A nyelvkontúrkövető algoritmusok és a gépi tanulás összekapcsolhatóságának vizsgálata

Trencsényi Réka

Debreceni Egyetem, Villamosmérnöki Tanszék

Kivonat A publikáció a digitális beszédszintézis tárgykörébe tartozik, és ötvözi a vizuális információkra épülő artikulációs beszédszintézis, illetve a gépi tanulóalgoritmusok eszköztárának alkalmazását. A vizuális információkat dinamikus MRI- és UH-felvételek automatikusan illesztett nyelvkontúrjaiból kinyerve gépi tanítást valósítunk meg, melynek célja a nyelvkontúr hiteles rekonstrukciója. A neurális hálózat be- és kimeneti paramétereinek különböző beállításával módosítható a tanulóalgoritmus jellege. Ennek megfelelően három különböző irányvonal mentén történik tanítás: MRI-adatokból MRI-nyelvkontúrt, UH-adatokból UH-nyelvkontúrt, illetve UH-adatokból MRI-nyelvkontúrt hozunk létre.
Kulcsszavak: artikulációs beszédszintézis, nyelvkontúrkövetés, gépi tanulás

1. Bevezető

A beszédkutatás egyik legdinamikusabban fejlődő, ugyanakkor egyre összetettebb technikai és módszertani kihívásokat rejtő területe a beszédfelismerés mellett a digitális beszédszintézis, ami már napjainkban is szerves részét képezi az ember-gép kapcsolatnak. Ebben a vonatkozásban kulcsfontosságú a gép kommunikációs szerepe, hiszen alapvető rendeltetése a szöveg-beszéd transzformáció megvalósítása, azaz a természetes emberi beszéd közben kialakuló akusztikai produktum élethű utánzása. Ennek kiterjesztett változatában a beszédet jellemző szupraszegmentális elemek (beszédritmus, hangerő, hangmagasság, hangszín, hanglejtés, hangsúly) figyelembevételével tovább finomítható a modell, aminek a beszédfelismerés területén is igen nagy jelentősége lehet (Czap és Pintér, 2015). Napjainkban a beszédszintézis területén zajló kutatások főként a szövegfelolvasó rendszerek megalkotására és tökéletesítésére fókuszálnak, ami olyan alkalmazások elterjedését teszi lehetővé, mint például az utastájékoztató rendszerek, a beszélő okoskészülékek, a szépirodalmi felolvasók, a képernyőolvasók vagy a telefonos tudakozó szolgáltatás. A kutatások hagyományos irányvonalát képviselő szövegfelolvasók esetén a beszédépítés emberi hangminták közvetlen vagy közvetett felhasználásával történik. A törekvések sikerességét a szakirodalom számos közleménye bizonyítja (Olaszy, 1999; Olaszy és mtsai, 2000; Németh és mtsai, 2006; Sproat, 1997; Schröder és Trouvain, 2003; Besacier és mtsai, 2014), melyek igen gazdag tudásanyagot és sokrétű tapasztalatot tükröznek. A klasszikus koncepciók mellett azonban olyan területek is kezdenek kibontakozni, melyek

kevésbé kidolgozottak, és rengeteg nyitott probléma vár még megoldásra. Ide sorolható például az artikulációs (Zappi és mtsai, 2016; Czap és mtsai, 2019) vagy a gépi tanuláson alapuló beszédszintézis (Wu és mtsai, 2015; Arik és mtsai, 2017).

