# HALLÁSSÉRÜLTEK BESZÉDÉNEK AUTOMATIKUS MINŐSÍTÉSE

**Czap László**

**Bevezetés**

A Magyarországon a jelnyelvi törvény (2009. évi CXXV. törvény) elfogadása óta eltelt idő a jelnyelv elismertségét illetően hozott sikereket. Ugyanakkor a siket és nagyothalló emberek mind iskoláikban, mind munkahelyükön, mind pedig a hétköznapi élet különböző színterein könnyebben találnak megértést, ha a többiek is egyre jobban értik mindazt, amit a hallássérültek nekik, a beszédpartnereknek mondanak. Az utóbbi cél akkor valósulhat meg, ha a siket gyerekek tökéletesebb (mert tudatos, pl. vizuálisan is többszörösen megerősített) hangképzési technikával rendelkeznek, akkor is, ha a halló társadalom tagjai közül is sokan törekszenek a jelnyelv elsajátítására (Bodnár et al. 2013).

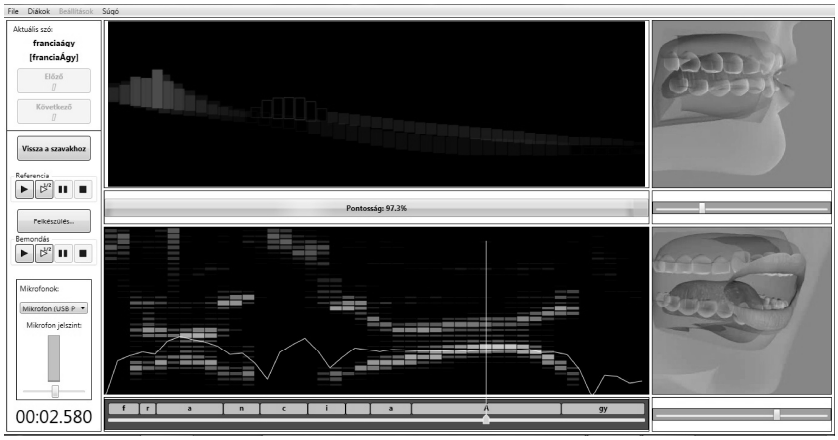
A beszédkutatókat a rendelkezésre álló korszerű eszközök szinte kötelezik a hallássérültek beszédével való foglalkozásra, annak élettani és akusztikai tanulmányozására. A számítógép az eddig láthatatlan jellemzőket is láthatóvá tudja tenni (McGarr–Whitehead 1995).

Ennek jegyében születtek úttörő hazai alkalmazások. Az egyik a Budapesti Műszaki és Gazdaságtudományi Egyetem Beszédakusztikai laboratóriumában Vicsi Klára vezetésével létrehozott Varázsdoboz csomag, amely spektrogramképek segítségével nyújt lehetőséget a szókiejtés korrekciójára (Vicsi 2002). A másik a Szegedi Tudományegyetem és Hégely Gábor gyógypedagógus fejlesztése, a hallássérültek kaposvári intézményének munkatársainak közreműködésével megalkotott Beszédmester program, amely biztosítja a beszédhangok elhangzással azonos időben történő vizuális megjelenítését (Paczolay et al. 2004).

A számítógépek kapacitásának növekedése, valamint a háromdimenziós modellezés és animáció fejlődése bővítette az artikuláció vizuális megjelenítésének eszköztárát. Az *Alap- és alkalmazott kutatások hallássérültek internetes beszédfejlesztésére és az előrehaladás objektív mérésére* címet viselő projekt a siket és nagyothalló személyek számára – az eddigi eszköztár kibővítésével – a beszédtanulás egy segédeszközének megalkotását szolgálta, amit Beszédasszisztensnek neveztünk el (1. ábra). A kutatás alapját a Miskolci Egyetemen kifejlesztett „beszélő fej” és a Debreceni Egyetemen kialakított audiovizuális transzkóder jelentette. A projekt gyakorlatban is hasznosítható célja egy komplex rendszer létrehozása, amely a beszédfolyamat audiovizuá-

DOI: 10.15775/Beszkut.2016.24.11

lis megjelenítését szolgáltatja, egyrészről a beszéd hangképének grafikus ábrázolásával, másrészről az artikuláció vizuális megjelenítésével, egy oktatási keretrendszerbe foglalva. A háromdimenziós fejmodell transzparens arcával a nyelvmozgást a természetes beszélőnél jobban meg tudja jeleníteni.




1. ábra
A Beszédasszisztens gyakorló felülete

A hallássérültek beszélni tanítását segítő Beszédasszisztens alkalmazásunk egyik szolgáltatása a gyakorlás során bemondott szavak és mondatok automatikus minősítése és visszajelzés a gyakorlást végző személy számára. Az automatikus minősítés a szubjektív értékelés eredményét próbálja leképezni.

Ezek mellett számos olyan funkciót tartalmaz a rendszer (prozódia megjelenítés, rögzítősorok és oppozíciós szópárok gyakorlása, tudásalapú rendszer implementálása), amely lehetővé teszi az egyéni gyakorlást nem csak számítógépen, hanem mobil eszközön is. A kifejlesztett technológia audiovizuális transzkódolását végző modulja nyelvfüggetlen, a beszélő fej és az automatikus minősítés adaptálható más nyelvekre (Bodnár et al. 2013).

