Multidiszciplináris tudományok, 4. kötet. (2014) 1. sz. pp. 67-76.

AZ AKUSZTIKUS ÉS VIZUÁLIS JEL ASZINKRONITÁSA A BESZÉDBEN

Dr. Czap László
egyetemi docens, Miskolci Egyetem, Villamosmérnöki Intézet, Automatizálási és Infokommunikációs Tanszék
Pintér Judit Mária
PhD hallgató, Miskolci Egyetem, Villamosmérnöki Intézet, Automatizálási és Infokommunikációs Tanszék
3515 Miskolc, Miskolc-Egyetemváros, e-mail: czap@uni-miskolc.hu
3515 Miskolc, Miskolc-Egyetemváros, e-mail: pinterjm@uni-miskolc.hu

Összefoglalás

Ismert megfigyelés, hogy az akusztikus és a vizuális jel időbeli elcsúszása befolyásolja a beszéd érthetőségét. Számos publikáció tárgyalja, hogy az érthetőséget eltérően befolyásolja, hogy a hang késik vagy siet a képhez viszonyítva. Ezek az eredmények rendszerint szubjektív teszteken alapulnak, és nem adnak magyarázatot a különbségre. Nem világos, hogy a jelenség percepciós vagy produkciós eredetű. Ebben a cikkben egy kétmódusú, gépi beszéd felismerési kísérletben tanulmányozzuk az audiovizuális aszimmetriát, kiküszöbölve a kísérleti alanyok percepciós tapasztalatát. Az eredményeket az audiovizuális beszédszintézis természetességének fokozására használjuk.

Kulcsszavak: *beszéd percepció, beszéd produkció, multi modalitás, audiovizuális beszédszintézis*

Abstract

The temporal synchrony of auditory and visual signals is known to affect the perception of audio visual speech. Several papers have discussed the asymmetry of acoustic and visual timing cues. These results are usually based on subjective intelligibility tests and the reason is remained obscure. It is not clear that the observation is perception or production origin. In this paper the effect of audio-visual asynchrony is studied in an automatic bimodal speech recognition task, eliminating the perception expertise of observers. Results are utilized to improve naturalness of audio visual speech synthesis.

Keywords: *speech perception, speech production, multimodality, audio visual speech synthesis*

1. Bevezetés

Az emberi beszédfelismerést nagymértékben befolyásolja a beszélő látványa. A beszélő szájmozgásának megfigyelése javítja a beszéd érthetőségét különösen zajos környezetben és a hallássérültek esetében. Amikor kétmódusú – akusztikus és vizuális – beszédfelismerést végzünk, jobban toleráljuk az akusztikus jel késését, mint a sietését, mivel ez utóbbi nem fordul elő természetes körülmények között. Magyarázhatja a különbséget percepciós

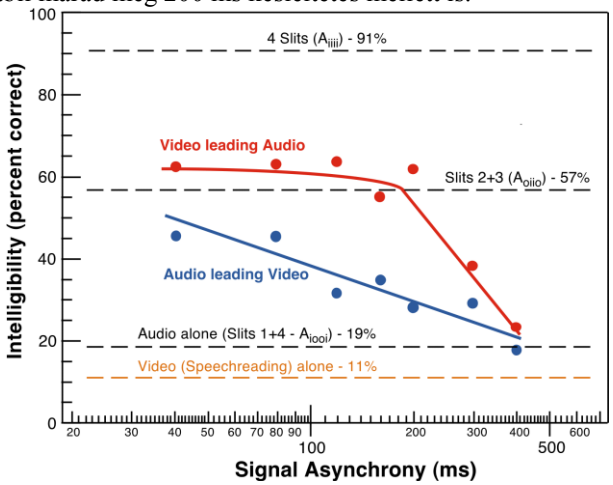
tapasztalatunk: A hang terjedési sebessége messze elmarad a fényétől. Figyelembe véve a hang 330 méter/szekundum terjedési sebességét, 13,2 méter távolságról hallgatva a beszélőt, egy TV képkocka (40 ms) a kép és a hang között az időkésleltetés. A hangerősítők és kivetítők világában ennél nagyobb késleltetésű hanggal is gyakorlatra tehetünk szert a kétmódusú felismerésben. Ez a percepciós tapasztalat magyarázhatja, hogy a hang késése kevésbé rontja a beszédértést, mint a sietése.

