddig tárgyalt alrendszereket implementáltuk, adatbázist rögzítettük és megtanítottuk a rendszert. Az eredményt siket emberekkel teszteltük érthetőség szempontjából, rögzített szövegkörnyezetben kulcsszó felismerési feladattal, amihez számokat, hónapokat és a napok neveit használtuk. Referenciaként egy valódi arc videófelvételét használtuk, és megmértük az adatbázisban rögzített vizuális beszéd érthetőségét is[3]. Az eredményt a 1. táblázat mutatja. A továbbiakban tárgyaltak szerint az eredmények legfontosabb részlete az, hogy a rendszer hangból nem teljesít sokkal rosszabbul, mint a lehető legjobb szintetikus fej.

Tábla 1: Felismerési arányok az alaprendszerben.

Videó származása	Felismerési pontosság
valódi videó	97%
referencia arcmozgás a fejmodellen	55%
hangból számolt arcmozgás a fejmodellen	48%

Mint később kiderült, ezt a rendszert mások is használták korábban, csak gyenge eredményei miatt nem publikálták nemzetközi szinten. A mi eredményeink szignifikánsan jobbak ezeknél a rendszereknél, aminek az oka az, hogy jeltolmács segítségével rögzítettük az adatbázist. Fontos, hogy jeltolmács alatt nem jelbeszéddel, hanem professzionális artikulációval kommunikáló embert értek, ami Magyarországon ugyanaz a szakma, és így az angol “lip-speaker” kifejezésnek nincs igazán jó megfelelője.

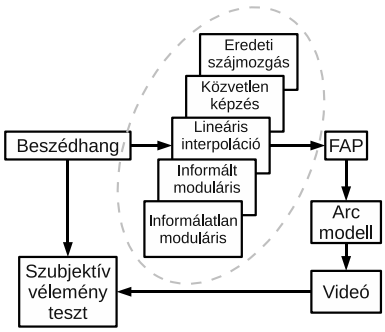

Fig. 3: Különböző AV képzés módszerek egy környezetben. Az informált és informálatlan beszédfelismerőket külön teszteltük a moduláris megközelítésekben.

3 Új tudományos eredmények

3.1 A közvetlen AV képzés természetessége

Összehasonlító tanulmányt készítettem az ATVS rendszerekről. A mi alaprendszerünket összevetettem beszédfelismerő alapú megközelítésekkel, amikhez a magyar nyelvhez létező legjobb alrendszerek összegyűjtésével láttam hozzá. Az AV képzéseket vetettem össze, aminek érdekében a többi komponens változatlan volt: ugyanaz az akusztikus környezet, ugyanazok a beszélők, ugyanaz a fejmodell, és ugyanazok az értékelő személyek a tesztben. Az eredmény azt mutatta, hogy a közvetlen AV képzés igen jó minőségű természetesség tekintetében.

> ***I. Megmutattam, hogy a közvetlen AV képzés, ami természeténél fogva gyorsabb a beszédfelismerővel kombinált megoldásoknál, természetesség tekintetében is jobb azoknál, amennyiben jeltolmácstól származnak a tanítóminták.*** [8]

3.1.1 Módszerek

Az egyik résztvevője a méréseknek a közvetlen AV képzésünk. A többi alapja egy magyar fejlesztésű beszédfelismerő Mihajlik Péter és munkatársai laboratóriumából, ami egy súlyozott véges állapotú automata alapú rejtett Markov modell (WFST-HMM), ami a mai beszédfelismerési módszerek legnépszerűbbike. Konkrétan a VOXerver [9] rendszer ez, ami informált és informálatlan működésre is képes. Informált esetben egy szótár megadható, ami a felismerést segíti, ám le is szűkíti. Az informálatlan felismerésnél egy általános nyelvi modellt használ a rendszer.

A beszédfelismerőt használó rendszerek legegyszerűbbike a lineáris interpoláció, amelyik megfelelteti a fonémákat a vizémákkal, és szomszédsági viszonyaiktól függetlenül lineárisan interpolál.

A vizuális koartikulációs hatások modellezéséhez szükség van egy ennél kifinomultabb módszerre. Különösen a szomszédos beszédszakaszok egymásra

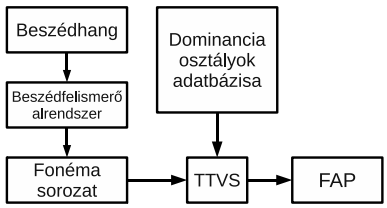

Fig. 4: A moduláris ATVS tartalmaz egy beszédfelismerő és egy szövegalapú vizuális beszéd szintetizátort.

Tábla 2: A vélemények átlaga és szórása.

módszer	átlagos pont	szórás
Eredeti szájmozgás	3.73	1.01
Közvetlen képzés	3.58	0.97
Informálatlan moduláris	3.43	1.08
Lineáris interpoláció	2.73	1.12
Informált moduláris	2.67	1.29

hatásának fonéma- és vizémafüggőségeinek kezeléséről van szó. Nem mindegy, hogy a fonéma képes-e dominálni a szomszédságát, ahogy két magánhangzó például teljes mértékben meghatározza a közöttük előforduló "h" fonéma vizuális megjelenését. Ezt a feladatot oldja meg a szövegalapú vizuális beszédszintetizátor, amely szövegből az audio beszédszintetizátorok mintájára fonémasorozatot állít elő, amelyhez időzítési információkat is generál, majd erre az adatsorra alkalmazza a vizéma dominancia szabályokat, így állítva elő élethű vizuális beszédet. Mi Czap László és munkatársai TTVS rendszerét használtuk[10], ami a ma létező legjobb minőségű ilyen rendszer magyar nyelvre.

