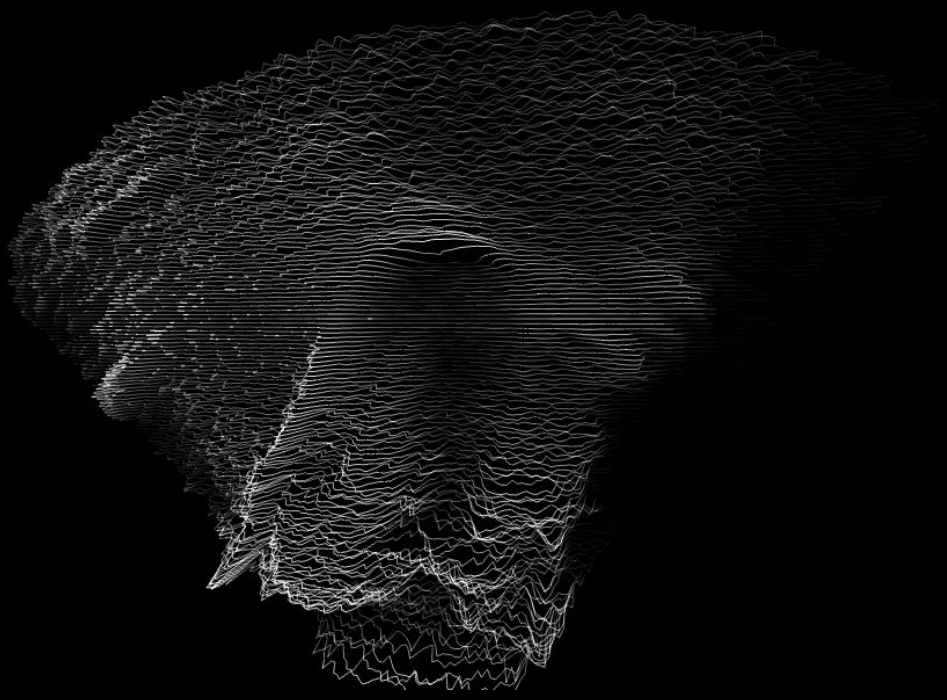
Mély neuronhálók alkalmazása ultrahangos nyelvkontúr követésre

TDK dolgozat



Készítette:

Csopor Dávid

csopor.david@gmail.com

Konzulens:

Csapó Tamás Gábor

csapot@tmit.bme.hu

MŰEGYETEM 1782

2014. október

Tartalomjegyzék

1. Bevezetés

Az emberi kommunikáció egyik alapvető eszköze a beszéd, mely hanghullámok (longitudinális hullámok) formájában közvetíti az információt. A beszéd összefoglaló neve mindannak, amit egy nyelvi közösség tagjai érintkezésük során hangosan közlésként mondanak [1]. A Földön mintegy 7000 élő nyelv található meg, melyek változatosak hangkészletükben, szókincsükben és nyelvtanukban [1].

Az artikuláció (a beszélő szervek mozgása) és az akusztikum (a keletkezett beszédjel) kapcsolata régóta foglalkoztatja a beszédkutatókat. A beszéd közbeni nyelvmozgást többek közt ultrahang, EMMA (elektromágneses artikulográf), MRI (mágnesesrezonancia-képalkotás) és röntgen felvevőkészülékekkel lehet rögzíteni. Az eljárások közül az ultrahangos felvétel ideális, mivel egyszerűen használható, elérhető árú, valamint nagyfelbontású (800x600 pixel) és nagysebességű (akár 100 kép/sec) felvétel készíthető vele. Az ultrahangos technológia hátránya viszont ebben a témakörben, hogy a rögzített képsorozatból ki kell nyerni a nyelv körvonalát ahhoz, hogy az adatokon további vizsgálatokat lehessen végezni [7]. A nyelvkontúr követés hagyományosan manuális vagy fél automatikus módon történt, azonban az elmúlt időszakban automatikus megoldások is megjelentek erre a célra (pl. AutoTrace: [2]).

A kutatás során a legújabb automatikus nyelvkontúr követő módszerek közül a nemzetközi szakirodalomban is előtérbe került mély neuronháló alapú technikákat vizsgáltuk [3]. Az Indiana University beszédkutató laboratóriumában rögzített két beszélő (egy magyar és egy amerikai angol) ultrahangos felvételein az AutoTrace különböző mély neuronháló elrendezéseit elemeztük annak eldöntésére, hogy melyik architektúra legalkalmasabb a feladatra [2], [3]. Emellett meghatároztuk, hogy a tanítóadat mennyiségének függvényében milyen mértékben tudja az automatikus nyelvkontúr követés a manuálist közelíteni. A tipikus hibák (például eltávolodás az eredeti nyelvkontúrtól; hiányzó nyelvkontúr szakaszok) számszerűsítésére több hibamértéket hasonlítottunk össze.

Az automatikus nyelvkontúr követés a beszédkutatás alapkérdéseinek (pl. a nyelv 3D-s mozgása milyen mértékben járul hozzá az akusztikai kimenet formálásához?) megválaszolása mellett hasznos lehet nyelvoktatásban, beszéd rehabilitációban illetve beszédtechnológiában, audiovizuális beszédszintézisben is [4].

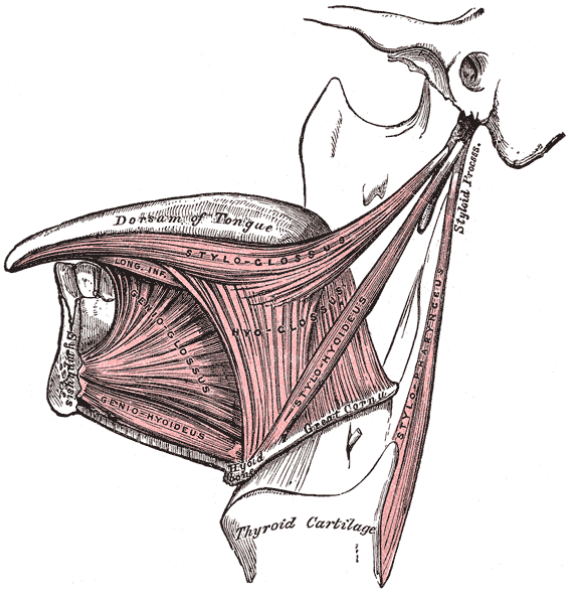
2. Elméleti háttér

Ebben a fejezetben áttekintjük a dolgozat megértéséhez szükséges alapfogalmakat. Szó lesz a nyelv felépítéséről, az ultrahang technológiáról és a mi esetünkben keletkezett ultrahangkép értelmezéséről. Áttekintünk egy szabadon használható automatikus nyelvkontúr követő alkalmazást (AutoTrace) és az abban implementált neurális háló felépítését.

2.1. Beszéd anatómiai háttere

A beszédet az ember a hangképző szerveivel hozza létre, ami megvalósulási formájában egyfajta levegőrezgés, így személyre szabott hangzást kap. Ezt befolyásolhatja az egyén egészségügyi állapota, arcmérete, a hangszalagok hossza, tömege, illetve a beszéd tudatos módosítása. A beszédképzés fiziológiai szervei a tüdő, a légcső, a szubglottális tér, a gége, a glottális tér, a garat, a száj– és orrüreg illetve a szupraglottális tér [1]. A szervrendszer működésének folyamatát az agy irányítja, melynek lényege, a tüdőből kiáramló levegőáramból és a gégeműködésből kialakuló zöngés vagy zöngétlen hang képzése, mely továbbhaladva módosul a gége feletti üregekben. Ezt a változást befolyásolják a passzív szervek, mint a fogak és a szájpad, illetve az aktívak mint a nyelv, nyelvcsap, állkapocs és az ajkak [1].

Az ultrahangos felvételek megértéséhez érdemes áttekinteni a nyelv felépítését. A nyelv különböző izomrostokból álló tömött szerv, melynek hátsó része a nyelvgyök, középső része a nyelvtest és az elülső része a nyelvcsúcs [5]. Nem csak a beszédben van fontos szerepe, hanem a falatképzésben, rágásban, nyelésben, de ízlelőszerv is [5]. A nyelv fő tömegét és mozgékonyságát adó harántcsíkolt izmok részben kívülről sugároznak a nyelvbe, részben a nyelv saját izmai. A kívülről jövők a nyelv helyváltoztatását a saját - belső - izmok a nyelv alakváltoztatását biztosítják [5].



1. ábra. Oldalmetszetben látható az állkapocs, a nyelv felülete, harántcsíkolt izomszövet rendszere és az alsó metszőfogak [6]. Az ultrahangos képeken a jobb oldalon lesznek láthatóak a metszőfogak.

2.2. Ultrahangos felvétel technológia

A nyelv monitorozása közben olyan eszközt kell használni ami nem érzékeny a hőmérsékletre és a nedvességre, illetve nem akadályozza a nyelvmozgást. Az utolsó feltétel miatt a nyelv vizsgálata a beszédtechnológiában csak az 1980-as évektől indult el, ekkor még az ultrahang gépeket csak egészségügyi intézményekben nagy kihasználtság mellett lehetett igénybe venni, viszont a technológia árcsökkenésével, javuló képi és felvételi minőségével vált elérhetőbbé a kutatók számára is [7]. Ultrahanggal történő információ szerzés a természetben sem ismeretlen, denevérek illetve delfinek magas frekvenciájú hangokat bocsájtanak ki, így érzékelve a tárgyakat.

2.2.1. Technológia

Esetünkben az ultrahang egy, vagy több piezoelektromos kristályból kibocsájtott nagyfrekvenciájú (MHz) longitudinális hullám, melynek a közeghatárról és a szöveti változásokról történő visszaverődését használjuk fel a képalkotás során [7]. A nem szövetromboló és valósidejű képmonitorozást biztosító nagyfrekvenciás hanghullámok

előállítását Pierre és Jacques Curie (1880) a piezoelektromos kristályokkal kapcsolatos vizsgálatai tették lehetővé [7].

A piezoelektromos kristály a mechanikus energiát elektromos energiává, illetve az elektromos energiát mechanikus energiává alakítja át a kristályban lévő molekulák elhelyezkedésének változtatásával. Ultrahang előállításánál a hanghullám frekvenciája függ a kristály vastagságától és a kristályt oldalról bezáró fémlapokra adott feszültségtől. Egy adott felületről visszaverődő hanghullám a kibocsájtás és az érzékelés között eltelt időt hordozza információként, amiből a hang adott közeghez tartozó terjedési sebességének segítségével kiszámítható a visszaverődés pontos helye.

