# Sclerosis Multiplex hangalapú felismerése akusztikai alapú beágyazások használatával

Gosztolya Gábor[1,2], Tóth László[1], Svindt Veronika[3],
Bóna Judit[4], Hoffmann Ildikó[3,5]

[1]Szegedi Tudományegyetem, Informatikai Intézet
[2]ELKH-SZTE Mesterséges Intelligencia Kutatócsoport, Szeged
[3]ELKH Nyelvtudományi Kutatóközpont, Budapest
[4]Eötvös Loránd Tudományegyetem,
Alkalmazott Nyelvészeti és Fonetikai Tanszék, Budapest
[5]Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék
ggabor @ inf.u-szeged.hu

**Kivonat** A sclerosis multiplex (SM) a központi idegrendszer krónikus gyulladásos megbetegedése. Mivel az SM nyelvi zavarokkal is együtt járhat, az automatikus beszédelemzés hasznosnak bizonyulhat akár az alig észrevehető beszédprodukciós változások detektálására is. Egy lényeges technikai kérdés ugyanakkor, hogy milyen jellemzőket érdemes kinyerni az alanyok beszédéből, melyeket aztán a gépi tanulási osztályozási lépés során használhatunk. Jelen cikkünkben HMM/DNN hibrid modellek mély neurális hálós akusztikus modelljeiből nyerünk ki aktivációkat, majd ezeket a teljes hangfelvételen többféleképpen összegezve (pl. átlag, szórás) használjuk jellemzőként. Kísérleteinket 23 SM alany és 22 egészséges kontroll személy négy-négy hangfelvételén végezzük. Eredményeink alapján a javasolt eljárás hatékonyabb azonosítást tesz lehetővé, mint az ugyanazon beszédadaton tanított x-vektor technika. Az elért AUC értékek tapasztalataink alapján nagyban függenek a beágyazás forrásrétegétől és a felvételszintű összegzés módjától is. A legjobb értékek az egyes beszédfeladatokon 0,824 és 0,911 közé estek.
**Kulcsszavak:** sclerosis multiplex, beszédelemzés, mély neurális hálók, beágyazások

## 1. Bevezetés

A sclerosis multiplex (SM) a központi idegrendszer krónikus gyulladásos megbetegedése. A betegségnek három fő típusa különböztethető meg: relapszáló-remittáló SM (tünetes- és tünetmentes állapotok váltakozása), másodlagos progresszív SM (tünetek állandósulása és folyamatos romlás) és elsődleges SM (folyamatos romlás a tünetek első fellépésétől). Változatos idegrendszeri háttere miatt az SM tünetei jellegükben és súlyosságukban sokfélék lehetnek (Szirmai, 2006). Az SM a motoros és a kognitív funkciók érintettsége mellett nyelvi zavarokkal is együtt járhat.

A sclerosis multiplex-szel együtt járó nyelvi- és beszédzavaroknak a kognitív képességek változásaival való összefüggésének kutatása a hazai és a nemzetközi szakirodalomban is újszerű, korábban alig vizsgált terület. Ugyanakkor az SM-mel élő személyek több, mint egyharmada számol be ilyen jellegű zavarokról, emellett a betegség 60-70%-ban átmeneti vagy tartós mentális és/vagy kognitív korlátozottsággal jár együtt (pl. munkamemória és végrehajtó funkciók zavara, téri tájékozódás zavara, krónikus fáradtság). A nyelv- és beszédfolyamatok a kognitív, a szenzoros és a motoros működés dinamikus összehangolását kívánják meg. A meglassult információfeldolgozási sebesség hatással van a nyelvi- és beszédfolyamatokra, változások jelenhetnek meg a beszédpercepció és/vagy a beszédprodukció folyamatában, illetve valamennyi nyelvi szintet érintheti (Bóna és mtsai, 2020; Svindt és mtsai, 2020; Renauld és mtsai, 2016). Habár csak a betegek egyharmadánál diagnosztizálnak dizartriát, az automatikus beszédelemzés így is hasznosnak bizonyulhat a megelőző, alig észrevehető beszédprodukciós változások detektálására (Mulfari és mtsai, 2021), mellyel lehetővé válhat akár a betegség korai észlelése vagy súlyosbodásának detektálása is.

