Proiect component TADARAV

- Raport ştiinţific şi tehnic în extenso 2020 -

Lucian Georgescu, Alexandru Caranica, Cristian Manolache, Gheorghe Pop, Dan Oneaţă,

Horia Cucu, Dragoş Burileanu, Corneliu Burileanu

Program: *PNCDI III - Programul 1 - Dezvoltarea sistemului naţional de CD*

Proiect complex: *Resurse şi tehnologii pentru dezvoltarea interfeţelor om-maşină în limba română (ReTeRom)*

Proiect component: *Tehnologii pentru adnotarea automată a datelor audio şi pentru realizarea interfeţelor de recunoaştere automată a vorbirii (TADARAV)*

Data: *30.11.2020*

Etapa: *3 / 2020*

Activitatea / activităţile:

- *Activitatea 3.9 - Analiza impactului utilizării de RAV complementare pentru generarea de adnotări în contextul îmbunătăţirii sistemelor de RAV*

- *Activitatea 3.10 - Îmbunătăţirea soluţiei de filtrare şi aliniere a transcrierilor aproximative cu semnalul de vorbire*

- *Activitatea 3.11 - Îmbunătăţirea soluţiei pentru generarea de scoruri de încredere pentru RAV*

- *Activitatea 3.12 - Analiza impactului utilizării transcrierilor aproximative în vederea reantrenării sistemelor de RAV*

- *Activitatea 3.13 - Analiza impactului utilizării scorurilor de încredere pentru filtrarea transcrierilor RAV în vederea reantrenării sistemelor RAV*

- *Activitatea 3.14 - Diseminare*

Număr contract: *73PCCDI / 2018*

Acord de colaborare: *30/20.02.2018 ICIA, 4726/01.03.2018 UTCN, 3950/07.03.2018 UPB, 3805/06.03.2018 UAIC*

Autoritatea contractantă: *Unitatea Executivă pentru Finanţarea Învăţământului Superior, a Cercetării, Dezvoltării şi Inovării*

Conducător proiect component: *Universitatea POLITEHNICA din Bucureşti*

Conducător proiect complex: *ICIA*

Responsabil proiect component: *Conf. Horia Cucu*

Responsabil proiect complex: *Prof. Corneliu Burileanu*

Cuprins

1 Rezumatul etapei

A treia etapă a proiectului TADARAV a avut trei obiective principale ce au fost realizate în proporţie de 100%:

1. evaluarea globală, la nivelul întregului proiect TADARAV, a metodei ce presupune utilizarea sistemelor de recunoaştere automată a vorbirii (RAV) complementare pentru generarea automată de adnotări pentru date audio şi, ulterior, utilizarea acestor adnotări pentru reantrenarea unui nou sistem de RAV.
2. îmbunătăţirea manierei de utilizare a transcrierilor aproximative ale materialelor ce conţin vorbire, împreună cu un sistem de RAV iniţial, pentru a produce în mod automat transcrieri precise pentru o parte a semnalului de vorbire;
3. propunerea de noi metode de generare de scoruri de încredere pentru RAV şi utilizarea scorurilor de încredere rezultate pentru a produce în mod automat transcrieri precise pentru o parte a semnalului de vorbire.

Activităţile realizate în etapa 3/2020 au fost următoarele:

- Activitatea 3.9 - Analiza impactului utilizării de RAV complementare pentru generarea de adnotări în contextul îmbunătăţirii sistemelor de RAV. Această activitate a presupus un studiu comparativ asupra performanţelor sistemului de RAV iniţial şi a sistemelor de RAV rezultate în cadrul activităţilor A1.13/2018 şi A2.13/2019. De asemenea, în cadrul acestei activităţi s-a încercat aplicarea metodei proiectate în activităţile anterioare cu un sistem de RAV complet nou, bazat pe platforma ESPnet. Sistemul de RAV rezultat s-a dovedit a fi extrem de lent în transcriere, astfel că nu a putut fi utilizat corespunzător. Cu toate acestea, per ansamblu, sistemul de RAV iniţial (disponibil la începutul proiectului) a putut fi îmbunătăţit folosind metoda proiectată în activităţile A1.13/2018 şi A2.13/2019 cu aproximativ 12% pe vorbire citită, respectiv 16% pe vorbire spontană.
- Activităţile 3.10 şi 3.12 - Îmbunătăţirea soluţiei de filtrare şi aliniere a transcrierilor aproximative cu semnalul de vorbire şi Analiza impactului utilizării transcrierilor aproximative în vederea reantrenării sistemelor de RAV. Activităţile A3.10 şi A3.12 au avut ca scop îmbunătăţirea şi evaluarea metodei de generare de seturi de date de vorbire adnotată folosind materiale audio brute împreună cu transcrieri aproximative. Ideea principală a acestei metode este următoarea: un sistem de RAV iniţial este folosit pentru a genera transcrieri pentru materialul audio brut, iar ulterior aceste transcrieri sunt aliniate cu textele aproximative deja existente. Părţile aliniate sunt considerate corecte şi sunt folosite pentru reantrenarea sistemului iniţial de RAV. În urma realizării activităţii A2.11 din etapa anterioară am tras concluzia că între două secvenţe de text aliniate corect există transcrieri aproximative care sunt în mare proporţie corecte. Îmbunătăţirea propusă şi evaluată în cadrul activităţilor din anul 2020 constă în utilizarea acestor transcrieri aproximative aflate între două secvenţe de text aliniate corect. Rezultatele experimentale au arătat că îmbunătăţirea RAV rezultată este nesemnificativă (mai mică de 1%).
- Activităţile 3.11 şi 3.13 - Îmbunătăţirea soluţiei pentru generarea de scoruri de încredere pentru RAV şi Analiza impactului utilizării scorurilor de încredere pentru filtrarea transcrierilor RAV în vederea reantrenării sistemelor RAV. În cadrul acestor activităţi au fost propuse o serie de noi metode de generare de scoruri de încredere pentru RAV folosind o platformă modernă de tip end-to-end: ESPnet. Metodele propuse au fost evaluate şi comparate, iar metoda cea mai performantă a fost selectată pentru utilizarea ulterioară în vederea producerii de transcrieri precise pentru o parte a semnalului de vorbire. Sistemul de RAV ESPnet s-a dovedit însă a fi prea lent pentru a putea fi utilizat într-o astfel de metodă de adnotare automată de date: cu un factor de timp real de aproximativ 3, transcrierea celor 913 de ore de vorbire brută ar fi durat aproximativ 100 de zile. În consecinţă, noile metode de estimare a scorurilor de încredere nu au putut fi evaluate în contextul reantrenării sistemului de RAV iniţial, folosind datele adnotate rezultate.

În urma activităţilor A3.9 - A3.13 din etapa 3/2020 a proiectului TADARAV, au rezultat toate livrabilele asumate de consorţiu la începutul acestei etape:

- Raport de analiză a impactului utilizării de RAV complementare pentru generarea de adnotări în contextul îmbunătăţirii sistemelor de RAV;

- Soluţie îmbunătăţită de filtrare şi aliniere a transcrierilor aproximative cu semnalul de vorbire (TRL4);
- Soluţie îmbunătăţită pentru generarea de scoruri de încredere pentru RAV (TRL4);
- Raport de analiză a impactului utilizării transcrierilor aproximative în vederea reantrenării sistemelor de RAV;
- Raport de analiză a impactului utilizării scorurilor de încredere pentru filtrarea transcrierilor RAV în vederea reantrenării sistemelor RAV;
- Sistem de RAV actualizat (TRL5).

Diseminarea rezultatelor proiectului (activitatea A3.14) a fost realizată prin intermediul website-ului proiectului (https://tadarav.speed.pub.ro) şi prin publicarea mai multor articole ştiinţifice după cum urmează:

- A.-L. Georgescu, H. Cucu, A. Buzo, C. Burileanu, "*RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition,*" in the Proceedings of The 12th Language Resources and Evaluation Conference (LREC), pp. 6606-6612, 2020, Marseille, France.
- C. Manolache, A.-L. Georgescu, A. Caranica, H. Cucu, "*Automatic Annotation of Speech Corpora using Approximate Transcripts,*" in the Proceedings of the 43rd International Conference on Telecommunications and Signal Processing (TSP), 2020, Milano, Italy.
- D. Oneaţă, A.-L. Georgescu, H. Cucu, D. Burileanu, C. Burileanu, "*Revisiting SincNet: An Evaluation of Feature and Network Hyperparameters for Speaker Recognition,*" in the Proceedings of the 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 2020.
- G. Pop, H. Cucu, D. Burileanu, C. Burileanu, "*Cough Sound Recognition in Respiratory Disease Epidemics,*" in Romanian Journal of Information Science and Technology, vol. 23, no. S, pp. S77–S89, 2020, ISSN 1453-8245, ISI IF 0.661.
- A.-L. Georgescu,, C. Manolache, D. Oneaţă, H. Cucu, C. Burileanu, "*Data-filtering methods for self-training of automatic speech recognition systems,*" in the Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Virtual, 2021.
- D. Oneaţă, A. Caranica, A. Stan, H. Cucu, "*An evaluation of word-level confidence estimation for end-to-end automatic speech recognition,*" in the Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Virtual, 2021.

Dintre articolele listate mai sus, al patrulea este deja indexat în Web of Science (Thompson Reuters - ISI), al doilea este deja indexat IEEE Xplore şi în curs de indexare în Web of Science, primul şi al treilea au apărut în volumul conferinţelor respective şi sunt în curs de indexare în Web of Science, iar al cincilea şi al şaselea vor apărea în volumul conferinţei respective şi vor fi indexare în Web of Science.

2 Descrierea ştiinţifică şi tehnică a activităţilor

2.1 Seturi de date

2.1.1 Seturi de date de vorbire adnotată

Pentru antrenarea şi evaluarea sistemelor de RAV, au fost folosite două seturi de date de vorbire în limba română: *Read Speech Corpus* (RSC), ce conţine vorbire citită, colectată în condiţii de laborator, fără zgomot de fundal şi *Spontaneous Speech Corpus* (SSC), ce conţine vorbire continuă, spontană, preluată de la posturi de radio şi TV, uneori afectată de zgomot. Ambele corpusuri cuprind fişiere audio şi transcrieri corespunzătoare şi sunt divizate în seturi de antrenare şi seturi de evaluare. RSC-train este setul de antrenare din RSC, ce conţine 100 ore de vorbire citită, cuvinte izolate sau fraze de la 157 de vorbitori diferiţi. RSC-eval este setul de evaluare din RSC; acesta conţine vorbire de la 22 de vorbitori diferiţi, însumând 5.5 ore de vorbire. SSC-train1+2 este setul de antrenare din SSC şi conţine 130 ore de vorbire spontană, majoritatea din emisiuni de ştiri şi *talkshow*-uri. SSC-eval1 este setul de evaluare din SSC şi însumează 3.5 ore de vorbire. Informaţii detaliate despre seturile de date RSC [Georgescu, 2020] şi SSC se regăsesc în raportul TADARAV 2019 [Georgescu, 2019a], secţiunea 2.1.1.

Seturile de date de evaluare RSC-eval şi SSC-eval au făcut în acest an obiectul unei analize amănunţite în vederea corectării eventualelor erori de adnotare. Această analiză a relevat faptul că o serie întreagă de pronunţii din setul RSC-eval nu corespund transcrierii de referinţă. Practic, vorbitorii care au înregistrat acele propoziţii au rostit cuvinte în plus sau în minus, au pronunţat greşit anumite cuvinte etc. Câteva dintre erorile identificate şi corectate sunt listate în Tabelul 2.1.a. În setul de date SSC-eval, set ce a fost obţinut prin transcrierea unor materiale audio deja existente, erorile identificate şi corectate constau în mare măsură în omisiuni de semne diacritice, inversiuni de litere sau cuvinte scrise greşit în transcrierea de referinţă. Tabelul 2.1.a prezintă şi o serie de erori identificate şi corectate în setul de date SSC-eval.

Tabelul 2.1.a *Exemple de erori din seturile de evaluare RSC-eval şi SSC-eval identificate şi corectate*

Set/ ID audio	**Eroare şi corectură**
RSC-eval 018_12_0066	ERR: [...] meciul universitatea CLUJ cfr cluj nu trebuia oprit COR: [...] meciul universitatea CRAIOVA cfr cluj nu trebuia oprit
RSC-eval 018_12_0069	ERR: [...] cu scorul de unu LA zero [...] COR: [...] cu scorul de unu ** zero [...]
RSC-eval 003_01_0573	ERR: [...] mea culpa am fost UN dobitoc [...] COR: [...] mea culpa am fost ** dobitoc [...]
RSC-eval 003_01_0575	ERR: [...] nu mai e pe MOMENT disponibil [...] COR: [...] nu mai e pe MOMENTUL disponibil [...]
SSC-eval	ERR: jandarm aflat în POSTU de pază COR: jandarm aflat în POSTUL de pază
SSC-eval	ERR: [...] era ANAGAJAT la jandarmeria sibiu din două mii cinci [...] COR: [...] era ANGAJAT la jandarmeria sibiu din două mii cinci [...]
SSC-eval	ERR: în cele din URMA şerban ionescu a infirmat această ştire COR: în cele din URMă şerban ionescu a infirmat această ştire
SSC-eval	ERR: [...] răspunde tuturor înTREBARILOR puse în ultimele zile COR: [...] răspunde tuturor înTREBăRILOR puse în ultimele zile
SSC-eval	ERR: pentru a o ajuta pe soţia mea să înCETZE teatrul zilnic făţărnicia COR: pentru a o ajuta pe soţia mea să înCETEZE teatrul zilnic făţărnicia
SSC-eval	ERR: POMPIERII au încercat să ajungă la el cu o barcă pneumatică COR: POMPIERI au încercat să ajungă la el cu o barcă pneumatică

După corectare, seturile de date RSC-eval şi SSC-eval (versiunile v2) au fost utilizate pentru reevaluarea celui mai performant sistem de RAV pe limba română, disponibil la finalul anului 2019. Redăm în Tabelul 2.1.b rezultatele, măsurate în termeni de eroare la nivel de cuvânt (WER), acestui sistem de RAV pe versiunile v1, respectiv v2 ale seturilor de evaluare. În acest fel, dorim să punem în evidenţă diferenţa *artificială* de performanţă rezultată nu din îmbunătăţirea sistemului de RAV, ci din corectarea seturilor de date de evaluare. De această diferenţă artificială de performanţă ar trebui să ţinem cont ori de câte ori vom compara rezultatele de RAV raportate de SpeeD pe seturile de date RSC-eval şi SSC-eval înainte vs. după această corectură (înainte de anul 2020, respectiv după şi inclusiv în anul 2020).

