Különböző wav2vec 2.0 rétegekből nyert beágyazások használata sclerosis multiplex felismerésére

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

[1] HUN-REN–SZTE Mesterséges Intelligencia Kutatócsoport, Szeged
[2] Szegedi Tudományegyetem, Informatikai Intézet
[3] HUN-REN Nyelvtudományi Kutatóközpont, Budapest
[4] ELTE Eötvös Loránd Tudományegyetem, Alkalmazott Nyelvészeti és Fonetikai Tanszék, Budapest
[5] Szegedi Tudományegyetem, Pszichiátriai Klinika

ggabor @ inf.u-szeged.hu

Kivonat A sclerosis multiplex (SM) a központi idegrendszer krónikus gyulladásos megbetegedése, mely többek között az alanyok beszédére is hatással van. Emiatt az automatikus beszédföldolgozás relatíve egyszerû, olcsó és találkozásmentes (távoli) módot nyújthat arra, hogy követhessük a betegek beszédprodukiójának változását. Egy ilyen rendszer tanítása során problémát jelent a rendelkezésre álló beszédanyag mennyiségének és (főleg) változatosságának korlátozottsága (pl. kevés beszélő), ami miatt általában nem praktikus egyetlen (end-to-end) mély hálót alkalmazni az SM detektálására. Ugyanakkor az járható megközelítés lehet, hogy osztályozásra valamilyen hagyományosabb módszert (pl. SVM-et vagy véletlen erdőt) alkalmazunk, mély hálók beágyazásait használva jellemzőkként. Egy korszerû mély háló (pl. wav2vec 2.0) esetén kézenfekvő a konvolúciós és a finomhangolt blokkok utolsó rétegét választani, azonban a közbülső rétegek is hasznosnak bizonyulhatnak. Jelen tanulmányban azt vizsgáljuk meg, hogy a közbülső (belső) rétegekből vett beágyazások használatával javíthatunk-e az SM automatikus felismerésének pontosságán. Eredményeink alapján a 24 finomhangolt rejtett réteg mélyebben fekvő kb. egyharmada bizonyult a leghasznosabbnak, statisztikailag szignifikáns javulást is eredményezve a blokkok utolsó rejtett rétegéhez képest.

1. Bevezetés

A sclerosis multiplex (SM) a központi idegrendszer krónikus gyulladásos megbetegedése, mely számos különféle kognitív és nyelvi tünettel jár együtt (Szirmai, 2006). A betegség előrehaladása jelentősen eltérhet az egyes alanyoknál, valamint időben is változhat. A betegség előrehaladtával romlás következhet be a beteg járásában, mozgáskoordinációjában, az alany fáradékonyabbá válhat, valamint

romolhatnak a kognitív és nyelvi funkciói is (akár dizartria is kialakulhat (Orsolya és mtsai, 2020)). Mindezek alapján az automatikus beszédelemzés potenciálisan alkalmazható a betegség automatikus, kontaktmentes és (viszonylag) olcsó előszűrésére, vagy a beteg állapotának romlásának követésére.

Az elmúlt bő egy évtizedben az automatikus beszédelemzés a beszédtechnológia egy önálló területévé fejlődött. Ide tartozik a *paralingvisztika* (*computational paralinguistics*), amelynek célja a beszélő érzelmi vagy fizikai állapotának megállapítása a beszédjelből (például érzelemfelismerés (Grósz és mtsai, 2023; Kondratenko és mtsai, 2023), a beszélő korának és nemének azonosítása (Kumar és mtsai, 2016), az alany pillanatnyi álmosságának mértékének megbecslése (Huckvale és mtsai, 2020), vagy dadogás detektálása (Grósz és mtsai, 2022)).

Az automatikus beszédelemzés másik nagy területe az *orvosi célú beszédfeldolgozás* (*pathological speech processing*). Itt a feladat annak megállapítása, hogy a beszélő szenved-e valamely konkrét betegségben (pl. Parkinson-kór (Jenei és mtsai, 2022; Klumpp és mtsai, 2022), Alzheimer-kór (Ivanova és mtsai, 2023; Pérez-Toro és mtsai, 2022), depresszió (Fara és mtsai, 2023; Kiss és mtsai, 2016)). Az utóbbi években a mély neurális hálók alkalmazása is rutinszerűvé vált a területen (Fara és mtsai, 2023; Jenei és mtsai, 2022; Pérez-Toro és mtsai, 2022).

Az önfelügyelt tanulás (*self-supervised learning*) megjelenésével napjaink talán legnépszerűbb beszédfeldolgozó architektúrájává a wav2vec 2.0 vált (Baevski és mtsai, 2020). Természetesen bizonyos feladatokon közvetlenül is alkalmazhatóak ilyen mély hálók (pl. beszédfelismerésre (Mihajlik és mtsai, 2022) vagy érzelemfelismerésre (Chen és Rudnicky, 2023)), ugyanakkor korlátozottabb mennyiségű tanítóadatnál sajnos nem triviális a betanításuk. Ilyen esetekben is lehetőségünk van azonban wav2vec 2.0 hálók alkalmazására, például valamely más feladaton (pl. beszédfelismerés vagy beszélőazonosítás) tanított hálók kiértékelésével és a kapott aktivációk (*beágyazások* (*embeddings*)) jellemzőkként használatával; ekkor a szigorúan vett osztályozási feladatra pl. SVM-et vagy véletlen erdőt alkalmazhatunk (Pepino és mtsai, 2021). Az orvosi beszédfeldolgozási területen extrém módon kevés a tanítóadat mennyisége, így a mély hálók az ilyen feladatokon általában ez utóbbi módon, jellemzőkinyerésre vannak felhasználva (Egas-López és mtsai, 2023; Pérez-Toro és mtsai, 2022; Thienpondt és mtsai, 2023; Cai és mtsai, 2025).