Elsősorban Maier és szerzőtársai (Maier et al. 2008, 2009, 2010) foglalkoztak a különböző betegségek következtében fellépő hangképzési zavarok vizsgálatával. Rejtett Markov-modell (HMM) alapú eljárásukat sikeresen alkalmazták olyan felnőtteknél, akiknek gégerák miatt eltávolították a gégéjüket, és olyan gyerekeknél, akik ajak- és szájpadhasadékkal születtek. Ezeknél a betegeknél szoros összefüggés érhető el a szubjektív és az automatikus minősítés között. Ez alapján készült a PEAKS (Program for Evaluation and Analysis of all Kinds of Speech disorders) egy rögzítő és elemző rendszer hangképzési és beszédzavarok automatikus vagy manuális minősítéséhez.

Az automatikus minősítés a Beszédasszisztens rendszert önállóan használó hallássérültek számára nyújt visszajelzést. A gyakorlás során produkált beszéd minősítése nem csak a hangképzésben mutatott előrehaladást mutatja meg, hanem bemeneti adatként szolgál a tudásalapú rendszer részére, amely a pedagógusok tapasztalatai alapján segíti a következő gyakorlandó szó kijelölését (Kovács et al. 2014). A cikkben az automatikus minősítés kialakításának részletei olvashatók.

**A referencia-hangadatbázis**

A referenciaértékek felvételéhez hangadatbázist alakítottunk ki a beszédprodukció különböző fokán álló hallássérült gyerekekkel. Az adatbázisban 2421 szó szerepel (egyes szavak többször is előfordulnak, de a bemondók eltérőek, ezért azok érthetősége is). A rögzítésre került minták szókészletét a szurdopedagógusok készítették elő, körültekintően figyelembe véve az egyes diákok aktív szókincsét. A hangfelvételeket 13 szurdopedagógus és nyelvi képzésben nem részesült, 23 naiv egyetemi hallgató értékelte. Minden pedagógus csak a másik iskola diákjainak bemondását értékelte, hogy elkerüljük a beszélő felismeréséből eredő előítéleteket. A bemondást többször is meghallgathatták az értékelők és megjegyzéseket is fűzhettek a mintákhoz. (Néhány példa: *érthetetlen, ritmushiba, szótagol, hangbetoldások, hangsúlya rossz, o-ö csere*) Az eredményeket internetes alkalmazáson keresztül rögzítettük. A minősítés alapját a pedagógusok esetén az általuk meghatározott ötfokozatú skála képezte:

A skála értelmezése:

Érthetetlen (1): az artikuláció teljesen torz; felismerhetetlenek a magán- és mássalhangzók; a szótagszám visszaadása sem megfelelő vagy nem kivehető; a levegővétel, a levegővel való gazdálkodás helytelen; rossz a tempó, a ritmus; dallamtalan, dinamikátlan vagy túl feszített a hangadás.

Nehezen érthető (2): súlyos torzítások, hangelhagyások, hangcserék; csak a magánhangzók egy része kivehető; a légzés elégtelensége miatt létrejövő torzítások, pl. túl levegős vagy fojtott; eltérő, zavaró hangszín, ritmus, tempó jellemzi.

Közepesen érthető (3): a magánhangzók ejtése helyes, a szótagszám megfelelő; súlyos beszédhibák előfordulhatnak pl. diszlália (az a beszédzavar, mely szerint egyes hangzók hiányosan képeztetnek, orrhangzósság, fejhangzósság stb.), prozódiai elégtelenségek.

Jól érthető (4): csekély mértékű beszédhibák; enyhe prozódiai elégtelenségek.

Hallók beszédével azonos szinten érthető (5): legfeljebb 1-2 hanghiba fordulhat elő.

A naiv hallgatóknak a mindennapi nyelvhasználat alapján kellett 1-től 5-ig pontozniuk a bemondásokat.

Egy szűkített szóhalmazon kvantitatív elemzést készíttettünk egy szakértővel. Ezt az értékelést próbáltuk az automatikus minősítéssel megközelíteni. A részletes elemzéshez ebből a 2421 szóból választottunk ki 300 szót. A kiválasztott szókészlet elég változatos nem csak a szavak hosszúsága alapján, hanem a hangkapcsolatok előfordulásának szempontjából is, ami az egész szóadatbázisra is jellemző.

Várakozásunk szerint a pedagógusok – lévén szakemberek – egységesebben értékeltek, mint a laikus hallgatók. Ennek ellenőrzésére megvizsgáltuk az egyes értékelő személyek pontjainak a szórását. Az a pedagógus, aki a pedagógusok átlagához viszonyítva a legkisebb szórást érte el (0,54) az előbbi skálán átlagosan fél jeggyel tért el az átlagtól. A legnagyobb szórást mutató pedagógus pontjai átlagosan egy jeggyel különböztek az átlagtól. A hallgatók pontszámainak szórásai a hallgatói átlaghoz viszonyítva meghaladják a pedagógusok által adott pontszámok szórását. Az eredményeket az 1. táblázat mutatja.

1. táblázat: Az értékelést végzők pontjainak jellemző szórásai

| Szórás | Pedagógusok | Hallgatók |
|---|---|---|
| **Minimum** | 0,54 | 0,87 |
| **Maximum** | 1,03 | 1,18 |
| **Átlag** | 0,70 | 0,96 |

A beszédminőség értékelésének nehézségét mutatja, hogy a pedagógusi és hallgatói átlagok egyes mintáknál jelentős eltérést mutatnak. A 2. táblázatban azt láthatjuk, hogy a pedagógusok és hallgatók átlagolt pontszámai hány szó esetében maradnak a megadott toleráncián belül.