Tabelul 2.1.b *Diferenţa artificială de performanţă rezultată în urma corectării seturilor de date RSC-eval şi SSC-eval pentru cel mai performant sistem de RAV SpeeD la finalul anului 2019.*

WER [%]	**RSC-eval**		**SSC-eval**	
	v1	**v2**	**v1**	**v2**
Best RAV SpeeD 2019	2.41	1.83	12.45	11.04

În etapa anterioară a proiectului, ca parte a activităţii A2.11, au fost obţinute seturile de date de vorbire adnotată SSC-train3-trans-v4 şi SSC-train4-trans-v4 (vezi raportul TADARAV 2019 [Georgescu, 2019a], secţiunea 2.2). Seturile de date au fost adnotate folosind metoda corelaţiei cu transcrierile aproximative. Şi celelalte metode de adnotare automată au fost utilizate pentru generare de adnotări, pornind de la seturile de date neadnotate SSC-train3-raw şi SSC-train4-raw, însă s-a dovedit că aceste seturi de date de vorbire adnotată (SSC-train3-trans-v4 şi SSC-train4-trans-v4) sunt cele mai utile în reantrenarea sistemelor de RAV. În consecinţă, aceste seturi de date au fost înglobate în setul de date *Spontaneous Speech Corpus* (SSC) şi au fost folosite în continuare în etapa 2020 pentru antrenarea sistemelor de RAV. Diversele componente ale seturilor de date SSC şi RSC sunt prezentate în Tabelul 2.1.c.

În etapa curentă a proiectului ReTeRom, suplimentar faţă de activităţile din cadrul proiectului, grupul SpeeD a mărit numărul de seturi de date de evaluare după cum urmează. Din setul de date de vorbire brut SSC-train3+4-raw au fost selectate şi transcrise 100 de fişiere de aproximativ un minut fiecare. A rezultat astfel setul de date SSC-eval2 ce cuprinde aproximativ 91 de minute de vorbire provenind de la un post de radio şi două posturi de televiziune. Mai multe informaţii despre achiziţia setului de date brut SSC-train3+4-raw se găsesc în raportul TADARAV 2019 [Georgescu, 2019a], secţiunea 2.1.1. Din setul de date CDep-raw, ce conţine discursuri susţinute în Parlamentul României, au fost selectate şi transcrise 300 de fişiere de aproximativ un minut fiecare. A rezultat astfel setul de date CDep-eval ce cuprinde aproximativ 5 ore de vorbire. Setul de date de vorbire CDep-raw este descris succint în secţiunea următoare.

Suplimentar, în etapa curentă a proiectului a fost utilizat pentru antrenarea sistemelor de RAV şi setul de date de vorbire din corpusul CoRoLa [Barbu, 2018]. Textele citite din CoRoLa sunt în principal înregistrări profesionale din diverse surse (posturi de radio, studiouri de înregistrare) împreună cu transcrierile lor. O altă parte a corpusului oral este reprezentată de texte citite de la posturi de ştiri (radio) sau texte citite de vorbitori profesionişti, înregistrate în studiouri, respectiv texte extrase din Wikipedia şi citite de voluntari neprofesionişti, înregistrate în medii neprofesionale. O parte semnificativă din transcrierile pentru acest set de date de vorbire conţineau adnotări inconsistente de nume de vorbitori, pronunţii incomplete sau informale etc. Pentru a putea utiliza o parte cât mai mare din acest set de date de vorbire la antrenarea sistemelor de RAV, au fost realizate corecturi ortografice în mod semiautomat.

2.1.2 Seturi de date de vorbire brute

În etapa curentă metodele de adnotare automată au fost aplicate pe două seturi de date brute: CDep-raw şi CoBiLiRo-raw (Tabelul 2.1.d)

Setul de date brut Camera Deputaţilor (CDep-raw), a fost achiziţionat în cursul anului trecut şi cuprinde înregistrările video şi stenogramele şedinţelor de pe site-ul Camerei Deputaţilor extrase între ianuarie 2003 şi februarie 2019. Setul de date conţine 3.510 ore de vorbire de la ~2.500 de vorbitori, însoţite de transcrieri aproximative însumând ~25M de cuvinte. Setul de date CDep-raw a fost achiziţionat cu ajutorul mai unor aplicaţii Java ce efectuează mai multe procese: (i) extragerea fişierelor video şi HTML pentru fiecare şedinţă;

Tabelul 2.1.c *Seturile de vorbire adnotată folosite pentru antrenarea şi evaluarea sistemelor de RAV şi seturile de vorbire adnotată obţinute în etapa anterioară (2/2019)*

Setul de date	Subset	Durată
Spontaneous Speech Corpus (SSC)	SSC-train1+2	130h, 44m
	SSC-train3-trans-v4	41h, 00m
	SSC-train4-trans-v4	250h, 10m
	SSC-eval1	3h, 29m
	SSC-eval2	1h, 31m
Read Speech Corpus (RSC)	RSC-train	94h, 46m
	RSC-eval	5h, 29m
Contemporary Romanian Language (CoRoLa)	n/a	85h, 11m
Camera Deputaţilor (CDep)	CDep-eval	5h, 00m

(ii) conversia fişierelor video în fişiere wav; (iiia) extragerea transcrierilor fiecărui vorbitor în fişiere HTML separate; (iiib) tăierea fişierelor wav în segmente de fişiere audio corespunzătoare transcrierilor vorbitorilor; (iiic) actualizarea unui fişier text ce conţine lista vorbitorilor; (iv) extragerea transcrierilor în format text din fişierele HTML. Detalii suplimentare despre achiziţia setului de date CDep-raw se găsesc în [Manolache, 2019], capitolul 4.

Setul de date brut CoBiLiRo-raw conţine 76 de fişiere audio împreună cu transcrierile aproximative corespunzătoare, extrase din emisiuni şi interviuri. Acest set de date este împărţit la rândul său în 4 subseturi ce însumează aproximativ 714k de cuvinte pronunţate într-o durată de aproximativ 70 de ore. Numărul de cuvinte şi de ore pentru fiecare subset se regăsesc în Tabelul 2.1.d.

2.1.3 Seturi de date de vorbire adnotată rezultate în această etapă a proiectului

După aplicarea celor trei metode de adnotare automată au fost obţinute seturile de date din Tabelul 2.1.e.

Tabelul 2.1.d *Seturi de date de vorbire neadnotată (+ transcrieri aproximative) utilizate ca date de intrare pentru cele trei metode de adnotare automată.*

Setul de date	Sursa	Durată	Transcrieri aproximative	Număr de vorbitori
CDep-raw	Parlamentul României	3,510h, 13m	25.1M cuvinte	~2,500
CoBiLiRo-raw	Alma[1]	19h, 29m	127.5k cuvinte	n/a
	VBarbu[2]	7h, 30m	80.8k cuvinte	
	GCVC[3]	16h, 49m	324.5k cuvinte	
	Ro100[4]	25h, 35m	181.8k cuvinte	

[1] Seria de emisiuni Alma Mater Iassiensis
[2] Interviu cu prof. Viorel Barbu
[3] Seria de emisiuni Ghici cine mai vine la cină
[4] Seria de emisiuni România 100: Iaşul în arcul timpului

Tabelul 2.1.e *Seturile de vorbire adnotată rezultate în urma aplicării metodelor de adnotare automată*

Setul de date	Sursa	Durată		Eficienţă aliniere [% ore]	
CDep-trans-v4	Parlamentul României	878h, 48m		25.0%	
CoBiLiRo-trans-v4	Alma	10h, 24m	31h, 30m	53.3%	45.4%
	VBarbu	1h, 24m		18.7%	
	GCVC	6h, 36m		39.3%	
	Ro100	13h, 6m		51.2%	

2.2 Activitatea 3.9 - Analiza impactului utilizării de RAV complementare pentru generarea de adnotări în contextul îmbunătăţirii sistemelor de RAV

Activitatea 3.9/2020 a avut două subactivităţi principale. În primul rând ne-am propus dezvoltarea suplimentară a metodei de generare automată de adnotări folosind sisteme RAV complementare. Astfel am încercat adaptarea şi aplicarea metodei proiectate în activităţile anterioare folosind un sistem de RAV complet nou, bazat pe platforma ESPnet, ca fiind unul dintre cele două sisteme RAV complementare. Secţiunea 2.2.1 prezintă paşii urmaţi pentru dezvoltarea acestui sistem.

În al doilea rând am efectuat un studiu comparativ asupra performanţelor sistemului de RAV iniţial (la începutul proiectului ReTeRom) şi a sistemelor de RAV rezultate în cadrul activităţilor A1.13/2018 şi A2.13/2019, ca urmare a proiectării şi aplicării metodei de generare automată de adnotări folosind sisteme RAV complementare. Metoda se bazează pe folosirea a două sisteme RAV cât mai diferite, ce produc transcrieri pentru un corpus neadnotat, considerând ulterior părţile identice din transcriere ca fiind corecte. Această presupunere este certificată prin verificarea complementarităţii sistemelor; cele două sisteme diferă din punct de vedere constructiv astfel încât ele produc erori diferite şi necorelate. În final, setul de date adnotat este utilizat pentru reantrenarea celui mai performant sistem iniţial.

2.2.1 Dezvoltarea sistemului de transcriere de vorbire ESPnet pentru limba română

Pentru dezvoltarea noului sistem de RAV pentru limba română, am optat să folosim ESPnet [Watanabe, 2018], un utilitar de învăţare profundă modern ce permite realizarea unui sistem de RAV complet sub forma unei reţele neurale unice (arhitectură de tip end-to-end). Biblioteca de învăţare profundă ESPnet poate fi folosită pentru o gamă întreagă de aplicaţii de inteligenţă artificială pornind de la recunoaşterea automată a vorbirii şi sinteză de vorbire pornind de la text şi mergând, ceva mai recent, şi spre traducerea automată vorbirii (en: speech translation) între multiple limbi de circulaţie internaţională, dar şi conversia vorbirii (en: voice conversion).

Utilitarul include exemple de reţete (directoare de proiect) pentru limba engleză, de la care se poate porni dezvoltarea şi cercetarea în domeniu, cele mai multe fiind bazate pe resurse de antrenare audio şi text disponibile în licenţe de tip sursă deschisă (*e.g.* LibriSpeech [Panayotov, 2015], CommonVoice [Ardila, 2020]), astfel încât orice cercetător să poată uşor replica rezultatele unor articole state-of-the-art.

ESPnet foloseşte, la rândul ei, bibliotecile Chainer [Tokui, 2019] şi PyTorch [Paszke, 2019] pentru implementarea funcţionalităţilor legate de învăţarea profundă (en: deep learning), lucru care simplifică extinderea codului şi arhitecturii de procesare. Ambele biblioteci sunt folosite de cercetătorii din domeniul învăţării automate, pentru aplicaţii în procesarea imaginilor şi a limbajului natural.

De asemenea, un alt mare avantaj al ESPnet constă în procesarea datelor audio de intrare în stil Kaldi (cel mai popular utilitar de învăţare profundă pentru RAV [Povey, 2011]) astfel încât extragerea parametrilor vocali şi formatul fişierelor caracteristicilor extrase se păstrează, iar portarea directoarele ce conţin resurse audio şi text între cele două sisteme fiind astfel extrem de facilă.

Arhitectura sistemului RAV propus

Sistemul RAV propus este bazat pe arhitectura end-to-end din ESPnet, folosind atât clasificarea temporală conexionistă (CTC) cât şi reţeaua codor-decodor bazată pe Transformer (en: self-attention encoder-decoder) [Kim, 2017; Watanabe, 2017]. Această a doua metodă utilizează un mecanism de auto-atenţie, pentru a

efectua alinierea între cadrele acustice şi simbolurile recunoscute, în timp ce CTC foloseşte ipotezele Markov pentru a rezolva eficient problemele secvenţiale prin intermediul programării dinamice.

Astfel, am adoptat pentru RAV o reţea hibridă de tip CTC/Transformer end-to-end (Figura 2.2.a), care utilizează în mod eficient avantajele ambelor arhitecturi în antrenare şi decodare. În timpul antrenării, folosim framework-ul de învăţare multi-obiectiv pentru a îmbunătăţi robusteţea alinierilor problematice cât şi pentru a obţine o convergenţă mai rapidă.