Egy ilyen automatikus beszédelemző eljárás kulcskérdése általában az, hogy az alanyok beszédéből milyen gépi tanulási jellemzőket nyerünk ki. Az orvosi beszédfeldolgozásban bevett megközelítésnek számít, hogy jellemzőink nem elsődlegesen az adott betegségre specifikusak, hanem (ilyen értelemben) általánosnak számítanak (García és mtsai, 2018; Jenei és Kiss, 2020). A jellemzőkinyerési technika kiválasztásánál figyelembe kell venni a területen tipikus adatszűkösséget is: mivel a hangfelvételek gyűjtése nehézkes és nagyon munkaigényes (a felvételek rögzítése adott protokollt követve történik, az alanyokat diagnosztizálni kell, stb.), az adatbázisok terjedelme jellemzően a beszédfelismerésben megszokottak töredéke, legföljebb néhány órát tesz ki.

Ezt figyelembe véve az általános jellegű és statisztikai jellemzőkinyerési eljárások általában egy általános, beszédfelismerési feladatra szánt korpuszt is használnak. Például az i-vektorok (Dehak és mtsai, 2009) esetén, melyek Univerzális háttérmodellje (Universal Background Model, UBM) gyakorlatilag egy keretszinten tanított Gaussi keverékmodell (Gaussian Mixture Model, GMM), ez a GMM fönnakadás nélkül tanítható egy ilyen általános jellegű adatbázison. Hasonló a helyzet a d-vektorokkal (Variani és mtsai, 2014) és az x-vektorokkal (Snyder és mtsai, 2018) is: ezek a mély neurális hálón (Deep Neural Network, DNN) alapuló eljárások modelljeinek tanítása is jellemzően egy külső, jelentősen nagyobb korpuszon történik. (Bár az i-vektorok, d-vektorok és x-vektorok eredetileg a beszélőazonosítási feladat megoldására lettek kifejlesztve, számos tanulmányban használják ezeket az eljárásokat nemverbális, vagy kimondottan orvosi jellegű problémákon, jellemzőkinyerő eszközként (Grzybowska és Kacprzak, 2016; Huckvale és mtsai, 2020; Egas-López és mtsai, 2021).)

Ezek az alkalmazások tekinthetők úgy is, hogy valamiképpen fölépítjük a „standard beszéd" egy modelljét, és a jellemzőkinyerési lépés során azt próbáljuk kifejezni, hogy az adott hangfelvételen hallható beszédjel miben és mennyiben különbözik ettől a „standard beszédtől". Természetesen az alkalmazott modellben (akár gyökeresen is) különböznek a fölsorolt módszerek, mint ahogy a tanítás

módjában és az ahhoz szükséges annotáció jellegében is. Az i-vektorok háttérmodelljének tanításához (mivel keretszintű jellemzőkre illesztett GMM-ről van szó) semmilyen annotáció nem szükséges. Az x-vektorok esetén a tanítási cél az aktuális beszélő azonosítója (sorszáma); ennek rendelkezésre állása ugyanakkor a gyakorlatban nem igazán szigorú követelmény, ez szinte minden korpusznál megtalálható. Az x-vektorok modellje egy speciális struktúrájú neurális háló, melynek alsóbb rétegei keretszinten, fölsőbb rétegei az egész felvétel szintjén működnek, a kettő között pedig egy speciális összegző réteg található. Így a háló gond nélkül tanítható csupán keretszintű jellemzőkre (pl. MFCC-kre) és felvételszintű célértékekre. Maguk az x-vektor jellemzők valamely szegmensszintű réteg aktivációiként állnak elő. A d-vektorok bizonyos szempontból az x-vektorok előzményeinek tekinthetők: esetükben a neurális hálókat hagyományos módon, keretszinten tanítjuk, bár a keretszintű címkéket (az aktuális beszélő azonosítóját) a felvételszintű címkézésből vesszük.

Vegyük észre, hogy a fölsorolt jellemzőkinyerő eljárások mindegyikénél szükség van valamilyen speciális lépésre, például egy GMM vagy egy (akár egyedi struktúrájú) neurális háló tanítására. Ugyanakkor keretszintű DNN akusztikus modellek (a beszédfelismerési terület hagyományos HMM/DNN hibrid modelljéből) elég könnyen elérhetőek, illetve elég nagy tapasztalat halmozódott már föl abban is, hogy azokat hogyan érdemes tanítani. Egy ilyen akusztikus modell jellemzőkinyerési használata számos gyakorlati előnyt nyújtana (természetesen amennyiben az elért pontosságértékek versenyképesek). Jelen cikkünkben egy ilyen megközelítést mutatunk be. Az akusztikus modell tanítását egy általános beszédadatbázis (a BEA korpusz (Neuberger és mtsai, 2014)) 60 órányi részhalmazán végezzük, míg az orvosi feladat sclerosis multiplex beszédből történő fölismerése négy különböző (beszélői) feladatból. Kísérleteinkben a javasolt eljárás hatékonyabbnak bizonyult, mint az azonos adaton tanított, szintén jellemzőkinyerésre használt x-vektor technika.

