Magyar nyelvű beszédleiratozó tanítása sok (tíz)ezer óra beszéddel

Dobsinszki Gergely[1,2], Kádár Máté Soma[1,2] Fegyó Tibor[1,2], Mády Katalin[3], Mihajlik Péter[2,3]

[1]SpeechTex Kft.
[2]Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Mesterséges Intelligencia Tanszék
[3]HUN-REN Nyelvtudományi Kutatóközpont
{dobsinszki, kadar, fegyo, mihajlik}@tmit.bme.hu, mady@nytud.hu

Kivonat Jelen tanulmány célja a magyar nyelvű beszédfelismerés pontosságának növelése nagy mennyiségű tanítóanyag és önfelügyelt tanulás (Self-Supervised Learning, SSL) módszerének alkalmazásával. Kísérleteink során kisebb, nyilvános és nagyobb, nem publikus adathalmazokon is teszteltük a önfelügyelt előtanítás hatékonyságát. Az SSL technika alkalmazása kis méretű, magyar nyelvű tanító adathalmazok esetén látványos teljesítményjavulást eredményezett a beszédfelismerő modellek pontosságában. A nagy, több ezer órás adathalmazon történő finomhangolás esetén is felgyorsította a tanulási konvergenciát, azonban a jelenlegi eredmények alapján nem múlja felül az angol nyelvű felügyelt előtanítással elért pontosságot. Mindemellett a minden korábbinál nagyobb, tisztán magyar nyelvű halmazzal felügyelten finomhangolt modellek több független kiértékelő halmazon is minden korábbit megelőző pontosságot értek el.
Kulcsszavak: mélytanulás, akusztikus modellezés, ASR, SSL

1. Bevezetés

Korábbi kutatások bizonyítják, hogy a neurális hálók tanításánál a konvergálási sebességet javíthatja, ha nem véletlen súlyokkal kezdünk egy tanítást, hanem előtanított (pretrained) súlyokból indulunk ki Cho és mtsai (2020). Nincs ez másképp az automatikus beszédfelismerés (ASR, Automatic Speech Recognition) területén sem, ahol egy a célnyelvtől eltérő nyelvre tanított háló súlyai is jó kiindulási állapotot jelenthetnek Huang és mtsai (2020).

Az egyre nagyobb méretű neurális modellek egyre nagyobb adatigényének kiszolgálása nem lehetséges drága és lassan előállítható manuális(an ellenőrzött) cimkékkel. Erre a problémára jelenthet megoldást az önfelügyelt tanítás (SSL, Self-Supervised Learning), aminek célja kimeneti címkék használata nélkül olyan kiindulási modell/súlyok előállítása, melyek képesek magas fokú tudásreprezentációra és így egy relatíve sekély ráépülő réteggel együtt történő finomhangolás révén lerövidíti a tanulást és javítja a végeredményeket az adott konkrét (pl. beszéd-szöveg átalakítási) feladatra Yang és mtsai (2022), Lee és mtsai (2024).

Munkánk során az volt a célunk, hogy megvizsgáljuk, magyar nyelvű beszédfelismerésnél milyen mértékben segíthet a célzottan magyar nyelvű hanyaganyagokon végzett önfelügyelt előtanítás. Nagyméretű, közel húszezer órányi magyar nyelvű hanganyagon tanítottunk SSL akusztikus modelleket, majd különféle jellegű és méretű, leiratozott hanganyagokon finomhangoltuk és nyilvános adatokon értékeltük ki őket. Eredményeinket összehasonlítva más, magyar nyelvre elérhető nyílt modellekkel elértekkel, láthatóan általánosabban használható, esetenként lényegesen magasabb beszédfelismerési pontosságot sikerült elérni.

2. Adatok

Első lépésként audio adatok gyűjtésére volt szükség az SSL tanításhoz. Ennél a módszernél a felügyelt tanításhoz képest nagyságrendekkel több adatra van szükség. Az adatok volumenéből adódóan fontos szempont lett a hangfájlok tömörítésének kiválasztása. Az adatok könnyebb kezelése végett tarred [1] gyűjteményeket hoztunk létre az egyes fájlokból. Mivel ezek a gyűjtemények több fájlt foglalnak egyszerre magukba, így kevesebb fájlolvasásra van szükség, optimálisabb tanítási folyamatot biztosítva.