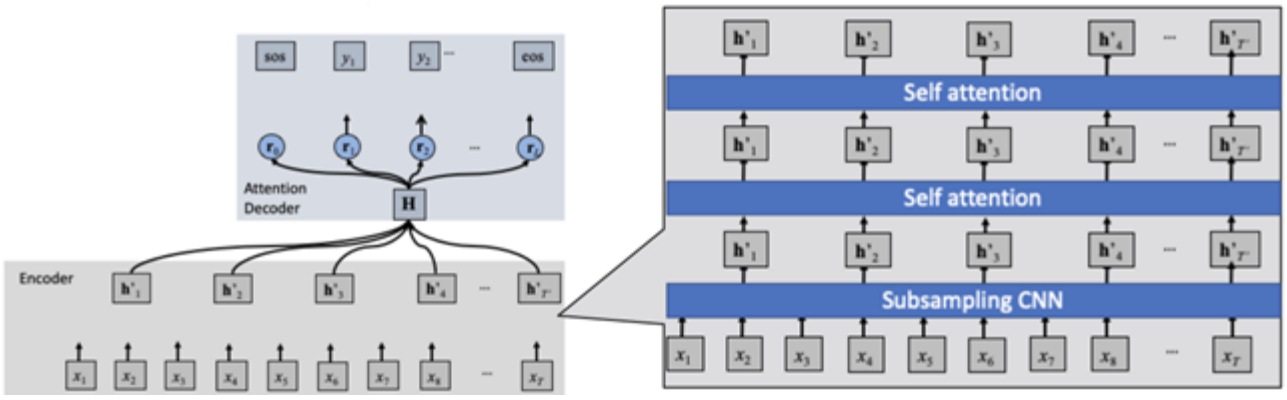

Figura 2.2.a. *Arhitectura reţelei CTC-Transformer utilizată, unde toate conexiunile recurente din codor-decodor-ul bazat pe atenţie sunt înlocuite cu un bloc de auto-atenţie (poate capta interdependenţe pe distanţe foarte lungi). [Hori, 2019]*

În timpul decodării, efectuăm inferenţa atât prin combinarea scorurilor bazate pe atenţie, cât şi a scorurilor CTC (Figura 2.2.b), într-un algoritm de căutare a fasciculului (en: beam-search) cu o singură trecere, pentru a elimina în continuare alinierile problematice. În plus faţă de arhitectura end-to-end prezentată mai sus realizată doar pe fişierele audio, am realizat o serie de experimente şi cu modele de limbă, de tip Transformer.

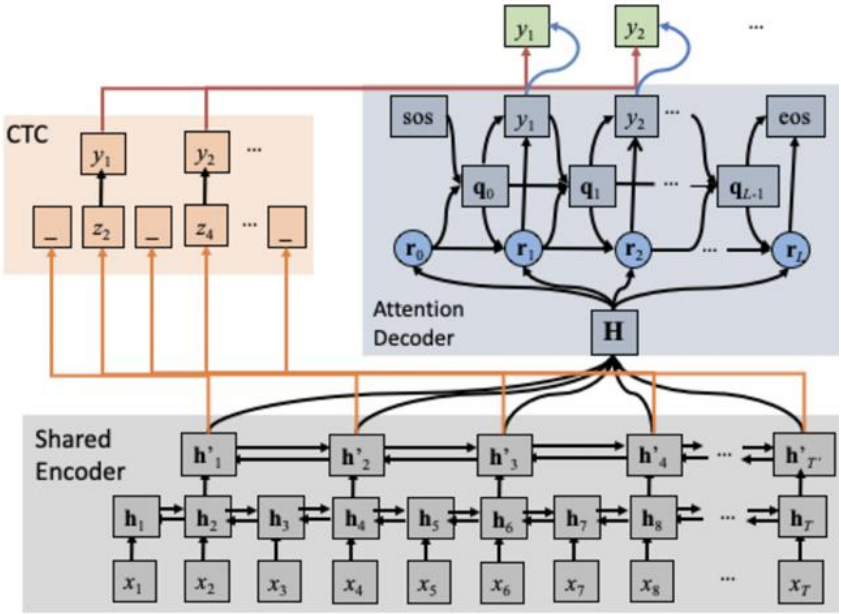

Figura 2.2.b. *Arhitectura reţelei CTC/attention, antrenarea fiind bazată pe învăţarea multi-obiectiv iar inferenţa bazată pe combinarea scorurilor.*

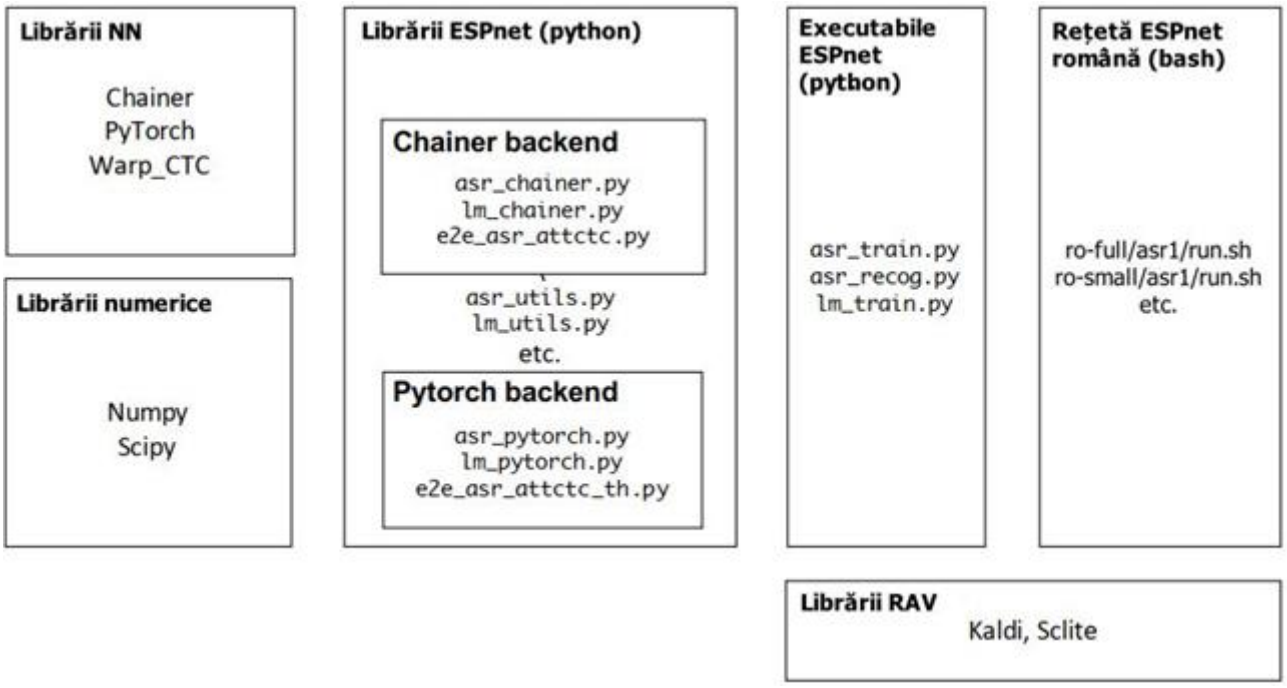

Figura 2.2.c. *Arhitectura software a sistemului RAV în limba română.*

Arhitectura software a sistemului ESPnet

Figura 2.2.c prezintă arhitectura software a sistemului de recunoaştere în limba română, construit în ESPnet. Componentele principale pentru antrenarea şi inferenţa unei reţele neuronale sunt scrise în Python, care apelează SDK-urile Chainer şi PyTorch, prin comutarea backend-ului în funcţie de opţiuni. Structura de directoare în care se preprocesează datele, după cum am menţionat anterior, urmează filozofia Kaldi, utilizând scripturi bash pentru lansarea diferitelor binare necesare preprocesării datelor şi a augmentării lor.

În continuare, Figura 2.2.d prezintă fluxul unei reţete (en: recipe) şi stagiile necesare executării cu succes a unei antrenări, apoi a unei decodări, pentru sistemul în limba română.

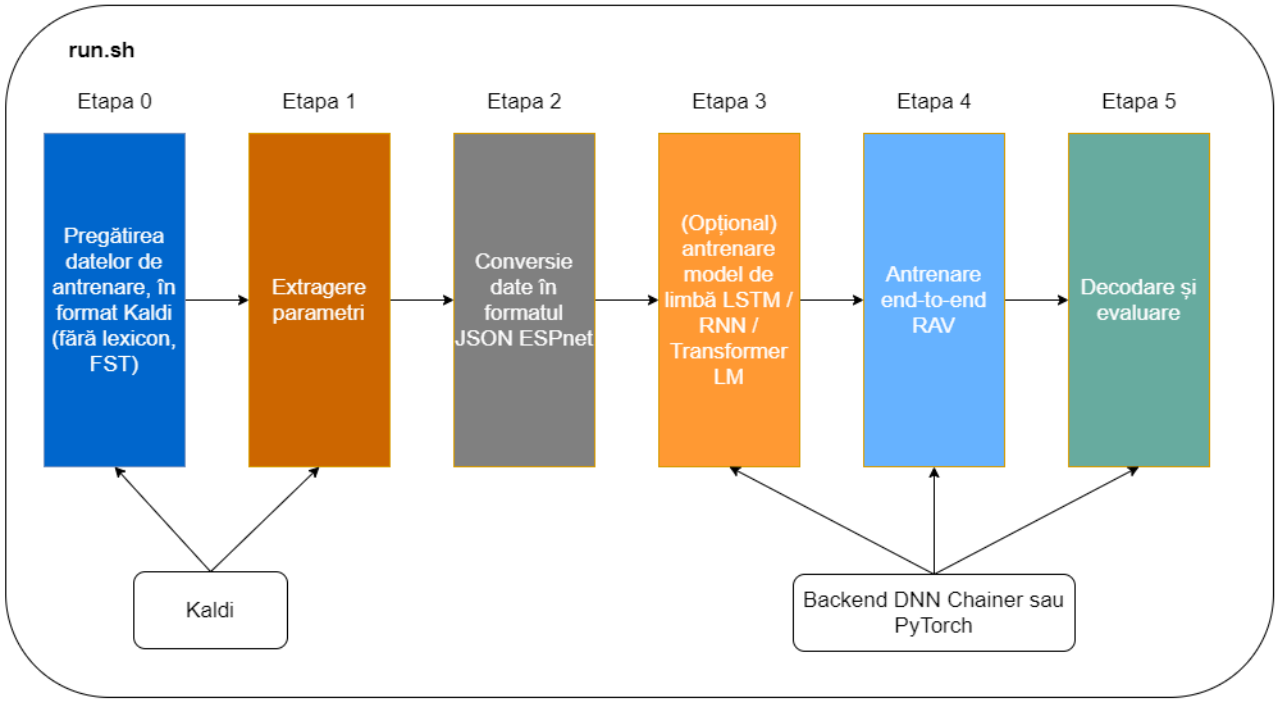

Figura 2.2.d. *Fluxul de execuţie al reţetei ESPnet în Limba Română.*

Reţeta este semnificativ simplificată graţie avantajului unui toolkit de RAV end-to-end, astfel, spre diferenţă de Kaldi, nu trebuie să includem un lexicon, compilarea transductorului de stare finită (FST), antrenarea / alinierea bazată pe modelarea mixturii HMM şi Gaussiene, precum şi generarea structurilor de tip latice pentru antrenarea secvenţială discriminativă.

Etapele principale necesare fluxului unui sistem în limba română sunt:

- Etapa 0. Pregătirea datelor. Pentru această etapă a fost utilizat formatul Kaldi de stocare a datelor de antrenare (acelaşi format utilizat şi pentru celelalte sisteme de RAV ale SpeeD).
- Etapa 1. Extragere parametri. Din datele de intrare se extrag parametri de vorbire tipici, de tip MFCC, însă de înaltă rezoluţie: 80 de coeficienţi cepstrali; vectorul de parametri corespunzător fiecărei ferestre din semnalul de vorbire va avea 80 sau 83 de dimensiuni în funcţie de cum la cei 80 de coeficienţi cepstrali se adaugă sau nu şi informaţii despre frecvenţa fundamentală.
- Etapa 2. Conversie formate în ESPnet. Această etapă converteşte informaţia din structura de directoare Kaldi (transcrieri, id-uri vorbitor şi limbă, lungimea vectorilor de intrare) într-un singur JSON (data.json). Caracteristicile extrase din fişierele audio rămân în format Kaldi.
- Etapa 3. Antrenare model de limbă. Pas opţional, necesar dacă se doreşte utilizarea unui model de limbă suplimentar pentru reevaluare lingvistică. În experimentele noastre, am încercat ambele variante: sistem RAV end-to-end şi sistem RAV cu reevaluare lingvistică.
- Etapa 4. Antrenare end-to-end. În această etapă se antrenează model codor-decodor hibrid de tip CTC-Transformer, folosind una dintre cele două biblioteci de învăţare profundă PyTorch sau Chainer.
- Etapa 5. Decodarea şi evaluarea. În această ultimă etapă se decodează setul de date de evaluare cu sistemul RAV nou antrenat format din modelul end-to-end (etapa 4) şi, eventual, modelul de limbă suplimentar (etapa 3). După decodare, se realizează alinierea transcrierii ipotetice cu cea de referinţă şi se calculează rata de eroare la nivel de cuvânt (WER).

Setup-ul experimental pentru experimentele de RAV

Experimentele realizate au urmărit dezvoltarea unui sistem de RAV bazat pe ESPnet în aceleaşi condiţii în care a fost dezvoltat cel mai bun sistem de RAV SpeeD din anul 2019. Motivaţia acestei alegeri a fost bineînţeles dorinţa realizării unei comparaţii relevante între sistemul de RAV actual şi noul sistem dezvoltat cu ESPNet.

Astfel, au fost utilizate seturile de date listate în continuare şi descrise pe larg în secţiunea 2.1.1 a acestui raport:

- RSC-train şi SSC-train1+2 - set de date de antrenare restrâns utilizat în etapa de calibrare a hiperparametrilor;
- RSC-train, SSC-train1+2, SSC-train3+4-trans-v4 - set de date de antrenare utilizat pentru antrenarea sistemului final, după calibrarea hiperparametrilor;
- RSC-eval, SSC-eval1 şi SSC-eval2 - set de date de evaluare;
- news2014 şi talkshows - set de date de text utilizat pentru antrenarea modelului de limbă.

