MISKOLCI EGYETEM
GÉPÉSZMÉRNÖKI ÉS INFORMATIKAI KAR

HATVANY JÓZSEF INFORMATIKAI TUDOMÁNYOK
DOKTORI ISKOLA

Vezető: PROF. DR. SZIGETI JENŐ

Az artikuláció geometriai és akusztikai jellemzőinek kapcsolata

PhD értekezés tézisei

Készítette: DR. TRENCSÉNYI RÉKA

Témavezető: PROF. DR. CZAP LÁSZLÓ

Miskolc
2024

A bírálóbizottság tagjai:

Elnök:

- Prof. Dr. Szigeti Jenő, egyetemi tanár,
  Miskolci Egyetem

Tagok:

- Prof. Dr. Vicsi Klára, egyetemi magántanár,
  Budapesti Műszaki és Gazdaságtudományi Egyetem
- Dr. Trohák Attila, egyetemi docens,
  Miskolci Egyetem
- Dr. Varga Attila, egyetemi docens,
  Miskolci Egyetem

Bírálók:

- Prof. Dr. Takács György, címzetes egyetemi tanár,
  Pázmány Péter Katolikus Egyetem
- Dr. Baksáné Dr. Varga Erika, egyetemi docens,
  Miskolci Egyetem

1. A kutatási téma és a kitűzött feladatok rövid ismertetése

A kutatási témám a beszédtudomány területeihez kapcsolódik. A beszédkutatás egyik legfontosabb tématerülete a beszédszintézis, ami elemi alkotóját képezheti az ember-gép kapcsolatnak. Ez esetben a gép kommunikációs szerepe abban nyilvánul meg, hogy kódoló adóvá válik, azaz beszédet produkál. Napjainkban a beszédszintézis legelterjedtebb irányzata a szövegfelolvasók készítése, melyek konkrét témakörre szűkített vagy általános témájú írott szöveget szólaltatnak meg. Ebbe az alkalmazási kategóriába sorolhatók például a szépirodalmi felolvasók, az utastájékoztató rendszerek, a hírolvasók, a hangos időjárás-jelentés vagy a telefonos tudakozó szolgáltatás. A beszédszintetizátorok megalkotásának célja a természetes emberi beszéd közben kialakuló akusztikai produktum élethű utánzása. Ebben a megközelítésben a beszéd hullámformája adja a kiindulópontot, amit kétfajta megoldásban alkalmaznak gépi beszéd előállítására. Az egyik csoportba az úgynevezett forráskódolású technikák tartoznak, melyek segítségével a beszédjelből kivonják a lényegi információkat és ezeket bemeneti adatsorozatként kezelik a szintézis során. A másik megoldás az emberi hangot közvetlenül használja fel a beszédépítéshez olyan módon, hogy a beszédjelből különböző hosszúságú hullámforma-részleteket vágnak ki és tárolnak el, majd az így kapott elemek megfelelő kiválasztásával és összefűzésével megkonstruálják a kívánt beszédhullámot. Ezeken túlmenően, tágabb módszertani szempontok alapján megkülönböztetünk még szabályalapú, illetve statisztikai elven működő

beszéd-előállítási eljárásokat. Az előbbi esetében megfigyelések és tapasztalatok szerint felállított szabályokkal koordinálják a szintézis egyes lépéseit, az utóbbi esetében pedig valószínűségeken alapuló belső rendszerállapotok révén jutnak el a beszédprodukcióhoz. A statisztikai elvű módszerek egyik tipikus válfaja a gépi tanulóalgoritmusok szerkesztése és alkalmazása, ami a jelenlegi tudományos kutatások egyik legaktívabban prosperáló irányzataként tartható számon.