2. táblázat: A pedagógusi és hallgatói értékelések átlagának különbségei

| Eltérés | A tűrésen belüli szavak száma |
|---|---|
| **≤ 0,1** | 47 |
| **≤ 0,2** | 85 |
| **≤ 0,5** | 189 |
| **≤ 1** | 274 |
| **≤ 1,5** | 292 |
| **≤ 2** | 298 |

A hallgatói pontszámoknak a pedagógusi pontszámokhoz legkisebb négyzetes hibát biztosító lineáris illesztéséhez regressziós egyenest vettünk fel, amelynek meredeksége: 0,93, a szükséges eltolás: 0,05. A meredekség közel van az egyhez, az eltolás a nullához, tehát a szakmai szempontok szerinti és a mindennapi nyelvhasználat szerinti értékelés hasonló eredményre vezetett.

**A szakértői elemzés**

A kiértékelést végző szakembert megkértük, hogy elemezze a 300 szóból álló mintahalmazt. Az eredmények hitelességét a beszédfeldolgozás területén szerzett több évtizedes szakmai tapasztalata támasztja alá. A távközlés nemzetközi szaktekintélye. A számítógépes beszédfeldolgozás egyik hazai megteremtője, munkája során már korábban is dolgozott hallássérültekkel.

A felkért szakértő feladata az volt, hogy kvantitatív értékelést adjon a kiejtésről. Azt feltételezte, hogy a beszéd öt fő tényezőjének együttese alapján születik meg a megítélt pontszám. Ezek a tényezők:

– a beszédtempó;
– a ritmus;
– a hangsúly;
– a dallam;
– a hanghibák.

Tesztelései során ezek közül a hanghibára és a ritmushibára adott számszerű értékelést. Elkészítette 294 szó (a szakértő 6 szóról úgy ítélte meg, hogy nem a címkefájlban megadott szó szerepel a hangfelvételen) hanghullámának átírását. Először úgy hallgatta meg a felvételeket, hogy nem nézte meg a szó leírását, azt jegyezte le, amit hallott. A lejegyzéskor nem törekedett érthető szavak megadására. A hanghibát a hibásan ejtett – más hang, hiányzó vagy feleslegesen betoldott – hangok számának és a szóban szereplő hangok számának az arányával írta le.

A szakértő a ritmushiba meghatározásához a jó kiejtés referenciájaként a szavak PROFIVOX szövegfelolvasó rendszerrel generált szintetizált hanganyagot használta (Németh–Olaszy szerk. 2010). A kézi szegmentáláskor a szabad felhasználású WaveSurfer programot használta (http://www.speech.kth.se/wavesurfer/). A szegmentálásnál már figyelembe vette, hogy mit kellene hallania, manuálisan végezte el a kényszerillesztést. Az időfüggvény, a spektrogram és hallás alapú elemzéssel (a szegmenshatártól vagy szegmenshatárig lejátszott hang együttese alapján) határozta meg a szegmenshatárokat. A címkefájlok adatait az eredeti és a szintetizált kimondásoknál átemelte egy-egy táblázatba szavanként (vö. 3. táblázat). A hangok időarányát a hang és a teljes szó időtartamának arányaként értelmezte:

$$r(i) = t(i) \,/\, \text{szumma}(t(i)),$$ ahol

– $i$ a szón belüli hang sorszáma,
– $t(i)$ az $i$-edik hang időtartama,
– $r(i)$ az időarány az $i$-edik hangra.

Az időarányokat a szintetizált referenciaszóra és a vizsgált kiejtésre is meghatározta. A vizsgált bemondás és a referencia időarányának hányadosa egy hangra (relatív időarány) megmutatja a ritmusbeli különbséget. Ha egy hang rövidebb, vagy egy másik, túl hosszan ejtett hang miatt megnő a teljes időtartam, az arány egynél kisebb lesz. Az aránytalanul hosszan ejtett han-

gokra egynél nagyobb érték adódik. A szóra vonatkozó ritmushibát a szakértő a szóban szereplő hangok relatív időarányainak szórásával írta le.

3. táblázat: Az *ablak* szó jellemző időtartamai másodpercben és számított időarányai

| 1. szó | Aktuális időtartam | Referencia időtartam | Aktuális időarány | Referencia időarány | Relatív időarány |
|---|---|---|---|---|---|
| ***a*** | 0,19 | 0,12 | 0,20 | 0,27 | 0,72 |
| ***b*** | 0,04 | 0,05 | 0,04 | 0,11 | 0,36 |
| ***l*** | 0,14 | 0,05 | 0,14 | 0,11 | 1,27 |
| ***a*** | 0,28 | 0,04 | 0,29 | 0,09 | 3,18 |
| ***k*** | 0,32 | 0,18 | 0,33 | 0,41 | 0,81 |

A 3. táblázatban példaképpen az első szó *(ablak)* jellemző értékeit követhetjük. A vizsgált kiejtés aktuális hangidőtartamait másodpercben látjuk. Az elemzés alatt álló bemondás időarányait a teljes szó időtartamához (0,97 s) viszonyítva kapjuk, a szintetizált referenciaszó teljes időtartama 0,44 s. A relatív időarány az aktuális és referencia időarányok hányadosa. A relatív időarányok szórása a szóra: 1,11.