A produkció oldali indokokat erősíti az ismert megfigyelés, hogy az artikulációs mozgások megelőzik a hang megjelenését, a hangképző szervek már előre készülnek a következő hang kimondására.

2. Korábbi kutatási eredmények

McGrath és Summerfield az akusztikus és a vizuális jel integrációját vizsgálta, a két modalitás közötti időeltolódás hatásaira koncentrálva [1]. Mondatok audiovizuális felismerését vizsgálták az akusztikus jel késleltetésének függvényében, normál hallású, gyakorlatlan megfigyelőkkel. Az eredeti hangot a hangszalagok záródásának pillanatához szinkronizált négyszögimpulzusokkal helyettesítették. Arra jutottak, hogy a kísérletben résztvevők nagy része – függetlenül attól, hogy a szájról olvasásban jeleskedtek-e – nem volt érzékeny a hang késleltetésére. A mondatok felismerése kb. 40 ms késleltetésig nem romlott.

Grant és Greenberg [2] kísérletekkel igazolta, hogy az akusztikus és vizuális modalitás integrálási képességünket erősen befolyásolja, hogy a hang késik vagy siet a videó jelhez képest. Amikor a hang megelőzi a képet, a modalitások integrációja meredeken csökken még kis időkülönbségek esetében is. Ha a hang lemarad a képtől, az audiovizuális integrálás viszonylag stabil marad még 200 ms késleltetés mellett is.

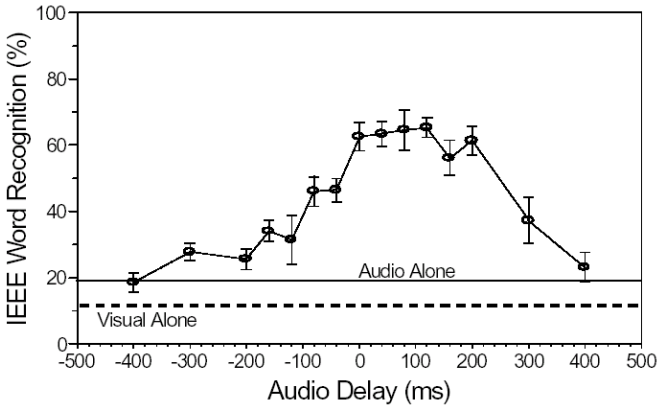


1. ábra. *9 fős kísérlet átlagos érthetőségi eredményei az akusztikus és vizuális jel időbeli elcsúszásának függvényében, az akusztikus jel sietése (Audio leading video) és késése (Video leading audio) esetén. A csak akusztikus és csak vizuális jel alapján mért felismerési eredményeket lent a szaggatott vonalak jelzik.[2]*

Grant és Greenberg az érthetőségi vizsgálat eredményeinek aszimmetriájára a csak akusztikus és csak vizuális jel absztrakciós szintjének különbségében keresi a magyaráza-

tot. Amikor az akusztikus jel siet a videó előtt, a beszédfeldolgozás időállandója viszonylag rövid, a hangok időtartamának nagyságrendjében van, 40-120 ms, mivel a hangzás és jelentés pusztán az akusztikus jelből is megfejthető. Amikor a video siet a hang előtt, a jelek integrációja valószínűleg hosszabb, mivel a vizuális információ csak a hangképzés helyének meghatározásában segít, így a feldolgozás ideje a szótagok időtartamának nagyságrendjébe esik, kb. 200 ms. Grant és Greenberg szerint az időben elcsúszott jelek esetében a korábban megjelenő gerjesztés határozza meg az absztrakció szintjét, végső soron a feldolgozás időtartamát.

