

A nyelvkontúr automatikus követése ultrahangos felvételeken

ZHAO LU – CZAP LÁSZLÓ

Miskolci Egyetem, Automatizálási és Infokommunikációs Intézet, Miskolc

qgezha@uni-miskolc.hu, czap@uni-miskolc.hu

Bevezetés

Az utóbbi évtizedekben a nyelv nem látható mozgásának követésére többféle eszköz is megjelent, ezek egyike az ultrahangos képalkotás. Lehetőséget kaptunk, hogy az MTA–ELTE Lendület Lingvális Artikuláció Kutatócsoport SonoSpeech rendszerén felvételeket készítsünk. A mérési módszer előnye, hogy gyors mozgás követésére is alkalmas, a kép és a hang szinkronizálható, minimális kényelmetlenséget okoz a beszélőnek, és nem éri őket káros sugárzás. Hátránya, hogy nem ad teljes háromdimenziós képet, csak hosszanti vagy keresztirányú metszetet mutat, és a nyelv hegye nem látható. A nyelvmozgás követésére a hosszanti (midszagittális) metszetet használtuk. A felvételeket kínai Shaanxi Xi'an nyelvjárású anyanyelvi női beszélővel készítettük.

Számos tanulmány hitelesen alátámasztja, hogy az emberi beszéd fiziológiai folyamatairól nyert vizuális információ nagymértékben elősegíti a beszédképzés komplex mechanizmusának megértését, és ezen keresztül a beszéd-szintézis módszereinek hatékony fejlesztését (Barnaud et al. 2019). A napjainkban rendelkezésünkre álló radiológiai és monitorozó eljárások – mint a mágneses rezonancia-képalkotás (MRI), a komputertomográfia (CT), az ultrahang (UH), az elektropalatográfia (EPG), az elektromágneses artikulográfia (EMA) vagy az elektrolottográfia (EGG) – nélkülözhetetlen szerepet játszanak az artikuláció dinamikus tulajdonságainak megismerésében. A képalkotó technikák segítségével kapott morfológiai és geometriai adatok felhasználásával ugyanis feltérképezhetők az adott beszédjelhez tartozó artikulációs mozgások, ami alapvető fontosságú egy artikulációt utánzó beszélő fej paraméterezésénél.

Az így megvalósított rendszer jól használható például a beszédterápiában, a nem anyanyelvi nyelvtanulási tréningek kialakításában vagy a néma beszéd megszólaltatásához szükséges szintetizátorok konstrukciójában.

Jelen publikációban ultrahangfelvételek feldolgozásával és részleges elemzésével foglalkozunk. Az ultrahang kényelmesen és egyszerűen hozzáférhető módszer, hiszen az orvosi központokban korlátozottan elérhető MRI és CT berendezésekkel szemben egy hordozható sisakhoz rögzített ultrahangfej elegendő a vizsgálatokhoz, így az elemzésekhez szükséges kép- és hanganyag rugalmasan elkészíthető anélkül, hogy a beszélőt káros sugárzás terhelné. Előnyös továbbá az is, hogy az eljárással nagy felbontású (közel ezer képpont egy radiális metszetben) és nagy sebességű (80–85 kép másodpercenként) felvételek produkálhatók. A jó térbeli felbontás elengedhetetlen ahhoz, hogy a nyelv alakja a lehető legélesebben kirajzolódjon, a jó időbeli felbontás pedig stabilan támogatja a folyamatos beszéd közben létrejövő gyors koartikulációs változások tanulmányozásának lehetőségét. Nem szabad figyelmen kívül hagynunk azt sem, hogy az ultrahang kiváltképp alkalmas a folyamatos beszéd elemzésére, hiszen a vokális traktus letapogatásához szükséges szkennelési idő töredéke annak, amit pl. az MRI képalkotás igényel. Mindezek mellett nehezíti a munkát az a körülmény, hogy az ultrahang a háromdimenziós alaki információk gyűjtését segítő MRI és CT felvételekkel ellentétben csak a nyelv helyzetéről ad információt a kétdimenziós szagittális síkban, így például a szájpad és a nyelv hegyének körvonala nem jelenik meg a képen. További technikai problémaként merül fel az, hogy a lehető legnagyobb pontossággal meg kell határoznunk a nyelv felszíni kontúrját, ami nem triviális feladat.