**A minősítési skála meghatározása**

A beszéd minőségének objektív értékelésére nem ismerünk a szubjektív tesztnél megbízhatóbb módszert. Természetes választásnak tűnik, ha az automatikus minősítés megalkotásához a szubjektív értékelés minden tesztelőt magába foglaló átlagát tekintjük referenciának. A szakértői elemzéssel arra kerestük a választ, hogy a számítógépes analízistől várható kvantitatív jellemzőkkel mennyire lehet megközelíteni a szubjektív tesztek eredményeit. Mivel a hanghiba, a ritmushiba és a szubjektív teszt pontszámai más-más tartományba esnek, a szakértői elemzés és a szubjektív tesztekből képezett átlagok összevetéséhez lineáris regressziót alkalmaztunk. A legkisebb négyzetes hibát a hanghiba és a ritmushiba figyelembevételével a hanghiba súlyozó együtthatójára −2,78, a ritmushibáéra −0,51 adódott. A lineáris illesztés során a részletes elemzésre került 294 szóra kapott szubjektív eredmények átlagára, valamint a hanghiba és ritmushiba együttes alkalmazására adódott optimális együtthatók:

$$y = ax + bz + c, \text{ ahol } a = -2{,}78;\ b = -0{,}51;\ c = 4{,}25$$

– $x$ a hanghiba (a hibás hangok és az összes hang hányadosa),
– $z$ a ritmushiba (a hangok relatív időarányainak szórása a szóra),
– $y$ a szubjektív értékekések átlaga.

A negatív szorzók azt fejezik ki, hogy minél nagyobbak a hibák, annál gyengébb a minőség. A tapasztalatok és az eredmények azt mutatják, hogy ha

feltételezzük, hogy a ritmushiba és a hanghiba mérőszáma egyformán jól fejezi ki a vonatkozó hiba mértékét, akkor helytálló az a megállapítás, hogy a hanghibára jóval érzékenyebb a szubjektív értékelő, mint a ritmushibára.

**Kísérletek hagyományos módszerekkel**

Irodalmi adatok alapján a beszédfelismerésre betanított rejtett Markov-modell bázisú felismerő valószínűségeiből próbáltunk a szubjektív értékekést megközelítő eredményt elérni (Maier et al. 2008, 2009, 2010).

**Rejtett Markov-modell alapú értékelés**

A felismerő betanítására a BABEL hangadatbázist használtuk, a hallássérült gyerekek felvételeihez adaptálva (Roach et al. 1998). A BABEL adatbázis három különböző részből áll: izolált és kapcsolt szavas számbemondásokból, CVC (mássalhangzó-magánhangzó-mássalhangzó) szótagokból, valamint folyamatos olvasott beszédből. Mind az olvasott mondatokat, mind a számjegysorozatokat oly módon tervezték, hogy jól lefedjék a magyar nyelvben előforduló hangkombinációkat. A folytonos részben a bemondások némelyike suttogó hangú. Az adatbázis egy része fonémákra van szegmentálva és fel van címkézve. Az adatbázisban összesen 30 férfi és 30 női beszélő hanganyaga, és mintegy 2000 mondat, valamint 14 000 kapcsolt szavas számjegysorozat szerepel.

A folyamatos beszédfelismerésre betanított HMM modellünk HTK implementációjából (Young et al. 2006) kiolvashatók, hogy egy adott hangot milyen valószínűséggel generál a hozzátartozó modell. A hallássérült gyerekek bemondásainak felismerési eredményeiből kiolvasott valószínűségek és a szó szubjektív értékelése között nem fedeztünk fel korrelációt. Ennek okát abban látjuk, hogy a hallássérültek kiejtési hibái nem tipizálhatóak. Következő kísérletünk a lényegkiemelési eljárások során kapott vektorok euklideszi távolságainak vizsgálatát célozta.

**Távolságon alapuló értékelés**

Amikor egy hang kiejtését vizsgáljuk, első gondolatként az merül fel, hogy hasonlítsuk össze egy referenciával és egy távolságfüggvény alapján értékeljük a hang megfelelését a referenciabemondáshoz képest. Megvizsgáltunk szokványos lényegkiemelési eljárásokat (MFCC, PLP, MEL) (Davis–Mermelstein 1980; Hermansky 1990; O'Shaughnessy 1987) a hallássérült gyerekek bemondásainak elemzésére.

A szegmentálási adatok alapján kijelöltük a hangok stacionárius (állandósult) szakaszát – amennyiben értelmezhető –, és elvégeztük a lényegkiemelést. A stacionárius szakaszt a hanghoz tartozó időintervallum közepére helyeztük. A minősíteni kívánt bemondást a pedagógusok választása alapján szépen beszélő gyerekek bemondásaival hasonlítottuk össze. A hangok stacionárius szakaszaira számított távolság erősen függött a bemondó személyétől. A következő kísérletben az aktuális jellemzővektorokat a teljes adatbázis hangjainak stacionárius szakaszaira kiszámított jellemzővektorok átlagával

vetettük össze. Így minden hangra kaptunk egy távolságértéket. A szó jellemzésére a hangokra kapott távolságértékek átlagát vettük. A szavakra kapott átlagokat a szubjektív értékelés alapján kapott, az adott csoportba tartozó szavakra átlagoltuk, a maximumra normálva (vö. Pintér 2015).

A 4. táblázatból látható, hogy a kapott távolságok nem követik következetesen a szubjektív minősítési osztályok besorolását. Gyakran csak ezredekben térnek el, nincsenek monoton összefüggésben az osztályok minősítésével, nem következetesek.