Pentru toate experimentele, setul de dezvoltare necesar în antrenarea reţelei neuronale ESPnet a fost selectat aleatoriu din setul de date de antrenare. Setul de dezvoltare reprezintă 10% din setul de date de antrenare şi a fost extras din acesta înainte de începerea procesului de antrenare.

Calibrarea hiperparametrilor sistemului de RAV

Utilitarul ESPNet include o serie reţete de antrenare pentru sisteme de RAV în limba engleză. Dintre acestea, cele mai populare sunt reţetele Librispeech, TED-LIUM v1, WSJ şi CommonVoice, reţete ce folosesc seturile de date cu acelaşi nume pentru antrenarea modelului de RAV. În ceea ce priveşte calibrarea hiperparametrilor sistemului, abordarea generală pe care am ales-o a fost următoarea:

- am analizat reţetele menţionate mai sus şi am identificat hiperparametri ce diferă de la caz la caz;
- hiperparametri care nu diferă la diversele reţete pentru limba engleză au primit aceleaşi valori şi în experimentul pe limba română;

- pentru hiperparametri ale căror valori diferă de la o reţetă la alta, am ales pentru limba română o valoare adaptată dimensiunii setului de date de antrenare pe care îl avem la dispoziţie, ţinând cont, bineînţeles, şi de particularităţile limbii române.

Pentru exemplificarea procedurii de mai sus, luăm ca exemplu dimensiunea vocabularului de subcuvinte (*i.e.* Byte Pair Encoding [Sennrich, 2016]). ESPnet nu transcrie vorbirea direct în cuvinte, ci în unităţi lingvistice mai scurte numite informal subcuvinte. Acestea sunt pur şi simplu secvenţe de litere ce apar în componenţa cuvintelor, însă fără să aibă vreo semnificaţie lingvistică anume (*e.g.* nu reprezintă prefixe, sufixe etc.). Subcuvintele se extrag din corpusul de date text de antrenare pe baza frecvenţei lor de apariţie. Dacă vocabularul de subcuvinte este limitat la dimensiunea N, atunci, folosind algoritmul prezentat în [Sennrich, 2016] se vor alege cele mai frecvente N subcuvinte cu care se pot genera toate cuvintele care apar în respectivul set de date.

În reţetele LibriSpeech, TED-LIUM v1, respectiv CommonVoice, dicţionarul de BPE-uri este generat pe baza transcrierilor fişierelor de vorbire, transcrieri ce cuprind 9.4M cuvinte, 6.6M cuvinte, respectiv 4.5M cuvinte. Dimensiunile vocabularelor de subcuvinte variază de la 5000 (LibriSpeech) la 365 (CommonVoice). Astfel, raportul (număr subcuvinte)/(număr cuvinte în setul de antrenare) variază între 0.05% pentru Librispeech şi 0.0075% pentru TED-LIUM. Am decis ca pentru limba română să păstrăm acest raport între limitele menţionate mai sus, însă am ţinut cont şi de faptul că limba română este o limbă cu multe forme flexionate. Astfel, am ales să folosim un vocabular de subcuvinte de dimensiune mare: 1000 de subcuvinte, raportul (număr subcuvinte)/(număr cuvinte în setul de antrenare) situându-se la limita superioară: 0.05%.

În mod similar au fost analizate valorile hiperparametrilor pentru reţetele pentru limba engleză şi au fost alese valori concrete pentru următorii hiperparametri:

- pentru straturile de atenţie:
  - adim (dimensiunea unităţilor neurale din straturile de atenţie) a fost aleasă 256;
  - aheads (numărul de capete pentru straturile de atenţie) a fost ales 4;
- pentru procedura de antrenare:
  - batch-size (numărul de rostiri per mini-batch) a fost limitat la 32, ca urmare a limitării memoriei RAM de doar 12 GB disponibile în procesoarele grafice Nvidia Tesla K40m, pe care s-a realizat antrenarea;
  - epochs (numărul de epoci de antrenare) a fost limitat la 60, deşi ar fi fost poate util să antrenăm modelul timp de 120 de epoci. Alegerea a fost influenţată de timpul mare de antrenare pe hardware-ul disponibil;
  - transformer-lr (rata de învăţare a modelului de tip transformer) a fost aleasă 5, ca un compromis între ratele folosite în reţetele pentru limba engleză
- pentru optimizarea obiectivului:
  - accum-grad (numărul gradienţi însumaţi înainte de o ajustare a ponderilor reţelei) a fost ales 3 pentru că antrenarea s-a realizat în paralel pe 3 plăci grafice, astfel că ne-am permis să ajustăm ponderile reţelei după 3 treceri forward-backward realizate în paralel.
  - batch-bins (numărul de ferestre de vorbire ce intră în componenţa fiecărui mini-batch) a fost setat la 12M, ca urmare a limitării memoriei RAM de doar 12 GB disponibile în procesoarele grafice Nvidia Tesla K40m, pe care s-a realizat antrenarea;
- pentru modelul de limbă (de tip Transfomer):
  - att-unit (numărul de unităţi de atenţie) a fost ales 256;
  - head (numărul de capete pentru straturile de atenţie) a fost ales 2;
  - layer (numărul de straturi transformer) a fost ales 4;
- pentru procedura de antrenare a modelului de limbă:
  - batchsize (numărul de propoziţii per min-batch) a fost ales 64;
  - epoch (numărul de epoci de antrenare) a fost limitat la 20, deşi ar fi fost poate util să antrenăm modelul timp de 50 de epoci. Alegerea a fost influenţată de timpul mare de antrenare pe hardware-ul disponibil.

După antrenarea folosind parametri evidenţiaţi anterior, au fost realizate un set de experimente pentru stabilirea valorilor optime pentru hiperparametri procesului de decodare:

- LMW (language model weight – ponderea scorului dat de modelul de limbă);
- CTCW (CTC weight – ponderea scorului dat de CTC).

Rezultatele experimentale

Rezultatele experimentale obţinute sunt evidenţiate în Tabelul 2.2.a. Putem observa că cele mai bune rezultate se obţin dacă ponderile date scorurilor provenind de la modelul de limbă suplimentar şi obiectivul CTC sunt egale (şi egale cu 0.5). De asemenea, putem observa că acest prim sistem RAV ESPnet se apropie de rezultatele obţinute cu sistemul RAV dezvoltat cu Kaldi, însă nu este încă la fel de performant. Bineînţeles, trebuie luat în considerare şi faptul că atât modelul end-to-end de transcriere de vorbire, cât şi modelul lingvistic suplimentar au fost antrenate un număr de epoci relativ mic (mai puţin de 50% din cât ar fi trebuit) din cauza timpului lung de antrenare pe sistemele hardware avute la dispoziţie. Într-un viitor apropiat, urmează să fie efectuate mai multe experimente cu sistemul ESPnet, iar aceste rezultate preliminare sunt îmbucurătoare şi ne permit să fim încrezători că vom putea egala performanţele sistemului dezvoltat cu Kaldi.

Din păcate, nu se poate afirma acelaşi lucru şi despre durata proceselor de antrenare şi decodare. Procesul de antrenare al sistemului RAV bazat pe ESPnet durează de aproximativ 5 ori mai multe decât un proces de antrenare similar pentru un sistem RAV Kaldi. În ceea ce priveşte procesul de decodare, experimentele noastre au arătat că sistemul ESPnet decodează o oră de vorbire în 2 până la 7 ore, în funcţie de cât de similară este acustica respectivei ore de vorbire raportat la setul de date de antrenare. De partea cealaltă, decodarea unei ore de vorbire cu sistemul RAV bazat pe Kaldi durează doar 1 minut. Acest lucru ne-a permis să folosim sistemele bazate pe Kaldi în cadrul proiectului ReTeRom pentru a transcrise seturi de date mari (peste 500 de ore) de vorbire neadnotată şi apoi să proiectăm şi să aplicăm diverse metode pentru a obţine un subset de date transcris corect (cu o acurateţe de peste 99%). Din cauza timpului extrem de lung de decodare pentru sistemul RAV bazat pe ESPnet acest sistem nu poate fi utilizat în mod similar în adnotarea automată a seturilor de date de vorbire.

Tabelul 2.2.a. *Comparaţie între sistemul RAV ESPnet şi sistemul RAV Kaldi. Optimizarea hiperparametrilor de decodare pentru sistemul RAV ESPnet în limba română.*

		Parametri decodare		**WER [%]**		
Set de antrenare de vorbire adnotată	**Model lingvistic**	**LMW**	**CTCW**	**RSC_eval**	**SSC_eval1**	**SSC_eval2**
RSC-train + SSC-train1+2	Nu	n/a	n/a	13.3	25.1	78.0
	Da	0.5	0.5	8.8	**21.3**	**38.9**
	Da	0.6	0.4	**8.7**	21.6	46.5
	Da	0.7	0.3	9.0	23.0	64.3
	Da	0.8	0.2	11.2	25.8	84.6
RSC-train + SSC-train1+2 + SSC-train3-trans-v4 + SSC-train4-trans-v4	Da	0.6	0.4	3.4	15.3	23.5
Sistem RAV baseline bazat pe Kaldi (antrenat pe acelaşi set de date ca mai sus, folosind model de limbă pentru reevaluare lingvistică)				1.8	11.0	14.0

2.2.2 Dezvoltarea metodei de adnotare bazate pe sisteme RAV complementare şi rezultatele obţinute pe parcursul proiectului

În această secţiune prezentăm un studiu comparativ asupra performanţelor sistemului de RAV iniţial (la începutul proiectului ReTeRom) şi a sistemelor de RAV rezultate în cadrul activităţilor A1.13/2018 şi A2.13/2019, ca urmare a proiectării şi aplicării metodei de generare automată de adnotări folosind sisteme RAV complementare.

Activitatea A1.13 din etapa 1/2018 a presupus utilizarea a două sisteme RAV iniţiale pentru obţinerea în mod automat de transcrieri cu grad ridicat de încredere, prin metoda sistemelor complementare. Sistemul iniţial RAV #1 [Georgescu, 2017; Georgescu, 2018] a fost dezvoltat cu utilitarul CMU Sphinx [Huggins-Daines, 2006]. Modelele sistemului sunt probabilistice, de tip HMM-GMM (modelul acustic), respectiv 2-gram (modelul lingvistic). Acest sistem avea un vocabular de 64k cuvinte. Sistemul iniţial RAV #2 [Georgescu, 2018] a fost dezvoltat cu utilitarul Kaldi [Povey, 2011], unde modelul acustic este unul de tip reţea neuronală cu întârziere în timp (time-delay - TDNN) specifică implementării NNET2. Modelul lingvistic este unul probabilistic, de tip 2-gram, folosit la decodare, în timp ce un model superior, de tip 4-gram, a fost folosit pentru reevaluarea lingvistică, corectând transcrierea iniţială.

Cele două sisteme au fost utilizate pentru a transcrie seturile de vorbire neadnotată SSC-train3-raw, ce însumează 136 ore, şi SSC-train4-raw, cu o durată de 777 ore. În urma transcrierii şi aplicării metodei, s-au obţinut seturile de date SSC-train3-compl-2018, cu o dimensiune de 49 ore, şi SSC-train4-compl-2018, cu o durată de 280 ore. Calitatea acestor noi seturi obţinute a fost verificată prin aplicarea metodei asupra unor seturi de date pentru care există transcriere de referinţă: RSC-eval şi SSC-eval. S-a constatat că selecţiile rezultate sunt greşite într-o foarte mică proporţie, numai 1.0% - 1.3%. Prin extrapolare, putem considera că şi noile date sunt corecte în mare măsură (aproximativ 99%), fiind astfel posibil ca acestea să fie folosite mai departe ca seturi de antrenare adiţionale.

Tabelul 2.2.a prezintă pe primele două linii caracteristicile şi performanţele sistemelor RAV iniţiale, în timp ce a treia linie prezintă caracteristicile şi performanţele sistemului RAV #2 îmbunătăţit, pentru a cărui antrenare au fost adăugate noile date obţinute. Acest nou sistem a obţinut o îmbunătăţire relativă de 3.81% pe setul de evaluare cu vorbire citită, respectiv 8.87% pe setul de evaluare cu vorbire spontană.

Activitatea 2.13 din etapa 2/2019 a presupus aplicarea metodei sistemelor complementare, de data aceasta, cele două sisteme iniţiale fiind sistemul RAV #2 prezentat în Activitatea 1.13 din 2018 şi sistemul RAV #3, având modelul acustic de tip time-delay (TDNN), creat cu utilitarul Kaldi folosind implementarea NNET3. Modelul de limbă folosit pentru decodare este un 2-gram, în timp ce modelul de limbă folosit pentru reevaluarea lingvistică este unul de tip reţea neuronală recurentă (RNN), cu un istoric de 5 cuvinte. Ambele modele de limbă folosesc un vocabular de 200k cuvinte.

Metoda a fost aplicată pe aceleaşi seturi de date de vorbire neadnotată: SSC-train3-raw şi SSC-train4-raw, cu o durată de 777 ore. În urma aplicării metodei, s-au obţinut seturile de date SSC-train3-compl-2019 cu o dimensiune de 79 ore, şi SSC-train4-compl-2019, cu o durată de 452 ore. Calitatea acestor noi date a fost verificată prin aplicarea metodei asupra seturilor de referinţă RSC-eval şi SSC-eval. Selecţiile rezultate au fost greşite în proporţie de 2.6% - 2.7%, mai mult decât în cazul sistemelor RAV #1 şi RAV #2. Acest fapt este pus pe seama diferenţelor mai mici dintre cele două sisteme, ambele fiind dezvoltate cu ajutorul utilitarului Kaldi, iar modelele acustice sunt asemănătoare din punct de vedere al tehnologiei folosite.