Egy neurális háló jellemzőkinyerésre használatához ki kell választanunk egy (vagy több) rejtett réteget, amelyből a beágyazásokat nyerjük. Egy wav2vec 2.0 architektúrájú háló két fő blokkból áll: egy alsó *konvolúciós* és egy fölső *finomhangolt* (vagy *kontextualizált*) blokkból, és kézenfekvőnek tűnik a blokkok utolsó rejtett rétegének használata (Gosztolya és mtsai, 2023; Kodali és mtsai, 2023). Jelen dolgozatunkban ennél alaposabb vizsgálatot végzünk: azt vizsgáljuk, hogy kaphatunk-e hasznosabb jellemzőket valamely közbülső rejtett réteg használatával. Az osztályozási feladat magyar nyelvű sclerosis multiplex alanyok megkülönböztetése lesz (szintén magyar nyelvű) egészséges kontroll alanyoktól, a beágyazásokat pedig két, magyar nyelvre finomhangolt wav2vec 2.0 hálóból nyerjük: ebből az egyik beszédfelismerésre, a másik pedig beszélőfelismerésre lett finomhangolva.

2. Hangfelvételek

A vizsgálatokra a budapesti Uzsoki Utcai Kórház Neurológiai Osztályán és az (akkor még) 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 beszélő (16 nő és 6 férfi) felvételein végeztük. Az SM alanyok mindegyike a relapszáló-remittáló (*relapsing-remitting*, RRMS) altípusba tartozott. A két csoport statisztikailag illesztett életkor, iskolázottság és nem szerint (azaz $p > 0,05$).

A felvételi protokoll során összesen 17 különféle feladatot rögzítettünk az alanyokkal. Jelen tanulmányunkban kísérleteinket a *szövegösszefoglalás* feladat hangfelvételein végeztük; ennek során az alanyoknak egy kétperces, számukra korábban ismeretlen tudományos ismeretterjesztő szöveg meghallgatása után minél pontosabban el kellett azt mesélniük. A feladat végrehajtása számos kognitív funkció (figyelemösszpontosítás, munkamemória, időbeli orientáció, rendszerezés) összehangolt működését igényli (Mar, 2004).

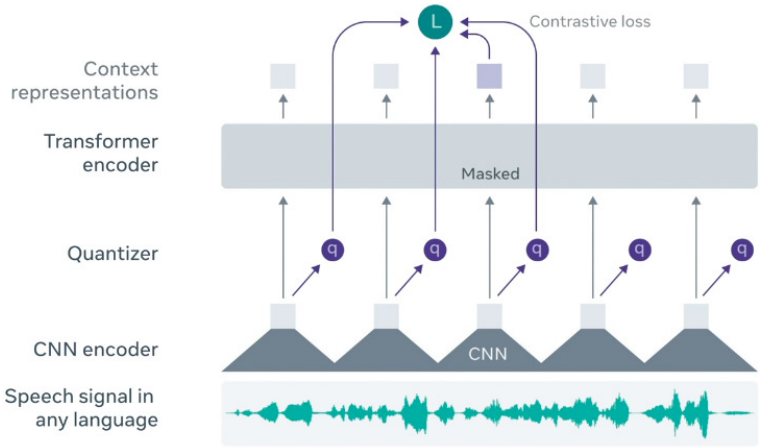
Az alanyok válaszait egy Sony PCM-A10 digitális diktafonnal, valamint csíptetős mikrofonnal rögzítettük. Az eredetileg sztereó, 48 kHz mintavételű felvételeket az automatikus feldolgozás előtt 16 kHz mintavételezésű, monó formátumra konvertáltuk.

3. wav2vec 2.0

A **wav2vec** architektúra egy konvolúciós neurális háló, melynek bemenete a nyers beszédhang-jel, kimenete pedig egy olyan reprezentáció, mely alkalmas egy beszédfelismerő rendszer bemenetének. Tanítása önfelügyelt módon történik, az adott felvétel következő szegmensének predikciójára (Schneider és mtsai, 2019). A **wav2vec 2.0** architektúra ezen a megközelítésen úgy javított, hogy a tanítás során – a modell zajtűrésének javítása érdekében – maszkolást (a bemenet egyes részeinek nulla értékkel helyettesítését) is alkalmazott. További különbség a kontrasztív tanítás (contrastive learning) alkalmazása, melynek során a modell célja annak felismerése is, hogy két különbözőféleképpen transzformált reprezentáció (esetünkben a finomhangolt réteg kimenete, valamint a konvolúciós réteg kimenetének diszkretizált (kvantált) formája) azonos bemenethez tartozik-e (Baevski és mtsai, 2020). A wav2vec 2.0 architektúra az 1. ábrán látható.