2.1. SSL

Az önfelügyelt tanításhoz tehát nagy mennyiségű magyar nyelvű beszédadatra volt szükség, aminek beszerzése önmagában nem jelentett kihívást, de a kísérletek könnyebb megismételhetőségének és az eredmények jobb hasonlíthatósága érdekében előnyben részesítettük a nyilvánosan elérhető adatbázisokat.

A tanítóadatok nagyját a Voxpopuli V2 Wang és mtsai (2021) halmaz tette ki. Ez a Meta/ Facebook Research által karbantartott halmaz több nemzet különböző nyelvű európai parlamenti felszólalását tartalmazza. Hosszabb (több perces) felvételek is akadnak, de mivel az ASR rendszerek tanításához rövidebb hangszeletekre van szükség, feldaraboltuk a hanganyagokat 20 másodpercnél nem hosszabb egységekre. Többnyire automatikusan detektált, csendes, beszédet nem tartalmazó részeknél igyekeztünk a vágásokat megtenni. Továbbá, a hosszabb szüneteket kivágva csökkentettük a teljes halmaz hosszát. A végső halmaz 17.470 (tizenhétezer-négyszázhetven) órányi magyar nyelvű felvételt tesz ki.

Egy relatíve kisebb, "házon belüli" (In House, IH) halmazt is használtunk az önfelügyelt tanításainkhoz. Az IH halmaz főképpen rádiós beszélgetőműsorokat tartalmaz. Előkészítésénél a Voxpopuli halmazhoz hasonlóan jártunk el, csend részeknél darabolva, vágva, így összesen 3.36 ezer órányi anyagot kaptunk. Bár terjedelemben lényegesen rövidebb a Facebook Research által rendszerezett halmaznál, a beszéd stílusát és jellegét tekintve más területhez tartozik, így jelentős hozzáadott értéket képviselt az SSL tanítási folyamat során, ahol a Voxpopuli + IH = SSL halmazon történt minden tanítás.

[1] Tarred adathalmazok - https://docs.nvidia.com/nemo-framework/user-guide/latest/nemotoolkit/asr/datasets.html#tarred-datasets

2.2. Mozilla Common Voice (CV-16)

Finomhangolás és kiértékelési kísérleteinkhez a szabadon hozzáférhető adatok közül a Mozzila Common Voice magyar nyelvű, 16.1 verziójú részhalmazát használtuk (CV-16) Ardila és mtsai (2019).

Az adott verzió összesen 92 órányi rögzített, ellenőrzött adatmennyiséget tartalmaz, melyből a tanító, validációs és teszt adathalmaz mérete órában kifejezve rendre 52.5, 16.8, 17.7.

2.3. BEA

A további finomhangolási kísérletekhez olyan adatbázis kerestünk, amiben társalgási beszéd található. A BEA (BEszélt nyelvi Adatbázis) Gósy (2013); Neuberger és mtsai (2014) a Nyelvtudományi Kutatóközpont által gyűjtött és karbantartott adathalmaz, mely spontán (monologikus és dialogikus), valamint olvasott beszédet is tartalmaz több egyedi beszélőtől. A BEÁ-n belül a BEA-Base részhalmazt Mihajlik és mtsai (2023) használtuk, ami kifejezetten beszédfelismerési modellek kiértékelése céljából készült. Az adathalmaz kutatási célokra ingyenesen hozzáférhető.

Az általunk tanításra használt (train-114) halmazának hossza 68 óra, validációs halmaza (dev-spont) 3.8, teszt halmaza (eval-spont) pedig 4.75 óra hosszú.