A nyelvkontúr követésének célja a vizémák – a fonémák vizuális megfelelői – statikus és dinamikus leírása. A statikus adatok a kulcspozíciók megadását jelentik. Az egyes jellemzők dominanciájára következtethetünk a különböző hangkörnyezetben mért alakzatok szórásából. Minél nagyobb a szórás, annál kisebb a jellemző dominanciája. A dinamikus leírás minden jellemzőre külön-külön meghatározza a kulcspozíciók közötti interpolációt. A dinamikus modellezés része az artikuláció alkalmazkodása a beszédtempóhoz, mivel a kulcspozíciók csak lassú beszédnél érik el névleges értékeiket. Gyors beszédnél a szélsőértékek lekerekednek. Az artikuláció modellezésének végső célja a magyar nyelvre több mint egy évtizedes finomítás után elfogadhatóan artikuláló beszélő fej adaptálása a közel negyvenmillió ember által beszélt kínai Shaanxi Xi'an nyelvjárásra. A dialektus kiejtését utánzó virtuális bemondó alapját képezheti egy hallássérültek beszélni tanulását segítő beszédasszisztens rendszernek, amely a magyar nyelvű változat adaptálásával kifejleszthető (Czap et al. 2019). A beszédasszisztens több szolgáltatása nyelvfüggetlen (pl.: a beszédhang vizualizálására alkalmas vonalas ábra, vagy a prozódia ábrázolása).

Az ultrahangos felvételek automatikus kiértékelése kulcskérdés, mivel a képkockák száma százazres nagyságrendű, manuális feldolgozásuk elképzelhetetlen. A nyelvkontúr követésének nehézségét az adja, hogy a képek háttere zajos, a kontúr vonala pedig nem folytonos.

A nyelvkontúr követésének pontosságát az ismert eljárásokkal összevetve vizsgáljuk. Külön kérdést vet fel a kapott határvonal leírása. A frekvencia-tartománybeli leírás alkalmas a kontúr megrajzolására és simítására, de az artikulációs jellemzőkkel nehezen hozható kapcsolatba. A görbe a vízszintes irányú 20, 40, 60 és 80 %-os pozíciójánál vett négy pontjával jellemezhetjük a nyelvhat aktuális helyzetét. A koordinátákból képezhető kombinációk alkalmasak lehetnek különböző artikulációs jellemzők becslésére (a nyelv vízszintes-függőleges pozíciója, koronális-dorzális jelleg), és alapadatokat szolgáltatnak a beszélő fej paraméterezéséhez.

A nyelvkontúr kijelölése történhet manuálisan vagy automatikus algoritmusok segítségével. Munkánk során MATLAB-környezetben megírt programjaink segítségével végeztük el az ultrahangfelvételek feldolgozását, amelynek keretében dinamikus programozásra alapozva illesztettünk segédgörbét a nyelv felszínére. A kontúrkövetési eredmények verifikálását Csapó Tamás és Steven Lulich cikke alapján végeztük (2015). A segédgörbe pozíciója a nyelvmozgás során dinamikusán változik, így olyan adathalmazhoz jutottunk, amelynek elemei hely és idő szerint is változnak.

A cikk első szerzőjeként megjelölt kínai nyelvjárást beszélő adatközlő ultrahangfelvétele mellett a hangzó beszéd amplitúdójának időfüggvénye is rendelkezésre állt, ahol a hanghatárok kijelölésével jól elkülöníthetővé váltak az egyes beszédhangok. Ennek megfelelően az adott ultrahangfelvétel és a hozzá tartozó amplitúdó-idő grafikon együttes analízise lehetővé tette, hogy a kiszámolt szagittális távolságokat konkrét hangokhoz, illetve hangátmenetekhez rendeljük hozzá. A kutatómunka végső célja, hogy az ultrahangfelvételekből nyert geometriai adatok alapján vizuális beszédszintézist valósítsunk meg háromdimenziós animációval (Czap–Zhao 2017).