Ugyanezeknek az eredményeknek egy másik interpretációját mutatja a 2. ábra.



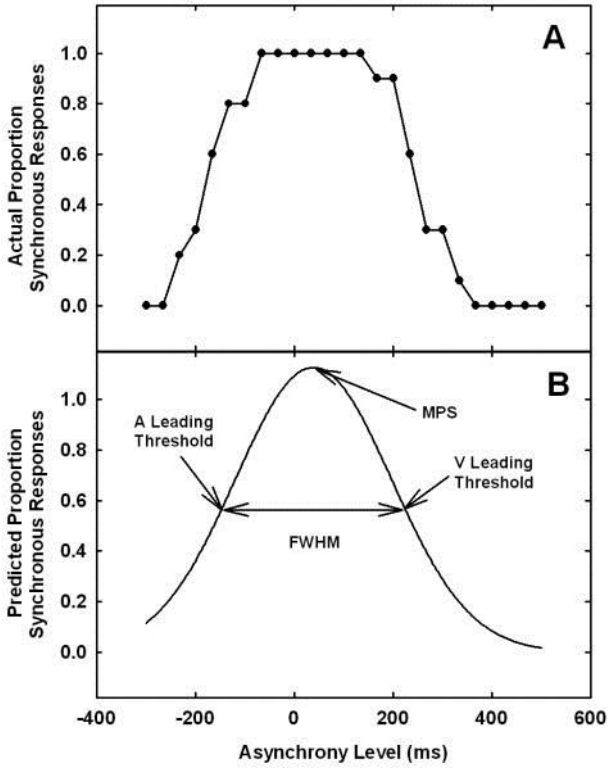
2. ábra. *IEEE mondatok átlagos audiovizuális érthetősége az akusztikus késleltetés függvényében. Megfigyelhető, hogy az 50 ms-os audió sietés és a 200 ms-os audió késletetés között az audiovizuális érthetőség jelentősen meghaladja a csak akusztikus és csak vizuális érthetőséget. [3]*

Más érvek is erősítik a feltételezést, hogy az audiovizuális aszimmetria percepció eredetű: Grant et al. [3] demonstrálta, hogy az aszimmetria jelen van függetlenül attól, hogy a feladat szavak, vagy szótagok felismerése, vagy csupán annak érzékelése, hogy az akusztikus és vizuális gerjesztés szinkronban van-e. Ez azt sugallja, hogy amint aszinkronitás lép fel, az információforrások integrálása romlik. A szerzők különböző sávszűrőket alkalmaztak a beszédre, és arra kérték a kísérlet alanyokat, hogy mondják meg, szinkronban van-e a hang a képpel.

Brungart et al. [4] azt állítják, hogy az egyik összetevője az audiovizuális jel integrálási képességének az akusztikus és vizuális modalitás feldolgozási ideje az érzékelő rendszerünk által. Neuro-fiziológiai megfigyelés szerint a vizuális visszajelzés információja 50 ms-mal az inger után éri el a hallókérget, míg az akusztikus gerjesztés 11 ms alatt. Ez azt sejteti, hogy a kétféle inger egyszerre ér a hallókéregbe, ha a kép 40-50 ms-mal megelőzi a hangot.