A szövegfelolvasó rendszerek a beszédszintézis klasszikus ágát képviselik. Emellett azonban olyan területek is kezdenek egyre élénkebben előtérbe kerülni, melyek kevésbé kidolgozottak, és rengeteg nyitott probléma vár még megoldásra. Ide sorolható például az artikulációs beszédszintézis, ami az akusztikai produktum utánzását emberi hangminták helyett a hangképzés és artikuláció gépi leképezése révén próbálja megvalósítani. Ennek egyik technológiai vonulata a robotok beszédének előállításához szükséges artikulációs elektromechanikus beszédkeltőkre irányuló kísérletezés. Szintén a jövő tendenciáinak kedvez a gégétől a száj-, illetve orrnyílásig terjedő artikulációs csatorna, más néven vokális traktus modellezésére épülő beszédszintézis, ami főként vizuális információkra támaszkodik. Az emberi beszéd fiziológiai folyamatairól nyert vizuális információk nagymértékben elősegítik a beszédképzés komplex mechanizmusának megértését, és ezen keresztül a beszédszintézis módszereinek hatékony fejlesztését. A napjainkban rendelkezésünkre álló radiológiai és monitorozó eljárások – úgymint mágneses rezonanciás képalkotás (MRI), komputertomográfia (CT), ultrahang (UH), elektropalatográfia (EPG), elektromágneses artikulográfia (EMA) vagy elektroglottográfia (EGG) – nélkülözhetetlen szerepet játszanak

az akusztikai-artikulációs konverzió problémájának kezelésében. A fentebb említett képalkotó és monitorozó technikák segítségével generált morfológiai és geometriai adatok felhasználásával maradéktalanul feltérképezhetők az adott beszédjelhez tartozó artikulációs mozgások. Nem triviális feladat azonban az artikuláció akusztikummal való összekapcsolása, azaz a vokális traktus morfológiai és geometriai adataira alapozott beszédprodukció megvalósítása. A problémafelvetés aktualitását mutatja, hogy az artikulációs-akusztikai kapcsolatrendszer feltárása, illetve gyakorlati leképezése alapvető fontosságú lehet például a klinikai célú beszédterápiában, a nem anyanyelvi nyelvtanulási tréningek kialakításában vagy a néma beszéd megszólaltatásához szükséges szintetizátorok konstrukciójában és fejlesztésében, ami szolgálhatja többek között a gégeeltávolításon átesett emberek rehabilitációját is.

A kutatómunka során érdemes figyelmet szentelni a különböző képalkotó eljárások segítségével előállított vizuális információk összehasonlítására és összehangolására is, hiszen a különböző forrásokból származó adatok szimultán elemzése tovább mélyítheti a beszéd artikulációs és akusztikai perspektíváihoz kapcsolódó tudást. A monitorozó technikák egyidejű komparatív alkalmazása egyáltalán nem triviális feladat, mivel a források megfelelő és hiteles összeegyeztetése nagyon komoly és szakszerű anatómiai, geometriai, mérnöki és informatikai ismereteket igényel. A törekvés azonban gyakorlatilag elengedhetetlen a vokális traktus működésének részletes feltárásához, illetve az artikulációs-akusztikai összefüggések mélyebb megértéséhez.

A fentebbiek tükrében doktori kutatómunkám egyik fő célkitűzése a beszéd közben készült kétdimenziós UH- és MRI-

felvételek radiális és négyszöges geometriáinak összehangolása volt, aminek alapját a felvételekre automatikus algoritmusokkal illesztett nyelv- és szájpadkontúrok képezték. Ezt a feladatot analitikus szabályszerűségekre és mesterséges intelligenciára támaszkodó megközelítésekben szándékoztam megvalósítani. Az analitikus irányvonalat követve olyan geometriai transzformációkat dolgoztam ki, melyek kölcsönösen egyértelműen és kétirányú módon összekapcsolják és egymásba ágyazzák a két forrás anatómiai környezeteit úgy, hogy a matematikai műveletek paramétereinek optimalizációja révén elérhető legyen az UH- és MRI-felvételek nyelv- és szájpadkontúrjai közötti lehető legjobb egybevágóság. A mesterséges intelligencia érvényesítéséhez gépi tanulóalgoritmusok alkalmazását irányoztam elő olyan neurális hálózatok megszerkesztésével, melyek az UH-nyelvkontúrokból kivont paraméterekre hagyatkozva realizálják az MRI-nyelvkontúrokból származó adatok betanulását.