A videókat a következőképpen állítottuk elő: a tanítási adatbázisban nem szereplő beszélő hangján futtattuk a közvetlen AV képzést. Ugyanezen beszélő hangján futtattuk a beszédfelismerőt, ami időzített fonémasorozatot adott. Ezen elvégeztük a lineáris interpolációt, illetve a TTVS rendszerbe beillesztve megkaptuk a vizémadominanciákat is figyelembe vevő paramétersorozatot. Ez utóbbit mind szótárral (informált), mind szótár nélkül (informálatlan) előállítottuk. Emellett referenciaként egy eredeti, videófelvételből nyert paraméterekkel hajtott vizuális beszédanyagot használtunk.

3.1.2 Eredmények

A tesztben 58 alany vett részt, akiknek osztályozniuk kellett a videók természetességét. Az eredmény a 2. táblázatban látható. Kétmintás t próba szerint az eredetitől való eltérés a közvetlen képzésnél nem szignifikáns ($p = 0.06$), a moduláris esetben pedig szignifikáns ($p = 0.00029$). A mérés egyéb eredménye, hogy az időbeli pontosság fontosabb a fonéma precizitásnál: az informálatlan felismerés időzítése technikai

okokból pontosabb, mint a minden fonémát jól felismerő szótáras informált eset.

Figyelemreméltó, hogy a lineári interpoláció (ami szintén informálatlan, tehát jól időzített, de fonémahibákat tartalmaz) jobb ugyan az informált modulárisnál, de a két moduláris rendszer átlagától szignifikánsan elmarad, ami a vizémadominancia kezelésének fontosságát bizonyítja.

3.1.3 Következtetések

Ez az első olyan közvetlen képzés, amelyet jeltolmáccsal rögzítettek. Összevetve más megközelítésekkel tehát fontos új eredmények birtokába jutottunk: nem csak számításigényben, de természetességben is jobb lehet a közvetlen képzés a modulárisnál.

A mérés szerint magyar nyelven a ma elérhető legjobb beszédfelismerő és szövegalapú vizuális beszédszintetizátor összekapcsolásából nyert rendszert kis előnnyel, de természetességben megelőzi a jeltolmáccsal tanított közvetlen képzés.

3.2 Időbeli aszimmetria

Az audiovizuális beszéd finom időszerkezetét vizsgáltuk, hogy meghatározzuk azt az időablakot, amin optimálisan lehet AV képzést végezni. Kölcsönös információ becslést alkalmaztunk a modalitások között annak kiderítésére, hogy a $\Delta t = 0$ szinkrontól mennyire érdemes még vizsgálni a hang tartalmát, ami befolyással lehet a képre. Azt az eredményt kaptuk, hogy az arc befolyása a hangra nem azonos módon csökken a múlt és a jövő felé.

> ***II. Megmutattam, hogy a látható beszédszervek jellemzői egy átlagos fonémahosszon belül szorosabban kötődnek a jövőbeli beszédhanghoz, mint a múltbelihez. A kapcsolatot kölcsönös információval mértem. A vizuális modalitás tehát megelőlegez információt az audio modalitásról.*** **[11]**

A jelenség, miszerint a száj korábban mozdul, mint a hang elkezdődne, közvetlenül megfigyelhető esetenként koartikulációban és levegővétel utáni szókezdésnél. Ennek tüzetesebb és általánosabb vizsgálatáról van szó.

Vannak tudományos eredmények időbeli aszimmetriáról a beszéd audio és video modalitások *érzékelése* között. Czap és társai különbséget tapasztaltak a kép-hang szinkron elcsúszásának irányában: a képet 200ms-el megelőző hang zavaróbb, mint a képtől 200ms-el késő hang. Az én eredményeim a *képződésben* mutatnak időbeli aszimmetriát, az irodalomban először. A mérések szerint a modalitások közötti információtartalom összefüggése aszimmetrikus a két, gondosan szinkronizált adatsor között. Ez a jelenség egy magyarázata lehet a már publikált percepcionális eredményeknek.

3.2.1 Kölcsönös információ

$$MI_{X,Y} = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

A kölcsönös információ X és Y között azt írja le, hogy mennyivel tudunk többet X-ről, ha Y-nak tudatában vagyunk (vagy fordítva). Ha X és Y függetlenek, akkor zéró a kölcsönös információjuk. Ezt úgy használtuk fel, hogy a modalitásokból képzett adatokat időben egymáson elcsúsztattuk, és 1ms gyakorisággal kölcsönös információt becsültük közöttük. $\Delta t = 0ms$ esetben szinkronban van a hang és a kép, a kölcsönös információ a közvetlen ATVS rendszerek pillanatnyi adatokon adható legjobb minőségének felső becslését adják. $\Delta t = 100ms$ pedig azt adja meg, hogy ha ismerjük a pillanatnyi képet, mennyit tudhatunk a 100ms elteltével érkező hangról (és viszont, ha birtokában vagyunk a jövőbeli hangnak, mennyit tudunk a jelenlegi képről) Nyilvánvaló, hogy egy fonéma hosszán túli eltoltásnak nincs gyakorlati haszna esetünkben.