## 2. Hangfelvételek

A vizsgálatokra a budapesti Uzsoki Utcai Kórház Neurológiai Osztályán és az Eötvös Loránd Kutatóhálózat Nyelvtudományi Kutatóközpontjában került sor. A vizsgálatot az Uzsoki Utcai Kórház etikai bizottsága hagyta jóvá, és a Helsinki Nyilatkozatnak megfelelően végeztük el. Kísérleteinket 23 SM alany (18 nő és 5 férfi) és 22 kontroll személy (16 nő és 6 férfi) felvételein végeztük. Az SM alanyok mindegyike a relapszáló-remittáló (relapsing-remitting, RRMS) altípusba tartozott. Az alanyok demográfiai adatait az 1. táblázat tartalmazza. A két csoport tagjainak jellemzőit az életkor és az iskolázottság esetében ANOVÁ-val, a beszélők nemének eloszlását $\chi^2$-próbával vizsgáltuk; látható, hogy a két csoport tagjai nem térnek el statisztikailag szignifikánsan egyik vizsgált jellemzőjükben sem.

A felvételi protokoll során sokféle feladatot rögzítettünk az alanyokkal; jelen tanulmányunkban (részben terjedelmi, részben technikai okok miatt) ezekből négyet használunk. Ezek a következők:

| | | Beszélőcsoportok | | Stat. |
|---|---|---|---|---|
| | | **SM** | **Kontroll** | $p$ |
| **Életkor** | átlag ± szórás | 39,00 ± 8,11 | 39,95 ± 7,22 | $p = 0,685$ |
| | terjedelem | [24, 56] | [28, 56] | |
| **Nem** | férfi / nő | 5 / 18 | 6 / 16 | $p = 0,536$ |
| **Iskolázottság (év)** | átlag ± szórás | 15,05 ± 2,17 | 16,09 ± 1,26 | $p = 0,100$ |
| | terjedelem | [12, 19] | [12, 19] | |

1. táblázat. A vizsgált csoportok demográfiai adatai.

– az SM alanyokat a betegségükről, a kontrollokat a munkájukról kérdeztük (**betegség / munka**),
– az alanyokat megkértük, hogy meséljék el részletesen az előző napjukat (**tegnapi nap**),
– egy kétperces, számukra korábban ismeretlen tudományos ismeretterjesztő szöveg meghallgatása után az alanyoknak minél pontosabban el kellett azt mesélniük (**szövegösszefoglalás**),
– végül föl kellett olvasniuk olyan mondatokat, amelyben CVCV hangkapcsolatú álszavak voltak (**fonetika**). Az első CV hangkapcsolat egy felpattanó zárhangból ([p, t, k]) és az [i:, a:, u:] magánhangzók egyikéből állt.

A felvételeket egy Sony PCM-A10 digitális diktafonnal, csíptetős mikrofonnal rögzítettük. Az eredetileg sztereó, 48 kHz mintavételű felvételeket a feldolgozás előtt 16 kHz mintavételezésű, monó formátumra konvertáltuk.

## 3. Akusztikus beágyazások

A bevezetőben fölsorolt okokból egy hagyományos előrecsatolt mély neurális hálót fogunk jellemzőkinyerésre alkalmazni. Ebből adódóan a javasolt eljárás első lépése egy ilyen modell tanításából áll, már amennyiben erre szükség van egyáltalán (ugyanis a javasolt megközelítés egyik előnye, hogy az ilyen neurális hálók elterjedtsége miatt egy ilyen jó eséllyel már eleve rendelkezésre áll). Természetesen ehhez a lépéshez szükséges, hogy rendelkezzünk valamely nagyobb méretű beszédadatbázissal, valamint hozzá tartozó annotált és időzített fonetikai címkékkel (vagy legalább szöveges átirattal). Véleményünk szerint azonban ez a gyakorlatban nem egy szigorú megkötés, az ilyen adatbázisok nagy száma és elterjedtsége miatt. Ezen lépés eredménye egy (keretszinten működő) DNN akusztikus modell.