2.4. BNC

Az általunk BNC-nek (Broadcast News and Conversations) nevezett adatgyűjtemény egy privát leiratozott, több ezer órányi, magyar nyelvű televíziós műsorok hangsávját tartalmazó halmaz. Túlnyomó többségben többrésztvevős beszélgetéseket tartalmaznak a felvételek, kisebb részben pedig olvasott híreket vagy beszédeket. Felhasználása előtt a hanganyagok darabolására és normalizálására volt szükség, így biztosítva a konzisztenciát és összehasonlíthatóságot a többi tanító és kiértékelő halmazokkal.

A BNC esetén az általunk létrehozott tanítóadat hossza 2703 óra, validációs halmaza 113 óra és teszt halmaza 41 órányi hanganyagot tartalmaz.

3. Kísérletek

Kísérleteink fókuszában az állt, hogy hogyan tudjuk a nagyméretű beszédadatokat a legjobban felhasználni beszédfelismerő tanításához. A nagy adatbázisokkal való manipuláció természetszerűleg nehézkes, ezért a hiperparamétereket a kisméretű CV-16 és BEA adatokon kíséreltük meg beállítani/optimalizálni. Ez esetekben tudatosan kerültük a kiértékelő halmazok használatát, azokat későbbre, a nagy (BNC) halmazon történő tanítás kiértékelésére tartogattuk.

3.1. Kísérleti környezet

Minden kísérletet az Nvidia NeMo Toolkit 1.23-as verziójával végeztünk Harper és mtsai (2023). A környezet több modellarchitektúrát is támogat, multi-gpu és multi-node tanítást tesz lehetővé. SSL tanítás is konfigurálható a segítségével.

A NeMo Toolkit támogatja a SLURM ütemező rendszert, ami segítségével az SSL tanítások történtek a Komondor szuperszámítógépen.

3.2. Modellek

A neurális architektúra kiválasztásánál arra ügyeltünk, hogy legalább egy angol nyelvű előtanított súlymátrix elérhető legyen. Ezen súlyokat felhasználva tudtunk olyan tanításokat végezni melyekhez az SSL alapú kísérleteinket hasonlítottuk. Önfelügyelt tanításnál először egy általános modellt készítettünk címkézetlen adatok segítségével. Ezen modellből kiindulva pedig beszédfelismerési feladatra finomhangoltuk a modellt különböző annotált adatokat felhasználva.

A kutatásainkhoz egy konkrét akusztikus modell architektúrát kerestünk. Mivel az SSL tanítás egy rendkívül számításigényes folyamat, igyekeztünk relatíve kis paraméterszámú, gyorsan tanítható modellt kiválasztani, szem előtt tartva a megfelelően magas pontosság mellett a lehetséges ipari felhasználási szempontokat is (véges GPU memória, számítási kapacitások, stb.). Választásunk így a NeMo toolkit-ben Harper és mtsai (2023) kényelmesen használható *FastConformer Large (CTC)* [2]-re esett, melynek paraméterszáma kb. 121 millió Rekesh és mtsai (2023). A Conformer architektúra Gulati és mtsai (2020) lényegében egy transformer encoder Vaswani és mtsai (2017), amiben a szokásos attention és előrecsatolt (fully connected) réteg mellett egy 1D konvolúciós réteg is megtalálható blokkonként a lokális kontextus hangsúlyosabb figyelembe vétele érdekében. Ez az architektúra talán a legnépszerűbb az end-to-end mélyneuronhálók beszédfelismerési alkalmazásában. A FastConformer csak annyiban tér el a klasszikus Conformer hálózattól, hogy hagyományos 1D konvolúció helyett ún. "time-channel separable" konvolúciót Kriman és mtsai (2020) alkalmaz, ami voltaképpen egy paraméter- és számításhatékony közelítő változata az utóbbinak.

A FastConformer hálózat kimenete a skálázható szótárméretű szótöredék (subwords unit) valószínűségeloszlása 80ms-onként, erre került a végső CTC Graves és mtsai (2006) réteg/költségfüggvény. A finomhangoláshoz használt hanganyagok szöveges leiratáak tokenizálására a SentencePiece Kudo és Richardson (2018) eljárást alkalmaztuk alapértelmezett 1024-es szótárérettel.