Kutatásaim másik nagyobb témaköre az artikulációs beszédszintézis kivitelezése volt, amihez szintén a fentebb említett UH- és MRI-felvételeket használtam fel. Első lépésben célul tűztem ki a szintézis alapjául szolgáló képalkotó források releváns geometriai adatainak dinamikus kinyerését, melyek birtokában önálló beszédhangok, illetve folyamatos beszéd előállítását vettem tervbe. Ennek során ismét a mesterséges intelligenciát kívántam segítségül hívni. Olyan gépi tanulóalgoritmusok felépítésére törekedtem, melyek a vizuális geometriai jellemzőkből kiindulva végrehajtják az UH- és MRI-felvételek akusztikai jeleiből eredeztetett artikulációs paraméterek betanulását.

2. Az alkalmazott módszerek és modellek

A kutatómunkám során kitűzött feladatokat a releváns elméleti tudás birtokában számítógépes módszerek segítségével valósítottam meg. Az UH- és MRI-felvételek kezelése nem nélkülözhette a képfeldolgozás bizonyos fogásainak bevetését, és a képi információkhoz kapcsolódó anatómiai kontúrvonalak megállapítását a dinamikus programozáson alapuló automatikus nyelvkontúrkövető algoritmusok felhasználása tette lehetővé. Az UH- és MRI-felvételek struktúráira és azok összehangolására irányuló vizsgálataimban kulcsfontosságú szerepet kaptak olyan matematikai megközelítések, melyek révén analitikus geometriai megfontolásokat és egzakt összefüggésekkel leírható transzformációkat érvényesítettem a két forrás vizuális elemeinek kölcsönösen egyértelmű megfeleltetésében. A geometriai transzformációk realizálását optimalizációs elvekre épülő algoritmusok kidolgozásával egészítettem ki. Kutatómunkám számos pontján alkalmaztam gépi tanulást, azaz mesterséges intelligenciát produkáló neurális hálózatokat konstruáltam az adott feladathoz tartozó paraméterek betanítására. A beszédszintézis végrehajtásakor a beszédtechnológia egyik leglényegesebb eszközeiként számontartott akusztikus csőmodellre, illetve lineáris predikció elvére támaszkodtam. Az elemzések egyes részfolyamatainak programozásához szükséges kódok mindegyikét MATLAB-felületen írtam meg, kutatásaim egyes fázisaiban azonban egyszerű manuális kalkulációk is helyet kaptak a vizsgálatokban.

3. Eredmények, tézisek

1. Az első szakaszban UH- és MRI-felvételek szimultán elemzésére és összehangolására fókuszáltam, melynek célja az UH-keretek radiális geometriájának, illetve az MRI-keretek négyszöges geometriájának kölcsönösen egyértelmű megfeleltetése volt. A két forrás összehangolásának egy lehetséges módja az artikuláció szempontjából releváns és azonos típusú szájüregi kontúrvonalak kétirányú konverziója. Ezt az elgondolást speciális geometriai transzformációk kivitelezésével valósítottam meg, melynek kézenfekvő eszközei a nyelv- és szájpadkontúrok voltak. A transzformációk vizuális sémáit és jellemző paramétereit magába foglaló matematikai keret kijelölését követően olyan optimalizációs technikákat alkottam meg, melyek eltérő perspektívák alapján megkeresik a transzformációk által definiált paraméterhalmaz legkedvezőbb értékeit. A transzformációk és az optimalizáció segítségével egymásra vetítettem a két forrás környezetét úgy, hogy relatív helyzetük ideális legyen, azaz az UH- és MRI-kontúrok közti globális távolság minimális legyen, biztosítva ezzel a görbék közti lehető legnagyobb átfedést. A transzformációs mechanizmus alapvetően három operációt ölel fel, ami a nyelv- és szájpadkontúrok által lefedett geometriai tartományokat érinti. A három művelet a radiális tartomány skálázása, a szögtartomány skálázása, valamint a szögtartomány forgatása által deklarált mozzanatokból tevődik össze. A sugár- és szögtartományok normálásáért egy-egy skálafaktor felel, ami nem más, mint a radiális nagyítás és a szög szerinti deformálás. A szögtartomány forgatása pedig egy transzlációs faktorral vehető figyelembe, ami a szögelfordulást adja. A szög szerinti deformá-

