Sclerosis multiplex felismerése spontán beszédből wav2vec 2.0 modellekből kinyert jellemzőkkel

Gosztolya Gábor[1,2], José Vicente Egas-López[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, Pszichiátriai Klinika
ggabor @ inf.u-szeged.hu

Kivonat A sclerosis multiplex (SM) a központi idegrendszer krónikus gyulladásos megbetegedése. Mivel az SM többek között az alanyok beszédét is befolyásolja, az automatikus beszédelemzés egyszerű, relatíve olcsó és találkozásmentes (távoli) módot kínálhat a beszédprodukció változásainak detektálására. Egy ilyen automatikus elemző eljárás fejlesztésének során azonban kritikusnak bizonyulhat, hogy milyen jellemzőket nyerünk ki a beszédproduktumból. Cikkünkben tíz wav2vec 2.0 modell segítségével számítunk jellemzőket, az így kapott osztályozási eredményeket pedig nagymennyiségű adaton tanított publikus, valamint kevesebb, de magyar nyelvű adaton magunk által tanított x-vektor neurális hálók használatával kapott eredményekkel is összevetjük. Kísérleteinkben a többnyelvű fonetikus készletre tanított wav2vec 2.0 modellek hatékonyabbnak bizonyultak, mint az alap („base") modellek. A legfontosabb attribútumnak ugyanakkor a modell paraméterszáma tűnik: a legjobb eredményt az egymilliárd tanítható paraméterrel bíró modell adta. Emellett azt találtuk, hogy a modell finomhangolása a célnyelvre (esetünkben a magyarra) javít az eredményeken, ugyanakkor (legalábbis kísérleti eredményeink alapján) más nyelvre finomhangolni nem érdemes. Meglepő módon nem sikerült viszont túlszárnyalnunk az x-vektorok teljesítményét, mely véleményünk szerint valószínűleg a keretszintű beágyazások bevett, de talán túlságosan egyszerű felvételszintű aggregációjának tudható be.
Kulcsszavak: sclerosis multiplex, wav2vec 2.0, beágyazások

1. Bevezetés

A sclerosis multiplex (SM) a központi idegrendszer krónikus gyulladásos megbetegedése. A motorikus képességek különféle károsodásai az SM egyik központ diagnosztikai jellemzőinek számítanak, míg a motorikus képességek változásai utalhatnak a beteg állapotának súlyosbodására is (Szirmai, 2006). Mivel a nyelvi, kognitív és motorikus funkciók az agyban egy elválaszthatatlan hálózatot

alkotnak, egyik tényező változása változást idézhet elő az összes többiben is. Emiatt az alanyok beszédprodukciójának változásának automatikus ellenőrzése, követése egy hatékony eszközt nyújthat egy SM beteg állapotának vizsgálatához, a betegség előrehaladásának monitorozásához is.

Az SM alanyok mintegy 60-70%-a számol be valamilyen kognitív károsodásról (csökkent információföldolgozási sebességről, krónikus fáradtságról vagy tájékozódási zavarokról). Emellett az SM betegek több mint egyharmadánál jelentkezik átmeneti vagy tartós beszédzavar is (Laakso és mtsai, 2000; Renauld és mtsai, 2016), ami szintén az automatikus beszédelemzés potenciális hasznosságát támasztja alá. A leggyakoribb nyelvi- és beszédtünetek a motoros beszédzavarok (dizartria, diszfónia), szótalálási nehézségek, a verbális fluencia csökkenése (Delgado-Álvarez és mtsai, 2021), mondatismétlési problémák és a magasabb szintű nyelvi folyamatok korlátozottsága (Darley és mtsai, 1972; Hartelius és mtsai, 2000). Az automatikus beszédelemzés használata fényt deríthet olyan tünetekre is, melyek a dizartria szintjét még el nem érő motorikus beszédzavarra utalhatnak (Mulfari és mtsai, 2021).