A jellemzőkinyerés második lépése során ezt az akusztikus modellt ki kell értékelnünk az alanyoktól rögzített hangfelvételeken. Ahelyett azonban, hogy (bevett módon) a kimeneteket rögzítenénk, valamely rejtett réteg aktivációit mentjük el. Mivel ezek az aktiváció-vektorok továbbra is keretszintűek, a harmadik lépésben a teljes hangfelvételen összegezzük azokat. Jelen cikkünkben négyféle

ilyen aggregációs lépést vizsgálunk meg: átlagot (mean), szórást (standard deviation), ferdeséget (skewness) és csúcsosságot (kurtosis) számítunk. Amellett, hogy külön-külön is teszteljük ezeket az összegző stratégiákat, a kapott (immár felvételszintű) beágyazásvektorokat össze is fűzhetjük. Az így kapott vektorokat jellemzőkként használjuk az osztályozási lépés során; ezek mérete így az adott rejtett réteg neuronszámának egy- és négyszerese között alakul.

## 4. A kísérletek technikai jellemzői

### 4.1. A DNN akusztikus modell

Mély neurális háló akusztikus modellünket a (magyar nyelvű) BEA Spontán-beszéd-adatbázis egy részhalmazán tanítottuk (Neuberger és mtsai, 2014). 165 beszélőt választottunk ki; a felvételekből automatikusan kivágtuk azokat a részeket, melyekben a felvételvezető hangja is hallható, így 10636 hangfelvételt kaptunk, összesen 60 órányi terjedelemben. Az eredeti sztereó, 44,1 kHz-en mintavételezett bemondásokat monó, 16 kHz-es formátumra konvertáltuk.

Mély neurális hálónk 5 rejtett rétegből állt, mindegyikben 1024 ReLU neuronnal, a kimeneti rétegben pedig a softmax aktivációs függvényt alkalmaztuk. Bemenetként az ún. FBANK jellemzőkészletet használtuk, amely 40 mel szűrősor energiáiból, illetve azok első- és másodrendű deriváltjaiból állt. Tanítás és kiértékelés során 15 keret széles mozgóablakot használtunk, így a háló bemeneteinek száma 1845 volt, míg a kimeneten 911 kontextusfüggő állapotot modelleztünk.

### 4.2. Jellemzőkinyerés

A beágyazásokat az akusztikus modell mindegyik rejtett rétegéből $(1 \dots 5)$ kimentettük; a keretszintű beágyazásvektorok mérete megfelelt a rejtett rétegek neuronszámának, így minden esetben 1024 méretű vektorokat kaptunk. Felvételszintű aggregálásra mind a négy korábban fölsorolt módszert (átlag, szórás, ferdeség és csúcsosság) kipróbáltuk külön-külön; emellett kísérleteztünk az átlag és szórás együttes használatával (2048 jellemző), valamint mind a négy összegző megközelítés alkalmazásával (4096 jellemző). Az összegzett értékeket (azaz a felvételszintű jellemzővektorokat) minden esetben standardizáltunk (azaz minden jellemzőt lineárisan nulla átlagra és egységnyi szórásra transzformáltunk).

### 4.3. Beszélőosztályozás

A jellemzőkinyerési lépés után a beszélőket Support Vector Machine (SVM, Schölkopf és mtsai, 2001) alkalmazásával osztályoztuk, a libSVM csomagot (Chang és Lin, 2011) használva. A túltanulás elkerülése érdekében lineáris kernelt használtunk, így egyetlen hiperparaméterünk az SVM $C$ (complexity) értéke volt; ezt a $10^{-5}, 10^{-4}, \dots, 10^{1}$ értékek közül választottuk ki. A tanítás beágyazott keresztvalidációval történt; minden csoportban (foldban) egy-egy SM beteg és egy kontroll alany volt (egy fold kivételével, amely egyetlen SM betegből állt), így

23 csoportot kaptunk. A $C$ hiperparamétert minden tanítás esetén egy további (belső, 22-szeres) keresztvalidációs lépés segítségével választottuk ki, a legjobb ROC görbe alatti terület (AUC) érték alapján.