Minél nagyobb a kibocsájtott hullám frekvenciája, annál jobb a felbontás, azaz kisebb szövetméreteket tudunk érzékelni, viszont a nagyobb frekvenciás hullámok gyorsabban nyelődnek el a szövetben, így kevésbé tudjuk a mélyebb szöveteket vizsgálni. Egy kristályból származó ultrahang hullám lehet fókuszálatlan illetve fókuszált, ami növeli az ultrahang hullámra merőleges felbontást, viszont behatárolja a mélységi képalkotás minőségét, mivel a fókusz-zónán túl gyors széttartás lép fel a hullámban [7].

Az alábbiakban röviden három szkennelési eljárást említenénk meg, amikkel kapcsolatban még több információ található meg Stone 2005-ös cikkében [7].

Az A típusú szkennelési módszer során egy kristályból adott irányba emittált hullámok az egyes közeghatárokon visszhangot hoznak létre, amit ha a vevő érzékel, az ultrahang gép jelében egy tüskeként jelenik meg, melynek amplitúdója összefüggésben van a közeghatár minőségbeli változásának nagyságával [7].

Az M típusú módszer nem egy, hanem több azonos kristályból történik a hullámok emittálása. A módszer a visszaverődést okozó felületek helyzetét ábrázolja az idő függvényében [7].

A valós idejű B szkenner ultrahang transzducer (adó és vevő) egy sor azonos piezoelektromos kristályból áll, felfogják a visszaverődött hullámokat, így a módszer több A típusú szkennert valósít meg [7]. A vevő az érzékelt visszhangokat elektromos jellé alakítja, majd a processzáló egység a jelet 2D-s szürkeárnyalatos, 90 és 120 fok közötti ék alakú képpé konvertálja [7]. A vevőtől távolabbról érkező visszhangokhoz kirajzolt pontok a megjelenített képen is messzebb helyezkednek el. A kirajzolt pontok fényességét a közegátmenetek minősége határozza meg. Ezzel a módszerrel készültek a dolgozat későbbi fejezeteiben a méréshez használt ultrahangos felvételek.

2.2.2. Felvétel analizálása

Az ultrahangos készülékek általában 10 -100 képet készítenek másodpercenként, amit úgy képesek tartani változó szkennelési sebesség mellett (30, 80, 90 Hz), hogy képeket duplikálnak, vagy éppen hagynak el. A sebességváltozás a vizsgálni kívánt szövet mélységétől függ, minél mélyebb, annál tovább tart a hullám adása és vétele között eltelt idő [7].

A legnagyobb határváltozást a nyelv felső határa okozza, ami így az ultrahangos képeken ideális esetben jól kivehető. Viszont mivel a hullámok nagy része nem jut tovább a nyelvhatáron, így a távolabbi szövetpontokról, a szájpadlásról kevesebb az információnk [7]. A 2. ábra bal oldalán egy ultrahangképen jól kivehető nyelvkontúr látszik, a jobb oldalon pedig vázlatosan szerepel az ultrahang fej az állkapocs alatt, zölddel a nyelv, illetve kékkel a szájpadlás.

2. ábra. A kép bal oldalán egy ultrahang felvételből kivehető nyelvkontúr, a jobb oldalán a fej metszete szerepel, zölddel a nyelv, kékkel a szájpadlás illetve piros határral az ultrahang felvételi tartománya. Az orientáció megegyezik az ultrahangos képen és a vázlaton.

Az ultrahang technológia nem mindig nyújt teljesen tökéletes nyelvkontúrt. A kép minősége függ a beszélőtől, általában fiatalabbaknál és nőknél jobb, de ez függ a száj hidratációjától is. A 2. ábrán látható, hogy a nyelvkontúr egy ponton megszakad, majd ismét folytatódik, előfordulhat olyan eset is, amikor csak egy része jelenik meg. Nem ritka a

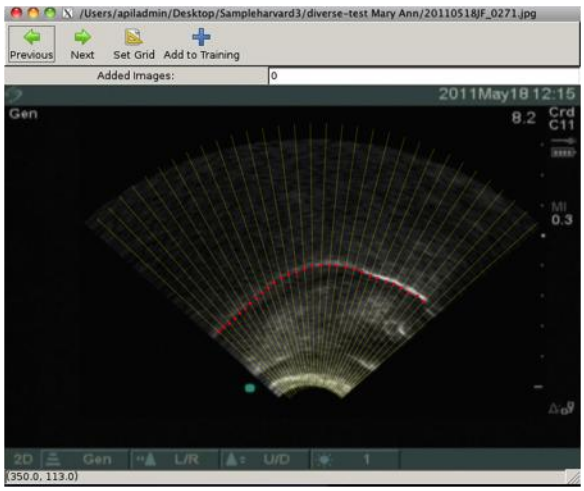
szakadás vagy az ugrás amikor a kontúr folytonossága megszakad, néha több kontúr is látszik egymás felett, amit a szájpadlás közelsége okoz.

2.3. Automatikus nyelvkontúr követő eljárások

A bevezetésben már említettük, hogy a nyelvmozgás monitorozására több módszert is felhasználtak már (EMMA, röntgen, MRI, ultrahang), de mindegyikből szükséges kinyerni a nyelv körvonalát. Minden képen egyesével berajzolni a nyelvkontúrt sokáig tart még egy szakértőnek is. Többféle automatikus nyelvkontúr követő eljárás is szabadon elérhető, mint az EdgeTrack [17], TongueTrack [18] és az AutoTrace [2]. A 2.3.1. alfejezetben az általunk használt eljárást, az AutoTrace-t mutatjuk be.

2.3.1 AutoTrace

A nyelvvonal követési idejét igyekszik lecsökkenteni az AutoTrace nevű nyílt forráskódú [8] Matlab és Python nyelven írt program, amivel a kézi nyelvkontúr vonalak bevitele illetve a betanított modell segítségével az automatikusan meghatározott kontúrvonal is megkapható [2]. Mivel az AutoTrace gépi tanulás alapú a tanítás során nagyméretű adathalmazt érdemes használni. Mint látszik az alábbi képen a nyelvkontúr egy polár-koordináta rendszerre kerül rá, mely 32 vonalból áll, így kvantálva az eredményt.



3. ábra. Az AutoTrace grid rendszere és a vonalakon lévő nyelvkontúrt jelölő pontok [8].

Az AutoTrace működéséhez a meglévő kézzel berajzolt nyelvkontúrral rendelkező képadathalmazból ki kell választani a neuronhálózatot betanító részhalmazt. A tanítóadat kiválasztására több módszert is kipróbáltak, mint például a nehezebben érzékelhető nyelvkontúr-képek elhagyását, illetve olyan képek keresését amikben a nyelvalakok különböznek egymástól [2]. A gyengébb minőségű ultrahangfelvételt okozhatta az alany dehidratáltsága vagy a rossz érintkezés az ultrahangfej és a bőr között.

A program a bemeneti tanítóadatot egy alanyokra (Subject) rendezett, statikusan kialakított mappaszerkezetben várja a következőket:

- Egy JPG mappában az összes képet az adott felvételből.
- Egy ROI (Region Of Interest) szöveges fájlt, ami az egész felvételre jellemző téglalap alakú területet adta meg, ahova a nyelvkontúrhoz tartozó pontok kerülhettek.
- Egy CSV fájlt, amiben az egyes képeken szereplő maximum 32 pont koordinátája szerepel.

A tanítóadat megadása után elindul a neuronháló tanítása, ami a tanítóadat függvényében néhány perctől néhány óráig tart. A modell elkészülte után a kiértékelést egy másik programrész végzi el.

Az AutoTrace egy folyamatos fejlesztés alatt álló program. Ebből adódóan nem volt minden problémamentes a használata során, például a grafikus interfészét nem sikerült rendeltetésszerűen használni, így külön Matlab scripttel kellett elindítani az AutoTrace egyes részfeladatait, mint a modell tanítást és a képek modellel történő nyelvkontúr lekövetését is.

2.4. Neuronhálók

Az idegrendszer és az idegsejt (neuron) tanulmányozása során jött az a gondolat, hogy érdemes lehet olyan számítási modelleket alkotni, melyek az élő szervezetben létező, bonyolult rendszeren alapulnak. Ezt az adaptív, párhuzamos számítási feladatokat végző eszközt mesterséges neurális hálózatnak nevezzük, melyek azonos felépítésű, egymással kapcsolatban lévő elemi egységből, neuronokból épülnek fel [9].

Bizonyos feladatok elvégzésére alapvetően jobbnak bizonyulnak a neurális hálók mint a hagyományos algoritmikus számítási rendszerek. Ilyen problémák a karakterek, számok, kézírás, kép illetve egyéb alakzatok felismerése és az olyan feladatok megoldása amelyekre hatékony algoritmikus alternatívát eddig nem sikerült találni [9].

A 2.4.1. alfejezetben a mély neuronhálók (DNN) felépítési tulajdonságairól lesz szó, a 2.4.2. alfejezetben pedig az AutoTrace-ben implementált DNN-t vizsgáljuk meg közelebbről.