3.3. SSL tanítás

AZ SSL tanításoknál Contrastive Baevski és mtsai (2020) és Masked Language Modeling (MLM) Hsu és mtsai (2021) hibafüggvényekkel kísérleteztünk. A kontrasztív megközelítések esetén a modellek úgy tanulnak, hogy különbséget tesznek

[2] https://github.com/NVIDIA/NeMo/blob/main/examples/asr/conf/fastconformer/fast-conformer_ctc_bpe.yaml

a valódi és a zavaró látens reprezentációk között, míg az MLM esetén a szöveges BERT modellek Devlin és mtsai (2019) tanításához hasonlóan a kimaszkolt audio részletet az ismert (nem maszkolt) kontextusból kell megbecsülnie a modellnek.

Az SSL tanítás során a NeMo adott verziójának alapértelmezett beállításait használtuk, csak azon paramétereket tüntetjük fel a továbbiakban, melyek ettől eltértek.

Az SSL tanításhoz 2 darab node-ot használtunk, összesen 16 darab A100 SXM4 GPU-val, GPU-nként 40 GB VRAM-al. Node-onként 64 CPU core és 256 GB RAM állt rendelkezésre. Egy epoch ideje nagyjából 1 óra volt. Tanításainknál mindenhol az AdamW optimizert Kingma és Ba (2014), Loshchilov és Hutter (2019) alkalmaztuk.

4. Eredmények

A következőkben a 2. fejezetben felsorolt halmazokon végzett SSL módon előtanított súlyokból kiindult tanítások eredményeit közöljük, melyek mellé az NVIDIA által előtanított angol nyelvű akusztikus modellel inicializált "cross-language transfer learning" eredményeket is megadjuk referenciának. A magyar nyelvű SSL modelleket és az angol nyelvű (felügyelten előtanított) modelleket minden esetben valamely magyar nyelvű adatbázison tanítottuk tovább (finomhangoltuk), mely során a validációs halmazt használtuk a tanítás monitorozására. A végső kiértékelést az előzőektől független tesztadatokon végeztük.

4.1. SSL + CV-16

Különböző SSL tanított modellek kipróbálása történt ebben a környezetben. A kezdeti súlyokat leszámítva, minden tanítási paraméter fix, megegyezik a NeMo FastConformer-Large tanítási receptjével. A modellek a CV-16 tanító halmazán tanultak (tovább), 100 epoch-on keresztül, $2 * 10^{-4}$-es LR, 256-os tokenizer mérettel.

Experiment Name	SSL Loss Type	SSL Epoch	Num Negatives	Val WER
en_weights	NA	NA	NA	16.49 %
ssl_weights_1	Contrastive	70	40	15.56 %
ssl_weights_2	Contrastive & MLM	100	40	13.20 %
ssl_weights_3	Contrastive & MLM	100	100	13.43 %

1. táblázat. Angol nyelvű és különböző SSL modellek finomhangolása CV-16-on.

4.2. SSL + BEA

Ebben a kísérletsorozatban már csak egyetlen SSL modellel dolgoztunk. Elsősorban learning rate (LR) hangolására fókuszáltunk. CV-16-os kísérletekkel ellentétben itt nem volt elég a kiindulási SSL modell-re lecserélni a kezdeti angol nyelvű

előtanított súlyokat jobb eredmények eléréséhez: ahogy a 2. táblázat mutatja, a tanulási ráta görbéjének alakítása jelentős hatással volt a validációs halmazon mért hibára.

Initial weights	Initial LR	Minimum LR	Warmup Steps	Val WER
en_weights	5e-4	5e-8	12500	17.11 %
ssl_weights_3	5e-4	5e-6	1000	17.47 %
ssl_weights_3	7e-4	5e-8	1000	18.56 %
ssl_weights_3	4e-4	5e-8	2000	16.46 %
ssl_weights_3	3e-4	5e-8	2000	15.99 %
ssl_weights_3	2e-4	5e-8	2000	15.32 %
ssl_weights_3	1e-4	5e-8	2000	14.81 %
ssl_weights_3	9e-5	5e-8	2000	14.53 %
ssl_weights_3	8e-5	5e-8	2000	14.76 %
en_weights	9e-5	5e-8	2000	26.73 %

2. táblázat. Különböző SSL modellek finomhangolása BEA-n és kiértékelésük BEA dev_spont-on.