Ha tehát a és v az audió and videó adatok:

$$\forall \Delta t \in [-1s, 1s] : MI(\Delta t) = \sum_{t=1}^{n} P(a_{t+\Delta t}, v_t) \log \frac{P(a_{t+\Delta t}, v_t)}{P(a_{t+\Delta t})P(v_t)} \tag{2}$$

ahol $P(x, y)$-t egy 2D hisztogrammal becsültük, amit Gauss ablakkal konvolváltunk, hogy a hisztogram 200x200 felbontását kitöltő mérésszámot csökkentsük. A definíció szerinti kölcsönös információ becslés két egydimenziós jel között is igen sok mintát igényel ahhoz képest, amennyi fáradtság egy audiovizuális adatbázis építése, nekünk pedig többdimenziós adataink vannak, ezért dimenziónként hasonlítjuk a modalitásokat, és az eredményeket összegezzük, amihez független komponens analízist (ICA) használunk.

3.2.2 Többcsatornás kölcsönös információ becslés

Az ICA bázistranszformációjának az a célja, hogy a minta pontjai az n dimenziós hiperkockában egyenletes eloszlást mutassanak. Ezzel az egyes dimenziók közötti együttes eloszlást kisimítva függetlenné teszi. Ez természetesen csak akkor lehetséges, ha valóban van n dimenziónyi független csatornányi adat a jelben. Hogy ezt garantáljuk, csak az első 6 főkomponens által kifeszített teret dolgoztuk fel.

A hangot MFC-ben írtuk fel, vettük ennek a 6 első főkomponensét (MFCPCA). A képet az MPEG-4 tartópontok koordinátáiban írtuk fel, majd vettük 6 első főkomponensét (FacePCA). Ezen adatsorokon számoltunk ICA-t. Az ICA számítás után ellenőrzésképpen mértük a modalitások csatornaközi kölcsönös információját, ami elegendően alacsony lett. Így már lehetővé válik az egyes modalitások csatornái között kapott kölcsönös információk összegzése, ami 15 intermodális kölcsönös információ összegét jelenti.

A hang-kép szinkronról többször is megbizonyosodtunk, hogy technikai hibák ne befolyásolják az eredményt. A szinkront a “papapa” szóval biztosítottuk a felvétel elején és végén, a felpattanó hang első gerjesztését az első nyitott szájat tartalmazó képkockához igazítottuk, így nem kell figyelembe venni olyan zavaró tényezőket, mint a mikrofon-száj távolság.

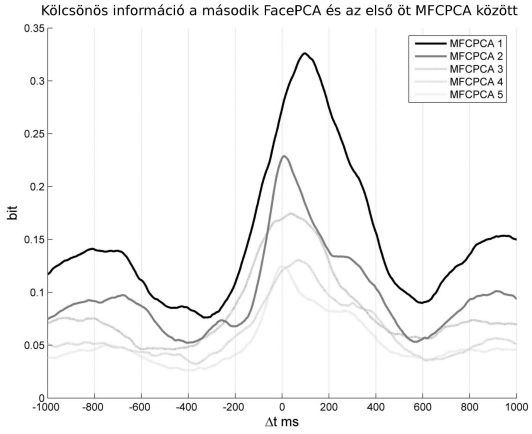

Fig. 5: Egy példa: a második FacePCA (a száj oldalirányú széthúzása) kapcsolata az egyes hangi főkomponensekkel. A pozitív Δt jövőbeli hangot jelent. A görbe láthatósa a fontosságát jelzi.

3.2.3 Eredmények

Kölcsönös információ becslése minden fontos (első 6 főkomponens által kifeszített térben készült) ICA komponens között megtörtént -1000 — 1000 ms tartományban. Emellett néhány kiemelt hang (MFCPCA) és kép (FacePCA) főkomponens közötti kölcsönös információ becslés eredményét is bemutatom, mivel ezekhez fizikai jelentés is kapcsolódik.

A kölcsönös információ görbéink mindegyike aszimmetrikus volt a jövőbeli hang javára. Ez azt jelenti, hogy a képi modalitás pillanatnyi állapota jelzést hordoz a hang elkövetkező állapotaira vonatkozóan, illetve hogy egy ATVS rendszer jól teszi, ha késést rak a rendszerbe, hogy az egy pillanathoz tartozó képi információ előállításához bevárja a következő 2-300ms hangot is.

A megfigyelések egybecsengenek azzal a gyakorlati megfigyeléssel, hogy az artikuláció esetenként hamarabb megkezdődik, mint ahogy hallanánk.

Általában a legmagasabb MI érték a szinkronpontban található, de néhány érdekes kivételt mértünk, mint az 5 ábrán is látható.

Az 6. ábrán látható, ahogy a FacePCA2 paraméter folyamatosan változik mialatt a hangi paraméterek változatlanok, tehát amikor a szó “ep” részletében közelítjük a száj becsukódásával járó “p”-t, a spektrális tartalom nem változik annyira, mint a száj, ami gyorsan csukódik. Ez valószínűleg a nyelv helyzetének változatlanságával magyarázható.