Az elmúlt évtizedben igen népszerű kutatási területté vált a valamilyen mentális vagy szervi betegséggel élők beszédének automatikus elemzése, számos tanulmány foglalkozott például Alzheimer-kórral (Pérez-Toro és mtsai, 2022), Parkinson-kórral (Moro-Velazquez és mtsai, 2020) vagy éppen depresszióval (Jenei és Kiss, 2020) élők beszédének automatikus földolgozásával. Mivel ezekben a problémákban egyetlen alany felel meg egy (gépi tanulási értelemben vett) példának, az alacsony elemszám miatt jellemzőbb Support Vector Machine (SVM) használata az osztályozási lépésre, míg a mély neurális hálók (Deep Neural Networks, DNN) inkább a jellemzőkinyerés során kapnak szerepet (Pérez-Toro és mtsai, 2022; Jenei és mtsai, 2022). Az ilyen orvosi jellegű korpuszok mérete ugyanakkor még azt sem igazán teszi lehetővé, hogy a jellemzőkinyerésre használt DNN-eket a vizsgált alanyok hangfelvételein tanítsuk, így ez a lépés a legtöbbször általános célú beszédadatbázisokon történik. Az így előálló DNN-ek úgy is tekinthetők, mint a „szokásos beszéd" valamiféle modelljei; a konkrét jellemzőkinyerési lépés pedig annak felelhet meg, hogy a „szokásos beszéd" és az aktuális alany hangfelvétele közötti különbséget számszerűsítjük.

Jelen dolgozatunkban nagyméretű korpuszokon előtanított wav2vec 2.0 neurálisháló-modellekkel végzünk kísérleteket. Ezeket a modelleket sclerosis multiplex és kontroll alanyok beszédfelvételein értékeljük ki, majd a kapott aktivációkból (*embedding*) számított felvételszintű jellemzők alapján, osztályozó eljárással próbáljuk besorolni a beszélőket a két kategória valamelyikébe. Mivel wav2vec 2.0 modellekből máris számos variáció érhető el, tíz különböző modellt tesztelünk: az általános „alap" (*base*) modellek mellett többnyelvi (*cross-lingual*) hálókat (melyek több nyelv fonetikai készletének felismerésére lettek tanítva) és egy-egy konkrét nyelvre finomhangolt modelleket is kipróbálunk. Az elért eredményeket előtanított és saját x-vektor neurális hálók (Snyder és mtsai, 2018) által kinyert jellemzők használatával kapott eredményekkel is összevetjük, melyek szintén népszerű jellemzőkinyerő módszernek számítanak az orvosi beszédfeldolgozás területén (ld. pl. Moro-Velazquez és mtsai, 2020; Jenei és mtsai, 2022).

2. Hangfelvételek

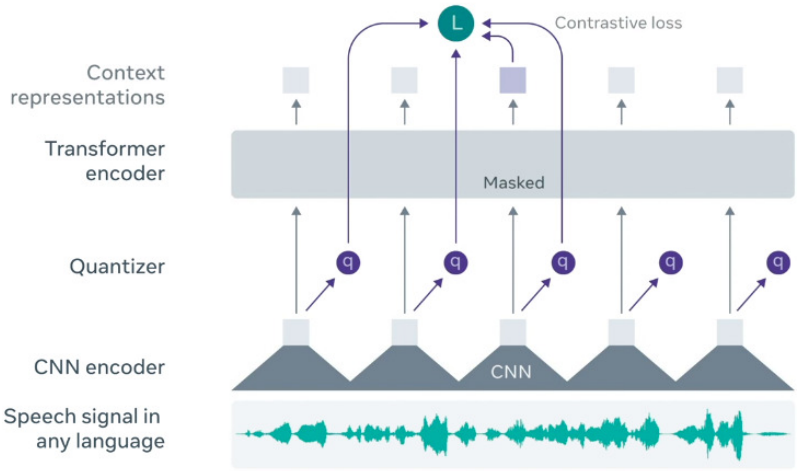
A vizsgálatokra a budapesti Uzsoki Utcai Kórház Neurológiai Osztályán és az Eötvös Loránd Kutatóhálózat Nyelvtudományi Kutatóközpontjában került sor. A vizsgálatot az Uzsoki Utcai Kórház etikai bizottsága hagyta jóvá, és a Helsinki Nyilatkozatnak megfelelően végeztük el. Kísérleteinket 23 SM alany (18 nő és 5 férfi) és 22 kontroll beszélő (16 nő és 6 férfi) felvételein végeztük. Az SM alanyok mindegyike a relapszáló-remittáló (relapsing-remitting, RRMS) altípusba tartozott. A két csoport tagjainak demográfiai jellemzőit az életkor és az iskolázottság esetében ANOVÁ-val, a beszélők nemének eloszlását χ^2-próbával vizsgáltuk; a két csoport tagjai egyik vizsgált jellemzőjükben sem tértek el statisztikailag szignifikáns mértékben (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 (figyelem-összpontosítás, 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 a feldolgozás előtt 16 kHz mintavételezésű, monó formátumra konvertáltuk.