A nyelvkontúr követésének módszere

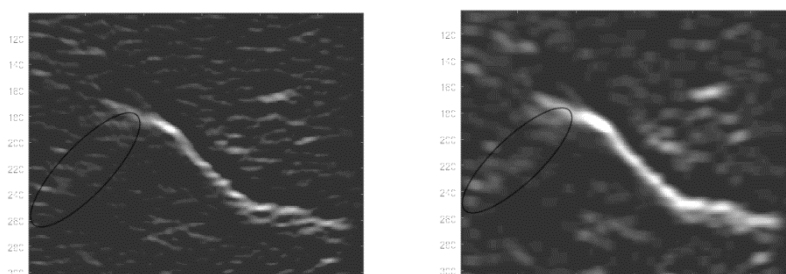
Dinamikus programozáson alapuló algoritmust dolgoztunk ki a nyelvhat mozgásának követésére. A nyelvkontúrt az ultrahangképen a nyelv és a fölötte lévő levegő határán visszaverődött sugárzás hozza létre. Az így kapott világos sáv alsó határán jelöljük ki a nyelvhat vonalát. Az elemzés a kép előfeldolgozását igényli. Az eljárás extrém méretű élkiemelő-átlagoló alakzattal egyszerre kezeli a folytonossági hiány és a zaj problémát. A szűrés után kapott képen azt a görbét keressük, amelynek kumulált világossága a kép bal szélétől a jobb széléig maximális. Az így kapott görbe pontjai követik

-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1

2. ábra: Átlagolást és éldetektálást végző operátor zajos képre

A képfeldolgozásban az éleket két merőleges gradiens képzésével és az eredő gradiens kiszámításával szokás végezni. A hamis élek elkerülése végett csak a függőleges irányú változást vizsgáljuk, és csak a felülről lefelé irányú csökkenést vesszük figyelembe.

A 3. ábra bal oldala az élkiemelés eredményét mutatja. Látható, hogy a kép bal oldali harmadában a nyelv kontúrvonala még mindig nem folytonos. Egy további átlagoló szűrővel (8x8-as mátrix) javíthatunk a folytonosságon. Az ábra jobb oldalán ennek eredményét látjuk.

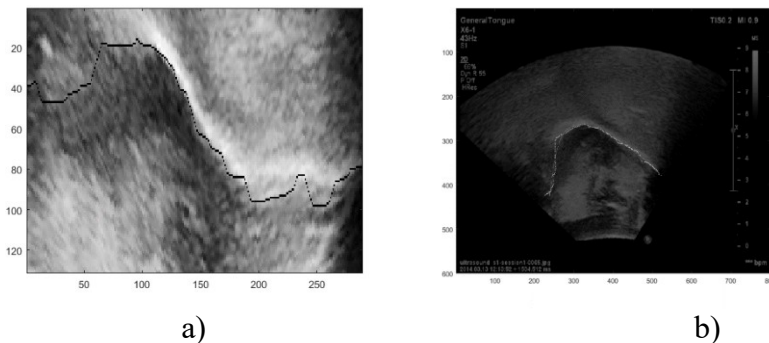


3. ábra: Az élkiemelés eredménye (bal oldalon), az élkiemelés további átlagolás után kapott képe (jobb oldalon)