Az artikulációs beszédszintézis az akusztikai produktum utánzását emberi hangminták alkalmazása helyett az emberi hangképzés és artikuláció gépi leképezése révén próbálja megvalósítani. Ennek egyik modern technológiai vonulata a robotok beszédének előállításához szükséges artikulációs elektromechanikus beszédkeltőkre irányuló kísérletezés. A szintézis kiindulópontja az artikulációs-akusztikai konverzió végrehajtása, ami a beszédhez kapcsolódó vizuális információra épül (Czap és Mátyás, 2005). Ennek folytán lényegi szerepet kapnak a különböző képalkotó eljárások (például mágnesesrezonancia-képalkotás (MRI), komputertomográfia (CT), ultrahang (UH)), melyek új információcsatornákat kapcsolnak be a tudományos kutatások folyamatába. Ennek megfelelően a beszéd közben készült MRI- vagy UH-felvételek potenciális forrásai lehetnek az emberi artikulációt jellemző paraméterek vizuális módon támogatott kinyerésének. Mivel a hangok képzésében legaktívabban a nyelv vesz részt, így elsősorban a nyelv mozgását célszerű a lehető legpontosabban monitorozni. Az utóbbi években az erre irányuló vizsgálatok közkedvelt eszközei a már említett MRI, CT és UH mellett az elektropalatográfia (EPG) vagy az elektromágneses artikulográfia (EMA). Az egyszerűbben hozzáférhető UH, EPG és EMA eljárások alkalmazásával csak bizonyos síkmetszetek mentén kaphatunk információt a beszéd dinamikai jellemzőiről, míg a klinikai körülményeket igénylő MRI és CT berendezések segítségével akár háromdimenziós morfológiai adatokra is szert tehetünk. A közelmúltban több tanulmány is foglalkozott dinamikus nyelvkontúr-követési algoritmusok kidolgozásával és fejlesztésével (Li és mtsai, 2005; Csapó és mtsai, 2017; Zhao és Czap, 2019), ami az egyik alappillérét képezheti az artikulációs beszédszintézis témakörében végzett kutatásoknak. A nyelvkontúr dinamikus letapogatását a szagittális síkban érdemes elvégezni, ahol egy kétdimenziós metszeten látható a nyelv fel-le, illetve előre-hátra irányú mozgása. A vizsgálatok legkényelmesebb kellékei UH- vagy MRI-felvételek lehetnek, melyek előnye a jó térbeli és időbeli felbontás, a kép- és hanganyag szinkronizálhatósága, illetve a beszélő alany sugárterheléstől való mentesítése. A nyelvkontúr kijelölése történhet manuálisan vagy automatikus algoritmusok segítségével, bár az adott felvételt alkotó képkockák számának százas vagy akár ezres nagyságrendje indokolttá teszi a dinamikus programozás favorizálását a kézi erővel szemben. A nyelvkontúr detektálásának hatékonyságát nagymértékben meghatározza a felvétel minősége, illetve a kontúrkövető algoritmus típusa (például AutoTrace3, EdgeTrak, TongueTrack, AutoTrace3.5) ezért gyakorlatilag elévülhetetlen ambíció a nyelvkontúrkövető programok finomítása.

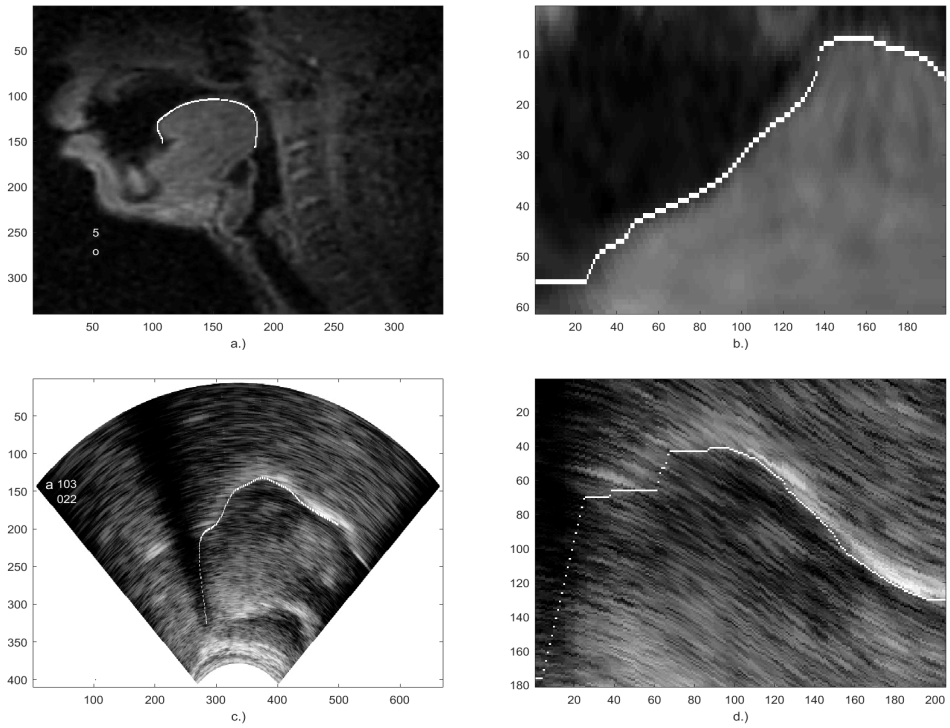
Ezen túlmenően perspektivikus irányvonalat jelöl ki a gépi tanulóalgoritmusok alkalmazása is, melynek során a gép bizonyos bemeneti paraméterek halmazából a környezetéből nyert információk alapján kimeneti eredményeket produkál, miközben javítja a teljesítőképességét. A gépi tanulóalgoritmus lényegében az emberi agy működését próbálja imitálni, így kulcsfontosságú szerepet játszik

a neurális hálózatok működésének ismerete, illetve élethű modellezése. A biológiai neurális hálózatok mintázatok alapján valósítják meg a tanulási folyamatot, ami a gépi tanulás esetében megfelelő algoritmusok megalkotásával képezhető le. A beszédszintézis területén a gép bemeneti paramétereinek halmazát képezhetik például emberi hangminták vagy vizuális forrásokból nyert adatok, melyekkel elvégezve a betanítást megszólaltatható az auditív produktum. A vizuális információkkal betanított neurális hálózat lehetősége tehát természetes módon kínálja fel az artikulációs beszédszintézis és a gépi tanulás módszereinek összekapcsolását. A lehetőségek jóformán korlátlanok, az eljárások, illetve ezek kombinációja pedig javarészt még nincs kimerítően feltárva. Jelen publikáció a nyelvkontúrkövetés és a gépi tanulóalgoritmusok együttes alkalmazhatóságának bizonyos vonatkozásait vizsgálja MRI- és UH-felvételek feldolgozásával.