Összehasonlító kísérleteinkben az AUC értéken kívül további kiértékelési metrikákat is kiszámítottunk: osztályozási pontosságot (classification accuracy, *Pont.*), pontosságot (precision, *Prec.*), fedést (recall) és $F_1$-értéket (F-measure). (Pontosság (precision), fedés és $F_1$ esetén az SM beszélőkategóriát tekintettük pozitív osztálynak; mivel csak két beszélőkategóriánk (SM és kontroll) volt, a két osztályra kapott AUC-értékek megegyeztek.) A két osztály közötti döntési küszöböt az irodalomban megszokott módon az egyenlő hibaértéknél (Equal Error Rate, EER) húztuk meg.

### 4.4. x-vektor DNN-ek

Összehasonlítási alapnak x-vektor neurális hálókat tanítottunk, a BEA adatbázis azonos (hatvanórányi) részhalmazán. Ehhez a Kaldi rendszert használtuk (Povey és mtsai, 2011) mind a hálók tanítása, mind az x-vektor jellemzők kinyerése során. Keretszintű jellemzőként mindhárom variációt kipróbáltuk, amit a Kaldi támogat: 23 MFCC-vel, 40 FBANK-kal, illetve spektrogramokkal is kísérleteztünk. A tanítás során szokásos eljárás a tanító adat méretét mesterségesen megnövelni úgy, hogy az eredeti hangfelvételekhez zajt adnak és/vagy visszhangosítják azokat (Snyder és mtsai, 2018). Emiatt minden keretszintű jellemzőtípusra két DNN modellt tanítottunk: egyet augmentációval, egyet pedig ennek a lépésnek a kihagyásával. (Az augmentáció 52636 felvételre (293 órányira) növelte a tanítóanyag méretét.) Meglepő módon mind a négy beszélői feladaton az MFCC-ket használó, zajjal augmentált modellek teljesítettek a legjobban.

## 5. Eredmények

Az elért AUC értékeket a 2. táblázat foglalja össze; minden beszédfeladatra és rétegre a legjobb értékeket **félkövérrel** jelöltük. Általánosságban elmondható, hogy az elért értékek elég magasak: bár néhány esetben (főleg a *tegnapi nap* beszédfeladat esetében) kimondottan alacsony (akár 0,565) AUC értékeket is mértünk, a legtöbb esetben 0,800 fölötti pontszámokat kaptunk. A négy összegző eljárás közül egyértelműen a szórás bizonyult a leghasznosabbnak: az összesen 20 esetből 15 alkalommal vezetett a legjobb (vagy közel a legjobb) eredményhez. Az átlag és a csúcsosság 5-5, a ferdeség pedig 3 esetben adta a legmagasabb (vagy ahhoz nagyon közeli) AUC értéket.

A 2. táblázat utolsó két sora a kombinált jellemzőkészletekkel elért eredményeket mutatja (itt a **félkövér** szám azt jelzi, hogy az eredmény magasabb, mint a kombinált módszerek önálló használatával kapott értékek közül a legmagasabb, tehát a kombináció javuláshoz vezetett). Ezen eredmények alapján ez a kombinációs megközelítés nem volt különösebben hatékony: még ha mértünk is javulást, az a legtöbb esetben minimális volt. Összesen három olyan esetet találhatunk,