3. Wav2vec 2.0

A beszédfelismerő rendszerek tanításához jellemzően nagymennyiségű címkézett (azaz szöveges átirattal is rendelkező) tanítóadat szükséges (Amodei és mtsai, 2016). Ennek áthidalására nyújt egy hatékony megoldást az, ha a rendelkezésre álló, de címkézetlen adaton előtanítunk egy mély neurális hálót (jellemzően felügyelet nélküli vagy önfelügyelt (self-supervised) módon). Ennek a hálónak a feladata gyakorlatilag egy hatékony reprezentáció megtanulása lesz közvetlenül a beszédjelből. Az így betanított háló súlyai ezután alkalmasak lehetnek arra, hogy egy másik DNN súlyait inicializáljuk velük; ezt a második hálót már a célfeladatra fogjuk tanítani (amelyhez általában korlátozott, de legalábbis lényegesen kisebb mennyiségű tanítóadat áll rendelkezésre, mint amit az első lépésben használhattunk).

A **wav2vec** architektúra egy konvolúciós neurális háló, melynek bemenete a nyers beszédhang-jel, kimenete pedig egy olyan reprezentáció, amely alkalmas egy beszédfelismerő rendszer bemenetének. Tanítása során a hálót az aktuális hangfelvétel következő szegmensének predikciójára tanítjuk (Schneider és mtsai, 2019); vegyük észre, hogy ehhez a fajta tanításhoz nem szükséges semmiféle annotáció (önfelügyelt tanítás). A **wav2vec 2.0** architektúra ezen a megközelítésen azzal javított, hogy a tanítás során – a modell zajtűrésének javítása érdekében – *maszkolást* (a bemenet egyes részeinek kinullázását) is alkalmazott. További különbség a *kontrasztív tanítás* (*contrastive learning*) használata: a tanítás során



1. ábra: Egy előtanított wav2vec 2.0 modell struktúrája. Forrás: `https://ai.facebook.com/blog`

a modell célja annak felismerése is, hogy két különbözőféleképpen transzformált reprezentáció (esetünkben a kontextualizált réteg kimenete, valamint a konvolúciós réteg kimenetének diszkretizált (kvantált) formája) azonos bemenethez tartozik-e.

A wav2vec 2.0 architektúra az 1. ábrán látható. Három részét különböztethetjük meg: a bejövő, nyers audiójelet konvolúciós rétegek (CNN) dolgozzák föl; kimenetüket látens reprezentációnak (*latent representation*) szokás nevezni. Ezt a reprezentációt transzformer rétegek dolgozzák föl, így kapjuk a kontextualizált reprezentációt (*contextualized representation*). Ebből a kimenetet a finomhangolással (*fine-tuning*), azaz a célfeladatra tanítással kapott lineáris projekciós réteggel kapjuk (Baevski és mtsai, 2020). Az önfelügyelt tanítási lépés során kizárólag az alsó két blokkot, tehát a konvolúciós és a kontextualizált rétegeket tanítjuk, míg az utolsó két blokk tanítása jellemzően egy beszédfelismerési feladaton történik (általában egy Connectionist Temporal Classification (CTC) veszteségfüggvénnyel, közvetlenül az átiratra). (Látható tehát, hogy a kontextualizált blokkot mind az előtanítás, mind a finomhangolási lépés során tanítjuk.)