4. táblázat: A különböző lényegkiemelési módszerekkel számított, hangokra kapott távolságértékek átlaga a minősítési intervallumokra

| Minősítés | MFCC | PLP | MEL |
|---|---|---|---|
| **[1-2]** | 0,997 | 1,000 | 0,979 |
| **[2-3]** | 0,999 | 0,982 | 1,000 |
| **[3-4]** | 1,000 | 0,910 | 0,913 |
| **[4-5]** | 0,998 | 0,944 | 0,875 |

**Neurális hálózatok kimeneti aktivitásán alapuló értékelés**

Ezek után a hangok lényegét neurális hálózatok kimeneti aktivitásával próbáltuk megragadni. Akusztikai-fonetikai osztályozásra tanítottunk be neurális hálózatokat, majd ezek kimeneteit is felhasználva az osztályon belüli megkülönböztetésre újabb neurális hálózatokat tanítottunk be a BABEL hang adatbázison. A tanítás során a helyes kimenetek a hangok saját időkeretében 1 értéket kaptak, a többiek 0-t. Az osztályozás jóságát a tanításban nem szereplő, a teljes hanganyag negyedét kitevő tesztelő alakzatokon ellenőriztük. A tesztelés során egy-egy hangra jósági kritériumként a saját kimenetek aktivitásának összegét osztottuk az idegen kimenetek aktivitásának összegével, az összes tesztelő időszegmensre számítva:

$$G_i = \sum_{\forall R} O_{NN} / \sum_{\forall F} O_{NN},$$ ahol

– $G_i$ a neurális hálózat jósága az $i$-edik hangra vagy a hangok egy osztályára,

– $O_{NN}$ neurális hálózat kimenet,

– $\forall R$ helyes kimenet, a hang összes saját időkeretében,

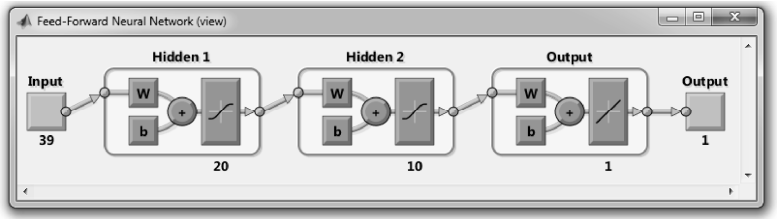
– $\forall F$ hibás kimeneti aktivitás az összes idegen időkeretben.

Azt a neurális hálózatot választottuk, amelynek a jósági tényezője minden hangra összegezve maximális volt, így rendelkezésünkre állt öt neurális hálózat az akusztikai-fonetikai osztályozáshoz, valamint négy neurális hálózat az osztályon belüli besoroláshoz. A vizsgált lényegkiemelési eljárások (MFCC, PLP, MEL sávenergia) közül a PLP mutatta a legnagyobb jósági tényezőket, így a PLP lényegkiemeléssel tanított neurális hálózatok kimeneteit használtuk fel. A tanítást elvégeztük többféle opcióval. A jósági tényező maximumát

nyújtó beállítás: az aktuális 40 ms-os keret 12 PLP adata és logaritmikus energiája mellé a megelőző 80 ms-os szakasz két keretének átlagát és a következő 80 ms két keretének átlagát is hozzávettük. A 3 × 13 jellemző írja le a 200 ms-os intervallum közepére eső 40 ms-os szegmenst. A fonetikai osztályozásra szánt öt neurális hálózat betanítása ezekkel a paraméterekkel történt. A neurális hálózatok által képezett osztályok:

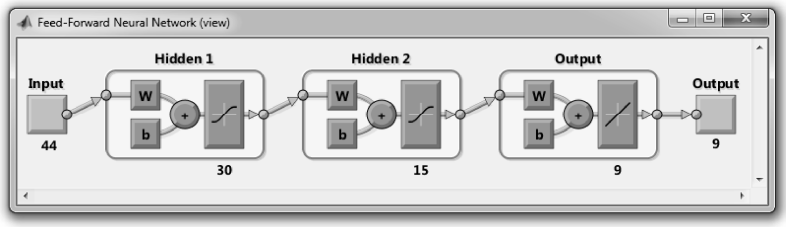
– szünet;
– magánhangzó *(u, o, a, á, e, ö, ü, é, i);*
– félmagánhangzó *(j, ny, n, m, r, l);*
– réshang *(v, z, zs, f, sz, s, h);*
– zárhang *(b, d, gy, g, p, t, ty, k, c, cs).*

A zárójelekben az osztályokra dedikált neurális hálózat kimeneteihez tartozó hangokat soroltuk fel. Próbáltuk az akusztikai-fonetikai osztályozást egyetlen neurális hálózattal is megvalósítani, de gyengébb eredményeket kaptunk, mint az egyes osztályokra dedikált neurális hálózatokkal (2. ábra).




2. ábra
Az akusztikai hangosztályt meghatározó neurális háló modellje

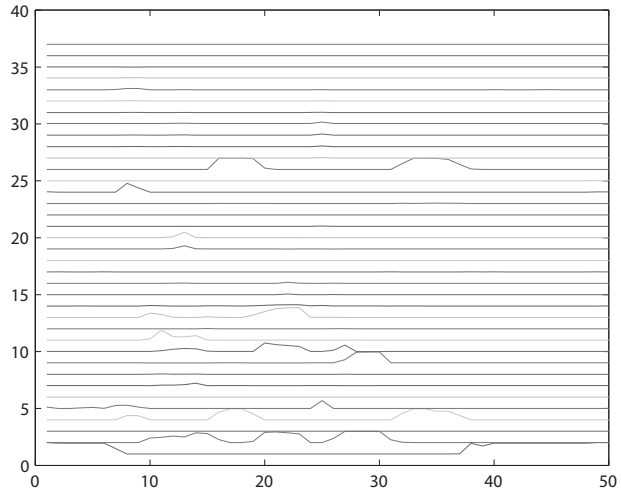
A fonetikai osztályokon belüli hangok felismerésére tanított neurális hálózatok a 39 PLP jellemzőn túl az öt osztályozó neurális hálózat kimeneteit is megkapták inputként. A 3. ábrán példaként a magánhangzók szortírozására szolgáló neurális hálózat felépítése látható.