3.1. Jellemzőkinyerés wav2vec 2.0 segítségével

Amennyiben egy wav2vec 2.0 hálót paralingvisztikai vagy orvosi célú beszédelemzési feladatban jellemzőkinyerésre szeretnénk használni, a legkézenfekvőbb választás vagy a konvolúciós, vagy a finomhangolt blokk utolsó rétegéből venni az aktivációkat (Fan és mtsai, 2021; Kodali és mtsai, 2023). Mivel ezek a beágyazások keretszintűek, a kinyert vektorok száma egyenesen arányos lesz a felvétel hosszával, így ebben a formában nem használhatóak jellemzőkként (legalábbis a



1. ábra: Egy finomhangolt wav2vec 2.0 háló struktúrája. Forrás: https://ai.meta.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio .

számunkra most releváns feladatban). Hogy felvételszintű jellemzőket kapjunk belőlük, valamilyen módon aggregálni kell őket a teljes hangfelvételen, például venni az átlagukat és a szórásukat (Gosztolya és mtsai, 2022; Pérez-Toro és mtsai, 2022; Vaessen és Van Leeuwen, 2021).

Jelen tanulmányunkban a finomhangolt blokk *belső* rétegeire koncentrálunk. Egy standard XLSR-53 hálóban (lásd (Babu és mtsai, 2022)) 24 ilyen transzformer réteg van, mely kiegészülve a konvolúciós blokk utolsó rejtett rétegével, 25 lehetőséget jelent. Közismert, hogy egy mély háló mélyebben fekvő rétegei jellemzően alacsonyabb szintű információkat tárolnak (beszédfeldolgozás esetén pl. csend, különböző zajok vagy további akusztikai paraméterek (pl. visszhang mértéke)), míg a magasabban található rétegekben elsősorban magasabb szintű (esetünkben például, a jellemzőkinyerésre használt háló függvényében, fonetikai vagy az aktuális beszélőre jellemző) információk találhatóak. Sajnos ezen kívül további kapaszkodó nem áll rendelkezésünkre a megfelelő rejtett réteg kiválasztására, így mind a 25 variációt végig fogjuk próbálni.

4. Kísérletek

4.1. Jellemzőkinyerés

Két wav2vec 2.0 modellt használtunk a jellemzőkinyerési lépés során, melyek a Facebook által előtanított `XLSR-53` modell finomhangolásával keletkeztek (Babu és mtsai, 2022). Ebben az architektúrában a konvolúciós réteg utolsó rejtett rétege 512 neuronból áll, míg az ezt követő kontextualizált blokk mint a 24 rejtett rétege 1024 neuront tartalmaz. A finomhangolt modellek közül az első a nyilvánosan elérhető `wav2vec2-large-xlsr53-hungarian`, melyet `jonatasgrosman`

finomhangolt a Mozilla Common Voice 6.1 adatbázis magyar részhalmazán (8 órányi adaton) (Grosman, 2021). A továbbiakban erre a modellre mint *leiratozó* modellre fogunk hivatkozni.

A másik fölhasznált modellünk saját tanítású volt, ugyanúgy a `XLSR-53` modellből kiindulva. Ezt a (magyar nyelvű) BEA Spontánbeszéd-adatbázis egy részhalmazán finomhangoltuk (Neuberger és mtsai, 2014). 165 beszélőt választottunk ki; a felvételekből automatikusan kivágtuk azokat a részeket, melyekben a felvételvezető hangja is hallható, így összesen körülbelül 60 órányi beszédanyagot használtunk. Az eredeti sztereó, 44,1 kHz-en mintavételezett bemondásokat monó, 16 kHz-es formátumra konvertáltuk. Az utolsó rejtett réteg aktivációit kiátlagoltuk (*average pooling*), az ezt követő kimeneti rétegben pedig szoftmax aktivációt használtunk 165 neuronnal. A betanult modell beszélőosztályozási hibája a teszthalmazon (de értelemszerűen ugyanezen beszélőkön) 1,92% volt. A továbbiakban erre a modellre mint *beszélő-azonosító* modellre fogunk hivatkozni.

A kinyert beágyazások mindkét háló és minden réteg esetén keretszintűek voltak, melyeket az időtengely mentén vett átlaggal és szórással konvertálunk felvételszintűvé. Mivel jelen munkánkban elsősorban a finomhangolt blokkból nyert beágyazásokra fókuszáltunk, a legtöbb teszt során 2048 felvételszintű jellemzőt használtunk; ez alól a konvolúciós blokk utolsó rétegét vizsgáló tesztjeink voltak kivételek, ahol 1024 jellemzőnk volt.

Természetesen a két wav2vec 2.0 hálóval kapott eredményeket nem lehet közvetlenül összehasonlítani, hiszen azok tanítása más-más adaton (és vélhetőleg más hiperparaméterekkel) történt. Véleményünk szerint azonban alapvetően így is alkalmasak arra, hogy demonstrálják a különböző célokra tanított hálók eredményességét, amikor jellemzőkinyerésre használjuk azokat.