2.4.1. Mély neuronhálók általánosan

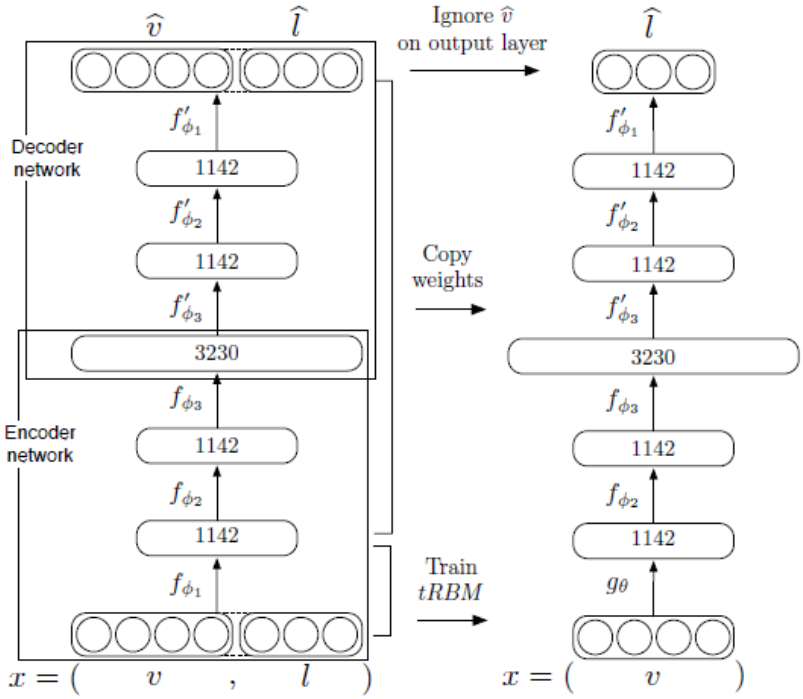
Egy egyszerű neuron modell egy többdimenziós bemenet komponenseinek lineáris kombinációját megvalósító hálózat, amelyet egy nemlineáris leképezés követ [9]. A neuronokon, a kívánt működés eléréséhez, iteratív tanulási eljárást végzünk, mely során a hálózat a működését reprezentáló összetartozó be- és kimeneti értékek alapján alakítja ki az átvitelét, illetve a hálózat bemenetére kerülő adatokban önmaga próbál hasonlóságot keresni [9].

A kezdeti neuronhálókkal kapcsolatos eredményeket egy kisebb szünet követte, mivel az új mély neurális hálózatokhoz nagyobb számítógépes kapacitás volt szükséges a gyorsabb működéshez [9].

A mély neuronhálok (DNN, Deep Neural Networks), abban különböznek a hagyományos neurális hálóktól, hogy a bemeneti (input layer) és a kimeneti (output layer) látható rétegeken kívül, egynél több rejtett réteg (hidden layer) is szerepel. A tanító modell minősége függ a rétegekben lévő neuronok számától. A bemeneti és a kimeneti látható rétegek neuron száma feladatfüggő. Kategorizálási feladatnál a kimeneti réteg a kategóriák számánál eggyel több szokott lenni. A bemeneti réteg képek analizálásánál, egy közbenső, kis méretű, fekete-fehér kép pixelszámával egyenlő.

2.4.2 Az AutoTrace-ben használt mély neuronháló

Az AutoTrace-ben használt neurális háló egy translational - Deep Belief Network (tDBN), mely lehetővé teszi, hogy egy szenzor adathalmaz (ultrahangos kép) és egy manuálisan bevitt adathalmaz (nyelvkontúr koordináták), segítségével egyszerre történjen meg a modell tanítása. A DBN egy altípusa a DNN-nek, ami több rejtett rétegből tevődik össze. A rétegek kapcsolódnak egymáshoz, az egyes rétegen belüli egységek nem.



4. ábra. Az AutoTrace-ben használt t-DBN vázlatos felépítése [16].

A 4. ábra bal oldalán található a neuronháló látható és rejtett rétegei. Legalul a bemeneti rétegben tanítás során kétféle adatot adunk meg, az ultrahangos kép egyszerűsített változatát (v-vel jelölve), és a manuálisan bevitt nyelvkontúrt (l-lel jelölve). Az efölött lévő részek a rejtett rétegeket jelölik. Az AutoTrace alapbeállításként két darab rejtett rétegű modellt használ, melyek neuron száma egy mátrixváltozó dimenziójától függ. A legfelső kimeneti rétegben a bemeneten megkapott kép közelítését, illetve a közelített nyelvkontúrt kaptuk.

Az iteratív tanítás során a belső rétegek neuronjaihoz tartozó súlyok úgy változnak, hogy a bemenő és kimenő nyelvkontúrok között minél kisebb legyen a különbség.

Magához a nyelvkontúr-követéshez a modell bemenete csak az ultrahang képeket veszi figyelembe, kimenetként pedig a nyelvkontúr koordinátáit adja meg, ez látható a 4. ábra jobb oldalán. Természetesen a tanítás során kialakult neuron súlyok átmásolódnak a letisztított modellbe.

3. Nyelvkontúr követő DNN modell tesztelése

Az irodalmi áttekintés végeztével a mérési feladat leírása következik. Bemutatjuk a vizsgálathoz használt ultrahangos adatbázisunkat (3.2), illetve az automatikusan lekövetett és a manuálisan bevitt nyelvkontúrok közötti eltérés mérésére meghatározott paramétereket (3.3).

3.1 Alapvető cél

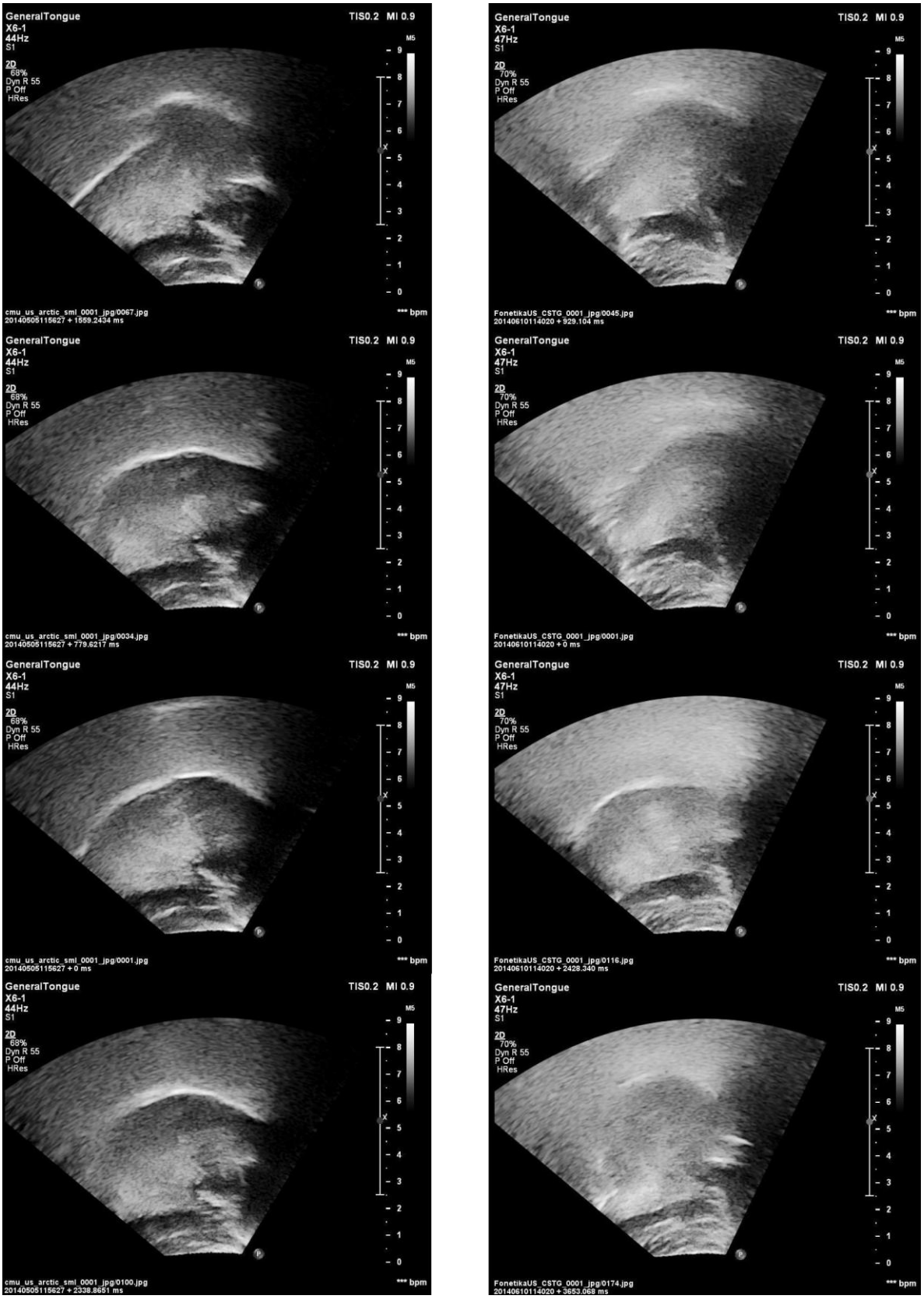
A méréssel alapvető célunk volt meghatározni a különböző mennyiségű ultrahangképpel tanított neurális háló tesztelés során mutatott pontosságát az emberi nyelvkontúr követéssel szemben és olyan paramétereket keresni melyek megmutatják a manuális és automatikus módszer közötti hibakülönbségeket. Cél volt továbbá megvizsgálni más neuronhálózat architektúrákat, és azok pontosságát kontúrkövetésben.

3.2 Adatbázis

A neuronháló tanításához és teszteléséhez használt ultrahang adathalmaz bemutatása következik.

3.2.1 Adatbázis alanyai

Az adatbázisban egy amerikai angol (jelölés: EN-FF) férfi és egy magyar anyanyelvű (jelölés: HU-FF) férfi beszélőtől találhatóak felvételek, melyek az Indiana University, Speech Production Laboratory süketszobájában készültek [10]. Az 5. ábrán a két beszélőtől látható pár ultrahang kép, amiből kivehető, hogy a bal oldali, EN-FF beszélő képei jobb minőségűek mint a jobb oldali HU-FF beszélő képei.



5. ábra. Bal oldalon EN-FF, jobb oldalon HU-FF beszélőtől származó ultrahang képek.

Ha az 5. ábrán található ultrahang képeket jobban megvizsgáljuk a következőket fedezhetjük fel:

- A bal oldali ultrahang képeken a nyelv belső struktúrája lényegesen jobban kivehető.
- A jobb oldali képeken általában rövidebb a látható nyelvszakasz.
- Az összes kép jobb oldalán található sötét folt az állkapocs csontja. A csontszöveten nem hatol át az ultrahang, ezért sötét.
- A képek minősége egy beszélőn belül is változhat, például a bal alsó képen már nem látszik a nyelvgyök.
- A bal felső képen egy nagyobb ugrás látható a nyelvkontúrban.
- Jobb oldalon a második képen két kontúr látható egymás fölött, a felső valószínűleg a szájpadlás.