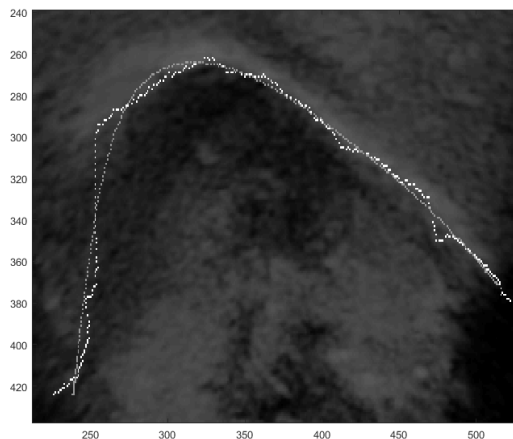
Mielőtt a nyelvkontúr követésére tennénk kísérletet, célszerű a kép oszlopait úgy eltolni, hogy átlagosan vízszintes vonalat kapjunk, vagyis az oszlopokat eltoljuk a sok képen számolt kontúrok átlagával. Ez a lépés nem csak a keresett élt közelíti a vízszintes irányhoz, hanem a kép alján található, inak által okozott zavaró reflexiókat is kitoljuk a képből. Az így kapott képen keressük a legnagyobb kumulált világosságú görbét.

A dinamikus programozás algoritmus a szabad kezdet és vég illesztésű, vagyis az első oszlop bármely soránál kezdődhet, és az utolsó oszlop bármely soránál végződik. A kumulált összeget minden egyes képpontra meg kell határozni balról jobbra haladva. Minden képpontban el kell tárolni a kumulált világosságot és azt, hogy melyik képpontból jutottunk ide. Minden pontban megvizsgáljuk tehát, hogy az előző oszlop előző két sora és következő két

sora között az összesen öt képpont közül melyik a legnagyobb összesített világosságú. Nyilván ehhez hozzáadva a vizsgált képpont világosságát lesz a vizsgált pont kumulált világossága maximális. Ha a képet reprezentáló mátrix előző oszlopának előző egy vagy két soráról lépünk tovább, egy vagy két pixelt léptünk lefelé. Ha az előző oszlop következő egy vagy két soráról továbblépve kapjuk a maximumot, egy vagy két pixelt léptünk felfelé. Ugyanazon a soron továbbhaladva vízszintesen léptünk tovább. Az utolsó oszlop maximumát megkeresve, ebből a képpontból visszafejthetjük a nyelvkontúr vonalát. A 4. a) ábrán az oszlopdiagramon megtalált nyelvkontúrt fekete vonal jelzi. A 4. b) ábrán a radiális ábrára visszavetített kontúrt fehér vonal ábrázolja. A kapott görbe követi az él egyenetlenségeit. Diszkrét koszinusz transzformációval (DCT) szűrhetjük a görbét. A simított görbét szürke vonal jelöli kinagyítva az 5. ábrán.



4. ábra: a) Az oszlopdiagramon kapott nyelvkontúr, b) A radiális képre visszavetített egyenetlen görbe (fehér) és a simított kontúr (szürke)