Tabelul 2.2.a *Performanţa sistemelor RAV iniţiale şi a sistemului RAV îmbunătăţit din A1.13, etapa 1.*

Model acustic		**Model linvistic**	**WER [%]**		**Îmbunătăţire relativă a WER [%]**	
Corpus antrenare	**Tip model**		**RSC-eval**	**SSC-eval**	**RSC-eval**	**SSC-eval**
RSC-train + SSC-train	HMM-GMM	Decodare RAV: 64k cuvinte, 3-gram	9.56	28.64	-	-
RSC-train + SSC-train	HMM-DNN (TDNN2)	Decodare RAV: 200k cuvinte, 2-gram Reev. lingv.: 200k cuvinte, 4-gram	3.41	17.91	-	-
RSC-train + SSC-train + SSC-train3-compl-2018 + SSC-train4-compl-2018	HMM-DNN (TDNN2)	Decodare RAV: 200k cuvinte, 2-gram Reev. lingv.: 200k cuvinte, 4-gram	3.28	16.32	3.81	8.87

Tabelul 2.2.b *Performanţa sistemelor RAV iniţiale şi a sistemului RAV îmbunătăţit din A2.13, etapa 2.*

Model acustic		**Model linvistic**	**WER [%]**		**Îmbunătăţire relativă a WER [%]**	
Corpus antrenare	**Tip model**		**RSC-eval**	**SSC-eval**	**RSC-eval**	**SSC-eval**
RSC-train + SSC-train	HMM-DNN (TDNN2)	Decodare RAV: 200k cuvinte, 2-gram Reev. lingv.: 200k cuvinte, 4-gram	3.41	17.91	-	-
RSC-train + SSC-train	HMM-DNN (TDNN3)	Decodare RAV: 200k cuvinte, 2-gram Reev. lingv.: RNN 5-gram	1.88	14.96	-	-
RSC-train + SSC-train + SSC-train3-compl-2018 + SSC-train4-compl-2018	HMM-DNN (TDNN3)	Decodare RAV: 200k cuvinte, 2-gram Reev. lingv.: RNN 5-gram	1.80	11.70	4.26	21.79
RSC-train + SSC-train + SSC-train3-compl-2019 + SSC-train4-compl-2019	HMM-DNN (TDNN3)	Decodare RAV: 200k cuvinte, 2-gram Reev. lingv.: RNN 5-gram	1.87	11.78	0.53	21.26

Noile date obţinute au fost adăugate la setul de antrenare deja existent, fiind creat sistemul RAV #3 îmbunătăţit. Tabelul 2.2.b prezintă pe primele 2 linii sistemele iniţiale RAV #2 şi RAV #3. Următoarele două linii prezintă două versiuni ale sistemului RAV #3 îmbunătăţit: (i) una antrenată după ce au fost adăugate datele SSC-train3-compl-2018 şi SSC-train4-compl-2018 la setul de date iniţial, respectiv (ii) una antrenată folosind seturile SSC-train3-compl-2019 şi SSC-train4-compl-2019 suplimentar faţă de setul iniţial. Urmărind rezultatele din acest tabel, putem trage două concluzii interesante:

- seturile de date obţinute în 2018 sunt mai mici, însă probabil mai precise (sistemele iniţiale din 2018 erau mai diferite decât sistemele din 2019) astfel că sistemul RAV reantrenat folosind seturile din 2018 este puţin mai bun decât sistemul RAV reantrenat folosind seturile generate in 2019.
- seturile de date nou obţinute sunt mult mai utile pentru dezvoltarea unui sistem ce se doreşte a transcrie corect vorbire spontană: îmbunătăţirea relativă a WER-ului este de peste 20% pentru vorbire spontană (SSC-eval) şi oarecum nesemnificativă pentru vorbire citită (RSC-eval).

Deşi activităţile din etapa 3/2020 nu prevedeau acest lucru, în cursul acestei etape s-a încercat dezvoltarea suplimentară a metodei utilizând ca sisteme iniţiale sistemul RAV #3 (2019) şi un nou sistem de RAV dezvoltat folosind utilitarul ESPnet (vezi secţiunea anterioară). În acest demers am întâmpinat însă o problemă ce nu a putut fi încă depăşită. Sistemul RAV bazat pe ESPnet s-a dovedit a fi foarte lent în transcriere, astfel încât nu a putut fi utilizat pentru transcrierea seturilor de date neadnotate SSC-train3-raw şi SSC-train4-raw. Transcrierea acestor seturi de date, cu o durată totală de 913 ore, ar fi necesitat aproximativ 2700 de ore, adică peste 100 de zile. Pentru comparaţie, menţionăm că transcrierea aceloraşi seturi de vorbire cu sistemul RAV bazat pe Kaldi a necesitat numai 16 ore. În urma acestei experienţe am tras concluzia că pentru aplicarea cu succes a metodelor proiectate în cadrul proiectului nu este suficient să existe sisteme de RAV iniţiale performante şi complementare, ci, în plus, acestea trebuie să transcrie vorbirea suficient de rapid.

2.3 Activităţile 3.10 şi 3.12 - Îmbunătăţirea soluţiei de filtrare şi aliniere a transcrierilor aproximative cu semnalul de vorbire şi Analiza impactului utilizării transcrierilor aproximative în vederea reantrenării sistemelor de RAV

În această secţiune vom aborda următoarele subiecte: (i) îmbunătăţirea metodei de utilizare a transcrierilor aproximative pentru generarea de adnotări pentru seturi de date de vorbire şi (ii) aplicarea metodei pe două noi seturi de date brute: corpusul CoBiLIRo şi corpusul CDep, urmată de evaluarea impactului seturilor de date nou adnotate asupra sistemului de RAV iniţial.

2.3.1 Îmbunătăţirea soluţiei de aliniere a transcrierilor aproximative cu semnalul de vorbire

Metoda de bază

Metoda de bază de filtrare şi aliniere a transcrierilor aproximative cu semnalul de vorbire a fost introdusă, prezentată şi evaluată în activitatea 2.11 din etapa 2019 a proiectului (vezi raport TADARAV 2019 [Georgescu, 2019a], secţiunea 2.2). În cele ce urmează este prezentată o descriere sumară a acestei metode. Obiectivul principal al metodei de bază de aliniere este obţinerea adnotărilor precise pentru o parte a unui

corpus de vorbire în mod automat. Noul corpus obţinut va fi folosit pentru reantrenarea sistemelor RAV existente crescând astfel variabilitatea modelului acustic, ceea ce ar trebui să implice mai departe o îmbunătăţire în acurateţea transcrierilor. Metoda de adnotare se bazează pe folosirea unui corpus de vorbire brut împreună cu transcrieri aproximative şi transcrierile produse de un sistem RAV iniţial. Cele două seturi de transcrieri ale corpusului de vorbire sunt aliniate folosind distanţa Levenstein. În urma alinierii, considerăm părţile identice dintre cele două transcrieri ca fiind corecte şi le selectăm ca parte a noului corpus.

Părţile de vorbire şi de transcrieri aproximative au fost extrase din mediul online mass-media (ştiri, interviuri, rapoarte). Deşi mediul online mass-media este o sursă bogată de vorbire şi text, transcrierile aproximative extrase sunt diferite de cele generate de sistemul RAV în ceea ce priveşte formatul. Transcrierile aproximative conţin titluri, nume proprii, semne de punctuaţie, numere, abrevieri etc., pe când transcrierile RAV sunt practic secvenţe de cuvinte cu litere mici însoţite de etichete de timp. Ca şi exemplu:

- Transcriere aproximativă: Bărbatul de 36 de ani povesteşte că muncise toată noaptea.
- Transcriere RAV: bărbatul(3.71, 4.02) de(4.02, 4.87) treizeci(4.87, 5.11) şi(5.11, 5.74) şase(5.74, 6.02) de(6.02, 6.35) ani(6.35, 6.71) povesteşte(6.71, 6.94) că(6.94, 7.58) muncise(7.58, 7.89) toată(7.89, 8.16) noaptea(8.16, 8.32)

Astfel, transcrierile aproximative trebuie supuse unor operaţii de procesare de text, anume: restaurare de diacritice, înlocuirea URL-urilor şi adreselor de email cu forma lor vorbită, înlocuirea numerelor cu text, expandarea abrevierilor, înlocuirea caracterelor speciale, transformarea literelor mari în litere mici.

După procesul de aliniere, segmentele de text obţinute sunt filtrate pe baza anumitor criterii cu scopul de a înlătura segmente foarte mici (în ceea ce priveşte durata vorbirii şi numărul de caractere) şi segmente ce conţin zone lungi de linişte (durata dintre cuvinte consecutive). În final, procesele de aliniere şi filtrare generează un set de transcrieri aliniate însoţite de etichete de timp. Etichetele de timp vor fi folosite mai departe pentru tăierea segmentelor audio corespunzătoare din materialele audio. Astfel, noul corpus de vorbire adnotată va conţine transcrieri aliniate şi segmentele audio corespunzătoare, ce pot fi folosite mai departe pentru reantrenarea sistemelor RAV.

Metoda propusă

Metoda descrisă anterior este foarte strictă în ceea ce priveşte corectitudinea transcrierilor aliniate, în sensul că sunt considerate a fi corecte doar secvenţele de cuvinte unde transcrierile aproximative şi transcrierile RAV se potrivesc perfect. Acest fapt duce la selectarea multor secvenţe scurte (*i.e.* ce cuprind puţine cuvinte). În urma unei analize manuale a transcrierilor s-a constatat că există multe secvenţe de câteva cuvinte pentru care procesul de aliniere a eşuat şi care se află între secvenţe lungi de cuvinte care s-au aliniat cu succes. Câteva astfel de exemple sunt prezentate în Tabelul 2.3.a.

Aceste secvenţe nu au fost aliniate din cauza mai multor motive precum zgomot de fundal în materialul audio, cuvinte care nu se regăsesc în vocabularul sistemului RAV (de regulă nume proprii), cuvinte adiţionale în fişierul audio care nu sunt pronunţate în materialul audio. Cel de-al treilea caz se întâlneşte de obicei în interviuri, unde transcrierea de pe website conţine text adiţional pentru a preciza cine vorbeşte.

***Tabelul 2.3.a**. Exemple de secvenţe de cuvinte pentru care procesul de aliniere a eşuat şi care se află între secvenţe lungi de cuvinte care s-au aliniat cu succes. Secvenţele pentru care procesul de aliniere a eşuat sunt marcate cu roşu/ verde având în vedere dacă transcrierea aproximativă este incorectă/corectă.*

Nr.	**Secvenţa anterioară (aliniată)**	**Secvenţă nealiniată**	**Secvenţa următoare (aliniată)**	**Observaţii**
1	… se îndreaptă către susţinătorii săi	o dronă	filmează evenimentul ...	confuzat cu "un pod" de către sistemul RAV din cauza zgomotului de fundal
2	… în zona	Egiptului	unde există ...	confuzat cu "edituri" de către sistemul RAV deoarece acest cuvânt nu se află în vocabular
3	… a obţinut fondurile	administrator bloc	la doi ani …	secvenţa de text nu este pronunţată în materialul audio, scopul acesteia în transcrierea aproximativă este aceea de a specifica cine vorbeşte

În urma unei analize a 50 de astfel de situaţii, am ajuns la concluzia că dacă am lua în considerare secvenţe de mai puţin de 6 cuvinte pentru care procesul de aliniere a eşuat, în aproximativ 68% din cazuri transcrierea aproximativă este corectă, pe când transcrierea RAV este incorectă. Astfel, în aceste cazuri am decis să unim cele 3 secvenţe de cuvinte. Adăugând aceste secvenţe de text sperăm să generăm un corpus de vorbire mai mare ce va fi folositor pentru reantrenarea sistemelor RAV. Noi credem că aceste secvenţe vor fi mai valoroase pentru reantrenare, deoarece sistemul RAV iniţial nu a reuşit să le transcrie. Desigur, există şi o parte negativă la acest compromis: 32% din secvenţele adăugate sunt incorecte şi pot afecta în mod negativ procesul de antrenare.

Evaluarea metodelor

O metodă de adnotare automată poate fi evaluată în funcţie de dimensiunea corpusului nou generat şi în funcţie de calitatea transcrierilor, exprimată în funcţie de rata de eroare la nivel de cuvânt (WER). Dimensiunea corpusului nou generat se doreşte a fi pe cat de mare posibilă. Desigur, limita superioară este dimensiunea corpusului audio brut (materialele audio). Astfel, o metrică mai potrivită este eficacitatea metodei, definită ca procentajul de material audio ce a fost selectat ca făcând parte din corpusul nou generat şi pentru care metoda a generat transcrieri corecte. Calitatea adnotării exprimată în rata de eroare la nivel de cuvânt (WER) şi/sau rata de eroare la nivel de caracter (ChER) nu poate fi măsurată pentru metoda de bază din cauza absenţei unei referinţe pentru transcrierile aproximative. Totuşi, evaluarea se poate face pe alt criteriu care se bazează pe această metrică. O adnotare de calitate înaltă ar trebui să ducă în mod normal la o performanţă mai bună a sistemului RAV reantrenat pe noul corpus de vorbire generat.

Rezultate alinieri

Procedura de aliniere şi filtrare prezentată în secţiunea *"Metoda de bază"* a fost aplicată pe seturile de date brute SSC-train3-raw şi SSC-train4-raw. În etapa 2/2019 am folosit ca punct de pornire transcrierile acestor două seturi de date generate cu un sistem RAV iniţial mai slab, şi anume cel prezentat în [Georgescu, 2017]. Au rezultat astfel seturile de date de vorbire adnotată SSC-train3-trans-v3 şi SSC-train4-trans-v3. Dimensiunile acestora, exprimate în număr de ore de vorbire şi eficienţa procesului de adnotare automată, exprimată sub forma procentului de date brute ce au putut fi adnotate, raportat la dimensiunea datelor brute sunt prezentate în Tabelul 2.3.b.