3.2.2. Felvételi körülmények

Az EN-FF beszélő a CMU - ARCTIC adatbázis [11] első 135 mondatát, míg a HU-FF beszélő a PPBA adatbázis [12] első 210 mondatát olvasta fel. A felvételek során párhuzamosan készültek beszéd és ultrahang felvételek, azonban jelen kutatásban csak az ultrahangos képeket használtuk fel. A nyelv mozgását Philips EpiQ-7G ultrahangos rendszerrel és xMatrix 6-1 Mhz fejjel rögzítették.

A felvételek során az ultrahang fej elmozdulásának elkerülése végett egy speciálisan erre a feladatra kialakított sisakot alkalmaztak (típusa: Ultrasound Stabilisation Headset, Articulate Instruments Ltd), mellyel az ultrahang felvevő szorosan a beszélő álla alá rögzíthető [13].

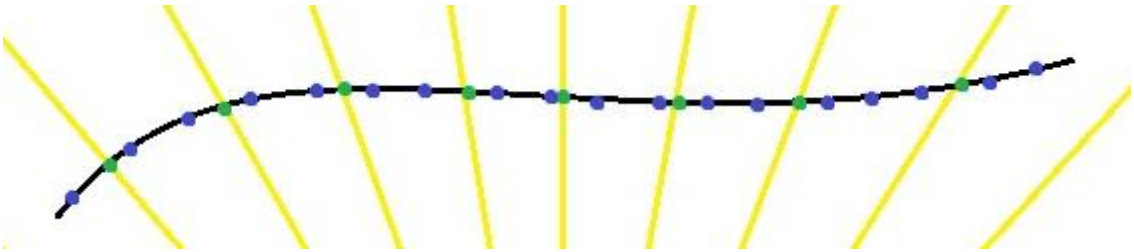
Az ultrahang adatok eredetileg DICOM formátumban készültek 800x600 pixel felbontásban és 40-50 kép/sec közötti sebességgel. A DICOM fájlokat az Image-J [14] programmal alakították JPG képekké.

A manuális nyelvkontúr követést a fenti felvételek egy kisebb részén (EN-FF beszélő: 8 mondat, HU-FF beszélő: 9 mondat) egy hallgató végezte egy speciálisan erre a feladatra készült weboldal segítségével. A manuális nyelvkontúr követés eredménye SQL adatbázisba került, melyből CSV formátumba exportálva lehet az adatokat feldolgozni.

Az EN-FF beszélőtől összesen 1140 ultrahangos képet, míg HU-FF beszélőtől összesen 1457 képet használtunk fel a kísérletekben.

3.3 Modellek pontosságát mérő hibamértékek

Az ultrahangos nyelvkontúr adatokat Descartes-féle koordinátában kaptuk meg, amit át kellett konvertálni egy meghatározott középpontú polár-koordinátává. Ez bizonyos adatvesztéssel jár, mivel 32 darab, meghatározott azimutális szöggel rendelkező egyenesre kellett illeszteni az adott pontokat, hogy az AutoTrace használni tudja őket. A 6. ábrán látható a kvantálás, kékkel az adatbázisból kapott pontok, zölddel az egyenesekre illeszkedő pontok.



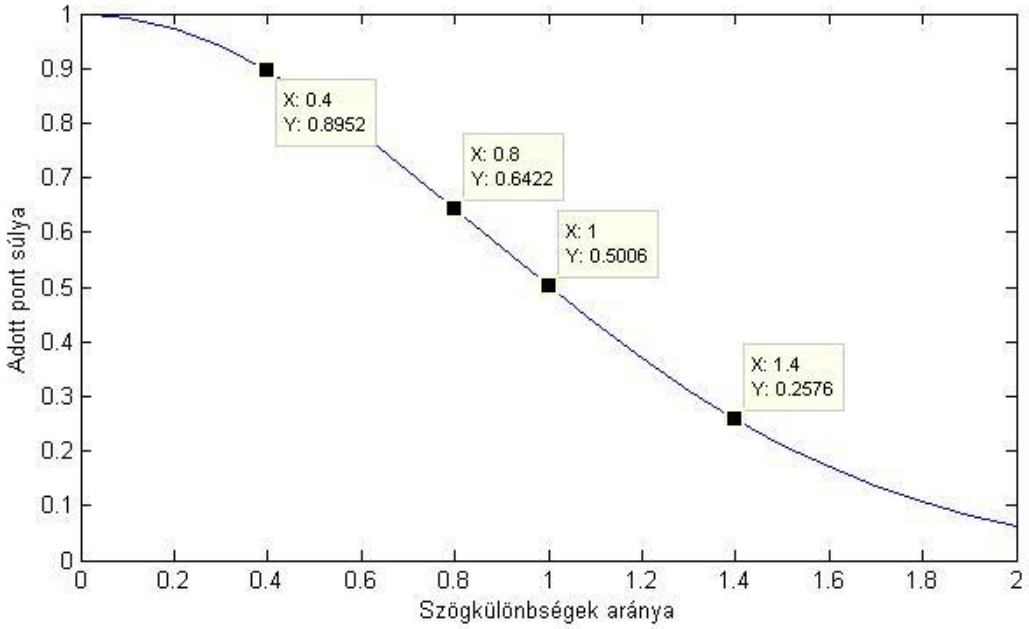
6. ábra. Adathalmaz (kék pontok) kvantálása a polár-koordináta adott szögű egyeneseire (zöld pontok).

A kvantálási hibát úgy próbáltuk minimalizálni, hogy adott polár-egyeneshez tartozó szomszédos két pontot, egy fél Gauss-függvény segítségével (1. képlet, 7. ábra) súlyoztunk aszerint, hogy a polár-koordinátára konvertált pontok szöge mennyire van közel az adott egyeneshez. Az első képletben ∂ értéke 0.85, c értéke pedig 0. Ezeket az értékek szükségesek ahhoz, hogy a függvény az egy bemenetre 0,5-öt adjon vissza, tehát azonos súllyal veszi figyelembe az egyenestől azonos szögtávolságra lévő pontokat (7. ábra).

A 7. ábrában szereplő szögkülönbségek arányát úgy kapjuk meg, hogy vesszük az egyik szomszédos pont és az adott polár-egyenes közötti szöget, majd elosztjuk a másik szomszédos pont és a polár-egyenes által bezárt szöggel. Ha az így kapott arányt átalakítjuk nulla és egy közötti értékre egy esetleges invertálással, akkor a számlálóban lévő érték

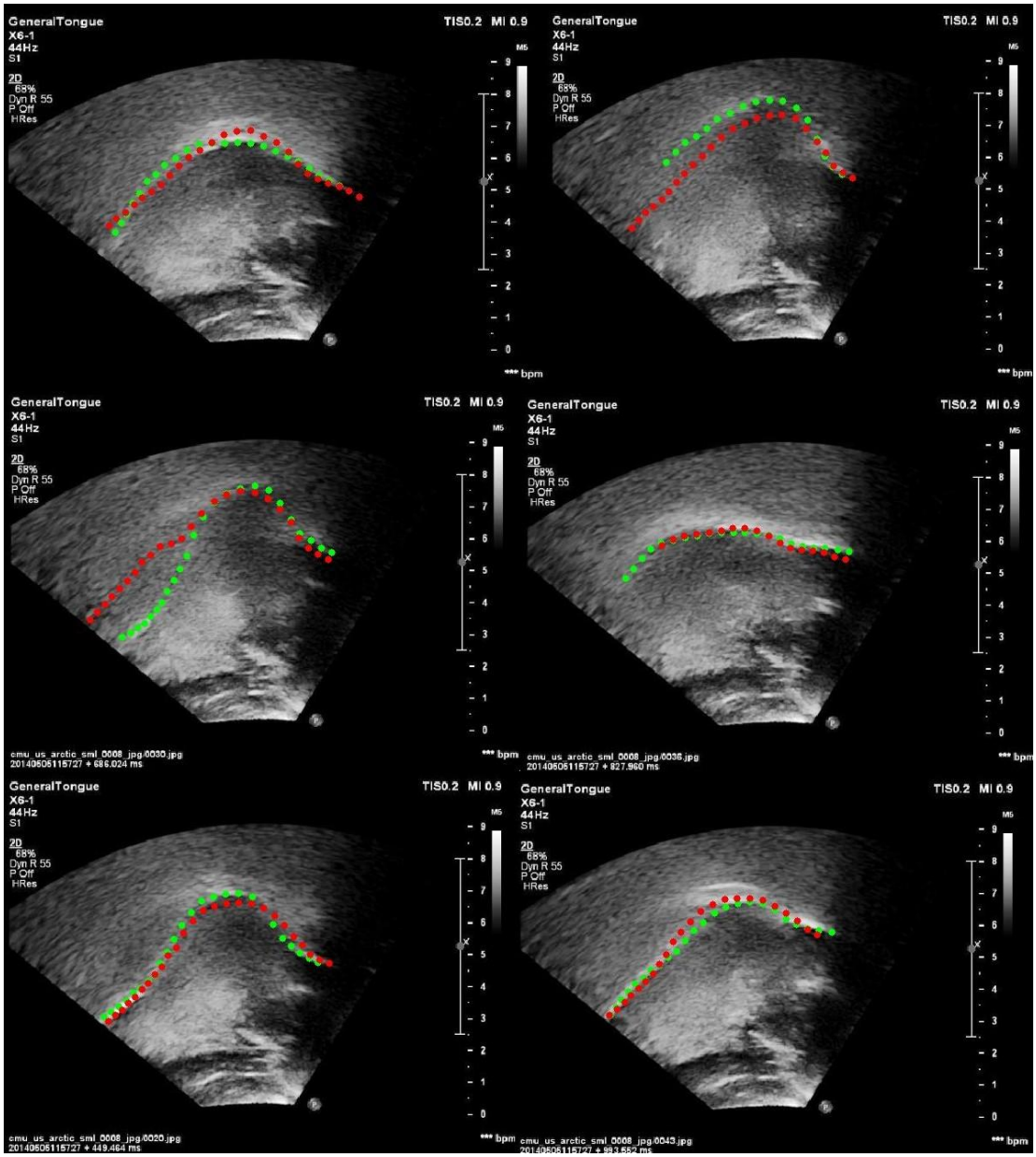
kisebb lesz, így a számlálóban felhasznált pont volt közelebb az egyeneshez, tehát ő kapja meg a Gauss-függvény kimenetét (P), míg a másik pont az egyhez képesti maradékkal (1-P) súlyozódik.