2. Automatikus nyelvkontúrkövetés

A vizsgálatok tárgyát beszéd közben készült MRI- és UH-felvételek képezték. Az MRI-felvételeket a Dél-kaliforniai Egyetem honlapján szabadon hozzáférhető multimédiás csomagból válogattam ki, az UH-felvételek pedig az MTA-ELTE Lendület Lingvális Artikuláció Kutatócsoport SonoSpeech rendszerével készült audiovizuális anyagok formájában álltak rendelkezésemre.

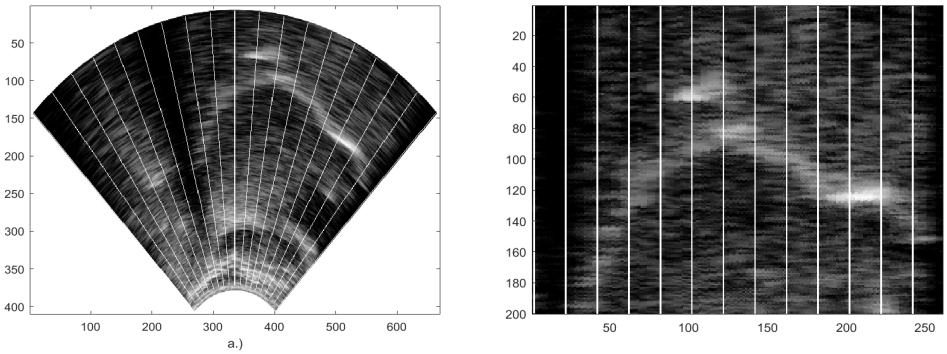
A nyelvkontúrkövetés elsődleges célja a beszédhangokhoz tartozó nyelvállások dinamikus leírása, illetve a koartikuláció során létrejövő hangátmeneteket jellemző nyelvmozgások tanulmányozása. A kvalitatív analízis mellett a nyelvkontúr a beszéd kvantitatív jellegű tanulmányozásának is jó kiindulópontja lehet, hiszen a nyelvkontúrból származtatható számszerű értékek elősegíthetik az artikulációs modellek mélyebb megértését és fejlesztését. A nyelvkontúr detektálására kidolgozott algoritmusok az alkalmazott eljárásoktól függően rendkívül változatosak lehetnek. A vizsgálataim segédeszközeként olyan algoritmust használtam fel, amely a dinamikus programozás technikáját alkalmazza. A nyelvhát az UH-felvételen világos sávként rajzolódik ki, az MRI-felvételen pedig a szájüregi levegő sötét tartománya és a nyelvszövet világos tartománya között létrejövő kontrasztként érzékelhető, így a kontúrkövetés mindkét esetben a nyelvhát vonalát meghatározó maximális világosságú képpontok megkeresését jelenti. Az algoritmus alkalmazását a felvételek előfeldolgozása előzi meg, ami a képalkotó eljárásokból adódó zajok és folytonossági hiányok megszüntetésére irányul. Az említett hibák redukálásának leghatékonyabb eszközei az élkiemelő és átlagoló operációk, amik matematikailag a konvolúció műveletével valósíthatók meg (Czap, 2007). A megkeresett maximális világosságú képpontok, igazodva a nyelvhát egyenetlen vonalához, egy nyers görbét hoznak létre, melynek simítása diszkrét koszinusz transzformációval oldható meg. Az 1. ábra képei automatikusan illesztett nyelvkontúrt mutatnak be egy-egy MRI- (a.) és b.)), illetve UH-kereten (c.) és d.)). Az 1.a ábrán az *o* hanghoz tartozó nyelvállás figyelhető meg, míg az 1.c ábra az *a* hangnak megfelelő nyelvállást jeleníti meg a simított nyelvkontúr kiemelésével. Az 1.b és 1.d ábrákon az 1.a, illetve 1.c kereteken megrajzolt nyelvontúrok simítás nélküli, kinagyított részletei láthatók.