Aceeaşi metodă de aliniere şi filtrare a fost aplicată în cadrul activităţii curente folosind ca punct de pornire transcrierile acestor două seturi de date generate cu un sistem RAV îmbunătăţit, şi anume cel prezentat în [Georgescu, 2019b]. A rezultat astfel versiunea v4 a acestor seturi de date adnotate, versiune ce cuprinde mai multe ore de vorbire, conform Tabelului 2.3.b.

În urma aplicării procedurii de aliniere şi filtrare prezentată în secţiunea *"Metoda propusă"* pe seturile de date brute SSC-train3-raw şi SSC-train4-raw, au fost obţinute versiunile v5 ale seturilor de date adnotate. În acest caz a fost utilizat pentru transcriere tot sistemul de RAV îmbunătăţit prezentat în [Georgescu, 2019b].

Tabelul 2.3.b. *Statistici pentru seturile de date SSC-train3-trans-v4 şi SSC-train4-trans-v4*

Set de date	Durată [# ore]	Eficienţă aliniere [% ore]
SSC-train3-trans-v3	37,5	27,4%
SSC-train3-trans-v4	41,0	30,0%
SSC-train3-trans-v5	46,1	33,7%
SSC-train4-trans-v3	228,8	29,4%
SSC-train4-trans-v4	250,2	32,2%
SSC-train4-trans-v5	281,6	36,2%

Aşa cum era de aşteptat, seturile de date v4 sunt mai mari (cu aproximativ 9%) decât seturile de date v3. Eficienţa alinierii a crescut de la 27.4% la 30.0% pentru primul set de date, respectiv de la 29.4% la 32.2% pe cel de-al doilea set de date. Ne aşteptăm ca această creştere a seturilor de date să contribuie pozitiv la reantrenarea sistemului de RAV iniţial, întrucât creşterea se bazează în totalitate pe nişte transcrieri RAV iniţiale mai precise.

De asemenea, seturile de date v5 sunt mai mari (cu aproximativ 12%) decât seturile de date v4. Eficienţa alinierii a crescut de la 30.0% la 33.7% pentru primul set de date, respectiv de la 32.3% la 36.2% pe cel de-al doilea set de date. Urmează să vedem, în secţiunea următoare, dacă aceasta creştere a seturilor de date contribuie pozitiv sau negativ la reantrenarea sistemului de RAV iniţial.

Sisteme RAV ce utilizează noile corpusuri obţinute

Sistemele RAV obţinute în urma reantrenării cu noile seturi de date sunt prezentate în Tabelul 2.3.d, alături de sistemul RAV iniţial (cel ce a fost utilizat pentru transcrierea datelor brute în anul 2020). Comparativ cu raportul precedent, evaluarea s-a realizat pe seturile de evaluare corectate RSC-eval-v2 şi SSC-eval-v2 (vezi secţiunea 2.1.1 pentru mai multe detalii).

***Tabelul 2.3.d**. Performanţa sistemelor RAV după reantrenare*

Sistem RAV	**Set antrenare**	**WER [%]**		**Îmbunătăţire relativă a WER [%]**	
		RSC-eval	**SSC-eval1**	**RSC-eval**	**SSC-eval1**
Sistem RAV iniţial	RSC-train + SSC-train	1.88	14.96	n/a	n/a
RAV reantrenat 2019 (metoda de bază)	RSC-train + SSC-train + SSC-train3-trans-v3 + SSC-train4-trans-v3	1.83	11.50	2.66%	23.1%
RAV reantrenat 2020 (metoda de bază)	RSC-train + SSC-train + SSC-train3-trans-v4 + SSC-train4-trans-v4	1.83	11.04	2.66%	26.2%
RAV reantrenat 2020 (metoda propusă)	RSC-train + SSC-train + SSC-train3-trans-v5 + SSC-train4-trans-v5	1.83	11.13	2.66%	25.6%

În ceea ce priveşte rezultatele de transcriere de vorbire continuă (setul RSC-eval), observăm că indiferent de metoda de aliniere şi filtrare utilizată şi indiferent de calitatea transcrierilor iniţiale, sistemele RAV reantrenate pe seturile iniţiale şi seturile nou adnotate prezintă aceeaşi îmbunătăţire faţă de sistemul RAV iniţial: WER de 1.83% vs. 1.88%.

Pentru vorbire spontană, metoda de bază aplicată anul trecut, folosind transcrieri iniţiale de calitate mai slabă, conduce la obţinerea unei îmbunătăţiri relative de WER de aproximativ 23%. Aceeaşi metodă, aplicată în acest an, pornind de la transcrieri iniţiale mai precise, conduce la obţinerea unei îmbunătăţiri relative de WER de aproximativ 26%. Se observă astfel importanţa calităţii transcrierilor iniţiale.

Metoda nou propusă în acest an conduce la obţinerea unei îmbunătăţiri relative de WER de aproximativ 25% faţă de sistemul RAV iniţial. Se observă astfel că propunerea de a utiliza în antrenarea RAV şi a secvenţelor de câteva cuvinte (maxim 6) pentru care procesul de aliniere a eşuat şi care se află între secvenţe lungi de cuvinte care s-au aliniat cu succes, nu conduce la rezultate RAV mai bune. Din nou rezultă faptul că este extrem de importantă calitatea transcrierilor pentru datele de antrenare, în detrimentul cantităţii acestor date.

2.3.2 Aplicarea metodei pe setul de date CoBiLiRo-raw

Analiză iniţială a setului de date CoBiLiRo-raw

Metoda de bază**,** descrisă în secţiunea 2.3.1, a fost aplicată şi pe setul de date CoBiLiRo-raw (vezi Tabelul 2.1.d din secţiunea 2.1.2), cu scopul îmbunătăţirii sistemului RAV. Această metodă a fost preferată faţă de metoda propusă în secţiunea anterioară deoarece cea din urmă a fost evaluată cu rezultate mai slabe. În primă fază, s-a aplicat metoda pe doar 2 fişiere din setul de date, în vederea analizării eventualelor probleme. Rezultatele alinierii se pot observa în Tabelul 2.3.e, unde sunt prezentate numărul de cuvinte aliniate, numărul de cuvinte ce se regăsesc în transcrierea aproximativă, numărul de cuvinte ce se regăsesc în transcrierea RAV, precum şi procentul de cuvinte aliniate raportat la numărul de cuvinte din transcrierea aproximativă.

***Tabelul 2.3.e**. Transcrieri RAV obţinute cu sistem RAV baseline [Georgescu, 2019b]*

Fişierul audio	Nr. cuvinte aliniate	Nr. cuvinte transcrieri aproximative	Nr. cuvinte transcrieri RAV	Aliniere [% cuvinte]
alma24	3.696	5.151	5.985	71,75%
interviu2	2.065	18.095	11.225	18,40%

După analiza rezultatelor din Tabelul 2.3.e, precum şi a fişierelor în cauză, s-au constatat următoarele: (i) în cazul fişierului alma24, transcrierea RAV conţine mai multe cuvinte decât cea aproximativă, deoarece în fişierul audio sunt pronunţate mai multe cuvinte decât există în transcrierea aproximativă, (ii) în cazul fişierului interviu2, transcrierea aproximativă conţine foarte mult text care nu este pronunţat în fişierul audio. Există reformulări, adăugiri, precum şi precizări suplimentare.

În Tabelele 2.3.f şi 2.3.g sunt prezentate câteva exemple de probleme întâlnite pentru cele 2 seturi de date. În Tabelul 2.3.f, unde sunt prezentate exemple de text lipsă pentru setul de date alma24, se poate observa pe prima coloană text care este întâlnit în transcrierea aproximativă dar care este şi pronunţat în fişierul audio, urmat de eticheta de timp la care se încheie pronunţia textului respectiv pe coloana a 2-a. Acest text este situat anterior faţă de cel de pe coloana a 3-a, text care este pronunţat în fişierul audio dar nu se regăseşte în transcrierea aproximativă. Pe coloanele 4 şi 5 din Tabelul 2.3.f se regăseşte text similar cu cel din primele 2 coloane cu excepţia faptului că acest text este situat ulterior faţă de cel de pe coloana 3, iar eticheta de timp marchează momentul de timp la care se începe pronunţia textului. În final, pe ultima coloană se regăseşte diferenţa dintre cele 2 etichete de timp, care marchează durata de timp pentru care lipseşte transcrierea de pe coloana 3. Analizând exemplele din Tabelul 2.3.f, se poate observa că există porţiuni din fişierul audio de până la aproximativ 20s pentru care nu există transcriere aproximativă.

***Tabelul 2.3.f**. Analiză probleme set de date alma24*

Context anterior (există atât în audio cât şi în transcrierea aproximativă)	Etichetă timp context anterior	Există în fişierul audio, dar nu există în transcrierea aproximativă	Context ulterior (există atât în audio cât şi în transcrierea aproximativă)	Etichetă timp context ulterior	Delta timp fără transcriere [s]
de când realizăm această emisiune	29,28	universală de dar în colaborare cu tvr iaşi sau tvr iaşi doar cu universitatea pentru a marca împlinirea a o sută cincizeci de ani de la crearea universităţii moderne a fost trimisă de variate până habar nici despre istoria universităţii emisiuni despre facultăţii şi cu diverse personalităţi	spuneam într-o discuţie anterioară	49,77	20,49
nouăzeci de ani	133,26	pentru ştiu că sunt program complex	am publicat	142,8	9,54
eu	311,43	sunt un pic mai sus	cred că	315,45	4,02
Curtea Constituţională	413,76	nu o să vă întreb despre legea pensiilor discutăm despre facultatea de drept	Domnule profesor	420	6,24

În Tabelul 2.3.g sunt prezentate exemple de text care se regăseşte în transcrierea aproximativă dar nu este pronunţat în fişierul audio. După cum se poate observa, există atât secvenţe scurte de text nepronunţat cât şi secvenţe mai lungi.

***Tabelul 2.3.g**. Analiză fişier interviu2*

Context anterior (există atât în audio cât şi în transcrierea aproximativă)	**Etichetă timp context anterior**	**Există în transcrierea aproximativă, dar nu există în fişierul audio**
universitatea din freiburg	110	în toate domeniile şi la toate palierele vasta sa activitate de construcţie a relaţiei dintre iaşi şi freiburg a început
foarte importante	165,99	pentru noi profesorii şi deplasări în germania pe atunci relaţia cu freiburg
realizările lui miron	205	poate cea mai importantă
îşi pot imagina	218,34	ce şansă enormă însemna pentru acei oameni simpla posibilitate de a ieşi fie şi pentru perioade scurte din închisoarea generalizată care era românia comunistă da acest lucru a însemnat ceva excepţional
îmi amintesc	223	că prin o mie nouă sute optzeci şi cinci a avut loc în germania un mare
o proprietate	337,08	unde au construit o casă unde continuă să îşi primească prietenii în plus au ctitorit acolo o frumoasă bisericuţă
înmormântat	341	în cimititul din imediata apropriere decizia de a-şi găsi odihna la malul mării nu poate fi desprinsă cred de simbolistica exilului atât de apropiată fostului mare exilat de altfel paul miron a şi scris o piesă de teatru despre ovidiu alt mare exilat dar pe ţărmurile pontului euxin

După analiza acestor probleme, s-a încercat o actualizare a modelelor de limba folosind transcrierile aproximative ale celor 2 seturi de date ca text adiţional, obţinând rezultatele din Tabelul 2.3.h.

***Tabelul 2.3.h**. Transcrieri RAV obţinute cu sistem cu model de limbă actualizat*

Set de date	**Nr. cuvinte aliniate**	**Nr. cuvinte transcrieri aproximative**	**Nr. cuvinte transcrieri RAV**	**Aliniere [% cuvinte]**
alma24	4.137	5.151	5.621	80,31%
interviu2	33	18.095	8.830	0,18%

Comparând rezultatele din Tabelul 2.3.h cu cele iniţiale din Tabelul 2.3.e, se poate observa o creştere a numărului de cuvinte aliniate pentru setul alma24, dar o scădere drastică pentru setul interviu2. După realizarea unor seturi de experimente, în care s-au aliniat anumite părţi de text din transcrierea aproximativă cu partea corespunzătoare a transcrierii RAV a setului interviu2 (prima jumătate a transcrierii aproximative cu prima jumătate a transcrierii RAV, a doua jumătate, primele 1000 de cuvinte, primele 2000 de cuvinte, … , primele 8000 de cuvinte) s-a constatat că modulul de aliniere din aplicaţia JavaNLP2 (NISTAlign, prezentat în raportul anterior, secţiunea 2.2.3) întâlneşte probleme când primeşte la intrare cantităţi mari de text. Ca soluţie pentru această problemă, partea de aliniere a fost mutată în exteriorul aplicaţiei JavaNLP2 şi executată cu toolkit-ul sclite, urmând însă ca operaţiile de filtrare să fie executate tot de javaNLP2. Rezultatele actualizate sunt prezentate în Tabelul 2.3.i.

Analizând rezultatele din Tabelul 2.3.i, putem observa o uşoară creştere a numărului de cuvinte aliniate în cazul setului alma24 comparativ cu rezultatele din Tabelul 2.3.h. De asemenea, pentru setul interviu2, problema generată de către aliniatorul NISTAlign din toolkit-ul CMU Sphinx a fost rezolvată folosind modulul alternativ de aliniere din toolkit-ul sclite. Deşi această problemă a fost rezolvată, procentul de cuvinte aliniate rămâne relativ mic.