3. ábra
A magánhangzók akusztikai hangosztályának neurális háló modellje

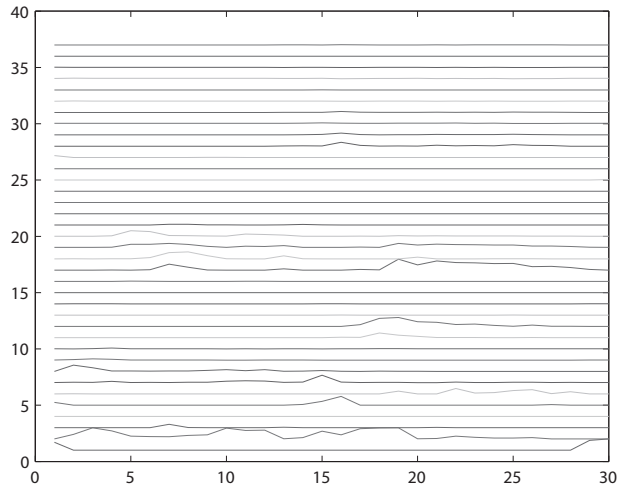
Az osztályozást elvégeztük a rövidebb PLP időkeretekkel betanított neurális hálózatokkal is, a legkisebb hibákat a fenti beállítással kaptuk. A számításokra a MATLAB programcsomag vonatkozó toolboxait használtuk.

A neurális hálózat ideális esetben az adott hangra egységnyi, merőben eltérő hangra 0 kimeneti aktivitással válaszol. A helyesen artikulált hangra nagy, a hibásan artikulált hangra kis kimeneti aktivitást produkál. A 4. és az 5. ábrán a hallók beszédével azonos minőségű *hűséges* és az alig érthető *valami* szavak kimeneti aktivitásait láthatjuk. Alulról felfelé az akusztikai-fonetikai osztályok felsorolásának sorrendjében az öt osztályhoz tartozó kimenet, majd szintén az osztályok felsorolása szerinti sorrendben az osztályokon belüli hangokhoz tartozó kimenetek aktivitása követhető. Jól látható, hogy a megfelelően artikulált hangok határozott kimeneteket produkálnak. Ezzel ellentétben, a gyenge minőségű beszéd esetén a neurális hálózat kimenetei lényegesen kisebb aktivitást mutatnak (vö. 4. és 5. ábra).




4. ábra
A neurális hálózat jelentős kimenetei a jól artikulált *hűséges* szó esetén

A hasonlósági mértéket az adott hanghoz tartozó kimenet aktivitásával azonosítottuk. Megvizsgáltuk a tanulmányozott szavak egyes hangjaihoz tartozó kimeneti aktivitások átlagát. Ha a 4. táblázatot kiegészítjük a neurális hálózatok egyes minőségi osztályokra kapott kimeneti aktivitásának átlagával (NN), látható, hogy csak az utóbbi mutat differenciált és monoton összefüggést a minőségi osztályokkal (5. táblázat).




5. ábra
A neurális hálózat gyenge kimenetei az alig érthető *valami* szó esetén

5. táblázat: A különböző lényegkiemelési módszerekkel számított, hangokra kapott távolságértékek átlaga a minősítési intervallumokra

| Minősítés | MFCC | PLP | MEL | NN |
|---|---|---|---|---|
| **[1-2]** | 0,997 | 1,000 | 0,979 | 0,401 |
| **[2-3]** | 0,999 | 0,982 | 1,000 | 0,557 |
| **[3-4]** | 1,000 | 0,910 | 0,913 | 0,816 |
| **[4-5]** | 0,998 | 0,944 | 0,875 | 1,000 |

A szubjektív tesztekkel az összehasonlítást részben korrelációs, részben a számított pontszámok különbsége szerint vizsgáltuk. Az összehasonlításhoz lineáris illesztést végeztünk a szubjektív tesztek pontjai és a hasonlóságmérték között. A hasonlóság 0 és 1 közötti kimeneteket produkál, a szubjektív tesztek pontjai 1 és 5 közé esnek. Gradiens módszerrel megkerestük azt a szorzót és eltolást, amellyel a hasonlóságértéket korrigálva a szubjektív pontszámokkal a legkisebb négyzetes hibát adja. Az automatikus minősítés jóságát a szakértői értékeléssel vetettük össze.

A legkisebb négyzetes hibát eredményező együtthatók meghatározása után megvizsgáltuk, hogy a szakértői minősítés korrigált pontszámai mennyiben térnek el a szubjektív minősítés eredményétől. A részletes elemzés alá vetett 294 szó közül megvizsgáltuk, hogy a pontszámok különbsége hány szónál kisebb az 1. táblázat oszlopaiban is szereplő értékeknél. Azt vizsgáljuk, hogy az automatikus minősítés a 6. táblázatban is szereplő toleranciákkal hány

szónál közelítette meg a szubjektív tesztek átlagát. Elosztottuk az automatikus és a szakértői minősítés szerint az adott tűréssel megegyező minősítésű szavak számát. Az egyes toleranciaosztályokra kapott eredményeket átlagolva 89 százalékos egyezőséget kapunk.