3.2.4 Következtetések

Bevezettünk egy többcsatornás kölcsönös információ becslő eljárást. Csökkentettük a csatornaközi kölcsönös információt ICA használatával. Hogy csak fontos adatokat

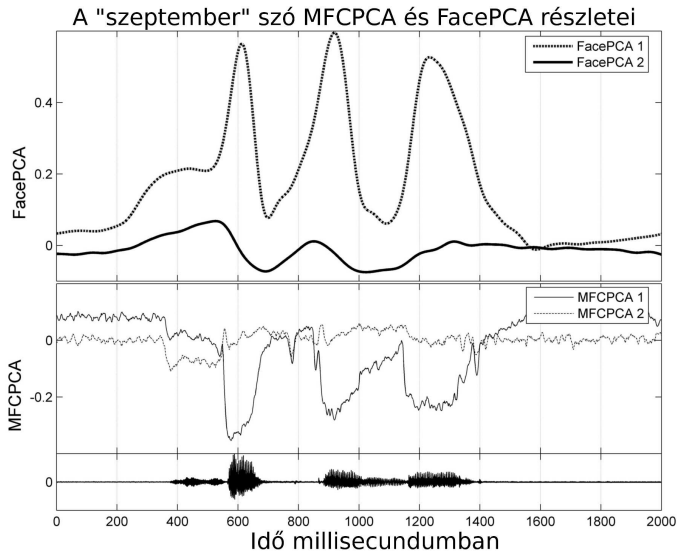

Fig. 6: A "Szeptember" szó, amin végigkövethető a főkomponensek változása modalitásonként.

dolgozzunk fel, ezt az első 6 főkomponens által kifeszített térben tettük meg. Így a hagyományos kölcsönös információ becslő eljárást használhattunk intermodális esetekben, majd az eredményeket összegezhettük. Az eredmény aszimmetrikus az időben, a jövőbeli hanghoz több köze van a szinkronpontbeli video modalitásnak, mint a múltbelihez. Az érvényességi tartománya a jelenségnek az adott beszéd átlagos fonémahossza. Gyors, hadaró beszédben nem jelenik meg a jelenség. Izoláltszavas adatbázison erősebben jelentkezik az aszimmetria, folyamatos olvasott szövegben is jelen van, csak gyengébben.

Az ATVS rendszerekre vonatkozó legfontosabb következmény, hogy érdemes kivárni 200ms időt a jövőbeli hangból. Más rendszerekre is vannak érdekes következményei, például multimodális beszédfelismerésnél javíthatja a válaszidőt az a tudás, amit a pillanatnyi kép alapján tudunk takarékoskodni a jövőbeli hang feldolgozásánál a lehetséges állapotok előszűrésével.

3.3 Beszélőfüggetlenség

A közvetlen ATVS rendszerek adatbázisához hang és kép adatpárok szükségesek[12]. A rendszer ezeket megtanulja, ha tehát egyetlen beszélő van az adatbázisban, akkor a kész rendszer beszélőfüggő lesz. Beszélőfüggetlenséghez olyan adatbázisra van szükség, amiben többféle ember beszédhangja szerepel. Ehhez azonban nehéz képi információt csatolni, hiszen ha mindenkinek a képi modalitását is rögzítenénk, elvesztenénk a szintetizált artikuláció következetességét, az esetleges artikulációs problémákkal pedig tovább rontanánk a helyzetet. A cél egy ATVS rendszer esetében a lehető legjobb (legtermészetesebb vagy legérthetőbb) vizuális artikuláció előállítása, nem pedig egy

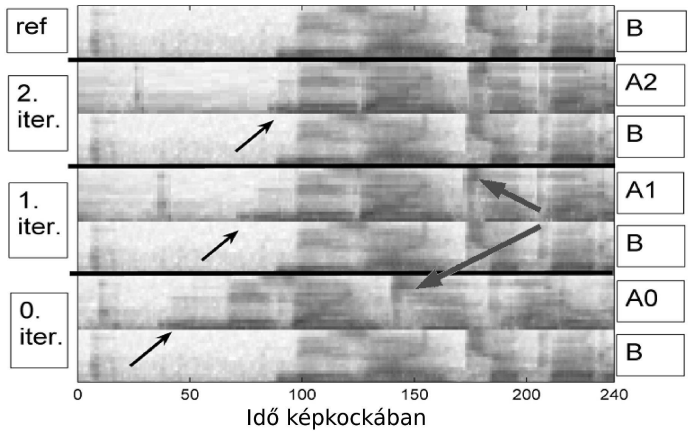

Fig. 7: Az illesztés iterációi. Megfigyelhető, hogy néhány jellemzőnek több iterációra van szüksége a pontos időzítéshez.

adott hang gazdájára leghasonlóbb viselkedés (ami a speech inversion területe, lásd pl. [13]). Ezért eleve csak tehetséges jeltolmácsok adataival érdemes dolgozni a képi modalitás oldalán, erről mérést is készítettünk érthetőség szempontjából: minden technikai paraméter másodrangú volt, az elsődleges hatás a vizuális artikuláció tisztasága.

Ezért olyan eljárásra van szükség a beszélőfüggetlen közvetlen átalakításhoz, ami sok ember hangjához a lehető legjobb vizuális beszédartikulációt rendeli.

> ***III. Kifejlesztettem egy idővetemítés alapú módszert arra a célra, hogy közvetlen AV képzéshez tanítómintákat szolgáltasson. Megmutattam, hogy a tanított rendszer pontossága a tanítóminták számának növelésével javul. A pontosságot olyan beszélővel mértem, amelyik a tanító adatbázisban nem szerepel.*** **[14]**

A módszer lényege, hogy kiváló artikulációjú jeltolmácsok képi modalitásához úgy rendelem hozzá más beszélők hangját, hogy az így elkészülő tanítóminta azt mutassa, hogy "hogyan mondta volna ezt egy professzionális jeltolmács". Így videófelvétel készítése nélkül lehet fejleszteni az adatbázist beszélőfüggetlenség irányába. A módszer korlátja az, hogy az új beszélőknek is pontosan azokat a mondatokat kell mondaniuk, amit a jeltolmács mondott, tehát spontán beszédből álló adatbázisra nem alkalmazható, csak izoláltszavas, illetve különböző olvasott adatbázisokra. Ez utóbbi nem erős korlátozás, az adatbázisok zöme a kontrollált fonetikus statisztikák miatt rögzített anyagú, olvasott adatbázis.