4.2. Osztályozás

Osztályozási kísérleteink egy hagyományos sémát követtek: mivel az orvosi beszédfeldolgozási területen egy beszélő felel meg egy gépi tanulási példának, adathalmazunk gépi tanulási szempontból nagyon kisméretűnek számít. Ezért nem definiáltunk külön tanító-, fejlesztési- és teszthalmazt, hanem keresztvalidációt alkalmaztunk. Minden csoportba (*fold*ba) egy SM és egy kontroll alany felvételei kerültek, így összesen 23 csoportot kaptunk. Az osztályozás minőségét a ROC görbe alatti terület (AUC) metrikával mértük, melynek használata szintén igen elterjedt a területen (Carvajal-Castaño és mtsai, 2022; Gosztolya és mtsai, 2022).

Osztályozó eljárásnak SVM-et használtunk (LibSVM implementáció (Chang és Lin, 2011)). A nu-SVM változatot használtuk lineáris kernellel; a C hiperparaméter értékét $10^{\{-5,\dots,1\}}$ között változtattuk. Beágyazott keresztvalidációt alkalmaztunk (Cawley és Talbot, 2010): 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.

Az AUC értékek robosztusságának ellenőrzése érdekében minden osztályozási kísérletet ötször végeztünk el, melyek a 23 beszélőcsoport (véletlenszerű) kialakításában tértek el egymástól. Az ábrákon és a táblázatokban az öt így kapott AUC érték átlagát adjuk meg, míg a modellek robosztusságának vizsgálatához

wav2vec 2.0 modell	Beágyazás	AUC		
		Átlag	Szórás	Terjedelem
Leiratozó	Utolsó konvolúciós	0,724	0,008	[0,712; 0,733]
	Utolsó finomhangolt	0,806	0,014	[0,787; 0,824]
Beszélő-azonosító	Utolsó konvolúciós	0,716	0,020	[0,690; 0,741]
	Utolsó finomhangolt	0,402	0,059	[0,324; 0,461]

1. táblázat. A mért átlagos AUC értékek, azok szórása és terjedelme ([min; max]) a konvolúciós és a finomhangolt blokkok utolsó rétegeiből nyert beágyazások használatával.

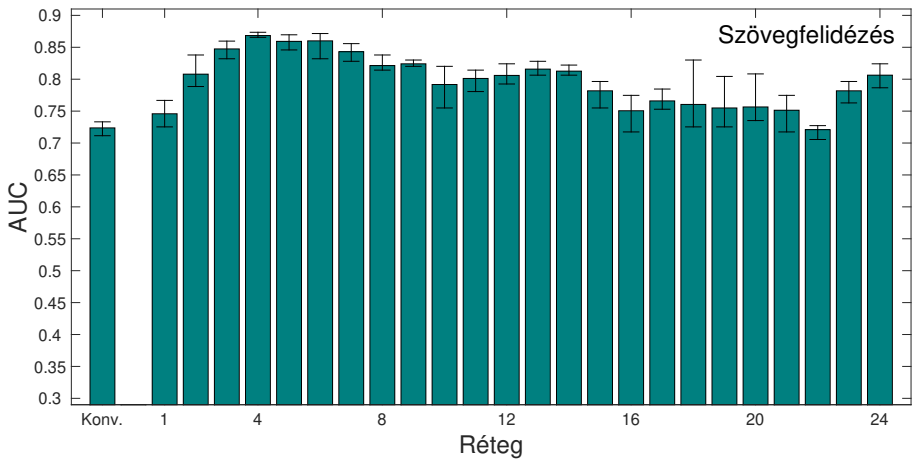
az AUC értékek szórását, valamint az értékek terjedelmét ([min; max]) tüntetjük föl.

5. Az utolsó rétegekkel kapott eredmények

Az 1. táblázat mutatja a konvolúciós és finomhangolt blokkok utolsó rejtett rétegeinek használatával kapott eredményeket. A **leiratozó** modellel kapott eredmények elfogadhatónak tekinthetőek, és igen hasonlóak a korábban a szakirodalomban megjelent AUC értékekhez (Gonzalez-Machorro és mtsai, 2023; Gosztolya és mtsai, 2023). A szórás is elég alacsony mindkét típusú beágyazás esetén, amely elég robosztus felismerést jelez. A **beszélő-azonosító** háló esetén a konvolúciós beágyazásokkal nagyon hasonló értékeket kaptunk, mint a leiratozó modellel: a 0,716-os átlagos AUC érték vállalhatónak tekinthető, bár a modellek AUC értékeinek szórása és terjedelme valamivel nagyobb. A finomhangolt blokk utolsó rejtett rétege esetén azonban az osztályozási teljesítmény igen rossznak bizonyult: a 0,402-es átlagos AUC érték még a véletlen találgatást is alulmúlja. Ez valószínűleg betudható annak, hogy a két háló eltérő feladatra lett tanítva, bár, mivel a tanítás számos paramétere (adatbázis, keretrendszer, hiperparaméterek) biztosan vagy vélhetően eltér, ezt érdemes fenntartással kezelni.