1.képlet: $$f(x;\ \partial, c) = e^{\frac{-(x-c)^2}{2*\partial^2}}$$



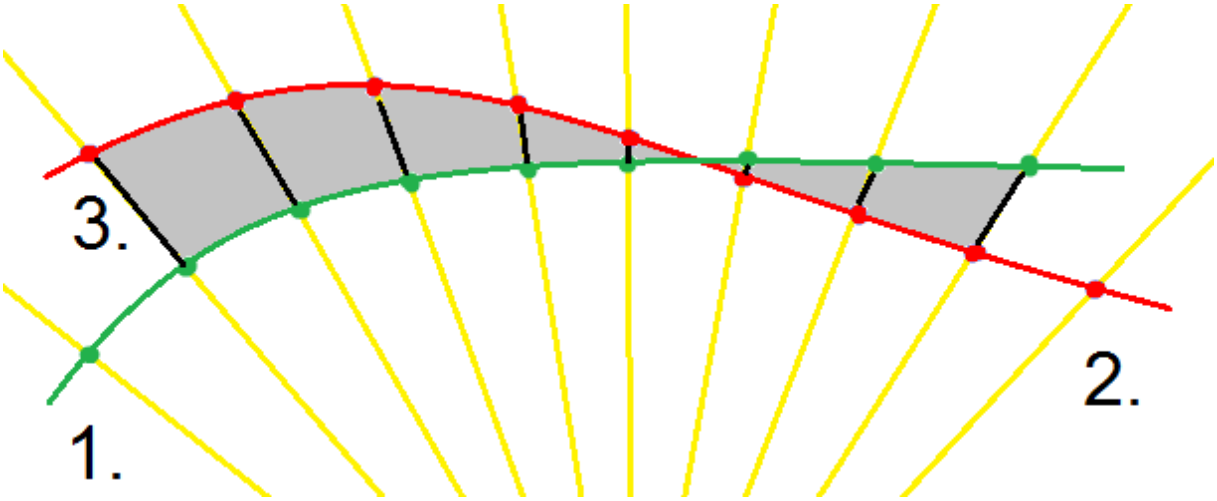
7. ábra. Az 1. képlettel megvalósított függvényt ábrázolja. A vízszintes tengely a két szomszédos pont és az adott polár-egyenes közötti szögek arányát, a függőleges tengely, pedig a kimeneti súlyt jelöli.

A 8. ábrán egy betanított neurális modell alapján elvégeztük a nyelvkontúr követését és az eredményt egyszerre ábrázoltuk az ultrahang képeken, zöld pontokkal a manuálisan bevitt, piros pontokkal a modell kimenetét ábrázoltuk.



8. ábra. Egy neuron modell kimenete (piros) és a manuális nyelvkontúr (zöld) több ultrahangképen.

A 8. ábrán látható, hogy előfordult olyan eset, amikor túltanítás, azaz a modell olyan helyen is érzékelt nyelvkontúrt, ahol a manuálisan jelölt pontok nem jelzik, például a jobb felső képen.



9. ábra. Zölddel a manuálisan bevitt, a pirossal jelölt a modell kimeneteként kapott nyelvkontúrt ábrázolja. A számok különböző hibamértékekre utalnak.

Az első paraméterünk az **RMSE** (Root-Mean-Square Error), avagy a négyzetes hibaátlag volt. A 2. képletben a nevező "n" értéke annyival egyezett meg, ahány polár-egyeneshez tartozott tanított és manuálisan bevitt pont is. Azaz, azokat a pontokat nem vettük figyelembe az RMSE számításánál, ahol nincs automatikusan és manuálisan meghatározott pont. Emiatt az RMSE hibamérték nem mutatja a követés teljes hibáját, kiegészítésére később más paramétereket is bevezetünk.

$\hat{y}_k$ a tanított nyelvkontúr vonalhoz tartozó k. polár-egyenesen lévő pont sugár irányú koordinátájának, y_k pedig a manuálisan bevitt nyelvkontúrhoz tartozó k. polár-egyenesen lévő pont sugár irányú koordinátájának értéke (9. ábrán fekete vonalak jelzik).

2. képlet: $$RMSE_{image} = \sqrt{\frac{\sum_{k=1}^{n}(\hat{y}_k - y_k)^2}{n}}$$

Második paraméterünk az **AREA**, azaz a két vonal közötti terület (9. ábrán szürkével jelölve). A pontokat a számítás során egyenes szakaszokkal kötöttük össze, így meghatározva a két kontúr közötti területet. Az AREA bevezetésének motivációja az volt, hogy az RMSE-vel szemben ez várhatóan pontosabban méri a manuális és automatikus nyelvkontúr közötti

hibát. Itt is figyelembe kellett venni azt, hogy csak akkor lehet területet számolni, ha adott polár-egyenesen van tanított és manuális pont is.

Az AREA hibamérték implementálása során figyelni kellett a szürkével jelölt (9. ábra), két polár-egyenes által közbezárt területre, ez lehetett négyszög, mivel a nyelvkontúr görbe pontjait szakaszokkal kötöttük össze, ekkor a két kontúrvonal nem metszi egymást. Illetve lehetett egy vagy két háromszög, ekkor viszont van metszéspont. Az első terület számítása egyértelmű, a nagyobbik háromszög területéből ki kell vonni a kisebb területét. A második esetben a háromszögek területének kiszámításához meg kellett határozni a kontúrvonalak metszéspontját, ami egy több ismeretlenes, geometriai egyenletrendszer megoldása volt.

Figyelembe szerettük volna venni az olyan hibákat is, mint amikor egy adott polár-egyenesen van manuálisan felvitt pont, de a tanítási adatban nincs, illetve amikor olyan egyenesre kerül tanítási adat, ahol nem volt manuális. A beszédfelismerésben használt Word Error Rate-hez [15] hasonlóan kialakítottuk a **Tracking Error Rate** (TER) változó hármast, mely tartalmazza az Insertion-t (beillesztés), a Deletiont (törlés) és a Substitution-t (helyettesítés).

Az **Insertion** (9. ábra 2. pontja) értéke egy képre megegyezik azon polár-egyenesek és a manuálisan meglévő pontok számának arányával, melyekre *került* tanított pont, de *nem került* manuális pont.

Deletion (9. ábra 1. pontja) értéke egy képre megegyezik azon polár-egyenesek és a manuálisan meglévő pontok számának arányával, melyekre *nem került* tanított pont, de *került* manuális pont.

Substitution (9. ábra 3. pontja) értéke egy képre megegyezik azon polár-egyenesek és a manuálisan meglévő pontok számának arányával, melyekre *került* tanított pont és *került* manuális pont is, de a kettő messzebb van egy adott távolságnál egymástól, mely távolságot mi 7 pixelben állapítottunk meg. Ezen távolság választásának oka, hogy egy friss kutatás szerint a manuális nyelvkontúr követés átlagos hibája a 7 pixelt közelíti [10].

4. Mérési eredmények

Az alábbiakban a mérésünk leírását és a hibaparaméterek eredményeit közöljük.

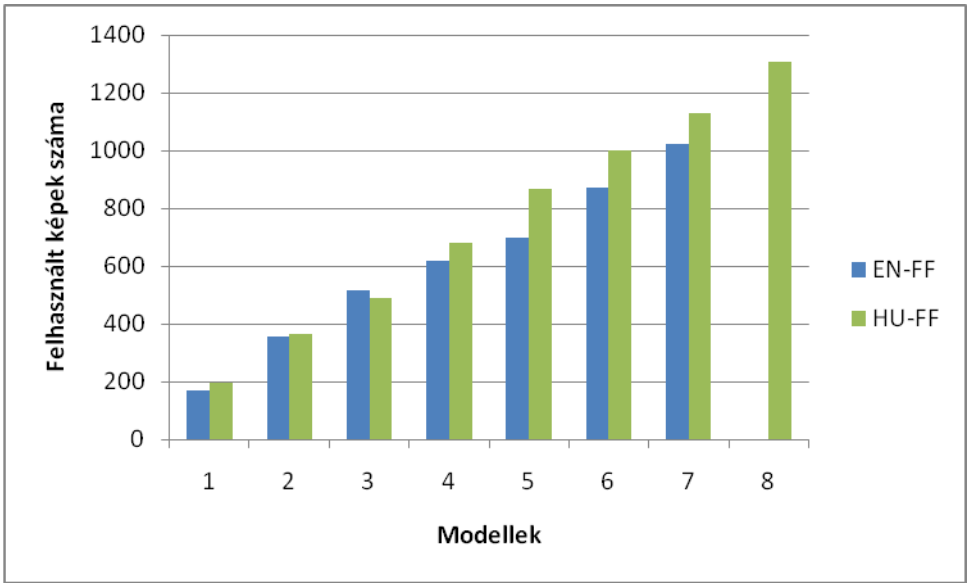
4.1. Tanítóadat leírása

A mély neurális háló tanítása során folyamatosan növelve akartuk megfigyelni hogy változnak az egyes hibamértékek. Ezért a két beszélőtől kiválasztottunk egy-egy mondatot és azokat jelöltük teszt adatoknak, azaz a kialakított modellek minőségét ezeken ellenőriztük. A két beszélő ultrahang képeit nem kevertük egymással, mivel eltérő minőségük és az ultrahang fej eltérő szöge a felvétel során, téves adatokat eredményezhetett volna.

A 3.2.2. alfejezetben szó esett a két beszélő adathalmazáról. Az első beszélő (EN-FF) esetén hét modellt tanítottunk be, az első modellt csak egy, az utolsót már hét mondattal. EN-FF alany nyolcadik mondata tesztelésre lett felhasználva.

A második beszélő (HU-FF) esetén nyolc modellt alakítottunk ki, hasonlóan az első beszélőhöz, modellenként növelve a tanításhoz felhasznált mondatok számát. HU-FF 9. mondata tesztelésre lett használva.

A 10. ábrán látszik, hogy az egyes személyekhez tartozó modellek tanítására használt képek száma körülbelül lineárisan növekszik, mivel a mondatok közel azonos hosszúságúak voltak.



10. ábra. Az EN-FF és HU-FF alanyok képeiből tanításra felhasznált képek száma.

4.2 AutoTrace neuronhálójának módosítása

A változó mennyiségű tanítóadattal létrehozott modell hibamértékeinek kiszámításán kívül, azt is megvizsgáltuk, hogy mennyire változnak meg ezen eredmények, ha megváltoztatjuk az AutoTrace neurál hálójának az architekturális felépítését.