Levitin et al. [5] szerint a kísérleti pszichológia egyik legrégebbi kérdése az egyidejű események érzékelése, különösen, ha eltérő érzékszervektől (látás/hallás, hallás/tapintás) érkeznek. Milyen mértékű időeltolódás vezet a két esemény egymásra következő jellegének érzékeléséhez? Cikkükben adatokat közölnek modalitások közötti szinkron érzékeléséről. A sorrendiség érzékelésére alacsonyabb küszöböt állapítottak meg kevésbé mesterkélt körülmények között. Megállapításaikat az idő, a sorrendiség és az észlelés tekintetében tartják meghatározónak

Hay-McCutcheon et al. [6] azt találta cochleáris implantátumot használó kísérleti alanyokkal végzett kísérletek alapján, hogy a kísérletben résztvevők kora erősebben befolyásolta az aszinkron audiovizuális beszédértést, mint a halláskárosodás foka, amit az implantátummal részben korrigáltak. Vizsgálataikból arra is következtettek, hogy a középkorú és idős emberek beszéd felismerési és szinkronitás észlelési eredményei attól is függnek, hogy a felismerendő egység szó, vagy mondat.



3. ábra. *Audiovizuális aszinkronitás függvény. Az A panel mutatja a megfigyelés eredményét, a B panel az erre fektetett Gauss görbét. A szinkron válaszok arányát látjuk az időeltolódás függvényében. Az "A-leading threshold" azt az időeltolódást mutatja, ahol az akusztikus jel sietése esetén a minimum és a maximum számtani közepét metszi a görbe. A "V-leading threshold" azt az időeltolódást mutatja, ahol a vizuális jel sietése esetén a minimum és a maximum számtani közepét metszi a görbe. FWHM jelzi a kettő közötti tartományt. A szinkronitás középértéke az MPS.[6]*

3. A kísérlet módszere

Audiovizuális beszéd felismerési kísérletet végeztünk az akusztikus és vizuális jel időeltolásának függvényében automatikus rejtett Markov modell (HMM) alapú beszédfelismerővel.

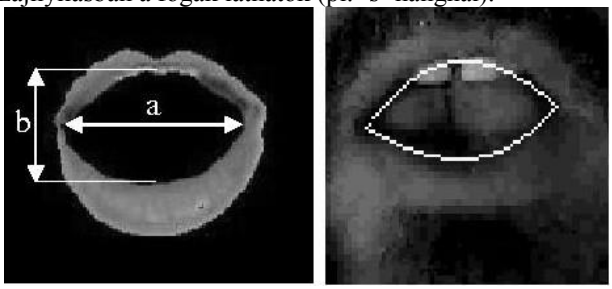
3.1. Az adatbázis

Házi videó berendezés és közönséges PC mikrofon felhasználásával felvett hang- és videó anyagon folyt az elemzés. Az audiovizuális adatbázis számok és dátumokban előforduló szavak, illetve szófűzérek egy beszélő által bemondott mintáit tartalmazza. A felismerés alapja a hangpár – az egyik hang közepétől a következő hang közepéig terjedő beszédszakasz. Az audiovizuális adatbázisban 163 hangpár fordul elő, mindegyik legalább ötször. 536 szó és szófűzér alkotta a tanító alakzatokat. A tesztelésre 35 szófűzér szolgált, amelyek összesen 1243 hangpárból álltak.

3.2. Vizuális előfeldolgozás

Az audiovizuális adatbázis videó felvételei egyféle beállítással, az ülő helyzetben természetes fejmozgás mellett, speciális világítási előírások nélkül készültek. Az általánosabb alkalmazhatóság érdekében a színinformációt nem akartuk felhasználni, ezért a képek feldolgozása a színes képek intenzitás képpé alakításával kezdődött. A vizuális előfeldolgozás keretében a videó felvételből ki kell vágni a szótárelemhez tartozó szakaszt. A hangot hangfájlba másolva, a videó jelet képkockánként elmentve szétválasztjuk a két modalitást. A váltottsoros letapogatású video egy képkockája két félképet tartalmaz – a páros és páratlan sorokba elmentve a félképeket. A képfeldolgozás eszközeinek igénybe vételével a kép félképekre alakítható vissza. A képkeretek félképekre bontásának következtében a vizuális jel mintavételi időköze 40 ms-ról 20 ms-ra csökken, vagyis az eredeti, másodpercenkénti 25 kép sebesség 50-re nő.