3.1. Többnyelvi reprezentációk

Az XLSR (Cross-lingual Speech Representations) a wav2vec 2.0 többnyelvi továbbfejlesztése (Conneau és mtsai, 2021). Az XLSR megközelítésben az előtanítás többnyelvű korpuszon (vagy több, eltérő nyelvű adatbázison) történik. A különböző nyelvekhez tartozó tanítópéldák az előtanítás során egy-egy batchben is vegyesen (és az adott nyelvhez rendelkezésre álló tanítópéldák gyakoriságát fi-

gyelembe véve) fordulnak elő. Így a konvolúciós blokk utáni kvantálási lépésben előálló beszéd-reprezentációkat több nyelv modellezése során is használja, amely fokozott mértékben jelentkezik rokon nyelvek esetében (Conneau és mtsai, 2021). Az XLSR előtanítás használata kimondottan hatékonynak bizonyult olyan nyelvek esetén, amelyekre még címkézetlen adatból is kevés áll rendelkezésre.

3.2. Jellemzőkinyerés wav2vec 2.0 hálókból

Egy wav2vec 2.0 struktúrájú hálóból beágyazásokat kinyerni két ponton kézenfekvő: egyrészt a konvolúciós blokk utolsó rétegéből (erre *konvolúciós* jellemzőkként fogunk hivatkozni), másrészt pedig a második (a kontextualizált) blokk utolsó rejtett rétegéből (erre a névhasonlóságot elkerülendő nem *kontextualizált*, hanem *rejtett* aktivációként fogunk hivatkozni). Mindkét beágyazásvektorsorozat hordozhat fontos információt mind a beszélőről (Lin és Mak, 2020), mind a beszédjelben jelen lévő további jelenségekről (Fan és mtsai, 2021). Technikai problémát okoz ugyanakkor, hogy a wav2vec 2.0 architektúrát beszédfelismerési célból fejlesztették ki, így a beágyazásokból nyert (keretszintű) jellemzővektorok száma arányos a hangfelvétel hosszával. Hogy felvételszintű osztályozásra alkalmassá tegyük őket, valamiféle aggregációs stratégiát kell alkalmaznunk. Erre a szokásos, bevált megközelítést követtük: az időtengely mentén kiátlagoltuk a vektorokat, illetve az egyes értékek szórását is kiszámítottuk.

4. Kísérleti paraméterek

A wav2vec 2.0 modellek beágyazásait a HuggingFace keretrendszer segítségével nyertük ki. A jellemzőkinyerési lépés után a beszélőket SVM (Schölkopf és mtsai, 2001) alkalmazásával soroltuk a két beszélőcsoport (SM és kontroll) valamelyikébe, a libSVM csomagot (Chang és Lin, 2011) használva. Korábbi tapasztalataink alapján (ld. pl. Egas-López és Gosztolya, 2021; Gosztolya és mtsai, 2022) lineáris kernelt használtunk; az így adódó egyetlen hiperparamétert (C, complexity) a $10^{-5}, 10^{-4}, \ldots, 10^{1}$ értékek közül választottuk ki. Beágyazott keresztvalidációt alkalmaztunk (Cawley és Talbot, 2010); 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. A modellek összehasonlítására is az AUC metrikát használtuk.