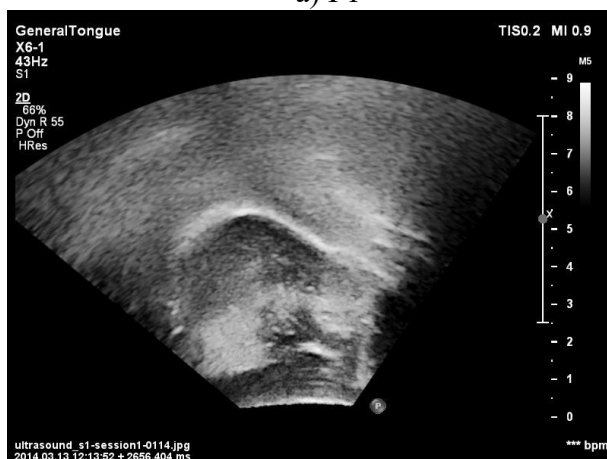


5. ábra: A 4. ábra b) rész érdekes területe kinagyítva

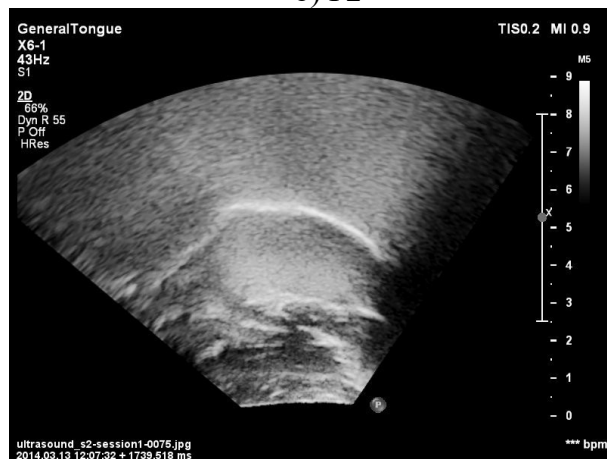
Eredmények

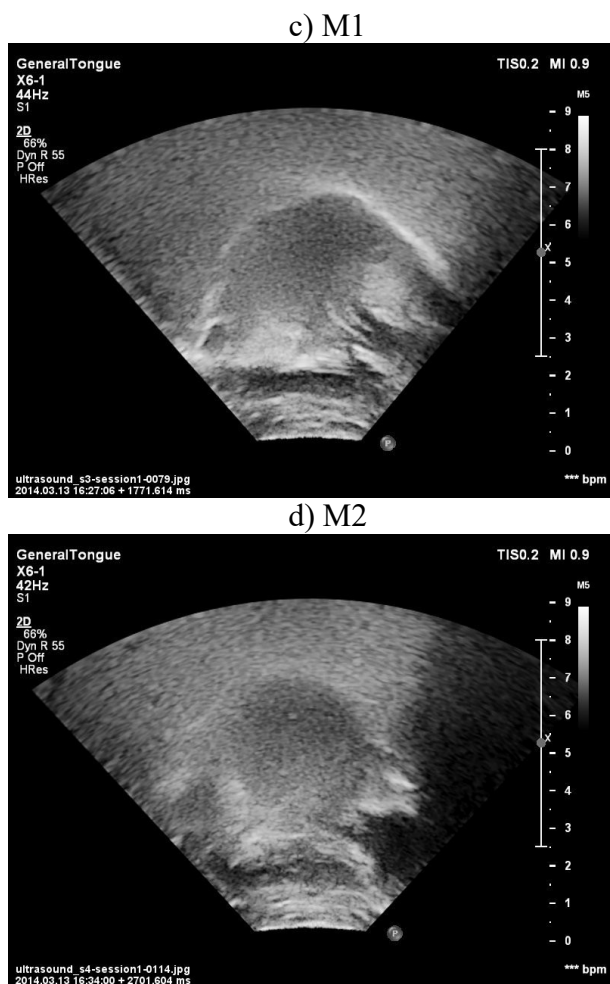
A nyelvkontúrkövetés eredményeinek verifikálását ugyanazon az ultrahang felvétel sorozaton végeztük, amelyen Csapó Tamás és Steven Lulich (2015) összehasonlította a kontúrkövető eljárásokat. Két női beszélő (F1, F2) és két férfi beszélő (M1, M2) egy-egy felvételen kétszer egymás után mondta be az *'I owe you a yoyo'* mondatot. A 6. ábra tanúsága szerint nyelvkontúr élessége láthatóan eltérő a négy beszélőnél.

a) F1



b) F2





6. ábra: A négy beszélő ultrahangos képe a 'yoyo' szó első 'o' hangjának közepén

A manuális és automatikus kontúr követésének sikerét az él láthatósága erősen befolyásolja. Csapó és Lulich (2015) hét önkéntessel manuális kontúr kijelölést végzett. A manuális kontúr kijelölést az átlagos abszolút hibával jellemezték (1. táblázat).

1. táblázat: A manuális kontúr kijelölés átlagos abszolút hibája (mm)