lást egységnyinek választottam, azaz a transzformációt szögtartó leképezésként értelmeztem, így az optimalizálandó paramétereket a radiális nagyítás és a szögelfordulás alkotta, melyekhez hozzávettem a transzformációk origójául szolgáló középpont koordinátáit is.

$\underline{T_1}$ **tézis**: **A kétdimenziós UH- és MRI-felvételek radiális és négyszöges struktúrái egymásba ágyazhatók olyan geometriai transzformációk révén, melyek megvalósítják a felvételek nyelv- és szájpadkontúrjai által definiált geometriai tartományok kétirányú konverzióját: a radiális tartomány skálázását, illetve a szögtartomány forgatását. Az operációkhoz kapcsolódó radiális nagyítás, szögelfordulás, valamint a transzformációk középpontjának optimalizálásával elérhető az UH- és MRI-felvételek anatómiai tartományai közötti legjobb átfedés. [1,2,4,5,9,11]**

2. A második szakaszban továbbra is az UH- és MRI-felvételek összehangolását céloztam meg, melynek kellékei a nyelvkontúrok voltak. Ezúttal azonban az analitikus geometriai szemléleten és az abból adódó egzakt matematikai transzformációkon alapuló optimalizációs megközelítést mesterséges intelligencia alkalmazásával váltottam fel. Ez azt jelenti, hogy a két forrás nyelvkontúrjait gépi tanulóalgoritmusok segítségével kapcsoltam össze, melynek során a neurális hálózatot eltérő konstrukciók szerint alkottam meg, és a kontúrok alaki sajátságait két külön-

böző jellegű paraméterrel vettem figyelembe. Egyrészt kijelöltem egy véges számú elemből álló diszkrét pontsorozatot a görbék mentén, létrehozva ezzel a tanítópontok halmazát. Másrészt pedig a görbék simítására alkalmazott diszkrét koszinusztranszformáció (DCT) végrehajtásával DCT-együtthatókat származtattam a kontúrokból. A neurális hálózatok bemenetét az UH-nyelvkontúrokból kinyert adatokkal gerjesztettem, a kimeneten pedig az MRI-nyelvkontúrokból kivont paramétereket állítottam be, tehát a rendszer végső soron a különböző beszédhangokhoz tartozó MRI-nyelvkontúrok alakjait tanulja be a tanítómintázatként alkalmazott UH-nyelvkontúrok alapján. Variálva a neurális hálózat bemeneti és kimeneti paramétereinek típusát és számát, illetve a rejtett rétegek és neuronok számát, különböző rendszerkonfigurációkat konstruáltam, és a kapott eredményeket kvalitatív és kvantitatív módon is analizáltam, kiválasztva a legjobb eredményt hozó rendszerbeállítást. A gépi tanítást kezdetben a bemeneti adatok forrásául szolgáló összes UH-nyelvkontúr felhasználásával vittem véghez, majd az UH-nyelvkontúrok seregén ötfokozatú szűrést végeztem, amivel kizártam a fals görbéket. Ennek eredményeképpen szelektáltam a hangátmenetekhez tartozó, a negatív iránytangensű, a szájpadkontúron túlcsorduló, a konvex és a kirívó kontúrokat. A betanított és az automatikus kontúrkövető algoritmussal illesztett MRI-nyelvkontúrok közötti kvalitatív és kvantitatív egyezés mértéke javul, ha a bemeneti adatok forrásául szolgáló UH-nyelvkontúrok halmazából kizárjuk a hangátmenetekhez tartozó, a negatív iránytangensű, a szájpadkontúron túlcsorduló, a konvex és a kirívó görbéket. A betanított és az automatikus kontúrkövető algoritmussal illesztett MRI-nyelvkontúrok közötti kvalitatív és kvantitatív egye-