5. Eredmények

5.1. Az alap wav2vec 2.0 modellekkel kapott eredmények

Első kísérleteinkben alap wav2vec 2.0 modelleket alkalmaztunk. Az első ilyen modell a `wav2vec2-base` (Baevski és mtsai, 2020) volt, melyet a LibriSpeech

wav2vec 2.0 modell	Beágyazás	Átlag	Átlag + szórás
wav2vec2-base	Konvolúciós	0,745	0,781
	Rejtett	0,731	0,802
wav2vec2-base-960h	Konvolúciós	0,745	0,781
	Rejtett	0,785	0,731
wav2vec2-large-960h	Konvolúciós	0,700	0,795
	Rejtett	0,771	0,763

1. táblázat. Az alap wav2vec 2.0 modellekkel kapott AUC értékek.

korpusz kb. 53 ezer órányi hangfelvételén előtanítottak, a finomhangolási lépés nélkül. A másik két tesztelt modellnek (`wav2vec2-base-960h` és `wav2vec2--base-960h`) mind az előtanítása, mind a finomhangolása 960 órányi (címkézett) adaton történt. Az utóbbi két modell között a fő különbség a tanítható paraméterek száma: a *base* modell 95 millió, míg a *large* modell 317 millió paraméterrel rendelkezik. Az ezen modellek felhasználásával kapott osztályozási eredmények az 1. táblázatban láthatóak.

A kapott AUC értékek versenyképesnek számítanak ugyan, viszont, mivel általában a $0,745 \ldots 0,802$ intervallumba esnek, semmiképpen sem kiemelkedőek. A kipróbált két felvételszintű aggregációs stratégia esetében a szórásértékek használata javította az osztályozási eredményeket (tehát az AUC értékek magasabbak lettek, mint amelyeket kizárólag az átlagvektorokkal kaptunk); ez alól kivételt képeztek a finomhangolt modellek (tehát `base-960h` és `large-960h`) kontextualizált („rejtett") beágyazásai. A két tesztelt beágyazástípus hasznosságát vizsgálva azt találjuk, hogy az „alap" modell (`wav2vec2-base`) esetében kontextualizált reprezentációk használatával kicsit jobb eredményeket kaptunk, mint a konvolúciós blokk kimeneteire támaszkodva. Ugyanakkor a két finomhangolt (960h) modell esetén a konvolúciós rétegek hasonló eredményre vezettek, mint korábban, míg a „rejtett" reprezendációk teljesítménye nagymértékben romlott az „Átlag + szórás" esetben. Ennek oka valószínűleg az, hogy a finomhangolási lépés eredményeképpen ezek a rétegek jobban fókuszálnak a beszéd fonetikai tartalmára, így aktivációik kevésbé voltak alkalmasak az SM detektálására.

5.2. A többnyelvi modellekkel kapott eredmények

Következőleg a többnyelvi modelleket vizsgáltuk meg. Az `XLSR-53` modell 53 különböző nyelvű (kb. 56 ezer órányi) adaton lett előtanítva, míg az XLSR utódja (*XLS-R*) esetén Babu és munkatársai körülbelül félmillió órányi felvételt használtak erre (mely 128 nyelvet ölelt föl) (Babu és mtsai, 2022). Három különböző méretű modell érhető el; technikai okokból eltekintünk a legnagyobb modell használatától (mely kétmilliárd súlyból áll), és a két kisebbre koncentrálunk (`wav2vec2-XLS-R-300M` 300 millió, míg `wav2vec2-XLS-R-1B` egymilliárd tanítható paramétert tartalmaz). A 2. táblázatban találhatóak az ezen modellek használatával kapott AUC-értékek.

wav2vec 2.0 modell	Beágyazás	Átlag	Átlag + szórás
XLSR-53	Konvolúciós	0,700	0,729
	Rejtett	0,763	0,787
XLS-R-300M	Konvolúciós	0,741	0,743
	Rejtett	0,773	0,706
XLS-R-1B	Konvolúciós	0,670	0,696
	Rejtett	0,862	0,872

2. táblázat. A többnyelvű wav2vec 2.0 modellekkel kapott AUC értékek.