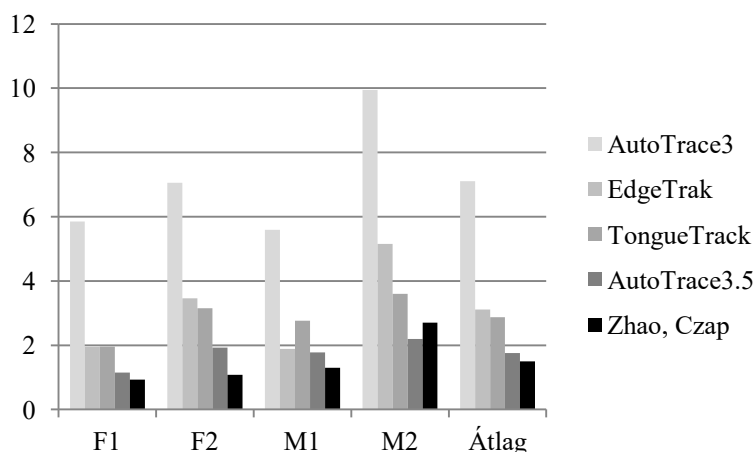
Beszélő	F1	F2	F3	F4
Átlagos abszolút hiba	0,95	1,09	1,17	2,11

Az automatikus kontúrkövetésre négy szabad hozzáférésű eljárást teszteltek. A legjobb eredményt az AutoTrace3.5 beállítással érték el. A felvételeket kettéválasztották, így a mintamondat minden felvételen egyszer szerepel. A négy beszélő két-két mintája összesen nyolc képsorozatot jelent. Az AutoTrace3.5 jelzésű esetben hét képsorozat manuális kontúrjelöléseinek átlaga képezte a tanító mintákat, a nyolcadik képsorozaton végezték a tesztelést. Amikor a saját képsorozat nem vett részt a tanításban (AutoTrace3, hat képsorozat a másik három beszélőtől szolgált a tanításra, a két saját képsorozat a tesztelésre), a kontúrkövetés pontossága drasztikusan romlott. A 2. táblázat a különböző eljárások négyzetes hiba átlagának négyzetgyökét tartalmazza (RMSE). Az átlagolást a négyzetes hiba logaritmusán végezték, majd az eredményt exponenciális függvényvel állították vissza (logRMSE). Az adatok mm-ben vannak megadva, a felvételeken szereplő nagyításban $1 \text{ mm} = 4,24 \text{ pixel}$.

2. táblázat: Az automatikus kontúrkövetés eredményei

Beszélő	F1	F2	M1	M2	Átlag
AutoTrace3	5,85	7,06	5,59	9,94	7,11
EdgeTrak	1,95	3,46	1,89	5,15	3,11
TongueTrack	1,96	3,15	2,76	3,6	2,87
AutoTrace3.5	1,15	1,93	1,78	2,19	1,76
Zhao, Czap	0,93	1,08	1,30	2,70	1,50

A 2. táblázat értékeit grafikusan ábrázolva mutatjuk a 7. ábrán.



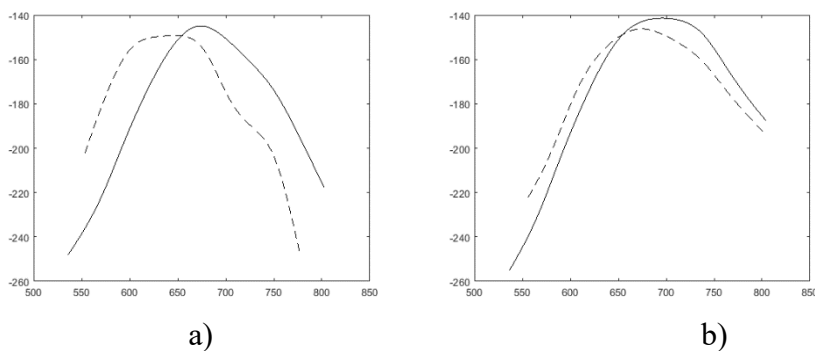
7. ábra: Az automatikus nyelvkontúr követés logRMSE hibái különböző programok és elemzések esetében (mm)

A logaritmusképzés a dinamikatartományt összenyomja, a kiugróan nagy hibák súlya csökken. A 3. táblázatban megadjuk a saját kontúrkövetés hibáinak különböző átlagait a manuális kontúr kijelölés átlagával összevetve (log(RMSE), lineáris RMSE, átlagos abszolút hiba).

3. táblázat: A saját kontúrkövetés különböző hibái.