zés mértéke javul, ha a bemeneten növeljük a tanítópontok számát függetlenül a kimeneti paraméterek típusától. A betanított és az automatikus kontúrkövető algoritmussal illesztett MRI-nyelvkontúrok közötti kvalitatív és kvantitatív egyezés mértéke javul, ha a bemeneten növeljük a DCT-együtthatók számát függetlenül a kimeneti paraméterek típusától.

$\underline{T_2}$ **tézis**: **A kétdimenziós UH- és MRI-felvételekre illesztett nyelvkontúrok gépi tanulóalgoritmusok segítségével összehangolhatók úgy, hogy UH-nyelvkontúrok paramétereire alapozva betaníthatók az MRI-nyelvkontúrok alakjai. A betanított és az automatikus kontúrkövető algoritmussal illesztett MRI-nyelvkontúrok közötti kvalitatív és kvantitatív egyezés mértéke javul, ha a kimeneten tanítópontok helyett DCT-együtthatókat alkalmazunk függetlenül a bemeneti paraméterek típusától. [3,6,7,8,10]**

3. A harmadik szakaszban gépi beszéd előállításával foglalkoztam, melynek kiindulópontját az UH- és MRI-felvételek képezték. Az volt ugyanis az alapvető elképzelésem, hogy a beszédszintézist olyan vizuális információkra támaszkodva valósítsam meg, amik az említett kétdimenziós képi forrásokból kinyerhetők. A szükséges vizuális adatokat egyrészt a felvételekre illesztett nyelv- és szájpadkontúrok segítségével származtattam úgy, hogy kidolgoztam két különböző algoritmust, melyek alkalmazásával a vokális traktusban dinamikus módon megmérhetők a

szájpad és a nyelvfelszín közötti szagittális radiális távolságok. Másrészt pedig a felvételekre illesztett nyelvkontúrok simítására használt diszkrét koszinusztranszformáció (DCT) együtthatóit is bevontam a vizsgálatokba. A kapott távolságadatokat és DCT-együtthatókat a gépi tanulás eszközei révén próbáltam meg összekapcsolni a beszédet jellemző különböző artikulációs paraméterekkel, melyeket az akusztikus csőmodell, valamint a lineáris predikció elvének (LPC) keretében értelmeztem. Ennek megfelelően a neurális hálózat bemenetét radiális távolságokkal vagy DCT-együtthatókkal gerjesztettem, a kimeneten pedig a beszédjelből közvetlenül kivont és az akusztikus csőmodell által definiált reflexiós tényezőket vagy a beszédjelből eredeztetett LPC-együtthatók közvetítésével kapott keresztmetszeteket állítottam be. A munkám során önálló, kitartott beszédhangokat, illetve folyamatos beszédet kívántam produkálni. A beszédhangok szintézise UH-MRI kombinációban történt, ami azt jelenti, hogy az UH-felvételekből származó vizuális adatokkal eszközöltem az MRI-felvételek beszédhangjainak betanítását. A folyamatos beszéd előállítását pedig UH-UH, illetve MRI-MRI párosításban kiviteleztem, tehát a rendszer mindkét oldalán ugyanazon forrásból kinyert paramétereket értelmeztem a tanításhoz.