6. táblázat: A szakértői és az automatikus minősítési pontszámok referenciához mért, tűrésen belüli szavak száma és aránya

| **Eltérés:** | **≤ 0,1** | **≤ 0,2** | **≤ 0,5** | **≤ 1** | **≤ 1,5** | **≤ 2** |
|---|---|---|---|---|---|---|
| **Szakértői** | 21 | 44 | 131 | 253 | 285 | 291 |
| **Automatikus** | 21 | 35 | 101 | 207 | 268 | 287 |
| **Automatikus/Szakértői** | 100% | 80% | 77% | 82% | 94% | 99% |

Felmerül a kérdés, hogy ha a szakértői és az automatikus minősítés szerinti elemzéshez hasonlóan egyetlen pontozást vetünk össze a referenciának tekintett teljes szubjektív átlaggal, hogyan alakul az egyes tesztelők előzőekben számított tolerancia szerinti találata. Ha elvégezzük a 6. táblázat szerinti elemzést mind a 36 tesztelésben részt vevő pedagógus és hallgató pontszámaira, és a szakértői eredményekkel hasonlítjuk össze, a szubjektív tesztelők átlaga 79%, a legjobb eredmény 111%, legrosszabb 39%, összesen egy volt jobb a szakértői eredményeknél. A legjobb és a legrosszabb egyezőséget is a pedagógusok produkálták.

Ha ugyanezt az elemzést a szakértői helyett az automatikus értékelés eredményéhez viszonyítva végezzük el, 90%-os átlagot kapunk, a legjobb érték 127%, a legrosszabb 45%. A 36 szubjektív tesztelő közül nyolc ért el az automatikus minősítésnél több találatot, akik közül egy pedagógus, hét pedig hallgató.

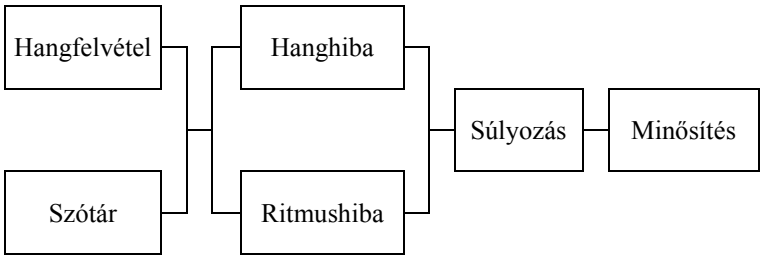
Mivel a szubjektív pontszámok elég nagy szórást mutatnak, azt is megvizsgáltuk, hogy a szakértői és az automatikus minősítés pontszáma hány szónál esik a hallgatói és a pedagógusi pontszámok átlagai közé. A szakértői hanghiba és a ritmushiba optimális illesztésekor a 294 szóból a hallgatói és a pedagógusi átlagok közé 54 szó esik. Ugyanez a neurális hálózatok kimeneteinek szavankénti átlagnál és a ritmushibánál, vagyis az automatikus minősítésnél 44 szó.

Az automatikus minősítéshez használt átlagos neurális hálózat kimeneti aktivitása és a ritmushiba együtthatói a legkisebb négyzetes hiba esetén: 2,92 és −0,76. Az együtthatókból megállapítható, hogy a szakértői hanghibánál kevésbé megbízható neurális hálózat kimeneteivel párosítva a ritmushiba nagyobb súlyozó együtthatót kap.

**A Beszédasszisztensbe integrált algoritmus**

A tapasztalatokat felhasználva alkottuk meg az automatikus minősítés algoritmusát (6. ábra). Első lépésként a gyakorlás során készített hangfelvételt a

projekt keretében kidolgozott, a hallássérültek gyakran alig érthető és akadozó beszédére adaptált dinamikus idővetemítési eljárással hangokra szegmentáljuk (Pintér–Czap 2015). A szegmentálásnál kihasználjuk, hogy az éppen gyakorlás alatt álló szót vagy mondatot ismerjük, fonémasorát fonetizátorral (Vicsi et al. 2005) előre meghatároztuk. A szegmentálás alapján a szakértő által javasolt ritmushibát számolni tudjuk. A stacionárius szakasz kijelöléséhez megkeressük a hangok időrésének közepét. Zárhangoknál és affrikátáknál az időrés utolsó szegmensét választjuk (burst). Az így kijelölt időkeretekre kiolvassuk a neurális hálózatok adott hanghoz tartozó kimeneteinek aktivitását. Ezek átlagát a ritmushibával kiegészítve a lineáris regresszió során kapott együtthatókkal súlyozott eredménye minősíti a vizsgált bemondást.




6. ábra
Az automatikus minősítés algoritmusa

**Összefoglalás**

Több módszert megvizsgáltunk a hallássérült gyerekek hangfelvételeinek elemzésére. Csak a hangfelismerésre betanított neurális hálózatok kimeneti aktivitására találtunk differenciált és monoton eredményeket a különböző minőségi osztályokra. Módszerünkkel a szubjektív értékelést a tolerancia tartományokban a szakértői becsléshez képest átlagosan 89 százalékos pontossággal közelítettük meg. Az automatikus minősítés a kísérletben részt vevő 36 szubjektív értékelőből 28-nál jobban megközelítette a szubjektív értékek átlagát. A 300 szó részletes értékelése a szakértőtől kéthetes elemző munkát igényelt. Az automatikus minősítés időigénye ezzel nem vethető össze, eredményei azonban megközelítik annak megbízhatóságát.