1. ábra: A nyelvkontúr követése MRI- és UH-felvételeken

Az 1.b és 1.d ábrák speciális transzformációval hozhatók létre az 1.a és 1.c ábrákból kiindulva. A transzformációs eljárás lényegét a 2. ábrán látható UH-keret segítségével érzékeltetem. Első lépésként a radiális geometriájú 2.a képen a kör középpontjából kiindulva sugárirányú metszeteket képzünk a felvétel által definiált -45° – 45°-os tartományban, melyek mentén lényegében újramintavételezzük a képet. Az így létrejövő metszeteket oszlopdiagramba rendezzük, melynek eredményeképpen egy olyan képmátrixot kapunk, ami a descartes-i x-y síkban jellemezhető a legkényelmesebben. A mátrixos szerkezet kialakítása nyomán áll elő a 2.b ábra. A vizsgálatok azt mutatják, hogy az 1/4°-onként végrehajtott mintavételezés a legideálisabb, hiszen ekkor a mátrix szomszédos oszlopai között nem fordul elő két pixelnél nagyobb változás a kontúrban. Az áttekinthetőség kedvéért a metszeteket csak 5°-onként ábrázoltam, amit a 2. ábra fehér vonalai szemléltetnek. Az eljárás MRI-keretek esetében ugyanilyen módon működik az MRI-kereten megfelelően kijelölt középpont és (a -45° – 45°-os tartománytól általában szélesebb) szögtartomány alkalmazásával.

Az MRI-felvételek adatközlője angol anyanyelvű férfi beszélő, aki VCV típusú hangsorokat szólaltat meg V magánhangzóval és C mássalhangzóval. A bemu-



2. ábra: Radiális és mátrixos geometriájú UH-keretek

tatott MRI-keret tanúsága szerint a kapott görbe hitelesen követi a nyelvhát vonalát. Az UH-felvételeken magyar, illetve kínai anyanyelvű női bemondótól származó hangsorok vannak rögzítve, melyek CVC, illetve VCV szerkezetűek. Összehasonlítva az 1. ábra képeit, feltűnhet, hogy az UH-felvételen a nyelvhát kevésbé éles határvonalként jelenik meg, ami egy elmosódott világos sávot eredményez. Ez a nyelv és a fölötte lévő levegő határán visszaverődő UH-hullámok következményeként alakul ki, így a nyelvkontúr a világos sáv alsó határán lokalizálható. Az UH-felvételek további sajátsága, hogy a nyelvgyök és az állcsont árnyékoló hatása miatt a nyelv hátsó része és a nyelvhegy nem látható a felvételen, így a nyelv alakjáról és mozgásáról csak részleges információt kaphatunk. A nyelvgyök és az állcsont árnyéka sötét sávként azonosítható az 1.c ábra bal és jobb oldali részén.

3. Gépi tanulás

Jelenlegi kutatómunkám célkitűzése az előző fejezetben bemutatott nyelvkontúrkövetés és a gépi tanulóalgoritmusok összekapcsolása, illetve az egymáshoz való viszonyuk bizonyos aspektusainak tanulmányozása. Programjaimat MATLAB-környezetben hoztam létre, és a gépi tanítást olyan algoritmussal valósítottam meg, amely a neurális hálózat súlyfaktorait a skálázott konjugált gradiens módszer (Moller, 1993) segítségével határozza meg. Ezen optimalizációs eljárás a problémához rendelt egyenletrendszert a bemeneti paraméterek ismeretében iterációs módszerrel oldja meg, miközben az eljárással számított kimeneti paraméterek értékei konvergálnak az előírt értékekhez. A módszer előnye, hogy az iterációs algoritmus lépésközeinek számát minimalizálva elég gyors konvergencia biztosítható, így a gépi tanítás viszonylag rövid idő alatt véghezvihető. Az iterációs lépések olyan irány mentén valósulnak meg, ami gyorsabb konvergenciát biztosít, mint a legmeredekebb ereszkedésnek megfelelő legnegatívabb gradiens, miközben megőrzi a korábbi lépésekben kapott hibaminimalizációt.

A neurális hálózatban két rejtett réteget helyeztem el, melyek egyenként 30 neuront tartalmaztak. A tanuláshoz szükséges bemeneti paramétereket a dinamikusan változó nyelvkontúr négy kiválasztott pontjának segítségével jelöltem ki, melyekhez kimeneti paraméterként a nyelvkontúr diszkrét koszinusz transzformáltját rendeltem hozzá. A négy kiválasztott pont relatív helyzete minden képkockán azonos olyan értelemben, hogy a négy pont minden nyelvkontúr esetében a görbe hosszának kb. 20%, 40%, 60%, 80%-ánál található.

A tanítást elsőként az MRI-forrásból származó be- és kimeneti paraméterek rögzítésével hajtottam végre, az eredményeket pedig ugyanazon MRI-kereteken teszteltem. A procedúrát hasonló elv alapján az UH-keretekre is megismételtem, végül az UH-forrásból kinyert bemeneti paraméterek, illetve az MRI-forrásból eredő kimeneti paraméterek kombinálásával újra lefuttattam az algoritmust, majd eredményeimet az MRI-kereteken teszteltem. A következő alfejezetek a három különböző megközelítést tárgyalják.