T_3 **tézis**: **A kétdimenziós UH- és MRI-felvételekből kiindulva artikulációs beszédszintézis valósítható meg mesterséges intelligencia felhasználásával. Szagittális radiális távolságokra, valamint DCT-együtthatókra alapozva neurális hálózatok segítségével betaníthatók a beszédjel artikulációs paramétereiként kezelhető reflexiós tényezők, illetve a vokális traktusbeli keresztmetszetek,**

melyekből az akusztikus csőmodell és a lineáris predikciós kódolás szerint is rekonstruálható az eredeti beszédjel. A szintézis során a keresztmetszetekkel szemben a reflexiós tényezők betanításával jobb minőségű gépi beszéd állítható elő, és a tanítóalakzatok szintjén a radiális távolságokkal szemben előnyt élveznek a DCT-koefficiensek. [12]

4. A publikációk jegyzéke

a.) A doktori disszertáció témájához kapcsolódó közlemények:

1. R. Trencsényi, *MRI- és UH-felvételek geometriai elemzése a beszédszintézisben*, Acta Medicinae et Sociologica 11(31), 55-65, 2020.

2. R. Trencsényi, L. Czap, *UH-és MRI-nyelvkontúrok optimalizációja*, In Speech Research Conference, Hungarian Research Institute for Linguistics, Budapest, Hungary, 14-15th December 2020, 86-88, 2020.

3. R. Trencsényi, *A nyelvkontúrkövető algoritmusok és a gépi tanulás összekapcsolhatóságának vizsgálata*, In XVI. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY 2020, Szeged, Magyarország, 2020. január 23–24., 233-244, 2020.

4. R. Trencsényi, L. Czap, *Possible methods for combining tongue contours of dynamic MRI and ultrasound records*, Acta Polytechnica Hungarica, 18(4), 143-160, 2021.

5. R. Trencsényi, L. Czap, *A possible optimisation procedure for US and MRI tongue contours*, In Proceedings of the 1st Conference on Information Technology and Data Science, CITDS 2020, CEUR Workshop Proceedings, Debrecen, Hungary, 6-8th November 2020, 259-269, 2021.

6. R. Trencsényi, L. Czap, *Machine learning applied in speech science*, In 23rd International Carpathian Control Conference, ICCC 2022, Piscataway (NJ), Amerikai Egyesült Államok: IEEE, Sinaia, Romania, 29th May - 1st June 2022, 309-314, 2022.

7. R. Trencsényi, L. Czap, *Articulatory data of audiovisual records of speech connected by machine learning*, In 2nd Conference on Information Technology and Data Science, CITDS 2021, Proceedings Piscataway (NJ), Amerikai Egyesült Államok: IEEE, Debrecen, Hungary, 16-18th May 2022, 297-301, 2022.

8. R. Trencsényi, L. Czap, *A neural network based approach for combining ultrasound and MRI data of 2-D dynamic records of human speech*, In 13th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2022, Piscataway (NJ), Amerikai Egyesült Államok: IEEE, Budapest, Hungary, 22nd-23rd September, 47-52, 2022.

9. R. Trencsényi, L. Czap, *Optimisation techniques in speech processing*, In Doktoranduszok Fóruma 2021, Miskolc-Egyetemváros, Magyarország, 98-104, 2022.

10. R. Trencsényi, *A gépi tanulóalgoritmusok hatékonyságának vizsgálata kétdimenziós ultrahang- és MRI-felvételek adatainak összekapcsolásában*, In Doktoranduszok Fóruma 2022, Miskolc-Egyetemváros, Magyarország, 84-89, 2023.

11. R. Trencsényi, L. Czap, *Association of relevant anatomic contours of ultrasound and MRI images by optimisation via different geometric parameters*, In 25th International Carpathian Control Conference, ICCC 2024, Krynica Zdrój, Poland, 22nd-24th May 2024, pp. 1-6.

12. R. Trencsényi, L. Czap, *Ultrasound- and MRI-based speech synthesis applying neural networks*, In 25th International Carpathian Control Conference, ICCC 2024, Krynica Zdrój, Poland, 22nd-24th May 2024, pp. 1-6.

b.) Egyéb közlemény:

– R. Trencsényi, L. Czap, *Artikulációs fonetikai jellemzők verifikálása kvantitatív adatokkal*, Beszédtudomány/Speech Science, 2(1), 243-260, 2021.