Meglepő módon, a konvolúciós rétegek aktivációinak használatával alacsonyabb AUC értékeket értünk el, mint az alap wav2vec 2.0 modellek esetén ($0,670 \ldots 0,743$). Az utolsó rejtett rétegből nyert beágyazások esetén az `XLSR-53` modellel hasonló eredményeket kaptunk, mint a `wav2vec2-base` modell esetén, és valamivel jobbakat, mint a két finomhangolt („960h") alap modellel (főleg amikor a beágyazások szórásait is fölhasználtuk). Ez alapján a többnyelvi wav2vec 2.0 megközelítés hatékonyabbnak tűnik, mint az egynyelvű előtanítás és finomhangolás, legalábbis mikor a célfeladatunk SM alanyok és egészséges kontrollok megkülönböztetése. Az `XLS-R` megközelítés nem bizonyult lényegesen hatékonyabbnak, mint az „eredeti" `XLSR-53`, legalábbis a 300 millió súlyból álló modell esetén. Ezzel szemben az egymilliárd paraméteres, nagyobb modell láthatólag jobb eredményekhez vezetett: a rejtett rétegek aktivációinak átlagát és szórását is fölhasználva $0,872$-es AUC-értéket értünk el.

5.3. Egy nyelvre finomhangolt XLSR modellekkel elért eredmények

A következő lépésben arra voltunk kíváncsiak, hogyan alakulnak az osztályozási eredményeink, ha egy konkrét nyelvre finomhangolt XLSR-53 modellt használunk a jellemzőkinyerési lépés során. Négy nyelvet (és így négy modellt) vizsgáltunk: mivel az SM adatbázisunk nyelve is magyar, első fölhasznált modellünk magyar nyelvű volt. A következő vizsgált modell (a nyelvrokonság miatt) a finn nyelvre finomhangolt XLSR-53 modell volt. A fennmaradó két nyelvet úgy választottuk, hogy azokra (relatíve) nagymennyiségű tanítóadat álljon rendelkezésre, így egy német és egy spanyol nyelvre finomhangolt wav2vec 2.0 hálót teszteltünk még. Mind a négy XLSR-53 modellt ugyanaz a csapat finomhangolta (`jonatasgrosman`) a Mozilla Common Voice 6.1 adatbázis megfelelő részhalmazán, ehhez 8 (magyar), 1 (finn), 777 (német) és 579 (spanyol) órányi adatot használva. Mivel a finomhangolás nem érintette a háló konvolúciós blokkját, csak a kontextualizált („rejtett") rétegből vett beágyazásokból nyert jellemzőkkel kapott osztályozási eredményeket adjuk meg.

A 3. táblázatban olvashatók a nyelvfüggő, finomhangolt wav2vec 2.0 modellekkel kapott eredmények (viszonyításként az eredeti XLSR-53 modellel elért értékeket is feltüntettük). A magyar modell egyértelműen jelentős javuláshoz

wav2vec 2.0 modell	Átlag	Átlag + szórás
XLSR-53	0,763	0,787
XLSR-Hungarian-53 (magyar)	0,816	0,820
XLSR-Finnish-53 (finn)	0,749	0,767
XLSR-German-53 (német)	0,791	0,783
XLSR-Spanish-53 (spanyol)	0,761	0,737

3. táblázat. A finomhangolt többnyelvű wav2vec 2.0 modellek kontextualizált („rejtett") rétegéből nyert beágyazásokkal kapott AUC-értékek.

x-vektor modell	Jellemzők	AUC
SRE-16 (angol)	MFCC	0,876
Saját (magyar)	MFCC	0,850
	FBANK	0,793
	Spektrogram	0,798

4. táblázat. Az x-vektor jellemzőkkel kapott AUC-értékek.

vezetett. Ugyanakkor a fennmaradó három modell esetében nem találtunk jelentős különbséget az eredeti többnyelvi XLSR-53 modellhez képest: a keretszintű beágyazásvektorok átlagával $0,749 \ldots 0,791$-es AUC-értékeket kaptunk (XLSR-53: $0,763$), míg a szórás aggregációt is használva $0,737 \ldots 0,783$-at (XLSR-53: $0,787$). Ez véleményünk szerint azt jelzi, hogy az alkalmazástól eltérő nyelvre finomhangolni az eredeti XLSR-53 wav2vec 2.0 modellt nem igazán kifizetődő, viszont a célnyelvre adaptálása hatékonyabb SM-detektáláshoz vezethet még akkor is, ha a célnyelvi tanítóadat mennyisége kimondottan csekély (esetünkben mindössze 8 órányi). Ez a stratégia mind az „Átlag" aggergáció esetén (ahol az AUC-érték $0,763$-ról $0,816$-ra nőtt), mind az „Átlag + szórás" esetben (ahol pedig $0,787$-ről $0,820$-ra emelkedett) tapasztalható.