Az 2. táblázatban látható ssl_weights_3 az 1. táblázat azonos nevű modell súlyait jelöli. A választás azért erre az előtanított modellre esett, mert stabilabban voltak alacsonyak a validációs WER értékek a CV-16-on végzett finomhangolás során (a táblázatban csak a végső értékek vannak feltüntetve) mint a többi próbálkozásnál.

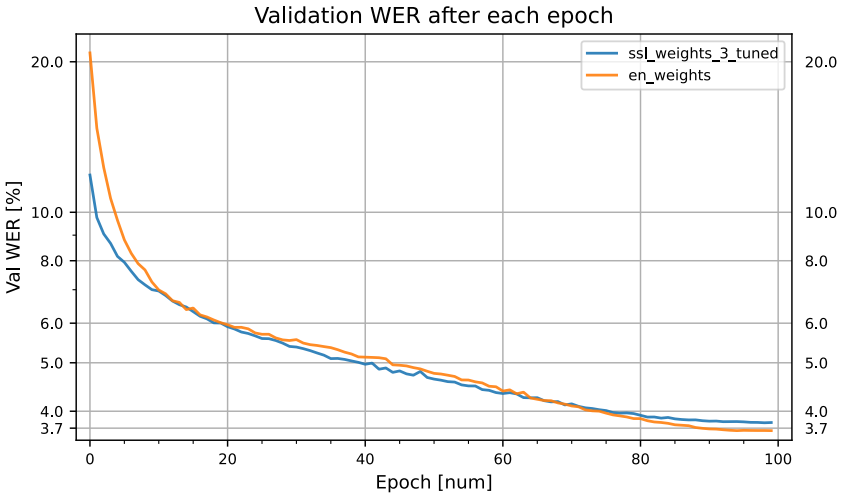
4.3. SSL + BNC

A harmadik, egyben legfontosabb kísérletként nagy mennyiségű leiratozott adattal, a BNC-el tanítottuk tovább a kiválasztott előtanított modellt.

A BNC-n történő finomhangolást a 20.000 órányi magyar nyelvű hanganyagon önfelügyelten előtanult, valamint az angol nyelven az NVIDIA által felügyelten előtanított modelleken is elvégeztük. Noha az előzetes tesztek során – lásd az 1. és 2. táblázatot – az angol súlyokkal történő inicializálást mindig felülmúlta a magyar SSL inicializálás, ezúttal más eredményt kaptunk.

Bár az SSL előtanítás a konvergenciát gyorsította, de végeredményben nem hozott javulást az angol nyelvű előtanuláshoz képest. Az 1. ábrán láthatóak a hiperparaméter-hangolt SSL alapú és angol nyelvű hálóból kiindult tanítások eredményei. Látható, hogy az SSL alapú gyorsabban konvergált, "jobban illeszkedett" a tanítási adathalmazra de végeredményben marginálisan gyengébb eredményt ért el a kiértékelő halmazon is (3. táblázat). Ennek oka a BNC adatokra szabott hiperparaméter-optimalizáció hiánya is lehet.

A BNC adatokkal finomhangolt modelleket összehasonlítottuk más, magyar nyelvű neurális háló alapú leiratozó rendszerekkel. Az egyik ilyen a BEAST2, amely limitált mennyiségű adaton, a BEA-n tanult, a BEA teszt halmazán legjobb eredményt elérve. A másik rendszer a szabadon hozzáférhető OpenAI által



1. ábra. BNC adatokon történő tanítás, angol (en_weights) és SSL módon tanított súlyokból (ssl_wights_3_tuned) kiindulva. Látható, hogy az SSL tanított gyorsabban konvergál a tanítás elején, végül az angol, felügyelten tanított súlyokkal éri el a pontosabb eredményt.