Az ajkak belső méretét reprezentáló ajakszélesség (a) és ajaknyílás (b) geometriai adatai és a szájnyílás világossága szolgálták a vizuális információt. A szájnyílás világosságát a belső szájsarkak és a szájnyílás felső illetve alsó széle közé rajzolt parabolák által határolt terület intenzitás értékével közelítettük, amelyet a k intenzitás faktor tartalmaz. A szájnyílás világossága minimális, ha a nyelv hátul helyezkedik el, közepes elülső nyelvállásnál, maximális, amikor a szájnyílásban a fogak láthatók (pl. ‘s’ hangnál).



4. ábra. *A vizuális előfeldolgozás paraméterei. a: szélesség, b: nyitás, k: a szájüreg átlagos világossága*

3.3. Akusztikus lényegkiemelés

Magától értetődik, hogy a szinkronitás végett a hangnál is célszerű a 20 ms-os keretidőt választani. A 22 050 Hz-es mintavételi frekvenciájú hang 512 mintája alkot egy feldolgozási egységet. A 20 ms-os lépésköz 441 mintának felel meg, ezzel tolódik el a feldolgozás minden lépésénél az időablak, ami 16%-os átlapolódást jelent.

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

Az akusztikus lényegkiemelés megvalósítására hozzáférhető, elfogadott megoldásként az MFCC (Mel-frequency cepstral coefficients) paraméterek meghatározásának MATLAB implementációját választottuk [8]. Keretenként 12 együttható mellett a differenciális jellemzőket (Δ) és ezek differenciáját (ΔΔ) használtuk fel.

3.4. Az akusztikus és vizuális modalitás integrálása

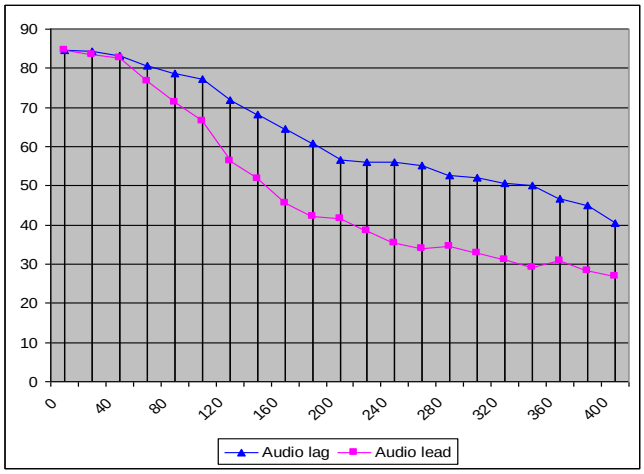
Az audiovizuális beszédfelismerés másik kulcskérdése a vizuális lényeg-kiemelés mellett, hogy az akusztikus és vizuális jel hogyan integrálható a legjobb felismerés érdekében. A humán beszéd felismerési kísérletek azt mutatják, hogy az integrált eredmények minden körülmények között felülmúlják mind az akusztikus, mind a vizuális egymódusú arányokat [9]. Az audiovizuális beszédfelismerés az akusztikus beszédfelismerésből fejlődött ki, annak eredményeire épít. Kézenfekvő megoldásnak tűnt, hogy az akusztikus és vizuális jellemzőket konkatenálva a bevált rejtett Markov modell alapú felismerőt a megnövelt dimenziójú paraméterekkel tanítjuk és teszteljük.

3.5. A mérés

A szófűzér felismerési kísérletben a helyes hangpár felismerési arányt mértük az akusztikus és vizuális jel egymáshoz képest változtatott késleltetésének függvényében. A késleltetés lépésköze 20 ms volt, és -400 ms-tól +400 ms-ig terjedt.