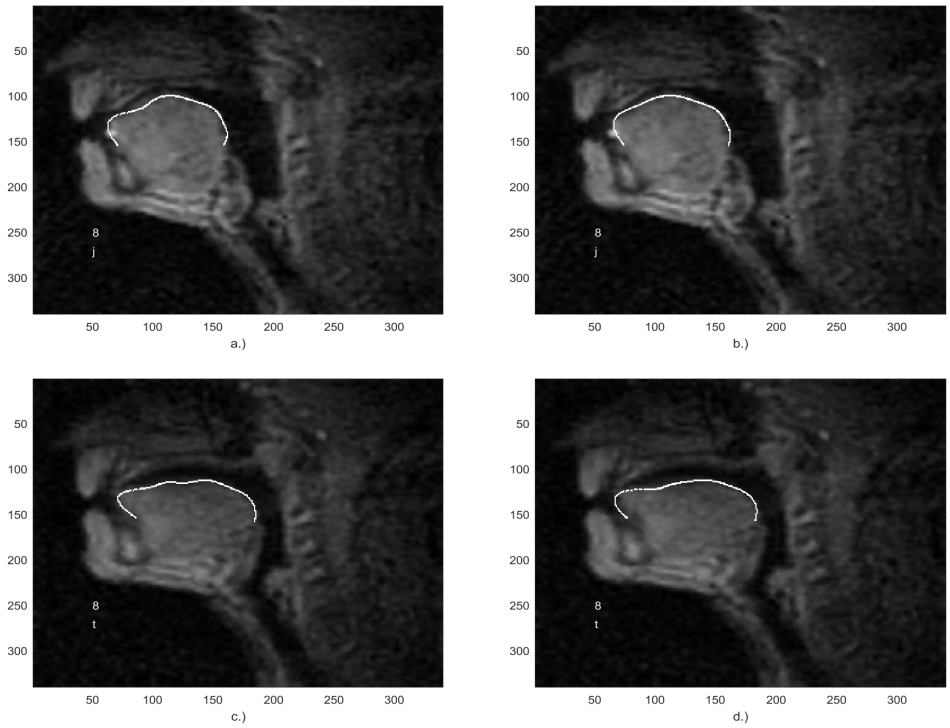
3.1. MRI-nyelvkontúr tanítása MRI-adatokkal

Az alfejezet az MRI-felvételek esetében elvégzett gépi tanítás eredményeit foglalja össze. A tanítás alapját az *a,á,c,cs,d,dz,dzs,e,é,g,i,j,k,l,n,o,ö,r,s,sz,t,u,ü,z,zs* beszédhangokhoz tartozó fonemikus konfigurációk képezték. A bemeneti paramétereket a nyelvkontúr négy kiválasztott pontjának képsíkban mért y koordinátája adta, a kimeneti paraméterek halmazát pedig a nyelvkontúr diszkrét koszinusz transzformáltjának első húsz együtthatója határozta meg, melynek alapján a tanulóalgoritmus futtatását követően inverz diszkrét koszinusz transzformációval rekonstruálható a betanított nyelvkontúr. Ez lényegében azt jelenti, hogy mindössze négy pont felhasználásával történik a teljes görbe előállítása. Eredményeimet a *j* és *t* hangok példáján keresztül mutatom be.

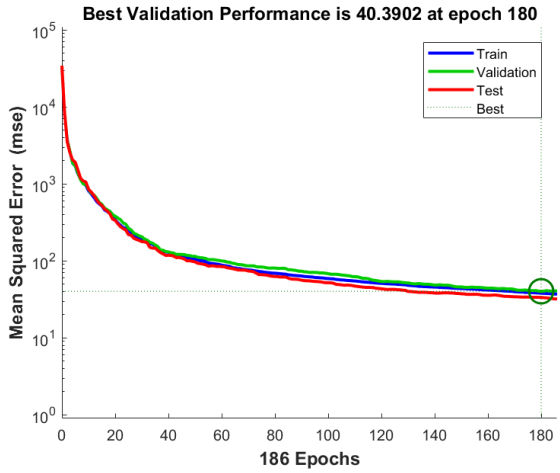
A 3.a és 3.c ábrák a *j*, illetve a *t* hangnak megfelelő nyelválláshoz illesztett nyelvkontúrokat prezentálnak. A 3.b és 3.d ábrák ugyanazon *j*, illetve *t* hanghoz tartozó betanított nyelvkontúrokat jelenítenek meg. Az illesztett és a betanított nyelvkontúrok összehasonlításakor nem mutatkozik figyelemreméltó vizuális különbség, minimális az eltérés a két görbe között. A 3. ábrán szemléltetett eredmények azt tükrözik, hogy a tanulóalgoritmus hatékonyan működik, amit a 4. ábra grafikonjai is alátámasztanak. Az ábrán a tanítás, a tesztelés és a validálás átlagos négyzetes hibája követhető nyomon. Látható, hogy gyors csökkenés mellett a tanítás és a tesztelés hibája lényegében azonos.