6. A közbülső rétegekkel kapott eredmények

A 2. ábra mutatja a **leiratozó** wav2vec 2.0 hálóból nyert beágyazások használatával kapott átlagos AUC értékeket. (*Konv.* jelöli a konvolúciós blokk utolsó rétegéből nyert beágyazásokhoz tartozó eredményeket, míg 1…24 a finomhangolt blokk megfelelő rejtett rétegét. A 24. réteg a blokk utolsó rétege, amelyhez tartozó eredmények az 1. táblázatban is szerepelnek.) A hibasávok a legkisebb és legnagyobb kapott értékeket mutatják. Meglepő módon az *összes* belső rejtett réteggel jobb eredményeket kaptunk, mint az utolsó konvolúciós rétegre támaszkodva. A különbségeket a Mann-Whitney U teszttel (lásd (Mann és Whitney, 1947)) elemezve azt látjuk, hogy a javulás a 23 esetből 20-ban bizonyult statisztikailag is szignifikánsnak (csak a 16., 18. és 22. rétegek képeznek kivételt).



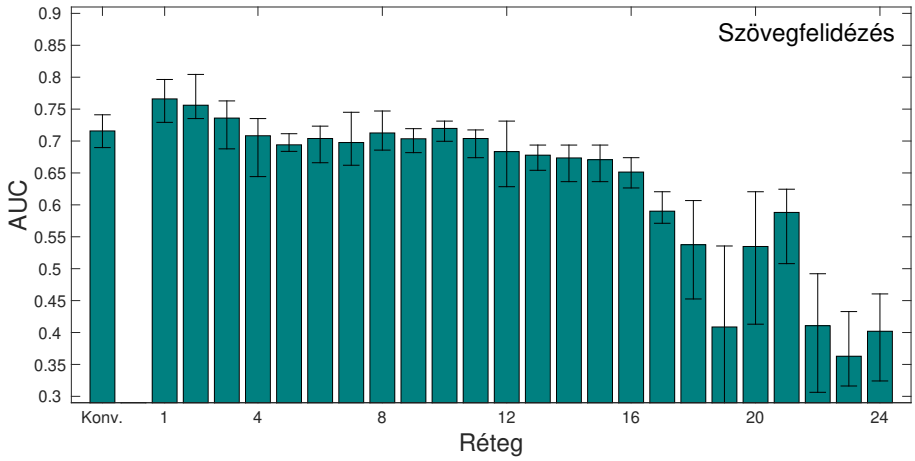
2. ábra: Az átlagos AUC értékek és a [min; max] terjedelem (hibasávokkal jelölve) a konvolúciós blokk utolsó rejtett rétege (*Konv.*), valamint a finomhangolt blokk összes rejtett rétege esetén, a *leiratozó* wav2vec 2.0 hálót használva.

A finomhangolt blokk utolsó rejtett rétegéhez hasonlítva viszont már csak hat esetben kaptunk statisztikailag szignifikáns (vagy időnként bármilyen) javulást; ezek mindegyike a finomhangolt blokk mélyebben fekvő részén (1. . .9. rétegek) található. Ez alapján érdemes lehet egy wav2vec 2.0 háló belső rejtett rétegeiből nyerni az osztályozáshoz használt beágyazásokat, de tanácsos lehet a mélyebben fekvő rétegeket választani, legalábbis sclerosis multiplex felismerése esetén.

A **beszélő-azonosító** háló esetén (ld. 3. ábra) az átlagos AUC értékek teljesen másként alakultak. A konvolúciós blokk utolsó rétegét ugyan néhány belső rejtett réteg túlszárnyalja (ebből az első három réteggel kapott javulás bizonyult szignifikánsnak), a későbbi rétegek használatával azonban csak azonos SM-azonosítási teljesítményt tudtunk elérni. A finomhangolt blokk magasabban található rejtett rétegei pedig egyenesen a pontosság csökkenéséhez vezettek. (Ebben az esetben nem láttuk sok értelmét, hogy a finomhangolt blokk utolsó rétegéhez is hozzámérjük a kapott eredményeket.)

A 2. táblázatban tüntettük föl a néhány kiválasztott rejtett réteg esetén kapott AUC értékeket; a szignifikáns javításokat „*” ($p < 0,05$) és „**” ($p < 0,01$) jelzi, míg „—” esetén nincs ilyen különbség. A per szimbólum („/”) előtti és utáni szimbólumok a konvolúciós, illetve a finomhangolt blokk utolsó rejtett rétegéhez viszonyított javulásra vonatkoznak. A **leiratozó** hálóból származó beágyazások esetén mind a négy kiválasztott réteggel szignifikáns javulást tudtunk elérni a konvolúciós blokk utolsó rejtett rétegéhez képest[1], míg az utolsó finomhangolt réteget a 4. és 6. rejtett réteggel tudtuk jelentősen túlszárnyalni. Ezen két réteg között nincs nagy különbség az átlagos AUC értékben, bár a 4. réteg esetén ez

[1] Valamint még további 16 réteggel



3. ábra: Az átlagos AUC értékek és a [min; max] terjedelem (hibasávokkal jelölve) a konvolúciós blokk utolsó rejtett rétege (*Konv.*), valamint a finomhangolt blokk összes rejtett rétege esetén, a *beszélő-azonosító* wav2vec 2.0 hálót használva.