***Tabelul 2.3.i**. Transcrieri RAV obţinute cu sistem cu model de limbă actualizat + aliniere cu sclite*

Set de date	**Nr. cuvinte aliniate**	**Nr. cuvinte transcrieri aproximative**	**Nr. cuvinte transcrieri RAV**	**Aliniere [% cuvinte] din transcrierea aproximativă**	**Aliniere [% cuvinte] din transcrierea RAV**
alma24	4.157	5.151	5.621	80,70%	73,95%
interviu2	3.811	18.095	8.830	21,06%	43,16%

Aplicare metodă de bază set de date CoBiLiRo-raw

În urma testării metodei de bază pe cele 2 fişiere de test, metoda de bază a fost aplicată pe întreg setul de date CoBiLiRo-raw, cu menţiunea că procesul de aliniere a fost efectuat de toolkit-ul sclite şi nu de CMU Sphinx (NISTAlign) pentru a evita eventualele probleme similare cu cele întâlnite la setul de date interviu2, menţionate în subcapitolul anterior. Sistemul RAV folosit pentru transcrierea întregului set de date CoBiLiRo-raw a fost sistemul RAV baseline [Georgescu, 2019b]. Sistemul RAV cu model de limbă actualizat putea conduce, în final, la obţinerea unui set de date adnotat mai mare, dar fără garanţia că acel text ar fi fost era corect. Din acest motiv sistemul RAV cu model de limbă actualizat nu a fost folosit. În urma aplicării metodei de bază s-a obţinut setul de date CoBiLiRo-trans-v4 ce conţine secvenţele de cuvinte aliniate pentru fiecare subset de date din CoBiLiRo-raw, împreună cu etichete de timp pentru fiecare cuvânt din secvenţă ce indică poziţia şi durata lor în fişierul audio corespunzător. Detalii legate de setul de date CoBiLiRo-trans-v4 se găsesc în Tabelul 2.1.e.

Sistemul RAV reantrenat cu seturile de date iniţiale şi setul de date nou adnotat (CoBiLiRo-trans-v4) a obţinut o rată de eroare la nivel de cuvânt de 1.8% pe setul RSC-eval ce conţine vorbire citită, respectiv 14.0% pe setul SSC-eval1 ce conţine vorbire spontană. Astfel, adăugarea setului CoBiLiRo-trans-v4 de 31 ore la setul iniţial de antrenare de 225 ore, a condus la o îmbunătăţire relativă faţă de sistemul iniţial cu 5.2% pe vorbire citită, respectiv 6.6% pe vorbire spontană.

2.3.3 Aplicarea metodei pentru setul de date CDep-raw

Metoda de bază prezentată în secţiunea 2.3.1 a fost aplicată şi pe setul de date CDep-raw (vezi Tabelul 2.1.d). Fişierele audio din setul de date CDep-raw au fost segmentate la 1 minut, obţinând setul de date CDep-raw-1min. Setul de fişiere audio CDep-raw-1min a fost transcris utilizând sistemul RAV care a obţinut cele mai bune rezultate în analiza făcută în 2019 [Georgescu, 2021]. A rezultat astfel setul de transcrieri RAV necesare pentru efectuarea procesului de aliniere din metoda de bază. După transcrierea fişierelor audio de 1 minut, 300 dintre acestea au fost separate, împreună cu transcrierile lor, în vederea formării unui set de date de evaluare CDep-eval. Transcrierile RAV ale celor 300 de fişiere audio au fost apoi aliniate cu transcrierile aproximative. Deoarece secvenţele aliniate nu conţin întreg textul vorbit din fişierul audio a fost necesară intervenţia manuală prin analiza şi corectarea fişierelor text în vederea obţinerii unor transcrieri de referinţă ce pot fi folosite mai departe pentru evaluarea viitoarelor sisteme RAV. Setul de transcrieri obţinute, exceptând cele din setul CDep-eval, au fost alipite astfel încât timpii să se refere la fişierul audio original şi nu la fişierul audio tăiat. După efectuarea procesului de aliniere între transcrierile RAV obţinute după alipire şi transcrierile aproximative CDep s-a obţinut setul de date CDep-trans-v4, ce conţine 21M de cuvinte şi 878.8 ore de vorbire (84,2% din cuvinte, respectiv 25% din numărul de ore). Analizând rezultatele se poate observa o posibilă neconcordanţă între numărul mare de cuvinte aliniate şi numărul mic de ore, ceea ce a dus spre efectuarea unor analize suplimentare. Aceste analize au constat în evaluarea fişierelor în funcţie de energie şi putere. Prin efectuarea acestor analize, ne-am aşteptat să întâlnim un număr mare de fişiere cu un nivel de putere sub un anumit prag. Ne-am fi aşteptat să găsim fişiere ce conţin linişte pe întreaga sau pe majoritatea duratei fişierului audio, ceea ce ar fi justificat numărul mic de ore raportat la numărul mare de cuvinte aliniate. După efectuarea evaluărilor şi a unei analize manuale a mai multor fişiere audio, am observat că există şi fişiere cu nivel relativ mic de putere ce conţin în mare parte vorbire decât linişte, însă nu am putut obţine încă nişte rezultate cantitative care să justifice neconcordanţa menţionată mai sus. Această problemă urmează să mai fie analizată în cadrul etapei 2021 a proiectului.

Sistemul RAV care a fost utilizat pentru generarea transcrierilor setului CDep-raw a fost reantrenat cu seturile de date iniţiale şi noul set adnotat CDep-trans-v4. Sistemul RAV reantrenat a fost evaluat pe seturile de evaluare RSC-eval şi SSC-eval1 şi SSC-eval2, iar rezultatele sunt prezentate în Tabelul 2.3.j alături de cele ale sistemului RAV iniţial.

Tabelul 2.3.j *Sistem RAV baseline vs sistem RAV*

Sistem RAV	**Seturi de antrenare**	**WER [%]**		
		RSC-eval	**SSC-eval1**	**SSC-eval2**
Sistem RAV iniţial [Georgescu, 2021]	RSC-train + SSC-train1+2 + SSC-train3+4-trans-v4	1.8	11.0	14.0
Sistem RAV reantrenat	RSC-train + SSC-train1+2 + SSC-train3+4-trans-v4 + CDep-trans-v4	1.7	12.3	15.4

Analizând rezultatele din tabelul 2.3.j, putem observa o uşoară scădere a WER de la 1.8% la 1.7% în cazul vorbirii citite (RSC-eval) şi o degradare de aproximativ 10% a performanţei pentru vorbire spontană (seturile SSC-eval1 şi SSC-eval2). Din aceste rezultate putem trage concluzia că, în ciuda adăugării setului de date CDep-trans-v4 (879h) la antrenarea modelului acustic, noul sistem RAV prezintă o performanţă mai scăzută per total faţă de sistemul iniţial. O posibilă explicaţie pentru această performanţă mai scăzută este faptul că prelegerile din şedinţele camerei deputaţilor, ce se regăsesc în setul CDep-trans-v4, conţin în mare parte vorbire citită (similară cu cea din RSC şi diferită faţă de SSC) şi cuvinte dintr-un domeniu diferit faţă de cel întâlnit în seturile de evaluare (ştiri, interviuri, radio).

2.4 Activităţile 3.11 şi 3.13 - Îmbunătăţirea soluţiei pentru generarea de scoruri de încredere pentru RAV şi Analiza impactului utilizării scorurilor de încredere pentru filtrarea transcrierilor RAV în vederea reantrenării sistemelor RAV

În această secţiune descriem soluţia îmbunătăţită pentru estimarea scorurilor de încredere. Scopul nostru este de a înzestra sistemele automate cu estimări cât mai fiabile a încrederii în predicţiile pe care acestea le generează. Ideal, cu cât sistemul este mai încrezător într-o predicţie (adică un scor de încredere mai mare), cu atât ne aşteptăm ca rezultatul prezis să fie corect. Alternativ, putem rezolva problema complementară a estimării incertitudinii — în acest caz, cu cât este mai mare scorul de incertitudine, cu atât este mai probabil ca predicţia să fie eronată.

În contextul proiectului de faţă, şi anume, al recunoaşterii automate a vorbirii (RAV) şi al adnotării automate a datelor audio, estimarea încrederii este de o importanţă crucială pentru că ne semnalează corectitudinea datelor transcrise automat şi ne permite o filtrare pe baza acestui scor. Dar metodele pentru estimarea scorurilor de încredere pentru RAV au aplicaţii multiple, care merg dincolo de sarcina noastră de interes: îmbunătăţirea robusteţii sistemelor în sarcini critice de siguranţă, evitarea erorilor în sistemele de dialog om-maşină sau facilitarea corecţiilor manuale în sarcinile de transcriere audio prin semnalarea erorilor. Mai mult, lucrări anterioare de specialitate au valorificat estimările scorurilor de încredere pentru o serie de sarcini care depind de RAV: selectarea predicţiilor cu grad ridicat de încredere pentru reantrenarea sistemului de bază [Sperber, 2017], propagarea incertitudinilor în traducerea automată a vorbirii [Vesely, 2013], adnotarea manuală a predicţiilor mai puţin sigure pentru învăţarea activă [Yu, 2010].

Metoda pe care o propunem vizează estimarea încrederii pentru sistemele RAV de la un capăt la altul (en., end-to-end) [Hadian, 2018] — spre deosebire de soluţia de bază care consideră sistemele RAV de tip hibrid, DNN-HMM. Modelele RAV de tip end-to-end au câştigat masiv în popularitate în ultima perioadă nu doar datorită performanţei lor (care o egalează şi chiar depăşeşte pe cea a RAV-urilor de tip clasic), dar şi pentru avantajele suplimentare de a fi simple din punct de vedere conceptual şi de a permite un proces de antrenare unitar [Lüscher, 2019; Tüske, 2019; Karita, 2019]. Cu toate acestea, surprinzător de puţine lucrări tratează sarcina de estimare a încrederii în sistemele RAV de la un capăt la altul.

Metoda noastră adresează două provocări principale ale dezvoltării metodelor de estimare a încrederii pentru sistemele RAV: *(i)* ieşirea structurată (textul) şi *(ii)* predicţiile granulare (la nivel de grafem sau token în loc de cuvinte). Sistemele RAV sunt modele structurate (care mapează secvenţe la secvenţe) spre deosebire de reţelele obişnuite de recunoaştere (cum ar fi clasificarea imaginilor) a căror ieşire este o singură etichetă. Natura secvenţială a ieşirii impune o etapă de decodare, ceea ce complică nu numai predicţia, cât şi algoritmul de estimare a încrederii, deoarece acesta trebuie să opereze într-un context auto-regresiv (folosind secvenţa deja prezisă). Din acest motiv, obţinem predicţiile pe baza unui RAV preantrenat şi apoi aplicăm metodele de estimare a încrederii peste probabilităţile următorului simbol, care sunt condiţionate de transcrierea fixă.

Pentru a evita constrângerile impuse de utilizarea unui dicţionar de cuvinte fix şi pentru a permite predicţii de cuvinte noi, sistemele RAV de tip end-to-end de obicei folosesc la ieşire token-uri de subcuvinte. Dar având în vedere că token-urilor le lipseşte semantica, pentru multe aplicaţii finale suntem interesaţi să estimăm încrederea la nivelul cuvintelor şi nu a token-urilor. În acest scop, explorăm metode de agregare a măsurilor de incertitudine de la nivel de token la unităţi mai mari, corespunzând cuvintelor. Dar tehnicile prezentate în continuare pot fi aplicate şi la structuri chiar mai ample, cum ar fi nivelul propoziţiei sau al enunţului.

2.4.1 Legături cu starea artei

În această subsecţiune facem o scurtă prezentare a lucrărilor de specialitate relevante pentru sarcina şi metoda propusă. Considerăm două direcţii: metode pentru scoruri de încredere pentru recunoaşterea automată a vorbirii (RAV) şi metode pentru scoruri de încredere pentru sisteme de tip end-to-end.

Scoruri de încredere pentru recunoaşterea automată vorbirii. Cele mai multe lucrări anterioare privind scorurile de încredere pentru RAV vizează sisteme clasice, bazate pe paradigma HMM-GMM. Aceste metode extrag mai întâi un set de caracteristici din reţeaua de decodare, modelul acustic sau de limbă, şi apoi antrenează un clasificator pentru a prezice dacă transcrierea este corectă sau nu. Exemple tipice de caracteristici includ probabilitatea realizării acustice, scorul modelului de limbă, durata cuvântului, numărul de alternative din reţeaua de confuzie [Kemp, 1997; Weintraub, 1997; Hazen, 2002]. Mai recent, Swarup *et al.* [Swarup, 2019] a mărit setul de caracteristici folosind embedding-uri profunde ale semnalului acustic de intrare şi ale textului prezis, în timp ce Errattahi *et al.* [Errattahi, 2018] a arătat că adaptarea la domeniu a caracteristicilor extrase aduce beneficii de performanţă. Clasificatorii folosiţi de metodele de estimare a scorurilor de încredere variază de la conditional random fields [Seigel, 2013; Cortina, 2016] şi perceptron cu mai multe straturi [Kalgaonkar, 2015] la reţele neuronale recurente bidirecţionale [Ogawa, 2017; Del-Agua, 2018; Li, 2019].