Beszélő	F1	F2	F3	F4	Átlag
logRMSE	0,93	1,08	1,30	2,70	1,50
linRMSE	1,07	1,33	1,68	3,25	1,83
Abszolút hiba	0,43	0,48	0,52	0,74	0,54

A nyelvkontúr elemzése különböző hangokra és hangkapcsolatokra lehetővé teszi, hogy alapadatokat szolgáltatassunk a Shaanxi Xi'an nyelvjárású vizuális beszédsszintézishez (beszélő fej). A dominancia osztályok megállapítását segíti a nyelvállás vizsgálata különböző hangkörnyezetben. A 8. ábrán a kínai adatközlő ultrahang felvételeiből vett 'p' és 't' hangok nyelvkontúráját rajzoltuk meg 'e', illetve 'a' magánhangzók közötti kiejtéssel. A 'p' nyelvállása nagymértékben függ a környezetétől. Az ultrahang képen sajnálatos módon nem látható nyelvhegy pozíciója a 't' esetében dominánsnak látszik.



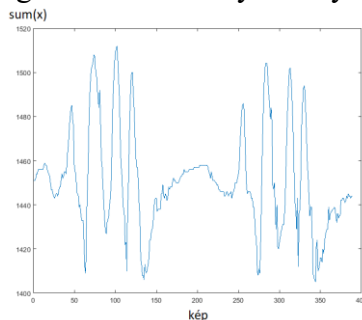
8. ábra: a) A nyelvkontúr az 'epe' (---) és 'apa' (- -) VCV szavak 'p' hangjának kezdetekor, b) A nyelvkontúr az 'ete' (---) és 'ata' (- -) szavak 't' hangjának kezdetekor

Következtetések

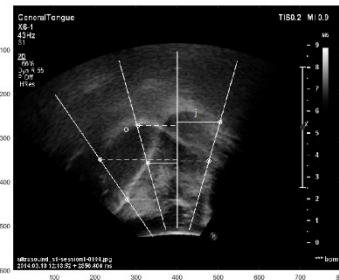
A dinamikus programozáson alapuló nyelvkontúr követési algoritmusunk az M2 beszélő kivételével kisebb hibát eredményezett, mint a többi Csapó és Lulich (2015) által tesztelt eljárás. Az M2 beszélőre kapott hiba a második legkisebb. A négy beszélőre számított átlagos hiba a mi módszerünkkel

adódott a legkisebbre. Az AutoTrace a tanító alakzatok manuális kontúrkijelölését követeli meg. A kutatók nem vizsgálták a megbízható eredményhez szükséges tanító minták számát. A dinamikus programozáson alapuló módszerünk csak a vizsgálandó terület (region of interest) kijelölését kívánja meg a hamis élek elkerülése végett.

A nyelvkontúr leírása és adatainak kiértékelése még nem tekinthető kidolgozottnak (Fabre et al. 2017). A kontúr simítására használt diszkét koszinusz transzformáció első hét együtthatóját felhasználva kaptuk a manuális kijelölés átlagához legközelebb eső eredményeket. Hét adattal tehát jellemezhető a teljes nyelvkontúr, a DCT együtthatók azonban nehezen kapcsolhatók össze a geometriai adatokkal. Kísérletet tettünk a nyelv leírására a radiális ábrán vett, a kontúr vonalának 20, 40, 60, és 80%-ánál vett koordinátaival. A 9. a) ábra a négy pont vízszintes koordinátáinak (abszcissza) összegét mutatja a Csapó–Lulich (2015) képsorozatok F1 beszélőjére. Az ultrahangos felvételeken a szájüreg elülső része a kép jobb oldalára esik. A nagyobb értékek a mátrix oszlopainak számozása szerint elől képzett hangra utalnak, a kisebb értékeket hátul képzett hangoknál kapjuk. Az ábrán jól látható a mintamondat (kétszer) négy 'j' hangjának maximuma, a magánhangzók vízszintes nyelvhelyzete ennél hátrébb esik.



a)



b)