wav2vec 2.0 modell	**Beágyazás**	**AUC**		
		Átlag	**Szórás**	**Terjedelem**
Leiratozó	2. finomhangolt$^{**/—}$	0,808	0,022	[0,789; 0,838]
	4. finomhangolt$^{**/**}$	0,868	0,004	[0,866; 0,874]
	6. finomhangolt$^{**/**}$	0,860	0,016	[0,832; 0,872]
	8. finomhangolt$^{**/—}$	0,821	0,010	[0,814; 0,838]
	Utolsó konvolúciós	0,724	0,008	[0,712; 0,733]
	Utolsó finomhangolt	0,806	0,014	[0,787; 0,824]
Beszélő-azonosító	2. finomhangolt$^{**/**}$	0,756	0,028	[0,735; 0,804]
	4. finomhangolt$^{—/**}$	0,708	0,038	[0,644; 0,735]
	6. finomhangolt$^{—/**}$	0,704	0,024	[0,666; 0,723]
	8. finomhangolt$^{—/**}$	0,713	0,026	[0,686; 0,747]
	Utolsó konvolúciós	0,716	0,020	[0,690; 0,741]
	Utolsó finomhangolt	0,402	0,059	[0,324; 0,461]

2. táblázat. Az átlagos AUC értékek, szórásuk és terjedelmük ([min;max]) néhány kiválasztott belső rejtett réteg esetén. A statisztikailag szignifikáns javulást „*” ($p < 0,05$) és „**” ($p < 0,01$) jelzi, míg „—” esetén nincs ilyen különbség.

valamivel magasabb (0,868 vs. 0,860), valamint az alacsonyabb szórás és szűkebb terjedelem robosztusabb működésre is utal.

A **beszélő-azonosító** wav2vec 2.0 háló esetén nem ennyire jók az eredmények: a konvolúciós blokk utolsó rejtett rétegével kapott osztályozási teljesít-

wav2vec 2.0 modell	**Beágyazás**	**AUC**		
		Átlag	**Szórás**	**Terjedelem**
Leiratozó	1...8. finomhangolt$^{**/*}$	0,832	0,040	[0,740; 0,872]
	9...16. finomhangolt$^{**/—}$	0,798	0,026	[0,741; 0,828]
	17...24. finomhangolt$^{**/—}$	0,762	0,033	[0,719; 0,817]
Beszélő-azonosító	1...8. finomhangolt$^{—/**}$	0,722	0,037	[0,664; 0,793]
	9...16. finomhangolt$^{—/**}$	0,686	0,029	[0,632; 0,730]
	17...24. finomhangolt$^{—/—}$	0,479	0,107	[0,311; 0,621]

3. táblázat. Az átlagos AUC értékek, szórásuk és terjedelmük (5. és 95. percentilisek) a finomhanglot blokk mélyebben, középen és magasabban fekvő régiói esetén. A statisztikailag szignifikáns javulást „*” ($p < 0.05$) és „**” ($p < 0.01$) jelzi, míg „—” esetén nincs ilyen különbség.

ményt csak a finomhangolt blokk 2. rejtett rétege tudta túlteljesíteni ($p = 0,003$), a többi kiemelt réteg már csak ekvivalens eredményekhez vezetett, ráadásul valamivel nagyobb szórás-értékek mellett. (A 2. finomhangolt rétegre épülő osztályozó modellek robosztussága megegyezik a konvolúciós rétegre építettekével.) A finomhangolt blokk utolsó rejtett rétegét sikerült mind a négy esetben szignifikánsan túlteljesíteni, de annak pontossága eleve nagyon alacsony volt. Összességében a beszélő-azonosító modellel kapott eredmények elmaradtak a leiratozó modellel kapottaktól, még a mélyebben fekvő rejtett rétegek esetén is.

Végül a 3. táblázatba szedtük össze finomhangolt blokk alsó, középső és legfölső egyharmadának összesített teljesítményét. (Ebben az esetben a terjedelmet a 40 (8×5) AUC érték 5. és 95. percentilisével számítottuk.) A **leiratozó** háló esetén mindegyik régió szignifikánsan jobb eredményeket adott, mint a konvolúciós blokk utolsó rejtett rétege ($p < 0,01$), a finomhangolt blokk utolsó rejtett rétegéből kapott beágyazásoknál azonban csak az alsó régió bizonyult jobbnak ($p = 0,038$), míg a legfölső régió szignifikánsan rosszabb eredményeket adott. A **beszélő-azonosító** háló esetén, meglepő módon, még a mélyebb régió is csupán azonos teljesítményt tudott nyújtani, mint a konvolúciós réteg (0,722 és 0,716 átlagos AUC értékek, $p = 0,691$), a középső és legfölső régiók pedig szignifikánsan rontottak ($p = 0,038$ és $p < 0,001$). A legtöbb esetben sikerült túlszárnyalni a finomhangolt blokk utolsó rejtett rétegét, azonban ez a referencia-érték eleve igen alacsony volt.

Összességében azt kaptuk, hogy mindkét háló esetén általában a finomhangolt blokk mélyebben fekvő rétegeiből érdemes jellemzőket kinyerni, mert azok alkalmasabbak a sclerosis multiplex felismerésére. A kétfajta hálótípus közül a beszédfelismerésre tanított (*leiratozó*) sokkal hatékonyabbnak bizonyult, mint a beszélők felismerésére tanított (*beszélő-azonosító*), azonban ezt fenntartással kell kezelnünk, hiszen a két háló eltérő (bár egyaránt magyar nyelvű) adaton és eltérő körülmények között lett betanítva.