5.4. Összevetés az x-vektor jellemzőkkel elért eredményekkel

Utolsó kísérletünkben, hogy az elért osztályozási eredményeket kontextusba helyezzük, összevetjük azokat az x-vektor jellemzőkinyerő eljárással kapottakkal. Ehhez egyrészt Snyder és munkatársai szabadon elérhető modelljét használjuk (`SRE-16 Snyder és mtsai, 2018`), mely a Switchboard (28 ezer felvétel) és a NIST SRE (63 ezer felvétel) korpuszok egy részén lett betanítva, másrészt saját x-vektor hálókat is tanítottunk a BEA Spontánbeszéd-adatbázis (Neuberger és mtsai, 2014) egy részhalmazán (165 beszélő, 60 órányi hangfelvétel). Utóbbi esetben keretszintű jellemzőként MFCC-t, FBANK-ot és spektrogramot is kipróbáltunk. Az SM-detektálási feladaton kapott AUC-értékek a 4. táblázatban találhatóak.

Bár az x-vektor neurális hálók architektúrája lényegesen egyszerűbb, mint egy wav2vec 2.0 hálóé, és sokkal kevesebb (kb. tízmillió) paraméterből is állnak, a kapott AUC-értékek igen magasak. Még a saját tanítású hálók használatával is olyan értékeket értünk el, melyeket csak két wav2vec 2.0 modell tudott túlszárnyalni (`XLS-R-1B` és `XLSR-Hungarian-53`). Ugyanakkor ezek a modellek véletlen súlyinicalizálással („from scratch") lettek tanítva, csupán 60 órányi hanganyag felhasználásával (mely különféle zajok hozzáadásával és visszhangosítással 240 órányira nőtt az FBANK és spektrogram jellemzők esetén). Az SRE-16 előtanított x-vektor modell által számított jellemzők pedig még ezeknél is magasabb AUC-értékhez vezettek (0,876), melyhez hasonlóan magas osztályozási teljesítményt kizárólag az egymilliárd súlyból álló wav2vec 2.0 modell volt képes nyújtani. Ez a kétségkívül meglepő eredmény véleményünk szerint a wav2vec 2.0 hálók esetén használt, bevett, mégis igen szimpla felvételszintű aggregációs technikának tudható be. Bár az x-vektor hálók összegző rétege szintén csupán a keretszintű aktivációs értékek átlagát és szórását számítja ki, az x-vektor struktúrában ezt még két további rejtett réteg követi, és az x-vektor jellemzőket általában az utolsó rejtett réteg aktivációi adják. Ezt a két utolsó, felvételszintű réteget ugyanúgy (sőt, együtt) tanítjuk, mint a mélyebb, keretszintű rétegeket, és így összességében a keretszintű aktivációk szofisztikáltabb aggregálási módját érhetjük el. Ezen felbuzdulva a közeljövőben az igen gyakran használt átlag- és szórásszámításon túlmutató aggregálási technikákat is meg tervezünk vizsgálni.