A módszer a dinamikus idővetemítésen (DTW) alapul, ami két időjel között egy szuboptimális illesztést ad. Eredetileg beszédfelismerési módszer, kötött szótáras felismerőkben alkalmazták. A módszert többféleképpen lehet paraméterezni azzal, hogy milyen atomi illeszkedési kiigazításokat engedünk meg. Az egyik legegyszerűbb

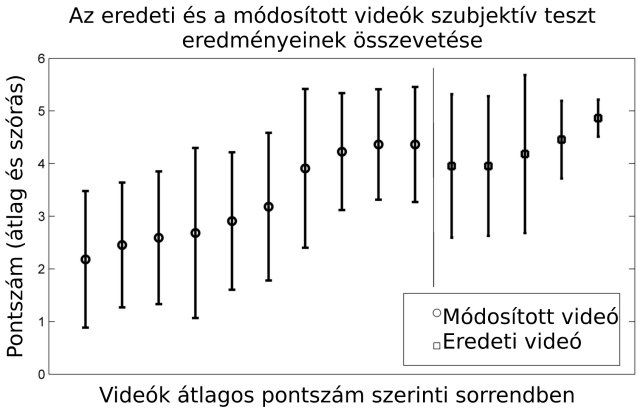

Fig. 8: Átlagok és szórások a szubjektív mérésben.

lehetőség az egy minta törlésének vagy beillesztésének lépésének engedélyezése, ami annyiban veszélyes, hogy elvetemült (értsd a vetemítés kiesik az értelmes tartományból) eredményeket adhat. Egy másik lehetséges paraméterezés tiltja beszúrás és törlés önálló alkalmazását, ezeket mindig egy mintányi közös lépés kell kövesse. Ez utóbbi paraméterezések nem tudnak nagyon változtatni a mintán, lévén legfeljebb felezhetik/duplázhatják a tempót szakaszonként, viszont nem adnak elvetemült eredményt. Én a két paraméterezés mindkét előnyét ötvöztem az iteratív megszorított módszeremmel, amely a második módszerrel igazít, majd újfent igazít az eredménnyel, amíg a változás nem elég kicsit, lásd 7. ábra.

A keletkező illesztés egy indexelés, hogy melyik időpillanatban hol tart a másik beszélő ugyanebben a tartalmi állapotban. Ilyen módon szinkronizálva a jeltolmács adott pillanatbeli szájállása jó választás a hozzáadandó beszélő adott hangi tartalmi állapotához. Ezekből készül a tanítóminta. Ezt minél több beszélővel elvégezve egyre jobban lefedhető a beszélők változatossága.

3.3.1 Szubjektív mérés

Az első mérés a tanítóminták validálása volt szubjektív méréssel. Elkészítettünk videókat, amelyeknek a hangját az egyes beszélőktől átvettük, a hozzá tartozó képet pedig a tanítóminták tartalmából olvastuk ki. Mivel a vetemítés képkockákat kihagy illetve megismétel, az így nyersen kapott videók akadozónak tűntek, ezért minden videót, a referenciának használt valódi jeltolmáccsal rögzített videót is időben simítottuk. A teszalanyok feladata az volt, hogy válogassák ki a “szinkronizált” videókat, ahol szerintük a kép és a hang más személytől származik. Ezt osztályozással tették meg, 5 jelentette a kétségkívül eredeti kép és hang viszonyt, 3 a bizonytalanságot, és 1 azt, hogy biztosan más forrásból származik a kép és a hang.

Ahogy az a 8. ábrán látható, a szórások átlapolódnak, és néhány tanítómintát eredetibbnek vélt a tesztközönség, mint a valódiakat. Az átlagos eredmény az

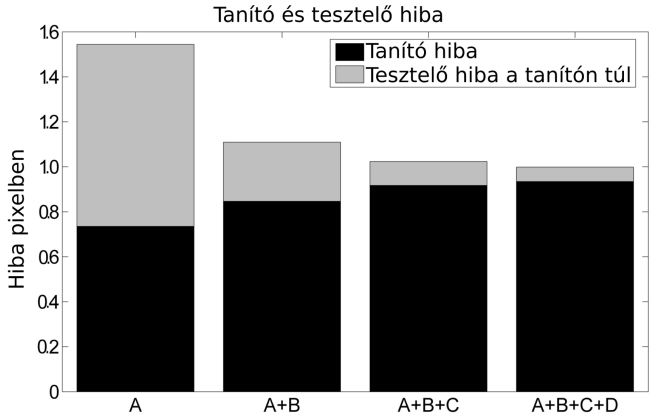

Fig. 9: Tanítás az A, majd A+B, A+B+C mintákkal. A hiba E beszélőre mondott neurális válaszfüggvény és a tanítóminták eltérése pixelben, az adatbázis rögzítésének felbontásában.

eredetiben 4.2, a tanítómintákon 3.2 volt, amit mi jó eredménynek értékeltünk, és megkezdtük a tanítóminták használatát.

3.3.2 Objektív mérés

A beszélőfüggetlenséget numerikusan úgy mértük, hogy a jeltolmács beszélő (A) mellé folyamatosan felvettünk az adatbázisba beszélőket (B,C,D) és közben mértük egy az adatbázisban nem található beszélő (E) eredményeit, összevetve E beszélő tanítómintáival. A mérés mértékegysége itt a pixel, az MPEG-4 tartópontok helyzete abban a felbontásban, ahogy az adatbázis készült. Ez egy jó mérték, mert az eredeti videófelvétel feldolgozási pontosságát össze lehet hasonlítani az eredményekkel.