7. Konklúzió

Jelen tanulmányunkban azt vizsgáltuk, hogy a sclerosis multiplex milyen mértékben detektálható automatikusan az alanyok beszédéből. Ehhez egy viszonylag hagyományos rendszert építettünk, ahol a jellemzőket egy wav2vec 2.0 hálóból nyertük, osztályozásra pedig SVM-et használtunk. 45 anyanyelvi magyar beszélő felvételeit használtuk (23 relapszáló-remittáló altípusú SM alanyttó és 22 demográfiailag illesztett egészséges kontrolltól). A wav2vec 2.0 háló konvolúciós és finomhangolt (kontextualizált) blokkjainak utolsó rejtett rétegeiből vett jellemzők mellett a finomhangolt blokk további 23 rétegét is megvizsgáltuk. Két különböző módon (beszédfelismerésre, valamint beszélőfelismerésre) tanított wav2vec 2.0 hálót is teszteltünk.

A beszédfelismerő (vagy leiratozó) háló esetén a konvolúciós blokk utolsó rejtett rétegéhez képest majdnem minden esetben statisztikailag szignifikáns javulást sikerült elérni, és a finomhangolt blokk utolsó rejtett rétegét is túlszárnyaltuk a mélyebben fekvő rejtett rétegekkel. A beszélő-azonosító háló esetén azonban, bár néhány mélyebben fekvő rejtett réteg javulást eredményezett, a legtöbb esetben azonos vagy egyenesen rosszabb eredményeket kaptunk. Ezekből a megfigyelésekből arra következtethetünk, hogy egy jellemzőkinyerésre használt wav2vec 2.0 hálót érdemesebb lehet a hagyományos beszédfelismerési feladatra tanítani, mint beszélőazonosításra. Ugyanakkor mindkét tanítási kritérium esetén a finomhangolt blokk mélyebben fekvő rejtett rétegekre érdemes támaszkodni.

Az elvégzett kísérletek hiányossága, hogy a két wav2vec 2.0 háló, bár architektúrájukban és a finomhangolás előtti súlyaikban megegyeztek, más adaton, eltérő keretrendszert használva és eltérő hiperparaméterekkel lettek tanítva. Bár a saját csapatunk által, beszélőazonosításra tanított háló (beszélő)osztályozási pontossága kellően alacsonynak tűnik (2% alatti), nem zárhatjuk ki, hogy valamilyen technikai probléma vezetett a látványosan alacsonyabb teljesítményhez, mikor a rejtett rétegek aktivációit használtuk jellemzőkként.

Megjegyeznénk még, hogy az egyes rejtett rétegek más jellegű információkat tárolnak, így a belőlük kinyert beágyazások kombinálása tovább javíthatja az osztályozási pontosságot. Sajnos az ilyen kombinációs kísérletek korrekt kiértékeléséhez valószínűleg több alanyra van szükség, mint a jelenleg rendelkezésünkre álló 45 fő (amely egyébként az orvosi célú beszédfeldolgozás területén elegendőnek számít). Ennek ellenére a közeljövőben tervezzük ilyen kombinációs kísérletek elvégzését.

8. Köszönetnyilvánítás

A kutatást részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta a NKFIH K-132460 pályázat keretében. A kutatást (amelyet a Szegedi Tudományegyetem valósított meg) az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal is támogatta a TKP2021-NVA-09 pályázat és a Mesterséges Intelligencia Nemzeti Laboratórium (RRF-2.3.1-21-2022-00004) keretében.

Hivatkozások

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., Auli, M.: XLS-R: Self-supervised cross-lingual speech representation learning at scale. In: Proceedings of Interspeech. pp. 2278–2282 (2022)

Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems 33, 12449–12460 (2020)

Cai, J., Song, Y., Wu, J., Chen, X.: Voice disorder classification using wav2vec 2.0 feature extraction. Journal of Voice p. to appear (2025)

Carvajal-Castaño, H.A., Pérez-Toro, P.A., Orozco-Arroyave, J.R.: Classification of Parkinson's Disease patients – a deep learning strategy. Electronics 11(17), 2684 (2022)

Cawley, G.C., Talbot, N.L.C.: On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research 11(Jul), 2079–2107 (2010)

Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 1–27 (2011)

Chen, L.W., Rudnicky, A.: Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In: Proceedings of ICASSP. Rhodes Island, Greece (2023)

Egas-López, J.V., Svindt, V., Bóna, J., Hoffmann, I., Gosztolya, G.: Automated multiple sclerosis screening based on encoded speech representations. In: Proceedings of Interspeech. pp. 3003–3007. Dublin, Ireland (Aug 2023)

Fan, Z., Li, M., Zhou, S., Xu, B.: Exploring wav2vec 2.0 on speaker verification and language identification. In: Proceedings of Interspeech. pp. 1509–1513 (2021)

Fara, S., Hickey, O., Georgescu, A., Goria, S., Molimpakis, E., Cummins, N.: Bayesian Networks for the robust and unbiased prediction of depression and its symptoms utilizing speech and multimodal data. In: Proceedings of Interspeech. pp. 1728–1732. Dublin, Ireland (2023)