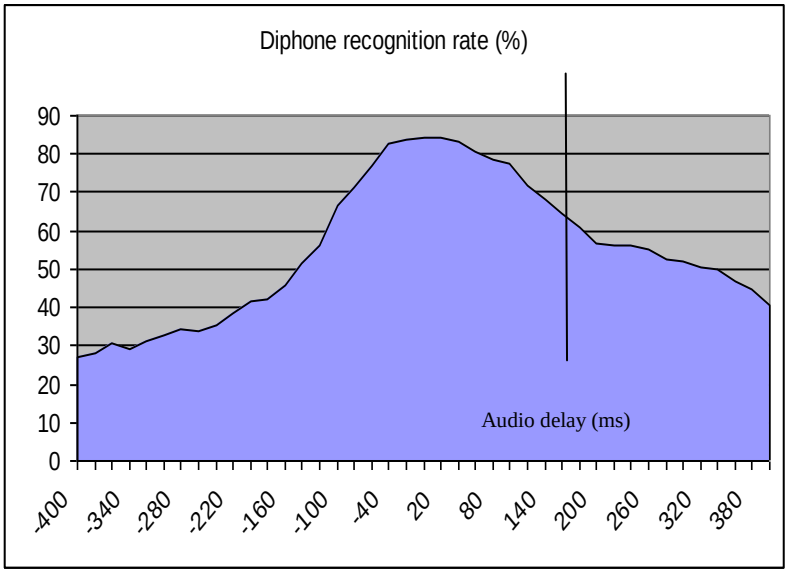
4. Eredmények

A hangpár felismerési arányok monoton csökkennek a késleltetés növekedésével és aszimmetriát mutatnak, ahogy az 5. ábrán látható.



5. ábra. *A hangpár felismerési arány csökkenése az akusztikus jel késése (lag ms), illetve sietése esetén (lead ms).*

A 6. ábra ugyanezeket az eredményeket más ábrázolásban mutatja, jobban láthatóvá téve az aszimmetriát.



6. ábra. *Hangpár felismerési arány az akusztikus jel késletetésének függvényében. A negatív érték a vizuális jel késését jelzi.*

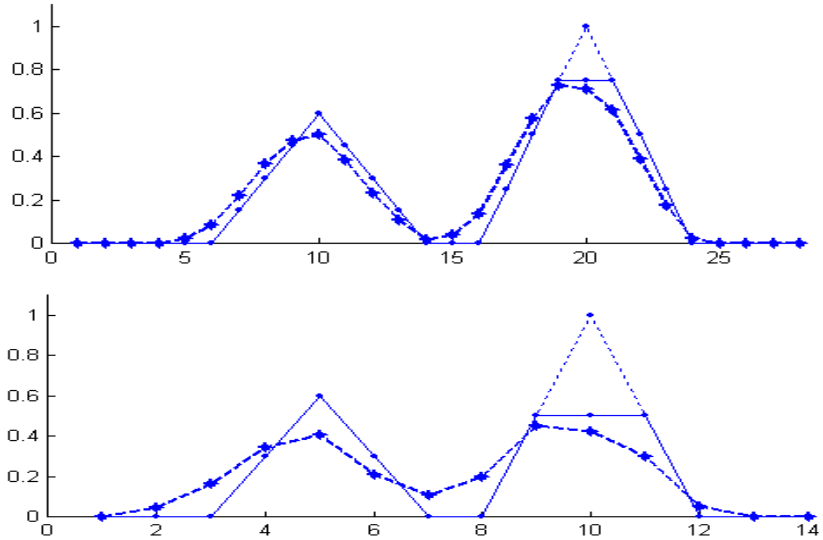
A hangpár felismerési kísérletben mért akusztikus és vizuális aszinkronitás beépítésével jelentősen javítottuk az audiovizuális beszédszintézis minőségét. A kimondás megkezdését kb. 300 ms időtartamú csend előzi meg. Ez alatt az idő alatt a levegővételt az ajkak megnyitása imitálja. Ezután az ajkak alaphelyzetéből elkezdődik az első domináns vizéma – ami a beszéd legkisebb akusztikus egységének a fonémának (hangzó) vizuális megfelelője – kialakítása. Ezzel a kiegészítéssel – amit előartikulációnak nevezhetünk – már az első hang megszólalása előtt kialakul az ajakforma, hasonlóan a természetes kiejtéshez.