Összesen 80 videó készült. Az eredmény látható a 9. ábrán. A pontosság megközelítette az 1 pixelt, ami a felvételnél alkalmazott pontkövetési eljárás pontossági határa. A mintákat többféle sorrendben hozzáadva mindig növekedett a pontosság, más-más mértékben.

3.3.3 Következtetések

Bemutattunk egy beszélőfüggetlen közvetlen ATVS rendszert. Szubjektív és objektív mérések bizonyítják, hogy a dinamikus idővetemítésen alapuló tanítóminta előállítás praktikusan használható, mert nincs szükség videófelvétel rögzítésére. A beszélőfüggetlenségre tett erőfeszítés csak offline, tanítás közben jelent többletterhelést, használat közben nincs költsége.

A moduláris megközelítésekben a beszélőfüggetlenség a beszédfelismerő alrendszer felelőssége. Ez ilyen helyzetekben könnyebben kezelhető, hiszen csak audio beszédadatbázisokból igen sok van, a beszédfelismerők beszélőfüggetlensége évtizedek

óta kutatott terület, és ennek megfelelő minőséget is hoznak. Ebben a tekintetben tehát nem a közvetlen AV képzés előnyéről beszélünk, hanem hátrányának csökkentéséről.

A módszer hátránya, hogy érzékeny a kiejtési hibákra, azokon az illesztés elcsúszhat. Ezért a beszélőknek mindent úgy kell mondaniuk, ahogy a jeltolmács tette, tehát a jeltolmács felvételénél gondosan kell a fonetikus tartalmat kezelni. Ugyancsak gondoskodást igényel a jeltolmács beszédanyagának összeállítása, lévén a fonetikus tartalom később állandó.

Összefoglalva megállapítható, hogy egy kevés hátránnyal járó módszerről van szó, ami nem jelent többletterhelést futásidőben a rendszerre, így a használata javasolható minden közvetlen AV képzésre alapuló módszernél.

3.4 Vizuális beszédre kiterjesztett audio átvitellel működő távjelenlét alkalmazások

Ahhoz, hogy fejmodellel lássunk el egy ATVS alkalmazást, ismerni kell a vizuális reprezentációt. Az alaprendszerben PCA paraméteket használtunk, ami a mérések statisztikáitól függ. Ha egy rendszerhez a fejmodellt grafikusnak kell megterveznie, akkor olyan reprezentációra van szükség, aminek a bázisvektoraihoz tartozó fejállásokat a tervező akadálytalanul lerajzolhatja. A grafikusok vizémákat könnyen rajzolnak, ezért érdemes megvizsgálni egy olyan ATVS rendszer lehetőségét, ami erre a reprezentációra képez.

> ***IV. Kifejlesztettem egy módszert, ami hangátvitelt tartalmazó távjelenlét alkalmazásokhoz csatolható vizuális beszéd szintézist tesz lehetővé vizéma alapú fejmodellel. Az így kapott rendszer jobb, mint az ilyen alkalmazásokban széles körben használt energia alapú kétvizémás interpolációs rendszer, futásidőben elhanyagolható többletterhelést jelent, és nem igényel beavatkozást a küldő oldalon.*** **[15]**

Korábban MFCC hangreprezentációt használtunk, a feladathoz azonban jobban illik az ilyen alkalmazások ad-hoc szabványos formátuma, a Speex.

3.4.1 Vizéma alapú dekompozíció

A grafikus tervező számára hiába lehetséges a modern 3D tervezőprogramokban az alakzatokhoz függvényeket rendelve paraméterfüggő arcot építeni, a gyakorlat azt mutatja, hogy sikeresebb egy fejmodellhez több alakot rendelni. Az alak a 3D modell egy módosítása, ami a modell vertexeinek elmozdításával keletkezik. Így egy fejmodell több alakját egy-egy vizémára lehet tervezni.

Ezek után minden arcállapotot a vizémák súlyozott átlagával közelíthető. A dekompozíció feladata a súlyok optimalizálása a minél pontosabb leírásért. A vizémákat MPEG-4 tartópontokban adjuk meg pixelben. A vizémaként kiválasztott paramétersorokat mint bázisvektorokat vesszük. Optimalizáláshoz parciális gradiens módszert használtunk konvexitáskényszer megtartásával, képkockánként. A

Fig. 10: Példák vizémákra, a dekompozícióban betöltött fontossági sorrendben.

konvexitáskényszer elégséges, de nem szükséges feltétele annak, hogy ne alakulhasson ki természetellenes állapot, így minden pont a lehetséges vizémák konvex burkában marad. A gradiens módszer stabilitásának növelésének érdekében a képkockák feldolgozásánál az iteráció az előző képkocka eredményéből indul.

$$\vec{G} = \sum_{i=1}^{N} w_i \vec{V_i} \tag{3}$$

ahol

$$\sum_{i=1}^{N} w_i = 1 \tag{4}$$

Tehát G állapot leírható V vizémák konvex összegével, ahol V tetszőleges lineáris reprezentáció, mint például tartópontkoordináta, vagy főkomponens vektor.

A módszer egyik feltételezése, hogy a 2D adatbázison tanított rendszer a vizémákon keresztül általánosítható a 3D fejmodellre. Ez a feltételezés vetítéssel indokolható.

Az alaprendszert tehát úgy alakítottam át, hogy az audio adatokat Speex paraméterekben, a video adatokat a dekompozíció eredményéből vettem.