Initial Weights	**Val WER (%)**	**Test WER (%)**
en_weights	3.66	3.83
ssl_weights_3	3.75	3.96

3. táblázat. Az magyar nyelvű adatokkal tanított SSL és az angol nyelven tanított súlyokból indított finomhangolás validációs és teszt eredményei az BNC halmazon.

Model	**LM**	**CV-16 WER (%)**	**FLEURS WER (%)**	**BEA WER (%)**
BEAST2	Van	19.53	25.63	**10.98**
Whisper large-v3	Implicit	13.4	12.9	21.70
FastConformer HU-SSL + BNC (saját)	Nincs	7.43	9.98	13.06
FastConformer EN + BNC (saját)	Nincs	**7.32**	**9.74**	12.70

4. táblázat. A BEAST2 Kádár és mtsai (2023) rendszer, a Whisper large-v3 Radford és mtsai (2022) és a saját megoldásaink eredményei (magyar SSL és angol előtanított hálókból kiindulva), három, magyar nyelvet is tartalmazó, szabadon vagy kutatási célra hozzáférhető kiértékelő halmazon.

fejlesztett Whisper architektúrának 3. iterációja. A Whisper több nyelven képes leiratozni, igény szerint az elhangzott hanganyag alapján nyelvdetektálást is végez, és fordítani is képes. A különböző megoldások pontosságát a CV-16, FLEURS Conneau és mtsai (2023) és BEA halmazokon vizsgáltuk, ezek a 4. táblázatban láthatóak.

Mindkét megoldásunk messze felülmúlja a másik kettő megvalósítást, teszi ezt nyelvmodell (LM: Language Model) használata, azaz előzetes írott nyelvi ismeretek nélkül. Mindemellett a legkisebb paraméterszámú architektúra a három megoldás közül. Egyedül a BEA halmazon teljesít rosszabbul a BEAST2-es modellnél, ez magyarázható többek közt azzal, hogy a tanításhoz nem lett felhasználva a BEA tanítóhalmaz, így "out of domain" ez a teszthalmaz számára. Ennek ellenére messze felülmúlja a Whisper large-v3 által elért eredményeket ezen a halmazon, mely számára szintén ismeretlen a BEA halmaz.

5. Összefoglalás

Az eredményeink nyomán megállapítható, hogy kis méretű adathalmazokkal való továbbtanításkor az SSL előtanítás jelentős javulást tud elérni a az idegen nyelven előtanított súlyokhoz képest. Ugyanakkor az általunk vizsgált modellnél ez az előny elveszik, amikor nagy mennyiségű leiratozott anyaggal történik a továbbtanítás.

Az önfelügyelt tanításhoz 20 ezer óránál is kevesebb anyagot használtunk, ami relatíve kevésnek tekinthető az ilyen jellegű megközelítéseknél. A jövőben szeretnénk ezt a halmazt bővíteni, és egy ilyen modellt finomhangolni felügyelten.

Szeretnénk továbbá nagyobb méretű modellekkel megvizsgálni az önfelügyelt tanítást. Ezek a modellek több VRAM-ot igényelnek, tovább tanulnak, így a jelenlegi munkánk kereteibe nem fértek bele. Viszont kutatások igazolják, hogy az SSL tanítás nagyobb méretű, nagy paraméterszámú modelleknél tudja igazán kifejteni hatását, így szeretnénk a FastConformer család egyel nagyobb modelljét megvizsgálni, a FastConformer XL-t, ami nagyjából 5-ször annyi paramétert tartalmaz, mint az általunk vizsgált architektúra.

Köszönetnyilvánítás

A munka részben az NKFI Alap által támogatott NKFIH K143075 és K135038, valamint az NKFIH-828-2/2021 (MILAB) projkektek támogatásával valósult meg. Köszönettel tartozunk továbbá a KIFÜ-nek (Kormányzati Informatikai Fejlesztési Ügynökség [3]), amiért hozzáférést biztosítottak a Magyarországon található Komondor szuperszámítógéphez, illetve az NVIDIA Akadémiai Hardver Támogatásnak (NVIDIA Academic Hardware Grant).