9. ábra: a) Az egyes képkockák kijelölt kontúr pontjai abszcisszáinak összege,
b) Az 'o' (---) és 'j' (---) hanghoz tartozó nyelvállás kijelölt pontjainak abszcisszái

Ez alapján jelöltük ki a 6. ábra képeit a négy beszélő képsorozatából. A 9. b) ábra egy képre másolva mutatja az 'o' és 'j' hangok középső képkockáinak kontúráját. A radiális ábrán minél hátrébb (balra) és feljebb esik egy pont, annál kisebb a hozzá tartozó mátrix oszlopának a sorszáma. Minél előrébb (jobbra) és feljebb kerül egy pont, annál nagyobb a hozzá tartozó oszlop sorszáma. Az ábrán véletlenül éppen a függőleges irányhoz tartozik a kontúr 60%-ánál vett minta, így a két görbe 3. kiemelt pontjának abszcisszája megegyezik. Az ábrán szaggatott vonal jelzi az 'o' kontúr kitüntetett pontjainak távolságát a függőleges egyenestől. A 'j' pontjainak távolságát

folytonos vonal mutatja. A nyelvkontúrok részletes elemzése további vizsgálatokat igényel.

Irodalom

- Barnaud, M. L. – Schwartz, J. L. – Bessière, P. – Diard, J. 2019. Computer simulations of coupled idiosyncrasies in speech perception and speech production with COSMO, a perceptuo-motor Bayesian model of speech communication. *PLOS ONE* 1– 34.
- Czap, L. – Zhao, L. 2017. *Phonetic Aspects of Chinese Shaanxi Xi'an Dialect*. 8th International Conference on Cognitive InfoCommunications: CogInfoCom. Debrecen, Hungary, 2017. 09. 11– 14. 51–56.
- Czap, L. – Pintér, J. M. – Baksa-Varga, E. 2019. Features and results of a speech improvement experiment on hard of hearing children. *Speech Communication* 106. 7–20.
- Csapó T. G. – Lulich S. M. 2015. Error analysis of extracted tongue contours from 2D ultrasound images. *Interspeech*, 2157–2161.
- Fabre, D. – Hueber, T. – Girin, L. – Alameda-Pineda, X. – Badin, P. 2017. Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract. *Speech Communication* 93. 63–75.

Köszönetnyilvánítás

Köszönjük az MTA–ELTE Lendület Lingvális Artikuláció Kutatócsoportjának, hogy lehetőséget adtak ultrahangos felvételek készítésére, Csapó Tamás Gábornak, hogy rendelkezésünkre bocsátotta a vizsgálatukhoz használt képsorozatokat és a módszerük reprodukálását segítő konzultációt.

A cikkben ismertetett kutató munka az EFOP-3.6.1-16-2016-00011 jelű „Fiatalodó és Megújuló Egyetem – Innovatív Tudásváros – a Miskolci Egyetem intelligens szakosodást szolgáló intézményi fejlesztése” projekt részeként – a Széchenyi 2020 keretében – az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Abstract

In recent decades, several types of devices have been constructed to track the invisible movement of the tongue. One of these is ultrasound imaging. The advantage of the measurement method is that it is suitable to track rapid movements, as well, image and sound can be synchronized, and it only causes the speaker, who is not affected by harmful radiation, minimal inconvenience. Its disadvantage is that it does not provide a full three-dimensional image but only shows a longitudinal or transverse section and

the tip of the tongue cannot be seen. The longitudinal (midsagittal) section was used to track tongue movements.

The automatic assessment of ultrasound pictures is a key issue as the number of frames amounts to several hundred thousand so it is impossible to process them manually. The difficulty of tracking tongue contour lies in the fact that the images have a noisy background and the contour line is discontinuous. An algorithm based on dynamic programming was elaborated to track the movement of the back of the tongue. With an extreme size edge enhancing kernel and averaging construction, the process simultaneously handles the problems of break of continuity and noise. In the image obtained after smoothing, the lightest line is searched for from the left to the right edge of the image. The points of the curve thus obtained follow the uneven line of the tongue contour. To smooth the curve, screening based on discrete cosine transformation was applied.