3.2. UH-nyelvkontúr tanítása UH-adatokkal

Az alfejezet az UH-felvételek esetében elvégzett gépi tanítás eredményeit foglalja össze. A tanítás ez esetben a "Most a CVCV, meg a CVCV volt." típusú bemondásokra épült. A bemeneti és kimeneti paraméterek értelmezése ugyanaz, mint az előző alfejezetben, és a lépéseket ezúttal a *g* és *s* hangok példáján keresztül vezetem végig.

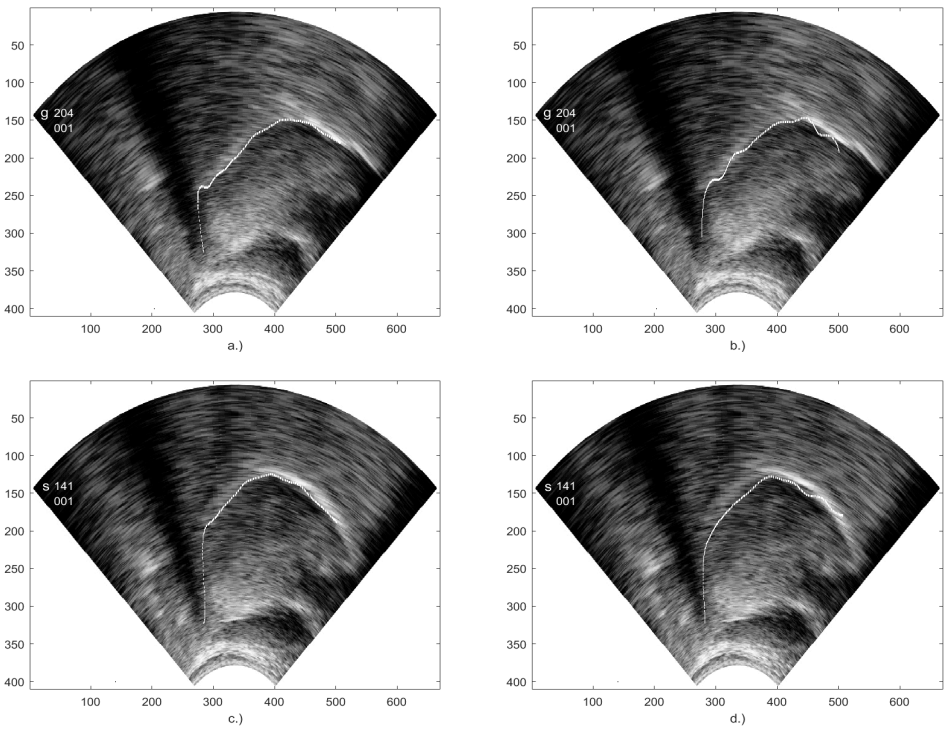


3. ábra: Az illesztett és betanított MRI-nyelvkontúr a *j* és *t* hangok esetében



4. ábra: A gépi tanítás átlagos négyzetes hibája MRI-MRI tanítás esetén

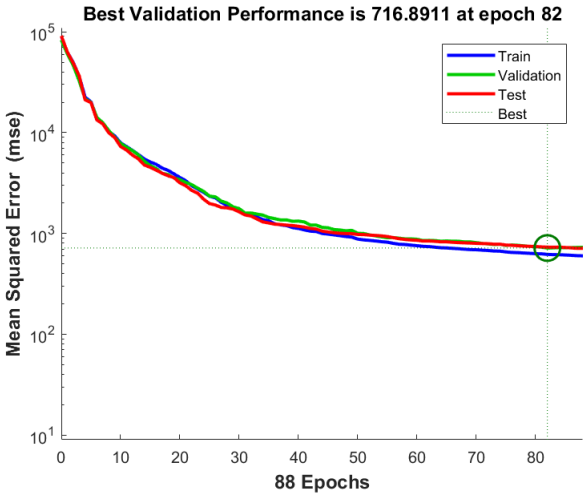
A 5.a és 5.c ábrák a *g*, illetve a *s* hangnak megfelelő nyelválláshoz illesztett nyelvkontúrokat demonstrálnak. A 5.b és 5.d ábrák ugyanazon *g*, illetve *s* hanghoz tartozó betanított nyelvkontúrokat mutatnak be. Összehasonlítva az illesztett és a betanított nyelvkontúrokat, ez esetben sem figyelhető meg számottevő különbség a két görbe között. A tanítás, a tesztelés és a validálás átlagos négyzetes hibájának alakulását az 6. ábra tünteti fel, melynek tendenciája hasonló az MRI-felvételekkel megvalósított tanítás során kapott görbékhez.