3.4.2 Eredmények

A tanítás eredménye látható a 11. ábrán. A fő mozdulatfolyam megmaradt, kisebb problémák láthatóak: a száj nem csukódik be teljesen a bilabiális nazálisnál, és a felpattanók gerjesztése külön látható. Sajnos itt nem alkalmazható a jövőbeli buffer használata, mivel nincs mód a hang késleltetésére.

Szubjektív teszteket végeztünk a dekompozícó eredményével tanított rendszereken. A tanítások minden esetben 1 millió epochosak voltak, így az alacsonyabb dimenziószámú reprezentációk pontosabbak voltak. Referenciaként a dekompozíció eredményét is használtuk videó előállításra, amiben nincs hang alapú számítás.

Az eredmények szerint a legjobb választás a két szabadsági fokú rendszer (lásd 12. ábra), ami egyébként egybevág Czap László eredményével. Érdekesség, hogy a legmagasabb numerikus hiba a neuronháló tanításánál a legjobb pontszámú eset.

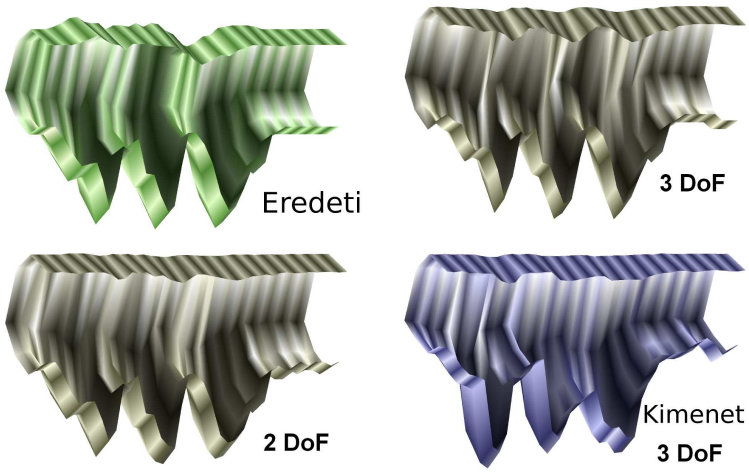

Fig. 11: A “Szeptember” szó szájkontúrjai időben. Az “Original” egy eredeti jel, a DoF jelölésűek dekomponált újraszintetizált jelek, és egy hang alapú szintézis látható.

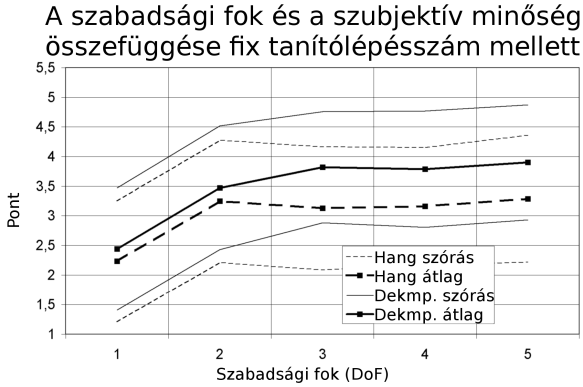

Fig. 12: Szubjektív véleménypontszámok a dekompozíció alapú rendszer eredményéről. A második szabadsági fok bevezetése szignifikáns előnyt jelent. A hangból számolt értékek nem maradnak el a dekompozíció eredményétől.

3.4.3 Következmények

A legfőbb kihívást a szokatlan képi reprezentáció jelentette, amit sikerrel alkalmaztunk. A bemutatott rendszer implementációjához elhanyagolható számítási teljesítmény szükséges, a használt legnagyobb szabadsági fok esetében is kevesebb mint 300 szorzás és összeadás kell. A képi megjelenítésben a lineáris kombináció indexazonos vertextömbök között hardveresen gyorsított, a grafikus kártya memóriájában kell helyet foglalni a vertextömböknek. A rendszerhez csak fogadó oldalon kell gondoskodni változtatásokról, a küldő oldalon nincs változás, annak közreműködése nélkül ki- vagy bekapcsolható.

A műfaj baseline megoldásánál szignifikánsan jobb eredményt ad a második szabadsági fok bevezetése. Erre a célra először hoztunk létre közvetlen ATVS rendszert, ami grafikusok igényeit is figyelembe veszi.

4 Publikációs lista

Nemzetközi folyóirat

- Gergely *Feldhoffer*, Tamás Bárdi : Conversion of continuous speech sound to articulation animation as an application of visual coarticulation modeling, Acta Cybernetica, 2007
- Gergely *Feldhoffer*, Attila Tihanyi, Balázs Oroszi : A comparative study of direct and ASR based modular audio to visual speech systems, Phonetician 2010 (elfogadva)

Nemzetközi konferencia

- Gyorgy Takacs, Attila Tihanyi, Tamas Bardi, Gergely *Feldhoffer*, Balint Srancsik: Database Construction for Speech to Lip-readable Animation Conversion, Proceedings 48th International Symposium ELMAR, Zadar, 2006
- G. Takács, A. Tihanyi, T. Bárdi, G. *Feldhoffer*, B. Srancsik: Signal Conversion from Natural Audio Speech to Synthetic Visible Speech, Int. Conf. on Signals and Electronic Systems, Lodz, Poland, September 2006
- G. Takács, A. Tihanyi, T. Bárdi, G. *Feldhoffer*, B. Srancsik: Speech to facial animation conversion for deaf applications, 14th European Signal Processing Conf., Florence, Italy, September 2006.
- Takács György, Tihanyi Attila, Bárdi Tamás, *Feldhoffer* Gergely,: Feasibility of Face Animation on Mobile Phones for Deaf Users, Proceedings of the 16st IST Mobile and Wireless Communication Summit, Budapest 2007
- Gergely *Feldhoffer*, Balázs Oroszi, György Takács, Attila Tihanyi, Tamás Bárdi: Inter-speaker Synchronization in Audiovisual Database for Lip-readable Speech to Animation Conversion, 10th International Conference on Text, Speech and Dialogue, Plzen 2007