A természetes vagy szintetizált beszédhez szinkronizálás folyamán különböző sebességű beszéddel szembesülhetünk. Lassú beszédnél a vizémák jellemzői megközelítik névleges értéküket, gyors beszédnél az artikuláció elnagyoltabb. A rugalmas csoportba sorolt jellemzőkre is igaz, hogy gyors beszédnél a lekerekítés erősebb. A rugalmas jellemzők kialakítására alkalmas a medián szűrés: A szűrésben résztvevő mintákat nagyság szerint sorba rendezve a középső lesz a szűrt érték. A szűrés három mintára történik. Egy jellemző időfüggvénye három lépésben alakul ki:

- A domináns és rugalmas vizémák kitüntetett pontjai között – a határozatlanok kihagyásával – lineáris interpoláció.
- A rugalmas vizémák környezetében medián szűrés. Ez kevesebb minta – gyors beszéd – esetén nagyobb csúcslevágást okoz.
- Az így kapott értékeken még egy simítás, amely az aktuális, a két megelőző és a követő mintákat érinti. A szűrt érték a négy minta súlyozott összege. A súlyozás állandó, nem függ a beszéd sebességétől. A simító szűrés egyrészt finomítja a mozgást, másrészt gyors beszédnél jobban lekerekíti a csúcsokat és előre csúszt-

atja a vizuális jelet. A szintetizált beszéd analízise alapján a szűrés hatása előre erősebb (két keret) mint hátra (egy keret).

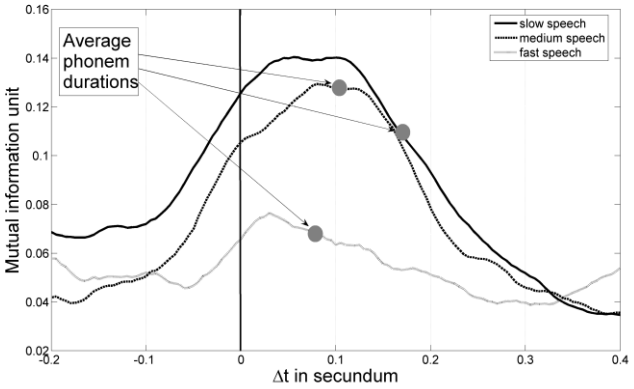
A 7. ábrán gyors és lassú beszédnél egy generált példán követhetjük a medián szűrés és a simítás hatását pl.: a nyelv vízszintes helyzetére. A példában a lassú beszéd kétszer annyi keretből áll, mint a gyors kimondás. Az ábrán jól követhető a gyors beszédnél érvényesülő lekerekítés, a medián szűrés és a simítás hatása egyaránt. A cikkünkben tárgyalt vizuális előkészítést a simítással implementáltuk az audiovizuális beszédszintetizátorban.



7. ábra. *Példa a domináns (1. csúcs) és rugalmas (2. csúcs) jellemző szűrésére és a lassú (fent) illetve gyors (lent) beszéd simítására. A lineáris interpoláció eredménye (...), a medián szűrés (___) és simítás (---) után.*