| Feladat | Jellemzők | Forrás rejtett réteg | | | | |
|---|---|---|---|---|---|---|
| | | 1. | 2. | 3. | 4. | 5. |
| Spontán beszéd (munka / betegség) | Átlag | **0,887** | **0,875** | 0,870 | 0,788 | 0,751 |
| | Szórás | **0,885** | 0,856 | **0,909** | **0,911** | 0,850 |
| | Ferdeség | **0,877** | 0,864 | 0,802 | 0,816 | 0,822 |
| | Csúcsosság | **0,879** | 0,832 | 0,796 | 0,824 | **0,879** |
| | Átlag + szórás | 0,836 | 0,870 | 0,907 | 0,836 | 0,781 |
| | Összes | 0,844 | 0,840 | 0,796 | 0,806 | 0,836 |
| Tegnapi nap | Átlag | 0,634 | 0,640 | 0,717 | 0,710 | 0,704 |
| | Szórás | **0,678** | **0,751** | **0,790** | **0,824** | **0,715** |
| | Ferdeség | 0,630 | 0,565 | 0,636 | 0,761 | 0,646 |
| | Csúcsosság | 0,575 | 0,601 | 0,660 | 0,767 | 0,642 |
| | Átlag + szórás | 0,593 | 0,704 | 0,769 | 0,769 | **0,723** |
| | Összes | 0,615 | 0,626 | 0,672 | 0,745 | 0,642 |
| Szöveg-összefoglalás | Átlag | 0,757 | 0,767 | 0,781 | 0,753 | **0,846** |
| | Szórás | **0,872** | 0,824 | 0,816 | **0,824** | **0,854** |
| | Ferdeség | **0,868** | 0,836 | **0,852** | 0,808 | 0,812 |
| | Csúcsosság | **0,866** | **0,848** | **0,850** | 0,781 | 0,802 |
| | Átlag + szórás | 0,836 | **0,850** | **0,854** | 0,792 | 0,842 |
| | Összes | 0,814 | 0,846 | 0,844 | **0,826** | 0,834 |
| Fonetika | Átlag | **0,737** | 0,719 | 0,830 | **0,846** | 0,838 |
| | Szórás | 0,713 | **0,767** | **0,850** | **0,854** | **0,864** |
| | Ferdeség | 0,652 | 0,717 | 0,721 | 0,810 | 0,810 |
| | Csúcsosság | 0,688 | 0,702 | 0,787 | 0,802 | 0,810 |
| | Átlag + szórás | 0,731 | **0,806** | 0,826 | 0,818 | 0,834 |
| | Összes | **0,739** | 0,700 | 0,759 | 0,783 | 0,816 |

2. táblázat. A beágyazás-alapú jellemzőkkel elért AUC értékek a vizsgált beszédfeladatokon.

ahol a keretszintű aktivációk átlagának és szórásának együttes használata lényegesen jobb volt, mint vagy csak az átlagokat, vagy csak a szórásokat használni jellemzőként (a *szövegösszefoglalás* feladat esetén a DNN 2. és 3. rétegéből, a *fonetika* feladat esetén pedig a DNN 2. rejtett rétegéből számítva a beágyazásokat), ezek 0,026-0,038 abszolút javuláshoz vezettek. Az összes jellemző használatának mérlege még rosszabb: mindkét esetben, ahol ez a megközelítés javított az AUC értékeken, a növekedés csupán (abszolút) 0,002 volt.

Az egyes beszédfeladatok eltérő mértékben voltak hasznosak. A legmagasabb értékeket $(0,751 \ldots 0,911$, átlag: 0,849) a *munka / betegség* feladatra kaptuk; ezt követte a *szövegösszefoglalás* $(0,753 \ldots 0,866$, átlag: 0,821) és a *fonetika* $(0,652 \ldots 0,864$, átlag: 0,776). A legalacsonyabb osztályozási értékekhez a *tegnapi nap* feladat vezetett $(0,565 \ldots 0,824$, átlag: 0,682).

| Feladat | Jellemző-kinyerési módszer | Pontosságértékek | | | | |
|---|---|---|---|---|---|---|
| | | **Pont.** | **Prec.** | **Fedés** | $F_1$ | **AUC** |
| Munka / betegség | 4. réteg | 82,2% | 82,6% | 82,6% | 82,6 | 0,911 |
| | x-vektorok | 73,3% | 73,9% | 73,9% | 73,9 | 0,775 |
| Tegnapi nap | 4. réteg | 68,9% | 69,6% | 69,6% | 69,6 | 0,824 |
| | x-vektorok | 60,0% | 60,9% | 60,9% | 60,9 | 0,725 |
| Szövegösszefoglalás | 1. réteg | 86,7% | 87,0% | 87,0% | 87,0 | 0,872 |
| | 4. réteg | 77,8% | 78,3% | 78,3% | 78,3 | 0,824 |
| | x-vektorok | 77,8% | 78,3% | 78,3% | 78,3 | 0,850 |
| Fonetika | 5. réteg | 82,2% | 82,6% | 82,6% | 82,6 | 0,864 |
| | 4. réteg | 82,2% | 82,6% | 82,6% | 82,6 | 0,854 |
| | x-vektorok | 77,8% | 78,3% | 78,3% | 78,3 | 0,775 |

3. táblázat. A legjobb és a 4. rejtett réteg szórás függvénnyel összegzett aktivációinak használatával, valamint a viszonyítási alapként megvizsgált x-vektorok használatával kapott kiértékelési metrikák. (Pont.: osztályozási pontosság; Prec.: pontosság (precision).)