[3] https://ror.org/01s0v4q65

Irodalomjegyzék

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G.: Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670 (2019)

Baevski, A., Zhou, H., Mohamed, A., Auli, M.: Wav2vec 2.0: A framework for self-supervised learning of speech representations (Oct 2020), https://arxiv.org/abs/2006.11477

Cho, E., Li, J., Kim, S., Jinyu, L.: Cross-language transfer learning and domain adaptation for end-to-end automatic speech recognition. In: Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7494–7498. IEEE (2020)

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., Bapna, A.: Fleurs: Few-shot learning evaluation of universal representations of speech. In: 2022 IEEE Spoken Language Technology Workshop (SLT). pp. 798–805. IEEE (2023)

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019), https://arxiv.org/abs/1810.04805

Gósy, M.: Bea–a multifunctional Hungarian spoken language database. Phonetica 105, 50–61 (2013)

Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning. pp. 369–376 (2006)

Gulati, A., Qin, J., Chiu, C.C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., és mtsai: Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100 (2020)

Harper, E., Majumdar, S., Kuchaiev, O., Li, J., Zhang, Y., Bakhturina, E., Noroozi, V., Subramanian, S., Koluguri, N., Huang, J., Jia, F., Balam, J., Yang, X., Livne, M., Dong, Y., Naren, S., Ginsburg, B.: Nemo: a toolkit for conversational ai and large language models (2023), https://nvidia.github.io/NeMo/

Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units (2021), https://arxiv.org/abs/2106.07447

Huang, J., Kuchaiev, O., O'Neill, P., Lavrukhin, V., Li, J., Flores, A., Kucsko, G., Ginsburg, B.: Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition (2020), https://arxiv.org/abs/2005.04290

Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., Zhang, Y.: Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6124–6128. IEEE (2020)

Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226 (2018)

Kádár, M.S., Dobsinszki, G., Mády, K., Mihajlik, P.: „feeding the beast" – a bea speech transcriber továbbfejlesztése és integrálása neurális nyelvmodellel. In: XIX. Magyar Számítógépes Nyelvészeti Konferencia. pp. 135–143 (2023)

Lee, Y., Willette, J.R., Kim, J., Hwang, S.J.: Visualizing the loss landscape of self-supervised vision transformer (May 2024), https://arxiv.org/abs/2405.18042

Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019), https://arxiv.org/abs/1711.05101

Mihajlik, P., Gráczi, T.E., Kohári, A., Tarján, B., Balog, A., Mády, K.: A bea továbbfejlesztése és alkalmazása kontrasztív gépi beszédfelismerési kísérletekre. In: Általános Nyelvészeti Tanulmányok XXXIV. Akadémiai Kiadó (2023)

Neuberger, T., Gyarmathy, D., Gráczi, T.E., Horváth, V., Gósy, M., Beke, A.: Development of a large spontaneous speech database of agglutinative Hungarian language. In: International Conference on Text, Speech, and Dialogue. pp. 424–431. Springer, Cham (2014)

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (Dec 2022), https://arxiv.org/abs/2212.04356

Rekesh, D., Koluguri, N.R., Kriman, S., Majumdar, S., Noroozi, V., Huang, H., Hrinchuk, O., Puvvada, K., Kumar, A., Balam, J., Ginsburg, B.: Fast conformer with linearly scalable attention for efficient speech recognition (2023), https://arxiv.org/abs/2305.05084

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., Dupoux, E.: VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In: Zong, C., Xia, F., Li, W., Navigli, R. (szerk.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 993–1003. Association for Computational Linguistics, Online (Aug 2021), https://aclanthology.org/2021.acl-long.80

Yang, H., Zhao, J., Haffari, G., Shareghi, E.: Self-supervised rewiring of pretrained speech encoders:towards faster fine-tuning with less labels in speech processing. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (szerk.) Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 1952–1959. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022), https://aclanthology.org/2022.findings-emnlp.141