Gonzalez-Machorro, M., Hecker, P., Reichel, U.D., Hammer, H.N., Hoepner, R., Pedrotti, L., Zmutt, A., Sagha, H., van Beek, J., Eyben, F., Schuller, D.M., Schuller, B.W., Arnrich, B.: Towards supporting an early diagnosis of multiple sclerosis using vocal features. In: Proceedings of Interspeech. pp. 1518–1522 (2023)

Gosztolya, G., Egas-López, J.V., Svindt, V., Bóna, J., Hoffmann, I.: Sclerosis multiplex felismerése spontán beszédből wav2vec 2.0 modellekből kinyert jellemzőkkel. In: Proceedings of MSZNY. pp. 33–43. Szeged, Hungary (2023)

Gosztolya, G., Tóth, L., Svindt, V., Bóna, J., Hoffmann, I.: Using acoustic Deep Neural Network embeddings to detect Multiple Sclerosis from speech. In: Proceedings of ICASSP. pp. 6927–6931. Singapore (May 2022)

Grosman, J.: Fine-tuned XLSR-53 large model for speech recognition in Hungarian. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-hungarian (2021)

Grósz, T., Porjazovski, D., Getman, Y., Kadiri, S., , Kurimo, M.: Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering. In: Proceedings of ACM Multimedia. pp. 7026–7029. Lisboa, Portugal (Oct 2022)

Grósz, T., Virkkunen, A., Porjazovski, D., Kurimo, M.: Discovering relevant subspaces of BERT, Wav2Vec 2.0, ELECTRA and ViT embeddings for humor and mimicked emotion recognition with integrated gradients. In: Proceedings of MuSe. pp. 27–34. Ottawa, Canada (2023)

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

Ivanova, O., Martínez-Nicolás, I., Meilán, J.J.G.: Speech changes in old age: Methodological considerations for speech-based discrimination of healthy ageing and Alzheimer's disease. International Journal of Language & Communication Disorders 59(1), 13–37 (2023)

Jenei, A.Z., Kiss, G., Sztahó, D.: Detection of speech related disorders by pretrained embedding models extracted biomarkers. In: Proceedings of SPECOM. pp. 279–289. Gurugram, India (2022)

Kiss, G., Tulics, M.G., Sztahó, D., Vicsi, K.: Language independent detection possibilities of depression by speech. In: Proceedings of NoLISP. pp. 103–114 (2016)

Klumpp, P., Arias-Vergara, T., Vásquez-Correa, J.C., Pérez-Toro, P.A., Orozco-Arroyave, J.R., Batliner, A., Nöth, E.: The phonetic footprint of Parkinson's disease. Computer Speech & Language 72(Mar), 101321 (2022)

Kodali, M., Kadiri, S.R., Alku, P.: Classification of vocal intensity category from speech using the wav2vec2 and whisper embeddings. In: Proceedings of Interspeech. pp. 4134–4138 (2023)

Kondratenko, V., Karpov, N., Sokolov, A., Savushkin, N., Kutuzov, O., Minkin, F.: Hybrid dataset for speech emotion recognition in Russian language. In: Proceedings of Interspeech. pp. 4548–4552 (2023)

Kumar, N., Nasir, M., Georgiou, P., Narayanan, S.S.: Robust multichannel gender classification from speech in movie audio. In: Proceedings of Interspeech. pp. 2233–2237. San Francisco, CA, USA (Sep 2016)

Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics 18(1), 50–60 (1947)

Mar, R.A.: The neuropsychology of narrative: Story comprehension, story production and their interrelation. Neuropsychologia 42(10), 1414–1434 (2004)

Mihajlik, P., Balog, A., Gráczi, T.E., Kohári, A., Tarján, B., Mády, K.: BEA-Base: A benchmark for ASR of spontaneous Hungarian. In: Proceedings of LREC. pp. 1970–1977 (2022)

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)

Orsolya, K., Alinka, T., Katalin, J., Péter, K.: A beszédsebesség vizsgálata parkinson-kór-, sclerosis multiplex, valamint stroke-eredetű dysarthriák ese-

tében. Rehabilitáció: A magyar rehabilitációs társaság folyóirata 30(1), 3–10 (2020)

Pepino, L., Riera, P., Ferrer, L.: Emotion recognition from speech using wav2vec 2.0 embeddings. In: Proceedings of Interspeech. pp. 3400–3404. Brno, Czechia (2021)

Pérez-Toro, P., Klumpp, P., Hernandez, A., Arias, T., Lillo, P., Slachevsky, A., García, A., Schuster, M., Maier, A., E.Nöth, Orozco-Arroyave, J.: Alzheimer's detection from English to Spanish using acoustic and linguistic embeddings. In: Proceedings of Interspeech. pp. 2483–2487 (2022)

Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: Unsupervised pre-training for speech recognition. In: Proceedings of Interspeech. pp. 3465–3469 (2019)

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

Thienpondt, J., Speksnijder, C.M., Demuynck, K.: Behavioral analysis of pathological speaker embeddings of patients during oncological treatment of oral cancer. In: Proceedings of Interspeech. pp. 3018–3022 (2023)

Vaessen, N., Van Leeuwen, D.A.: Fine-tuning wav2vec2 for speaker recognition. In: Proceedings of ICASSP. pp. 7967–7971 (2021)