5. ábra: Az illesztett és betanított UH-nyelvkontúr a *g* és *s* hangok esetében

3.3. MRI-nyelvkontúr tanítása UH-adatokkal

Az előző két alfejezetben a gépi tanulás be- és kimeneti paraméterei ugyanazon forrásból származtak, hiszen MRI-nyelvkontúrt MRI-adatokkal, UH-nyelvkontúrt pedig UH-adatokkal tanítottunk. Érdemes azonban azt is tanulmányozni, hogy milyen sikerrel kapcsolhatók össze a két különböző forrás paraméterei. Ebből a célból a neurális hálózatot úgy szerkesztettem meg, hogy bemeneti paramétereit az UH-nyelvkontúr négy kiválasztott pontja, kimeneti paramétereit pe-

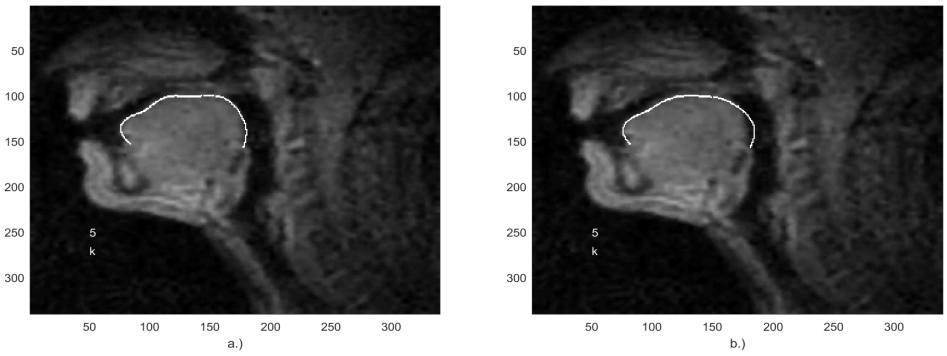


6. ábra: A gépi tanítás átlagos négyzetes hibája UH-UH tanítás esetén

dig az MRI-nyelvkontúr diszkrét koszinusz transzformáltja alkotta. Ezáltal egy olyan tanítási mechanizmus hozható létre, melynek során MRI-nyelvkontúrt alkothatunk UH-adatok felhasználásával. Eredményeim ismertetéséhez újfent az *a* hangot hozom fel példaként. Megjegyzem, hogy a felhasznált adatbázis mérete nagyságrendekkel elmarad a 3.1., illetve 3.2. alfejezetekben taglalt körülményekhez képest. Ennek oka, hogy az MRI- és UH-forrásokból származó felvételek nem minden esetben azonos típusú bemondásokat szólaltatnak meg, és emellett az egyes beszédhangokhoz rendelt képkockák száma sem egyezik meg, ami megnehezíti a tanulóalgoritmus paramétereinek összehangolását. A bemondások és mintaszámok szinkronizálása azonban jelenleg is folyamatban van.

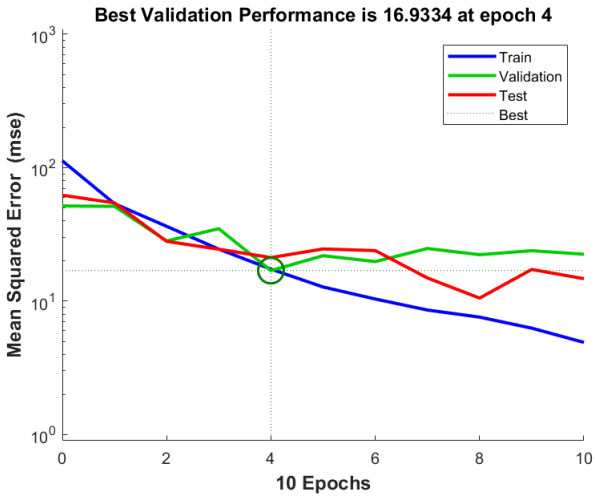
A 7.a ábra a *k* hangnak megfelelő nyelválláshoz illesztett nyelvkontúrt szemléltet. A 7.b ábra ugyanazon *k* hanghoz tartozó betanított nyelvkontúrt illusztrál. Az eredmény akár több szempontból is érdekes lehet, hiszen amellett, hogy különböző anyanyelvű, eltérő nemű adatközlők különböző képalkotó eljárással készített felvételeiből származnak a neurális hálózat révén összekapcsolt be- és kimeneti paraméterek, az sem elhanyagolható körülmény, hogy a tanítás szűkebb adathalmazból kiindulva produkál bővebb adathalmazt. A 2. fejezet végén ugyanis említettem, hogy az UH-felvétel nem képes megjeleníteni a nyelv hátsó részét és a nyelvhegy régióját, ami az MRI-felvételen természetesen akadályok nélkül látható. Ez pedig azt vetíti előre, hogy az UH-felvételekből származó részleges adatokkal tanulóalgoritmusok bevetésével hatékonyan becsülhető a teljes nyelvhát kontúrja.

A 8. ábrán a tanítás, a tesztelés és a validálás átlagos négyzetes hibájának futása elevenedik meg. Látható, hogy a tanítás és a tesztelés hibagörbéje nem mutat olyan mértékű együtthaladást, mint amit a 4. és 6. ábrák tükröznek. Ez a tanítóalakzatok fentebb említett csekély számának a következménye, a kezdeti



7. ábra: Az illesztett és betanított UH-nyelvkontúr a *k* hang esetében

adathalmaz bővítésével azonban javulás várható a görbék relatív lefutásának tekintetében.