Érdemes azt is megvizsgálni, hogy melyik rejtett rétegből kinyert beágyazások vezettek a legjobb osztályozási eredményekhez. Azt találjuk, hogy ez egyértelműen függ a beszélő feladatától: míg a *munka / betegség* és a *tegnapi nap* feladatok esetén a 3-4. rejtett réteggel kaptuk a legjobb eredményeket, és a többi réteg használatával kapott AUC értékek lényegesen alacsonyabbnak adódtak, a szövegösszefoglalásnál az 1. és az 5. (tehát a legalacsonyabban és a legmagasabban fekvő) réteg bizonyult a legjobbnak. A *fonetika* feladat esetén a fölső (3-5.) rétegek adták a legjobb eredményt, az alsóbb rétegek ennél lényegesen rosszabb AUC értékekhez vezettek. Véleményünk szerint ez azt tükrözi, hogy az adott feladatban jellemzően miben különbözik az SM és a kontroll alanyok beszédprodukciója. Ismert, hogy egy neurális háló alsóbb rétegei egyszerűbb, alacsonyabb szintű jellemzőkinyerést végeznek, míg a legfölső rétegek már kimondottan magas szintű információkat számítanak ki. Egy DNN akusztikus modell esetében már az alacsonyabb rétegeknél fölismerhető lehet pl. a csönd, míg a legfölső réteg már fonetikai szintű információkat tárol. Véleményünk szerint ennek tudható be, hogy a *fonetika* feladat esetén a fölsőbb rétegek (elsősorban a 4-5. rejtett rétegek) aktivációi bizonyultak hasznosabbnak, hiszen itt maga a feladat is bizonyos fonetikai kombinációk kiejtésére koncentrál. Ugyanakkor, mivel ez egy olvasási feladat volt, az alsóbb rétegek (elsősorban az 1-2. rejtett réteg) lényegesen alacsonyabb értékekhez vezettek. Ezzel szemben a *munka / betegség* vagy a *szövegösszefoglalás* feladatok spontánbeszéd-feladatok voltak, ahol az alany memóriájának is fontos szerep jutott. Itt a legalsó rejtett rétegekből kinyert aktivációk jóval pontosabb osztályozást tettek lehetővé (0,757 és 0,887 közé eső AUC értékek), mint a *fonetika* feladatban (0,652 . . . 0,737).

A 3. táblázatban néhány kiválasztott megközelítés több kiértékelési metrikája is látható. Eddigi eredményeink alapján minden feladathoz kiválasztottuk a legjobb AUC értékhez vezető rejtett réteget, valamint, amennyiben nem az bizonyult a legjobbnak, a 4. rejtett réteget is (mely minden feladatnál stabilan jól teljesített). Felvételszintű összegzésre a szórás függvényt használtuk. Viszonyításként föltüntettük továbbá a négy beszédfeladatra az x-vektorokkal kapott pontosságértékeket is. Összességében elmondható, hogy az osztályozási eredmények is kimondottan magasak: a *tegnapi nap* feladat kivételével (melyre 69% körüli értékeket kaptunk) 77,8% és 87% közé estek. Az x-vektor jellemzőkkel kapott értékek alapvetően rosszabbak voltak: egyedül a *szövegösszefoglalás* feladat esetében érték el a 4. rejtett rétegből kinyert aktivációk számait (illetve az AUC érték magasabb is volt az x-vektorok esetében). Azonban a legjobbnak bizonyult, 1. rejtett rétegből számított beágyazásokkal kapott pontosságértékek még ennél a feladatnál is meghaladták az x-vektorok használatával elérteket; ezek alapján a bemutatott technika hatékonynak és versenyképesnek minősíthető.