6. Összegzés

Jelen tanulmányunkban sclerosis multiplex (SM) beszédhangból történő azonosítását vizsgáltuk. Ennek érdekében tíz különböző, transzformer-alapú wav2vec 2.0 modellt használtunk jellemzőkinyerésre; a kapott keretszintű beágyazásvektorokat átlag és szórás segítségével aggregáltuk felvételszintű jellemzőkké. Azt találtuk, hogy az önfelügyelt tanulással előtanított hálókhoz képest a többnyelvi előtanítás nem hoz érdemi előrelépést az SM felismerésében, az igen nagyméretű (`XLS-R-1B`) `XLS-R` modell kivételével. Emellett a többnyelvi hálók finomhangolása sem javított az osztályozás teljesítményén, kivéve amikor a finomhangoláshoz használt tanítóadat nyelve megegyezett az alkalmazás nyelvével (azaz mindkettő magyar volt). Meglepő módon a wav2vec 2.0 modellek használatával elért osztályozási teljesítmény nem múlta felül a standard `SRE-16` előtanított x-vektor modellel kapott osztályozását sem, miközben az x-vektor neurális háló paramétereinek száma a wav2vec 2.0 háló paraméterszámának csupán töredéke. Mivel ennek oka véleményünk szerint a keretszintű beágyazásvektorok túlságosan egyszerű aggregálási módja lehetett, a közeljövőben a wav2vec 2.0 beágyazások kifinomultabb aggregálási módszereire tervezünk fókuszálni.

Köszönetnyilvánítás

A kutatást részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta (NKFIH, K-132460, NKFIH-1279-2/2020, ELTE TKP2020-IKA-06 és

TKP2021-NVA-09). A 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, RRF-2.3.1-21-2022-00004) keretében.

Hivatkozások

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L.J., Fougner, C., Hannun, A.Y., Jun, B., Han, T., LeGresley, P., Li, X., Lin, L., Narang, S., Ng, A., Ozair, S., Prenger, R., Qian, S., Raiman, J., Satheesh, S., Seetapun, D., Sengupta, S., Sriram, A., Wang, C.J., Wang, Y., Wang, Z., Xiao, B., Xie, Y., Yogatama, D., Zhan, J., Zhu, Z.: Deep speech 2: End-to-end speech recognition in English and Mandarin. In: ICML. pp. 173–182 (2016)

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.: XLS-R: Self-supervised cross-lingual speech representation learning at scale. In: Interspeech. pp. 2278–2282 (2022)

Baevski, A., Zhou, H., 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)

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)

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition. In: Interspeech. pp. 2426–2430 (2021)

Darley, F.L., Brown, J.R., Goldstein, N.P.: Dysarthria in multiple sclerosis. Journal of Speech and Hearing Research 15(2), 229–245 (1972)

Delgado-Álvarez, A., Matias-Guiu, J.A., Delgado-Alonso, C., Hernández-Lorenzo, L., Cortés-Martínez, A., Vidorreta, L., Montero-Escribano, P., Pytel, V., Matias-Guiu, J.: Cognitive processes underlying verbal fluency in multiple sclerosis. Frontiers in Neurology 11, 629183 (2021)

Egas-López, J.V., Gosztolya, G.: Using the Fisher Vector approach for cold identification. Acta Cybernetica 25(2), 223–232 (2021)

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

Gosztolya, G., Tóth, L., Svindt, V., Bóna, J., Hoffmann, I.: Sclerosis multiplex hangalapú felismerése akusztikai alapú beágyazások használatával. In: MSZNY. pp. 151–160. Szeged (2022)

Hartelius, L., Runmarker, B., Andersen, O.: Prevalence and characteristics of dysarthria in a multiple-sclerosis incidence cohort: Relation to neurological data. Folia Phoniatrica et Logopaedica 52(4), 160–177 (2000)

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

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

Laakso, K., Brunnegård, K., Hartelius, L., Ahlsén, E.: Assessing high-level language in individuals with multiple sclerosis: A pilot study. Clinical Linguistics & Phonetics 14(5), 329–349 (2000)

Lin, W., Mak, M.W.: Wav2spk: A simple DNN architecture for learning speaker embeddings from waveforms. In: Interspeech. pp. 3211–3215 (2020)

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

Moro-Velazquez, L., Villalba, J., Dehak, N.: Using x-vectors to automatically detect Parkinson's disease from speech. In: ICASSP. pp. 1155–1159 (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)

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

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

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

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: 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)

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