- Gergely *Feldhoffer*, Tamás Bárdi, György Takács and Attila Tihanyi: Temporal Asymmetry in Relations of Acoustic and Visual Features of Speech, 15th European Signal Processing Conf., Poznan, Poland, September 2007

- Takács, György; Tihanyi, Attila; *Feldhoffer*, Gergely; Bárdi, Tamás; Oroszi Balázs: Synchronization of acoustic speech data for machine learning based audio to visual conversion , 19th International Congress on Acoustics, Madrid, 2-7 september 2007

- Gergely *Feldhoffer*: Speaker Independent Continuous Voice to Facial Animation on Mobile Platforms, PROCEEDINGS 49th International Symposium ELMAR, Zadar, 2007.

Magyar publikációk

- Bárdi T., *Feldhoffer* G., Harczos T., Srancsik B., Szabó G. D: Audiovizuális beszéd-adatbázis és alkalmazásai, Híradástechnika 2005/10

- *Feldhoffer* G., Bárdi T., Jung G., Hegedûs I. M.: Mobiltelefon alkalmazások siket felhasználóknak, Híradástechnika 2005/10.

- Takács György, Tihanyi Attila, Bárdi Tamás, *Feldhoffer* Gergely, Srancsik Bálint: Beszédjel átalakítása mozgó száj képévé siketek kommunikációjának segítésére, Híradástechnika 3. 2006

- Takács György, Tihanyi Attila, Bárdi Tamás, *Feldhoffer* Gergely, Srancsik Bálint: MPEG-4 modell alkalmazása szájmozgás megjelenítésére, Híradástechnika 8. 2006

- *Feldhoffer* Gergely, Bárdi Tamás: Látható beszéd: beszédhang alapú fejmodell animáció siketeknek, IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2006.

Irodalomjegyzék

[1] J. Kewley J. Beskow, I. Karlsson and G. Salvi. Synface - a talking head telephone for the hearing-impaired. *Computers Helping People with Special Needs*, pages 1178–1186, 2004. 1.1.1

[2] M. De Smet S. Al Moubayed and H. Van Hamme. Lip synchronization: from phone lattice to PCA eigen-projections using neural networks. In *Proceedings of Interspeech 2008*, Brisbane, Australia, Sep 2008. 1.1.1

[3] T. Bárdi G. Feldhoffer Gy. Takács, A. Tihanyi and B. Srancsik. Speech to facial animation conversion for deaf customers. In *4th European Signal Processing Conf.*, Florence, Italy, 2006. 1.1.1, 2.1.1

[4] J. Yamagishi G. Hofer and H. Shimodaira. Speech-driven lip motion generation with a trajectory HMM. In *Proc. Interspeech 2008*, pages 2314–2317, Brisbane, Australia, 2008. 1.1.1

[5] O. N. Garcia R. Gutierrez-Osuna P. Kakumanu, A. Esposito. A comparison of acoustic coding models for speech-driven facial animation. *Speech Communication*, 48:598–615, 2006. 1.1.1

[6] V. Libal P. Scanlon, G. Potamianos and S. M. Chu. Mutual information based visual feature selection for lipreading. In *in Proc. of ICSLP*, 2004. 1.1.1

[7] A. Robinson-Mosher E. Sifakis, A. Selle and R. Fedkiw. Simulating speech with a physics-based facial muscle model. *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, pages 261–270, 2006. 1.1.1

[8] A. Tihanyi G. Feldhoffer and B. Oroszi. A comparative study of direct and asr based modular audio to visual speech systems (accepted). *Phonetician*, 2010. 3.1

[9] B. Németh P. Mihajlik, T. Fegyó and V. Trón. Towards automatic transcription of large spoken archives in agglutinating languages: Hungarian ASR for the MALACH project. In *Speech and Dialogue: 10th International Conference*, Pilsen, Czech Republic, 2007. 3.1.1

[10] L. Czap and J. Mátyás. Virtual speaker. *Híradástechnika Selected Papers*, Vol LX/6:2–5, 2005. 3.1.1

[11] Gy. Takács G. Feldhoffer, T. Bárdi and T. Tihanyi. Temporal asymmetry in relations of acoustic and visual features of speech. In *15th European Signal Processing Conf.*, Poznan, Poland, 2007. 3.2

[12] T. Bárdi-G. Feldhoffer B. Srancsik G. Takács, A. Tihanyi. Database construction for speech to lipreadable animation conversion. In *ELMAR Zadar*, pages 151–154, 2006. 3.3

[13] Hedvig Kjellström and Olov Engwall. Audiovisual-to-articulatory inversion. *Speech Communication*, 51(3):195–209, 2009. 3.3

[14] G. Feldhoffer. Speaker independent continuous voice to facial animation on mobile platforms. In *49th International Symposium ELMAR*, Zadar, Croatia, 2007. 3.3

[15] G. Feldhoffer and B. Oroszi. An efficient voice driven face animation method for cyber telepresence applications. In *2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, Bratislava, Slovak Republic, 2009. 3.4