**Irodalom**

Bodnár Ildikó – Czap László – Pintér Judit 2013. Kutatási projekt hallássérültek internetes beszédfejlesztésére. *Alkalmazott Nyelvészeti Közlemények* VIII/2. 19–32.

Davis, Steven B. – Mermelstein, Paul 1980, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28/4. 357–366.

Hermansky, Hynek 1990. Perceptual linear predictive (PLP) analysis for speech, *Journal of the Acoustical Society of America* 87/4. 1738–1752.

Kovács, Szilveszter – Tóth, Ágnes – Czap, László 2014. Fuzzy model based user adaptive framework for consonant articulation and pronunciation therapy in Hungarian hearing-impaired education. In: *5th IEEE International Conference on Cognitive Infocommunications: CogInfoCom 2014.* Vietri sul Mare, Olaszország, 361–366.

Maier, Andreas – Hönig, Florian – Hacker, Christian – Schuster, Maria – Nöth Elmar 2008. Automatic evaluation of characteristic speech disorders in children with cleft lip and palate. In: *Proceedings of 11th International Conference on Spoken Language Processing,* Brisbane, Australia. 1757–1760.

Maier, Andreas – Haderlein, Tino – Eysholdt, Ulrich – Rosanowski, Frank – Batliner, Anton – Schuster, Maria – Nöth, Elmar 2009. PEAKS – A system for the automatic evaluation of voice and speech disorders. *Speech Communication* 51/5. 425–437.

Maier, Andreas – Haderlein, Tino – Stelzle, Florian – Nöth, Elmar – Nkenke, Emeka – Rosanowski, Frank – Schützenberger, Anna – Schuster, Maria 2010. Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. In: *EURASIP Journal on Audio, Speech, and Music Processing.* http://asmp.eurasipjournals.springeropen.com/articles/10.1155/2010/926951 (A letöltés ideje: 2015. december 10.)

McGarr, Nancy S. – Whitehead, Robert 1995. A hallássérültek fonéma-ejtésének kérdései. In Csányi Yvonne (szerk.): *Tanulmányok a hallássérültek beszédérthetőségének fejlesztéséről.* Bárczi Gusztáv Gyógypedagógiai Tanárképző Főiskola, Budapest. 19–23.

Németh Géza – Olaszy Gábor (szerk.) 2010. *A magyar beszéd*. Akadémiai Kiadó, Budapest.

O'Shaughnessy, Douglas 1987. *Speech communication: human and machine*. Addison-Wesley, Reading, MA.

Paczolay Dénes – Kocsor András – Sejtes Györgyi – Hégely Gábor 2004. A „Beszédmester" csomag bemutatása: informatikai és nyelvi aspektusok. *Alkalmazott Nyelvtudomány* 1. 57–80.

Pintér Judit Mária 2015. *A beszédminőség automatikus értékelése*. PhD-értekezés. Miskolci Egyetem, Miskolc.

Pintér Judit Mária – Czap László 2015. Gyenge minőségű beszéd szegmentálása. In: *Proceedings of the XXth International Scientific Conference of Young Engineers*. Kolozsvár, 119–122.

Roach, Peter – Arnfield, S. – Barry, W. J. – Dimitrova, S. – Boldea, M. – Fourcin, A. – Gonet, R. – Gubrynowicz, E. – Hallum, L. – Lamel, L. – Marasek, K. – Marchal, A. – Meister, E. – Vicsi, K. 1998. BABEL: A database of Central and Eastern European languages. In: *Proceedings of the First International Conference on Languages Resources and Evaluation* 1. Granada, Spain, May 28–30. 371–374.

Vicsi Klára 2002. Varázsdoboz. Audiovizuális számítógépes fejlesztő program beszédhibás gyerekek részére. *Démosztenész Hírmondó* 13. 8–16.

Vicsi Klára – Kocsor András – Tóth László – Velkei Szabolcs – Szaszák György – Teleki Caba – Bánhalmi András – Paczolay Dénes 2005. A Magyar Referencia Beszédadatbázis és alkalmazása orvosi diktáló rendszerek kifejlesztéséhez. In Alexin

Zoltán – Csendes Dóra (szerk.): *III. Magyar Számítógépes Nyelvészeti Konferencia MSZNY 2005*. Szegedi Tudományegyetem, Szeged. 435–438.
Young, Steve – Evermann, Gunnar – Gales, Mark – Hain, Thomas – Kershaw, Dan – Liu, Xunying A. – Moore, Gareth 2006. The HTK book (for HTK version 3.4). http://htk.eng.cam.ac.uk/ (A letöltés ideje: 2015. december 10.)

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

**Automatic assessment of the speech of hearing impaired**

Under the framework of a project we have developed a ‘Speech Assistant System’ that aims at supporting the speech production enhancement of the deaf and hearing impaired children. One service of our application is the automatic assessment of words and sentences in the course of practice and providing feedback to the trainee. Neural networks were trained for speech sound classification and their output activities formed the basis of scoring the speech quality. Our results were almost as close to the average subjective scores as that of a detailed evaluation of a speech scientist. The scores of the automatic assessment approached the average ones better than 28 out of 36 subjects.