8. ábra: A gépi tanítás átlagos négyzetes hibája UH-MRI tanítás esetén

4. Összefoglaló

A cikk az artikulációs beszédszintézisben fontos szerepet játszó automatikus nyelvkontúrkövető algoritmusok, illetve a gépi tanítás együttes alkalmazását demonstrálja dinamikus MRI- és UH-felvételek feldolgozásával. A gépi tanulás a

neurális hálózat be- és kimeneti paramétereinek megfelelő kombinálásával MRI-MRI, UH-UH, illetve UH-MRI viszonylatban valósul meg. Megjegyzem, hogy a jelenlegi fázisban még igen korlátozott számú tanító- és tesztelőalakzat áll rendelkezésre, de a forrásadatok fokozatos bővítés alatt állnak. Az aktuális eredmények a folyamatban lévő kutatómunkából csupán egy keskeny szeletet, egy pillanatképet villantanak fel, hiszen az artikulációs beszédszintézis és a gépi tanulás területei önmagukban véve is rendkívül sok problémát vetnek még fel, amiknek jó része egyelőre csak részlegesen tekinthető megoldottnak. Ennek megfelelően a kutatások jövőbeli irányát meghatározhatja például a vizuális információkra alapozott, statisztikai elven működő vagy szabályalapú algoritmusokkal létrehozott beszédszintézis modelljeinek tökéletesítése, ami alapvető fontosságú lehet például a klinikai célú beszédterápiában, a nem anyanyelvi nyelvtanulási tréningek kialakításában vagy a néma beszéd megszólaltatásához szükséges szintetizátorok konstrukciójában és fejlesztésében.

Hivatkozások

Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Andrew, N., Raiman, J., Sengupta, S., Mohammad, S.: Deep voice: Real-time neural text-to-speech. In: Proceedings of the 34th International Conference on Machine Learning, 70, 195-204 (2017)

Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: A survey. Speech Comm., 56, 85-100 (2014)

Czap, L., Mátyás, J.: Virtual announcer. Infocommunications Journal, 60, 2-5 (2005)

Czap, L., Mátyás, J.: Virtual speaker. In: Ádám, T., Vásárhelyi, J., Varga, A. (szerk.): Proceedings of 6th International Carpathian Control Conference ICCC 2005 Miskolc, Magyarország: Miskolci Egyetem, 351-358 (2005)

Czap, L.: Képfeldolgozás. Miskolc-Egyetemváros, Magyarország: Miskolci Egyetem, 151 p. (2007)

Czap, L., Pintér, J. M.: Intensity feature for speech stress detection. In: Petras, I., Podlubny, I., Kacur, J., Vásárhelyi, J. (szerk.): Proceedings of the 16th International Carpathian Control Conference Miskolc, Magyarország: IEEE IAS/IES/PELS, 91-94. (2015)

Czap, L., Pintér, J. M., Baksa-Varga, E.: Features and Results of a Speech Improvement Experiment on Hard of Hearing Children. Speech Communication, 106, 7-20 (2019)

Csapó, T. G., Deme, A., Gráczi, T. E., Markó, A.,, Varjasi, G.: Szinkronizált beszéd- és nyelvultrahang-felvételek a Sono-Speech rendszerrel. In: Vincze V. (szerk.): XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017). Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, 339-346 (2017)

Li, M., Kambhamettu, C., Stone, M.: Automatic contour tracking in ultrasound images. Clinical linguistics and phonetics, 19, 545-554 (2005)

Moller, M. F.: A scaled conjugate gradient algorithm for fast supervised learning. Neural networks, 6, 525-533 (1993)

Németh, G., Olaszy, G., Fék, M.: Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei. Beszédkutatás, 183-196 (2006)

Olaszy, G.: Beszédadatbázisok készítése gépi beszédelőállításhoz. Beszédkutatás99, 68-89 (1999)

Olaszy, G., Németh, G., Olaszi, P., Kiss, G.: Profivox: a legkorszerűbb hazai beszédszintetizátor. Beszédkutatás 2000, 167-179 (2000)

Schröder, M., Trouvain, J.: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. Int. J. Speech Tech., 6, 365-377 (2003)

Sproat, R. W.: Multilingual text-to-speech synthesis. KLUWER Academic Publishers (1997)

Zappi, V., Vasuvedan, A., Allen, A., Raghuvanshi, N., Fels, S.: Towards real-time two-dimensional wave propagation for articulatory speech synthesis. In: Proceedings of Meetings on Acoustics 171ASA, 26, 045005 (2016)

Zhao, L., Czap, L.: A nyelvkontúr automatikus követése ultrahangos felvételeken. Beszédkutatás, 27, 331-343 (2019)

Wu, Z., Valentini-Botinhao, C., Watts, O., King, S.: Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), 4460-4464 (2015)