Az AutoTrace-be implementált DNN-ben alapértelmezetten két rejtett réteg található meg. A szakirodalomban nem találtuk meg a két rejtett réteg és a rétegeken alapbeállításként használt neuron szám indoklását [3]. Megvizsgáltuk mi történik a hibaparaméterekkel, ha a rejtett rétegek számát nullára, egyre és háromra változtatjuk, illetve az alapértelmezett architektúra két belső rejtett rétegének neuron számát kétszerezzük, háromszorozzuk, felezzük és negyedeljük. A módosított hálózatokat az EN-FF beszélő első négy mondatával tanítottuk és az utolsó mondatával teszteltük le, mivel később kiderül (4.3. fejezet), hogy négy mondatnál több adat nem okoz jelentősebb javulást a kontúrkövetésben.

4.3 Eredmények

4.3.1 Tanítóadat növelésének eredménye

Az EN-FF alanyra vonatkozó mérési eredmények a 11., 12. és a 13., míg a HU-FF alanyra vonatkozó eredmények a 14., 15. és a 16. ábrán tekinthetőek meg.

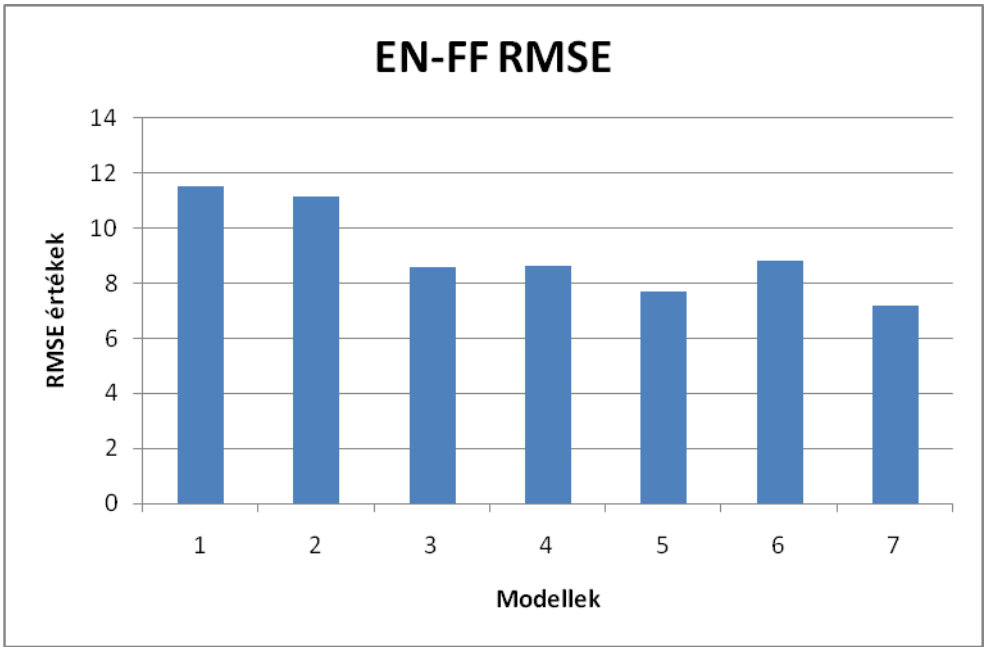
Érdemes megfigyelni, hogy az RMSE és az AREA mindkét beszélő második és harmadik modelljénél beáll egy bizonyos értékre és kevésbé tud tovább csökkenni, annak ellenére hogy több képet használtunk a többi modell tanítására (10. ábra). Elmondható tehát az, hogy a több tanítóadat pontosítja a neurális háló nyelvkontúr követő képességét, de csak egy bizonyos pontig.

További érdekesség mindkét beszélőnél, hogy az RMSE minimum értéke 7 pixel környékén van, amit korábban a manuális nyelvkontúr követés átlagos hibájának ismertünk meg [10]. Ez alapján az AutoTrace-es automatikus nyelvkontúr követés kb. 600 képnyi tanítóadatot felhasználva már közelíti a manuális nyelvkontúr követés hibáját.

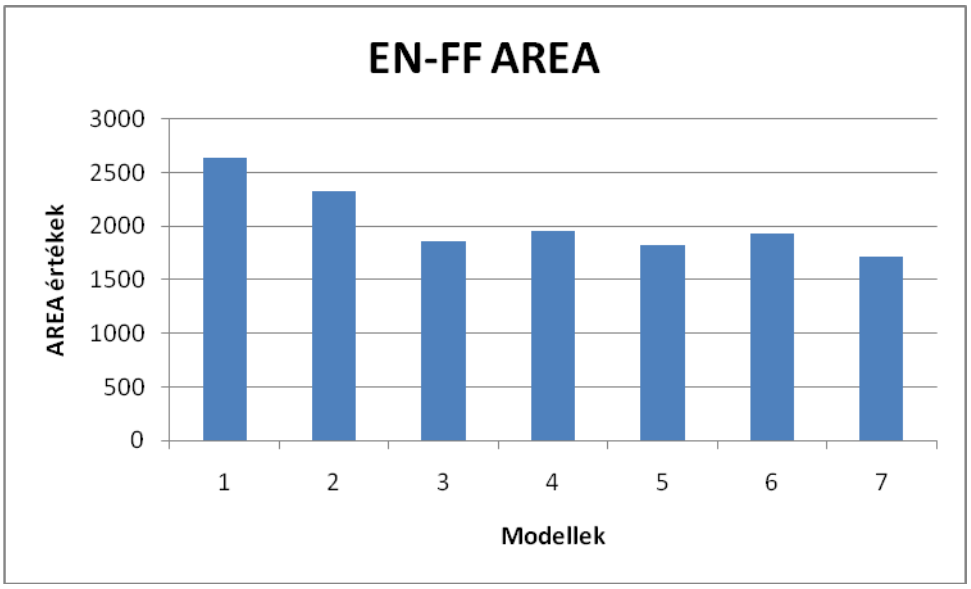
Ha megvizsgáljuk EN-FF és HU-FF RMSE és AREA hibaértékeit kiderül, hogy míg egy-egy beszélőn belül nagy (~0,9) , két beszélő teljes adatát nézve a korreláció kisebb (~0,6). Feltételezésünk szerint ez annak tudható be, hogy az RMSE számításakor (2. képlet) a nevező értéke "n" megegyezik azon polár-egyenesek számával, amin található manuális és automatikus pont is. Látszik (13. és 16. ábra), hogy átlagosan az Insertion és Deletion értékek összege EN-FF beszélőnél nagyobbak (több az olyan polár-egyenes, amin csak manuális vagy csak automatikus pont van), ami arra utal, hogy "n" értéke átlagosan kisebb, így az RMSE értékek nagyobbak lesznek EN-FF-nél, mint HU-FF beszélőnél.

A TER mérési eredményeket megvizsgálva látható, hogy a Substitution (helyettesítés) értéke egy pontig csökken, majd beáll, nem változik nagy mértékben, ami ismét arra utal, hogy a több tanítóadat nem pontosította a modellünket egy határ után.

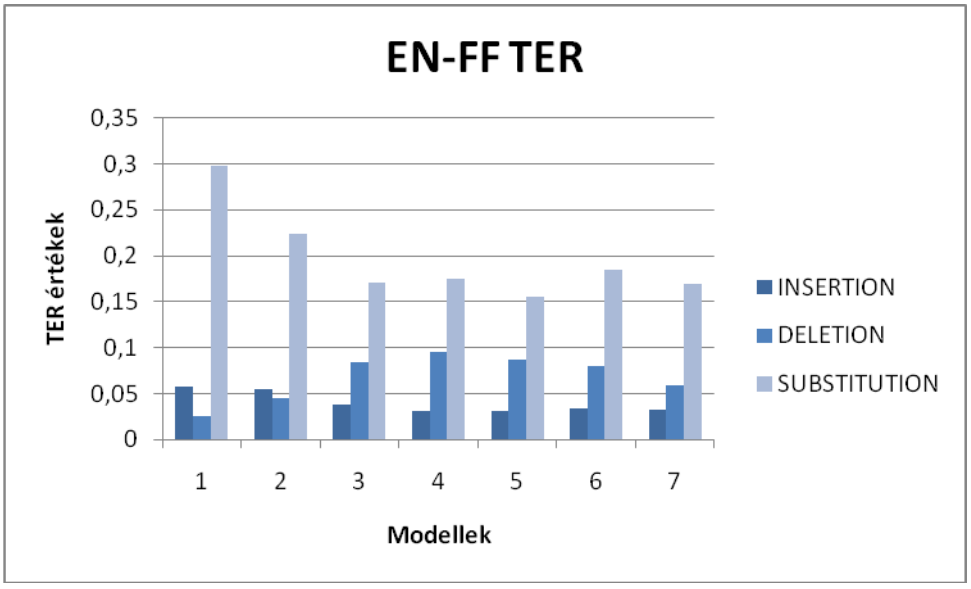
Viszont az Insertion (behelyettesítés) és Deletion (törlés) értékei nem változnak egyértelműen a tanítóadat változásával, így valószínűleg értékük közvetlenül nem függ a tanítóadat mennyiségétől. Az Insertion és Deletion hibák csökkentéséhez a tanítóadat előzetes válogatása vagy más neurális háló architektúra lehet szükséges, amit a 4.3.2. fejezetben vizsgáltunk meg.