## 6. Összegzés

Jelen cikkünkben relapszáló-remittáló sclerosis multiplex (SM) betegek és kontroll alanyok beszédfelvételeit vizsgáltuk. Minden alanytól három spontánbeszéd-felvételt és egy speciális, fonetikai feladat hanganyagát használtuk. Jellemzőkinyerési technikánkat egy standard HMM/DNN hibrid modell DNN akusztikus modelljére alapoztuk: a rejtett rétegek keretszintű aktivációit négy különféle módon (átlag, szórás, ferdeség és csúcsosság) összegeztük. Kísérleteink során ez a jellemzőkinyerési megközelítés hatékonynak bizonyult: az elért pontosságértékek minden esetben meghaladták az x-vektorokkal elért értékeket. Az összegző eljárások közül a szórás használata bizonyult a leghatékonyabbnak, a neurális háló jellemzőkinyerésre használt rejtett rétegének kiválasztásánál azonban figyelembe kellett vennünk az alanyok aktuális beszédfeladatát is: az olvasási, fonetikai feladat esetében az alsó rétegekkel lényegesen alacsonyabb értékeket kaptunk, mint a legfölső réteggel, míg a háromból két spontánbeszéd-feladatnál a legmélyebben fekvő rétegek aktivációinak használata is versenyképes eredményhez vezetett. A legjobb AUC értékek a beszédfeladattól függően 0,824 és 0,911 közé estek, demonstrálva az alkalmazott jellemzőkinyerési technika potenciálját.

## Köszönetnyilvánítás

A kutatást részben támogatta a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal – NKFIH, K-132460, NKFIH-1279-2/2020. Gosztolya Gábor kutatásait az MTA Bolyai János ösztöndíja és az Új Nemzeti Kiválóság Program Bolyai+ pályázata (azonosító: ÚNKP-21-5-SZTE) is támogatta. A publikációban szereplő kutatást (amelyet a Szegedi Tudományegyetem valósított meg) az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal is támogatta a Mesterséges Intelligencia Nemzeti Laboratórium (MILAB) keretében.

## Hivatkozások

Bóna, J., Svindt, V., Hoffmann, I.: Voice onset time of Hungarian voiceless plosives in Multiple Sclerosis. In: ISSP. pp. 202–205 (Dec 2020)

Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 1–27 (2011)

Dehak, N., Kenny, P., Dehak, R., Glembek, O., Dumouchel, P., Burget, L., Hubeika, V., Castaldo, F.: Support Vector Machines and Joint Factor Analysis for speaker verification. In: ICASSP. pp. 4237–4240 (2009)

Egas-López, J., Vetráb, M., Tóth, L., Gosztolya, G.: Identifying conflict escalation and primates by using ensemble x-vectors and Fisher vector features. In: Interspeech. pp. 476–480 (2021)

García, N., Vásquez-Correa, J.C., Orozco-Arroyave, J.R., Nöth, E.: Multimodal i-vectors to detect and evaluate Parkinson's Disease. In: Interspeech. pp. 2349–2353. Hyderabad, India (2018)

Grzybowska, J., Kacprzak, S.: Speaker age classification and regression using i-vectors. In: Interspeech. pp. 1402–1406 (2016)

Huckvale, M., Beke, A., Ikushima, M.: Prediction of sleepiness ratings from voice by man and machine. In: Interspeech. pp. 4571–4575 (2020)

Jenei, A.Z., Kiss, G.: Depresszió detektálása korrelációs struktúrán alkalmazott konvolúciós hálók segítségével. In: MSZNY. pp. 59–71. Szeged (Jan 2020)

Mulfari, D., Meoni, G., Marini, M., Fanucci, L.: Machine learning assistive application for users with speech disorders. Applied Soft Computing 103(May), 107147 (2021)

Neuberger, T., Gyarmathy, D., Gráczi, T., Horváth, V., Gósy, M., Beke, A.: Development of a large spontaneous speech database of agglutinative Hungarian language. In: TSD. pp. 424–431 (2014)

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., Veselý, K.: The Kaldi speech recognition toolkit. In: Proceedings of ASRU (2011)

Renauld, S., Mohamed-Said, L., Macoir, J.: Language disorders in multiple sclerosis: A systematic review. Multiple Sclerosis and Related Disorders 10, 103–111 (2016)

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R.: Estimating the support of a high-dimensional distribution. Neural Computation 13(7), 1443–1471 (2001)

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust DNN embeddings for speaker verification. In: ICASSP. pp. 5329–5333. Calgary, Alberta, Kanada (2018)

Svindt, V., Bóna, J., Hoffmann, I.: Changes in temporal features of speech in secondary progressive multiple sclerosis (SPMS) – case studies. Clinical Linguistics & Phonetics 34(4), 339–356 (2020)

Szirmai, I.: Neurológia. Medicina, Budapest (2006)

Variani, E., Lei, X., McDermott, E., Moreno, I., G-Dominguez, J.: Deep neural networks for small footprint text-dependent speaker verification. In: ICASSP. pp. 4080–4084 (2014)