5. Értékelés

A gépi beszéd felismerési eredmények, amelyeket a hang és a kép időbeli eltolódása függvényében kaptunk, mentesek a kísérletben résztvevők beszédérzékelési gyakorlottságától, ami az emberi beszéd felismerési kísérletekre elkerülhetetlenül hatást gyakorol. A tapasztalat egyezik a Feldhoffer et al. [10] által kapottakkal, akik az audiovizuális paramétereket kölcsönös információ vizsgálatával tanulmányozták. Analizálták az akusztikus és vizuális jellemzők finom szerkezetét, hogy a beszédet animációs adatokká konvertáló rendszerük hatékonyságát javítsák. A főkomponens analízis (PCA) célja dimenziók (attribútumok) olyan új halmazának a keresése, mely jobban tükrözi az adatok variabilitását. Az akusztikus (MFCPCA) és vizuális (FacePCA) főkomponensek kölcsönös információit különböző időeltolással vizsgálták. Azt találták, hogy a professzionális jeltolmács artikulációs mozgása akár 100 milliszekundummal is megelőzheti az akusztikus jel változását. A tapasztalt időeltolódás figyelembe vételével javítható a hangvezérelt animáció.



8. ábra. *Eltolt FacePCA1 és MFCPCA kölcsönös információ. Pozitív t hang késleltetést jelent [8].*

A különböző beszédtempó hatásaira vonatkozó megállapításokat is tesz a cikk, ahogy Brungart et al is [4].

6. Összefoglalás

Az akusztikus és vizuális jelek érzékelésének aszimmetrikus jellegére vonatkozó megállapításaink hasonlóak a korábbi audiovizuális beszéd felismerési kísérletek kimeneteihez. Az emberi beszédfelismerés sajátságaitól független eredmények arra utalnak, hogy az aszinkron jelleg a beszédprodukcióban is jelen van. Ellentétben a percepción alapuló közleményekkel, a gépi beszéd felismerési eredmények kiküszöbölik a humán kísérleti alanyok beszédérzékelési tapasztalatait. A korábbi percepciós kísérleti eredményekkel tapasztalt erős korreláció azt sugallja, hogy azokban a megfigyelésekben a beszédprodukció sajátosságai is jelen vannak.

7. Köszönetnyilvánítás

A kutató munka a Miskolci Egyetem stratégiai kutatási területén működő Mechatronikai és Logisztikai Kiválósági Központ keretében, a TÁMOP-4.2.2. C-11/1/KONV-2012-0002 jelű projekt részeként az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg.

8. Irodalom

[1] McGrath, M, Summerfield, Q.: *ntermodal timing relations and audio-visual speech recognition by normal-hearing adults*, Journal of the Acoustical Society of America, 678–685, 1985

[2] Grant, K. W., and Greenberg, S.: *Speech intelligibility derived from asynchronous processing of auditory-visual information*, AVSP 2001, 132-137, Scheelsminde, Denmark, September 7-9, 2001.

[3] Grant, K.W., van Wassenhove, V., Poeppel: *Discrimination of Auditory-Visual*

Synchrony AVSP 2003, 31-35, Jorioz, France, 2003

[4] Brungart, D. S., Iyer, N, Simpson, B. D, van Wassenhove, V.: *The effects of temporal asynchrony on the intelligibility of accelerated speech*, AVSP 2008, Moreton Island, 19-24, Australia, 2008

[5] Levitin, D. J., Mathews, M. V., and MacLean, K.: *The Perception of Cross-Modal Simultaneity*, International Journal of Computing Anticipatory Systems, 323-329, Belgium, 1999

[6] Hay-McCutcheon, M. J., Pisoni, D. B., and Hunt, K. K.: *Audiovisual Asynchrony Detection and Speech Perception in Hearing-Impaired Listeners with Cochlear Implants: A Preliminary Analysis*, Int J. Audiol. 48(6): 321–333, 2009,

[7] Grant, K. W., Greenberg, S., Poeppel D, van Wassenhove, V.: *Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing*, Seminars in Hearing, 3:241–255, 2004

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

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

[10] Feldhoffer, G., Bárdi, T., Takács, G., Tihanyi, A.: *Temporal Asymmetry in Relations of Acoustic and Visual Features of Speech*, 15th European Signal Processing Conf., 2341-2345, Poznan, Poland, September 2007