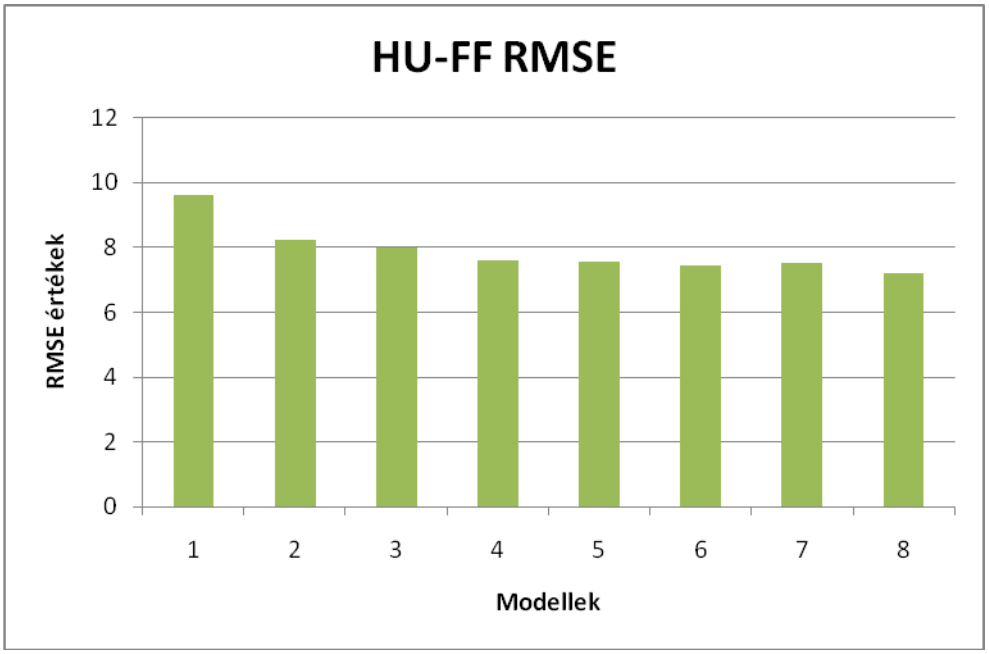
11. ábra. EN-FF beszélő modelljeinek átlagos RMSE értéke.



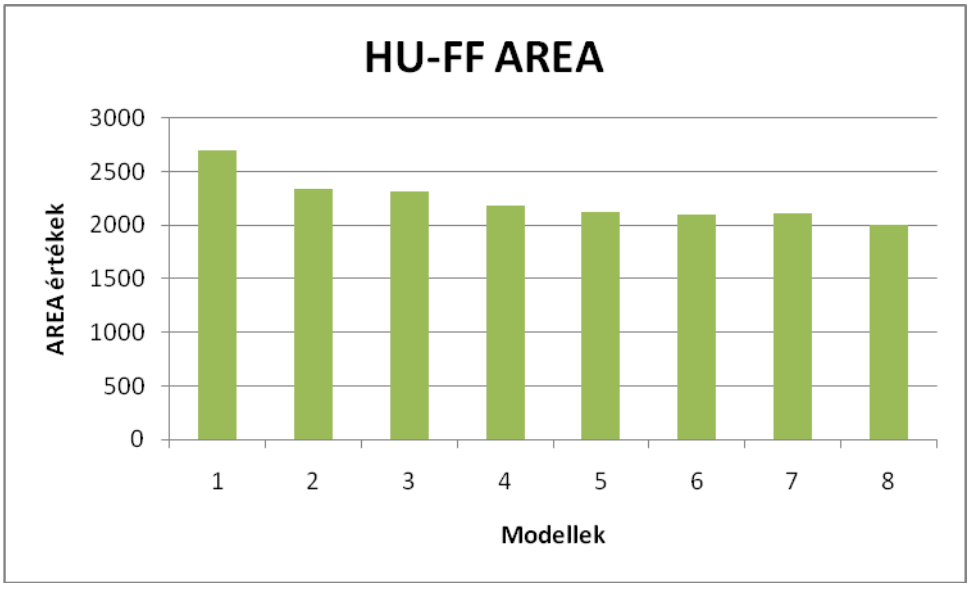
12. ábra. EN-FF beszélő modelljeinek átlagos AREA értékei.



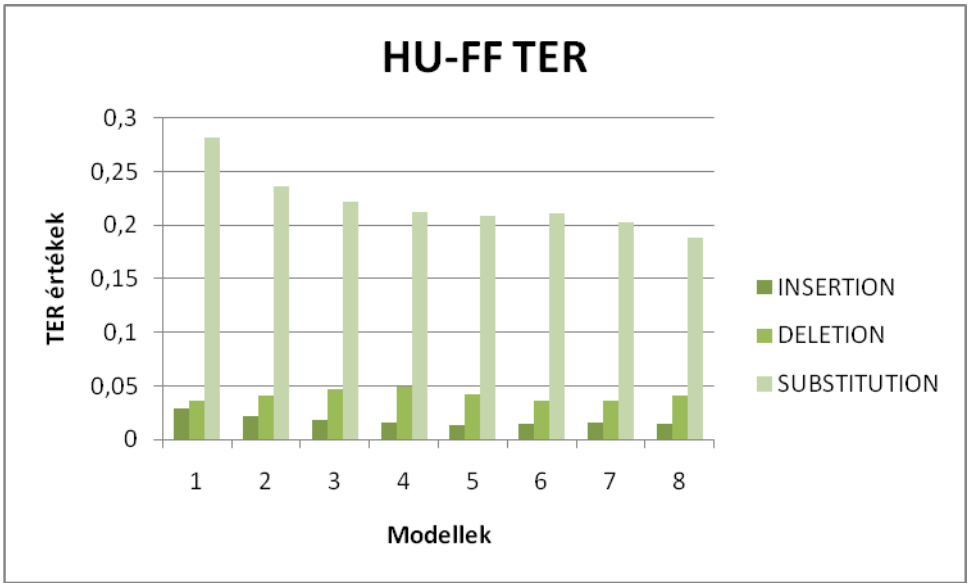
13. ábra. EN-FF beszélő modelljeinek átlagos TER (Insertion, Deletion és Substitution) értékei.



14. ábra. HU-FF beszélő modelljeinek átlagos RMSE értéke.



15. ábra. HU-FF beszélő modelljeinek átlagos AREA értékei.



16. ábra. HU-FF beszélő modelljeinek átlagos TER (Insertion, Deletion és Substitution) értékei.

4.3.2. Neuronháló módosításának eredménye

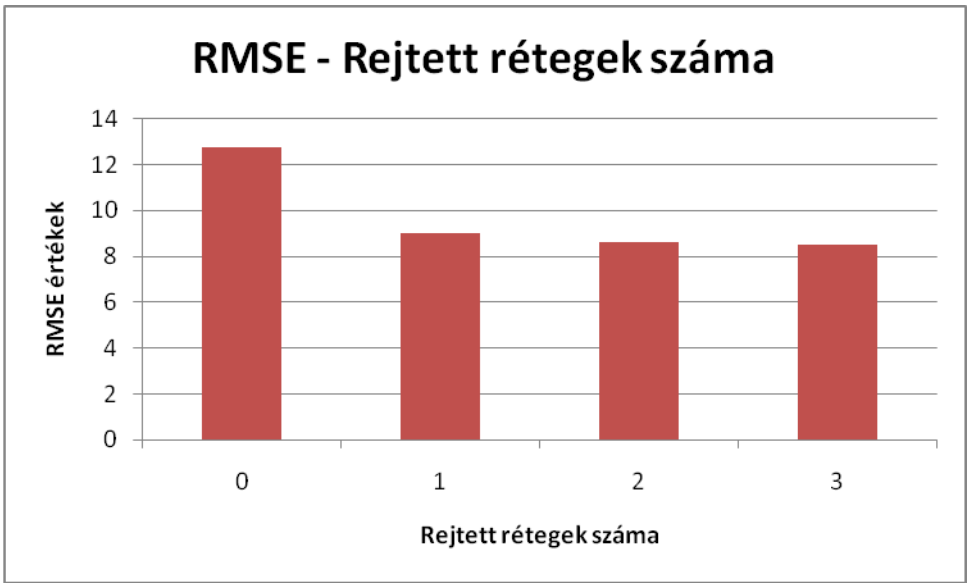
A következőkben bemutatjuk a különböző architektúrájú modelljeink kontúrkövetésére kapott hibaértékeket. A DNN tanításához az EN-FF beszélő első négy, teszteléshez az utolsó mondatát használtuk fel, mivel az alany ultrahang képein tisztábban

kivehető a nyelvkontúr. Egy grafikonon ábrázoltuk a nulla, egy, kettő és három belső rejtett réteggel rendelkező modelljeink RMSE (17. ábra), AREA (18. ábra) és TER (19. ábra) értékeit.

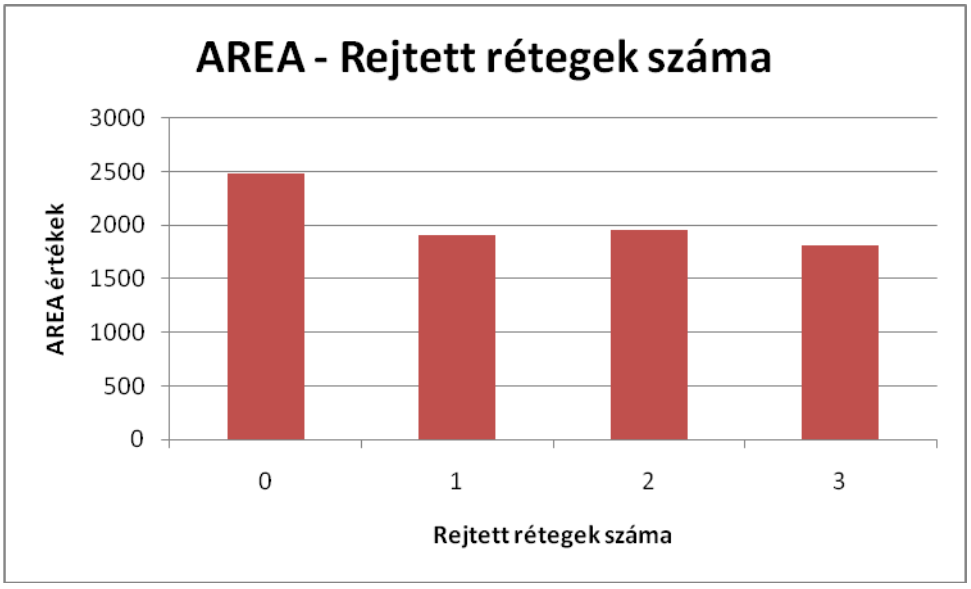
Az alapértelmezett két rejtett réteget tartalmazó hálózat és a megváltoztatott neuron számú azonos rejtett réteggel rendelkező hálózatok hibaértékeit az 1. táblázatban közöltük.

A 17., 18. és 19. ábrából (1. táblázat) megfigyelhető, hogy a rejtett rétegek számának növelése javítja a kontúrkövetést (RMSE, AREA, Substitution), de a hiba itt is egy határhoz tart. Az Insertion hibaérték folyamatosan csökken, a Deletion értékek növekednek a rejtett rétegszám növekedésével (19. ábra). Ez valószínűleg a modell minőségével kapcsolatos, egy rosszabb modell több helyre rak pontot mint kellene, így nagyobb Insertion, kisebb Deletion értékeket kapunk, míg egy jobb modell igyekszik csak oda rakni pontot ahol volt manuális is, ezért kisebb az Insertion, de nagyobb a Deletion hibaérték.

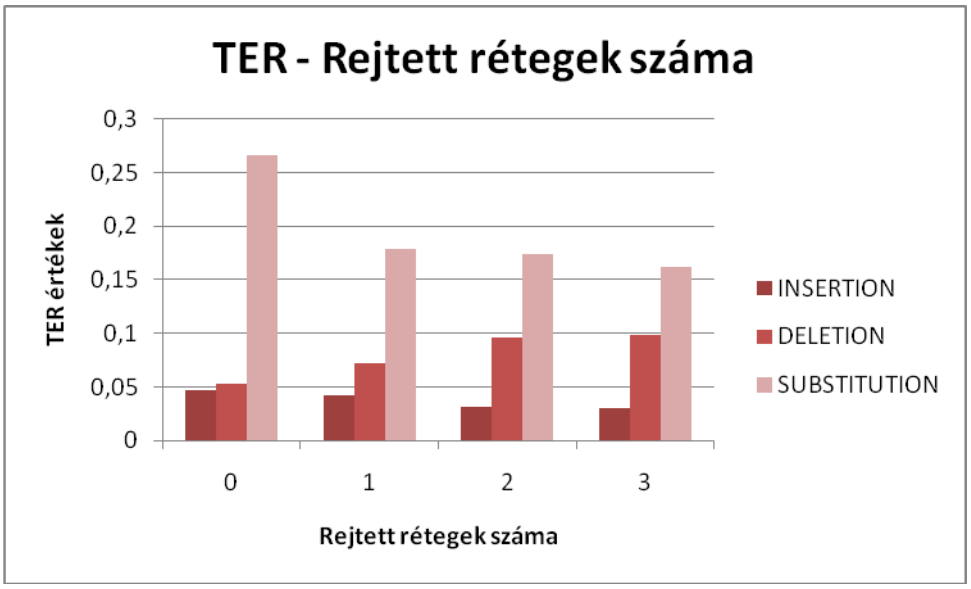
A két rejtett réteget tartalmazó hálózat neuron számának csökkentése (2. táblázat) javította a kontúrkövetés minőségét (AREA és Substitution csökenés), ez azt is jelenti, hogy érdemes lehet olyan DNN hálózatokat kialakítani, melyek több rejtett réteget tartalmaznak, de kevesebb neuront, így csökkentve a tanítási időt és a modell méretét. A kialakított fájlok mérete több neuron szám esetén nőtt (~70; 120; 350; 700; 1400 MB) a betanítási idővel együtt.



17. ábra. RMSE értékek különböző rejtett réteget tartalmazó neurális háló esetében.



18. ábra. AREA értékek különböző rejtett réteget tartalmazó neurális háló esetében.



19. ábra. TER értékek különböző rejtett réteget tartalmazó neurális háló esetében.

	RMSE	AREA	INSERTION	DELETION	SUBSTITUTION
0 db rejtett réteg	12,743	2484,653	0,0467	0,0528	0,266
1 db rejtett réteg	9,0159	1908,5	0,042	0,0725	0,1794
2 db rejtett réteg	8,6472	1959,339	0,0313	0,0959	0,1744
3 db rejtett réteg	8,4944	1818	0,0297	0,0987	0,1615

1. táblázat. Hibaértékek változó számú rejtett rétegek esetén. Alapértelmezett a 2 db rejtett réteg.

	RMSE	AREA	INSERTION	DELETION	SUBSTITUTION
Negyedszeres neuron szám	8,957	1929,43	0,0341	0,0888	0,1773
Félszeres neuron szám	8,7748	1926,868	0,0336	0,0993	0,1652
Alapértelmezett neuron szám	8,6472	1959,339	0,0313	0,0959	0,1744
Kétszeres neuron szám	8,3755	2094,5496	0,032	0,067	0,2067
Háromszoros neuron szám	8,774	2304,94	0,0252	0,1024	0,2201

2. táblázat. Hibaértékek két rejtett réteggel és változó neuron szám esetén.

5. Összefoglalás

Áttekintettük a beszéd folyamatát, kiemelve a nyelv szerepét és annak szervi felépítését. Ezután bemutattuk az ultrahangos technológia alapjait és a beszédkutatásban betöltött szerepét, illetve a nyelvről készült ultrahang képet közelebbről is megvizsgáltunk. Szabadon használható automatikus nyelvkontúr követő alkalmazások közül (EdgeTrack, TongueTrack, AutoTrace), az AutoTrace működését konkrétabban megvizsgáltuk. Bemutattuk az AutoTrace-ben használt neurális hálózatot. Bemutattuk a kialakított mérési eljárásunkat, a modell pontosságát meghatározó hibamértékeinket és a mérési eredményeinket is. Az eredmények alapján kiderült, hogy a több tanítóadat nem feltétlenül okoz javulást, érdemes lehet megváltoztatni a neuronháló architekturális felépítését.

Az ultrahang felvételek 2014 nyarán készültek, amiket szeptemberben kaptam meg, azóta foglalkozom a témával. Az AutoTrace működtetéséhez és az eredmények teszteléséhez több mint 3000 sor Matlab kód készült el.

6. Felhasználási, továbbfejlesztési lehetőségek

A dolgozatban elvégzett mérések hozzájárulhatnak egy hatékonyabb nyelvkontúr követő eljárás megvalósításához (3D-s nyelvfelület követés), így segítve azokat a kutatásokat, melyek a nyelvmozgás és a beszéd akusztikai kimenete között keresik az összefüggést. A pontosabb nyelvkövetés hasznos lehet a nyelvoktatásban, beszéd rehabilitációban illetve az audiovizuális beszédszintézisben is [4].

További lehetőségeket tartogat magában a több beszélővel történő tanítás és a modellnek idegen újabb beszélővel végrehajtott tesztelés, a tanításra használt adatok automatikusan történő válogatása (rossz minőségű képek elhagyása), olyan képekkel történő tanítás melyek egymástól lényegesen eltérő nyelvkontúr adatokat tartalmaznak.

7. Köszönetnyilvánítás

Ezúton mondok köszönetet Csapó Tamás Gábor konzulensemnek a munkám során nyújtott segítségéért, észrevételeiért és tanácsaiért, Gustave Hahn-Powell-nek az AutoTrace-szel kapcsolatos kérdések megválaszolásáért és a tanítási adatokkal kapcsolatos tanácsaiért. A méréshez használt ultrahangos képek és manuális nyelvkontúrok az Indiana University, Speech Production Laboratory-ban készültek, melyért köszönetet mondok Steven M. Lilich-nak [10]. Köszönetet mondok továbbá Elizabeth Mazzocco-nak az ultrahangos adatbázison a manuális nyelvkontúr követés elvégzéséért. Az adatokat a Budapesti Műszaki és Gazdaságtudományi Egyetem Távközlési és Médiainformatikai Tanszéke bocsátotta rendelkezésemre. A borítókép a http://airtightinteractive.com/demos/js/ruttetra/ oldalon készült.

8. Irodalomjegyzék

[1] G. Németh, G. Olaszy, "A magyar beszéd", Akadémia Kiadó, Budapest 2010.

[2] J.-H. Sung, J. Berry, M. Cooper, G. Hahn-Powell and D. Archangeli, “Testing AutoTrace: A Machine-learning Approach to Automated Tongue Contour Data Extraction,” in Ultrafest VI pp. 9–10., 2013.

[3] J. Berry, I. Fasel, L. Fadiga, and D. Archangeli, “Training Deep Nets with Imbalanced and Unlabeled Data,” in Proc. Interspeech, pp. 1756–1759., 2012.

[4] T. Hueber, E. Benaroya, B. Denby, and G. Chollet, “Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface,” in Proc. Interspeech,pp. 593–596., 2011.

[5] E. Tarsoly, "Funkcionális anatómia", Medicina Könyvkiadó Zrt., Budapest, 123. old., 2010.

[6] H. Gray, "Anatomy of the Human Body", Philadelphia: Lea & Febiger, 1918.

[7] M. Stone "A guide to analysing tongue motion from ultrasound images" in Clinical Linguistics & Phonetics, pp. 455-501., 2005.

[8] AutoTrace Github link 2014. október 21. : https://github.com/jjberry/Autotrace

[9] S. Russel, P. Norvig, "Mesterséges Intelligencia Modern Megközelítésben", Panem Kiadó, Budapest, 2006.

[10] T.G. Csapó & S.M., Lulich, "Tongue contour tracings from 2D ultrasound image sequences: quantification of measurement error using manual and automatic tracing methods", bírálat alatt

[11] J. Kominek & A.W., Black, "CMU ARCTIC databases for speech synthesis", Carnegie Mellon University, 2003.

[12] G. Olaszy, "Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai" Beszédkutatás 2013, pp.261–270., 2013.

[13] A. Wrench, "Articulate Assistant Advanced: Ultrasound module", In Ultrafest IV. New York, NY, USA., 2007.

[14] Image-J v1.46, National Institutes of Health, USA, 2014 október 21. : http://imagej.nih.gov/ij

[15] P. Mihajlik, "Spontán magyar nyelvű beszéd gépi felismerése nyelv specifikus szabályok nélkül", 10. o., PhD disszertáció, BME TMIT, 2014. október 21.: http://dokutar.omikk.bme.hu/collections/phd/Villamosmernoki_es_Informatikai_Kar/2011/Mihajlik_Peter/ertekezes.pdf

[16] I. Fasel and J.Berry "Deep Belief Networks for Real-Tima Extraction of Tongue Contours from Ultrasound During Speech", University of Arizona., 2010.

[17] M. Li, C. Kambhamettu, and M. Stone, "Automatic contour tracking in ultrasound images", Clinical Linguistics & Phonetics 19, 545-554., 2005.

[18] L. Tang, T. Bressmann, and G. Hamarneh, "Tongue contour tracking in dynamic ultrasound via higher-order MRFs and efficient fusion moves", Medical Image Analysis 16. 1503-1520., 2012.