Scoruri de încredere în sistemele end-to-end. Metoda de bază pentru estimarea încrederii în reţelele neuronale este de a utiliza în mod direct probabilitatea predicţiei celei mai probabile [Hendrycks, 2016]. S-a observat că reţelele neuronale tind să fie prea sigure şi estimările probabilităţii pot fi îmbunătăţite prin scalarea temperaturii [Hinton, 2015], ceea ce duce de obicei la o mai bună calibrare [Guo, 2017; Ashukha, 2020]. Cea mai promiţătoare direcţie în ceea ce priveşte simplitatea şi utilitatea implică estimarea Monte Carlo: Gal şi Ghahramani [Gal, 2016] folosesc tehnica dropout la momentul inferenţei pentru a obţine predicţii multiple, care sunt apoi mediate, în timp ce Lakshminarayanan *et al.* [Lakshminarayanan, 2017] fac media predicţiilor unui ansamblu de reţele de obicei antrenate cu iniţializări diferite. Această ultimă metodă s-a dovedit a fi foarte fiabilă pentru condiţiile dificile în care avem date din afara domeniului [Ovadia, 2019], dar este costisitoare din punct de vedere computaţional, deoarece implică antrenarea a multiple modele de RAV [Ashukha, 2020]. O abordare diferită a estimării încrederii este de a învăţa un clasificator (de obicei o altă reţea neuronală) direct deasupra activărilor reţelei de RAV [Corbière, 2019; Chen, 2019].

La intersecţia acestor două direcţii de cercetare, menţionăm lucrarea foarte recentă a lui Malinin şi Gales [Malinin, 2020], care similar cu noi abordează sarcina de estimare a încrederii pentru sistemele RAV end-to-end. Cu toate acestea, ei sunt preocupaţi de estimarea incertitudinii la nivel de token şi frază, în timp ce noi suntem interesaţi de estimarea la nivel de cuvânt, şi, în consecinţă, oferim mai multă atenţie asupra tehnicilor de agregare. Mai mult, ei folosesc ansambluri ca metodă principală de estimare a încrederii, în timp ce noi evaluăm şi metodele de reducere a temperaturii şi de scădere. Deşi tehnica de dropout a fost folosită anterior pentru obţinerea scorurilor de încredere pentru RAV [Vyas, 2019], metoda este diferită de abordarea noastră. În acea lucrare, autorii generează mai multe ipoteze prin dropout şi apoi atribuie confidenţe cuvintelor pe baza frecvenţei apariţiilor lor în ipotezele aliniate. În schimb, noi agregăm probabilităţile posterioare şi nu ipotezele, ceea ce simplifică procedura, deoarece evită pasul de aliniere.

2.4.2 Metodologia

În această subsecţiune prezentăm efectiv metodologia propusă pentru estimarea scorurilor de încredere şi modalităţi propuse de îmbunătăţire a acestora. Începem mai întâi cu o descriere generală a metodei şi a notaţiei implicate.

Considerăm un model de tip secvenţă-la-secvenţă (en. sequence-to-sequence) care mapează o secvenţă audio $\boldsymbol{a}$ la una de token-uri $\boldsymbol{t} = (t_1, \ldots, t_T)$. Modelul este specificat de parametri θ, care sunt învăţaţi prin minimizarea pe setul de antrenare a unei funcţii de pierdere, cum ar fi funcţia *connectionist temporal classification* (CTC) sau divergenţa Kullback-Leibler. La inferenţă, modelul generează probabilităţi pentru următorul simbol k într-o manieră autoregresivă $p(t_k|\hat{t}_{<k}, a; \theta)$, bazat pe token-urile deja prezise $\hat{t}_{<k}$. Aceste probabilităţi sunt utilizate pentru efectuarea decodării prin metoda beam search pentru a obţine cea mai probabilă secvenţă de token-uri. Având în vedere că probabilitatea de ieşire condiţionată este o distribuţie peste V token-uri din vocabular, o notăm cu un vector V-dimensional p_k.

Estimarea scorurilor de încredere

Scopul nostru este de a obţine un scor de încredere pentru fiecare cuvânt din transcrierea produsă de RAV. Realizăm acest lucru în doi paşi. Mai întâi, folosind probabilităţile aposteriori de la fiecare moment de timp p_k, extragem caracteristici pentru a reprezenta scorul de încredere al fiecărui token $s_k^{(t)}$. În cea de-a doua etapă, agregăm scorurile la nivel de token în scoruri de încredere la nivel de cuvânt $s_j^{(w)}$, pe baza token-urilor care aparţin fiecărui cuvânt. În continuare vom detalia aceşti doi paşi; vezi şi figura 2.4.a.

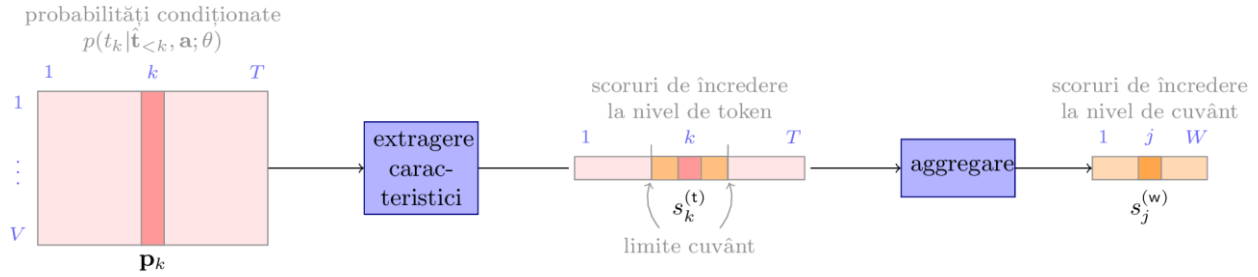

Figura 2.4.a *Prezentare schematică generală a metodei propuse de estimare a scorurilor de încredere. Pe baza unui sistem de recunoaştere a vorbirii (RAV) de tip end-to-end, obţinem probabilităţi p_k pentru fiecare token k condiţionate de o rostire a şi token-urile prezise anterior $\hat{t}_{<k}$. Pe baza acestor probabilităţi extragem scoruri de încredere la nivel de token $s^{(t)}$, pe care le agregăm apoi pentru a obţine scoruri la nivel de cuvânt $s^{(w)}$. Dimensiunea vocabularului de token-uri este notată cu V, numărul de token-uri este notat cu T şi numărul de cuvinte cu W.*

Extragerea caracteristicilor. Pentru a măsura încrederea într-o predicţie la nivel de token folosim două variante:

- Logaritmul probabilităţii (log-proba) celei mai probabile predicţii date de clasificator, adică $s^{(t)} = log\ max\ \boldsymbol{p}$. S-a observat empiric că acest tip de caracteristică oferă un baseline puternic pentru sarcinile de clasificare a greşelilor şi detectare a eşantioanelor din afara distribuţiei [Hendrycks, 2016].
- Entropie negativă (neg-entropie) calculată peste vocabularul token-urilor la fiecare moment de timp, adică $s^{(t)} = p^T log\ p$. O entropie mare înseamnă o incertitudine mare sau, invers, o entropie *negativă* mare implică o predicţie încrezătoare. Deşi entropia este de obicei utilizată ca măsură de încredere împreună cu tehnica de dropout [Gal, 2016], după cum vom vedea în curând, tehnica de dropout va putea fi folosită şi peste probabilităţile originale.

Agregarea. Pentru a obţine caracteristici la nivel de cuvânt din cele la nivel de simbol, experimentăm cu trei tipuri de funcţii de agregare: suma, media, minimul. Deoarece ambele caracteristici propuse sunt negative, însumarea a mai multe token-uri va duce la valori mai mici şi, prin urmare, la scoruri de încredere mai mici; acest comportament poate fi de dorit deoarece cuvintele mai lungi sunt mai susceptibile de a fi eronate. De asemenea, atunci când însumăm logaritmul probabilităţilor, obţinem un scor la nivel de cuvânt corespunzător probabilităţii log a întregii secvenţe. Utilizarea agregării cu funcţia de minim este justificată de faptul că am putea dori o încredere scăzută dacă cel puţin unul din token-uri are o încredere scăzută.

Îmbunătăţirea probabilităţilor la nivel de token

Propunem trei moduri de a face mai fiabile probabilităţile la nivel de token: scalarea temperaturii, tehnica de dropout şi ansamblurile de modele. Presupunerea noastră este că îmbunătăţind probabilităţile la nivel de token, vom obţine şi scoruri de confidenţă la nivel de cuvânt mai bune.

Scalarea temperaturii [Hinton, 2015; Guo, 2017] constă în împărţirea activărilor de tip logit (valorile de dinainte de stratul softmax) la un scalar τ (cunoscut sub numele de temperatură). Valoarea lui τvariază de la zero la infinit şi controlează forma distribuţiei: când $\tau \to 0$ obţinem o distribuţie uniformă, când $\tau \to \infty$ obţinem o distribuţie Dirac localizată pe cea mai probabilă ieşire. Pe baza temperaturii τ actualizăm probabilităţile la nivel de token la fiecare moment de timp, după cum urmează:

$$p'_k = softmax(log(p_k) \,/\, \tau)$$

Apoi extragem caracteristici $s^{(t)}$ peste probabilităţile actualizate, le agregăm în scorul la nivel de cuvânt $s^{(w)}$ şi, în cele din urmă, clasificăm cuvântul drept corect sau incorect:

$$P(correct) = \sigma(\alpha s^{(w)} + \beta)$$

Variabilele α, β, τ sunt parametri şi sunt învăţate prin optimizarea unei funcţii de pierdere de entropie încrucişată (en., *cross-entropy loss*) pe un set de validare. Etichetele sunt setate la nivel de cuvânt prin alinierea la textul de bază cu transcrierea. Remarcăm faptul că parametrii α şi β nu modifică ordinea predicţiilor, dar ne permit să învăţăm un model de estimare a încrederii calibrat.

Dropout [Srivastava, 2014] este o tehnică care maschează părţi aleatorii ale activărilor într-o reţea, făcând reţeaua mai puţin predispusă la supra-antrenare (en., *overfitting*). În [Gal, 2016] s-a observat că tehnica de dropout induce o distribuţie de probabilitate peste ponderile reţelei şi poate fi în consecinţă utilizată pentru inferenţa Bayesiană aproximativă. Folosim această idee şi calculăm probabilităţile la nivel de token obţinute prin mai multe rulări folosind dropout:

$$p'_k = \frac{1}{N}\sum_n \hat{p}_k$$

unde $\hat{p}$ specifică predicţia dropout-ului. Probabilităţile actualizate sunt apoi utilizate pentru a extrage oricare dintre caracteristicile de incertitudine propuse (log-proba sau neg-entropie).

Ansamblurile [Lakshminarayanan, 2017] se bazează pe aceeaşi idee de mediere a predicţiilor din mai multe surse (ca în cazul dropout-ului), dar în acest caz ponderile provin din reţele antrenate independent (folosind iniţializări diferite ale reţelei). În cazul nostru, calculăm media predicţiilor la nivel de token:

$$p'_k = \frac{1}{N}\sum_n p(t_k|\hat{t}_{<k}, a; \theta_n)$$

unde $\{\theta_n\}_{n=1\ldots N}$ specifică ansamblul de modele. Este important de subliniat că trebuie să avem acelaşi context pentru toate modelele din ansamblu, deci îl folosim pe cel dat de un model preantrenat.

Cele trei abordări prezentate pot fi combinate; de exemplu, mai întâi putem actualiza probabilităţile folosind scalarea temperaturii şi apoi media mai multe predicţii folosind tehnica de dropout.

2.4.3 Setup-ul experimental pentru experimente pe limba engleză

În această subsecţiune descriem bazele de date, sistemele de recunoaştere automată a vorbirii utilizate şi metricile de evaluare utilizate pentru evaluarea scorurilor de încredere pe limba engleză

Baze de date

Pentru experimentele pe limba engleză am ales mai multe seturi de date, toate disponibile public şi utilizate de comunitate mai ales pentru evaluarea sistemelor de recunoaştere a vorbirii.

LibriSpeech [Panayotov, 2015] este un corpus de aproximativ 1000 de ore de cărţi audio citite. Datele au fost derivate din proiectul LibriVox şi au fost atent segmentate şi aliniate. Folosim setul de date atât pentru antrenare, cât şi pentru evaluare. Pentru antrenare folosim cele trei părţi ale acestuia `clean100`, `clean360` şi `other500`, în timp ce pentru validare şi evaluare folosim părţile standard denumite `clean`, respectiv `other`.

TED-LIUM2 [Rousseau, 2014] constă în discursuri şi transcrierile acestora colectate de pe site-ul web TED. Utilizăm setul de date pentru evaluare şi, în consecinţă, folosim numai subseturile predefinite `dev` şi `test`.

CommonVoice [Ardila, 2020] este un set de date colaborativ cu transcrieri scurte care sunt citite de oameni din întreaga lume. Există mai multe versiuni ale setului de date şi am folosit prima versiune. Folosim setul de date pentru evaluare şi am definit subseturi de `dev` şi `test` prin alegerea a 10% din rostiri în mod aleatoriu pentru fiecare dintre ele.

Tabelul 2.4.a prezintă dimensiunile părţilor de `test` pentru fiecare set de date de evaluare.

Tabelul 2.4.a *Dimensiunea seturilor de date (partea* `test`*) care sunt folosite pentru evaluarea metodelor pentru scorurile de încredere pe limba engleză.*

Bază de date	Num. rostiri	Durată (ore)
Libri clean	2.6K	5.4
Libri other	2.9K	5.3
TED	1.1K	2.6
CommonVoice	66K	72

Sistemul de recunoaştere automată a vorbirii

Sistemul de recunoaştere automată a vorbirii (RAV) este implementat folosind librăria ESPnet [Watanabe, 2018]. Modelul implementează arhitectura de tip Transformer [Vaswani, 2017] şi primeşte la intrare un banc de filtre Mel 80-dimensionale (extrase cu utilitarul Kaldi [Povey, 2011]) şi produce la ieşire o secvenţă de token-uri. Vocabularul are dimensiunea de 5,000 de token-uri şi este obţinut prin segmentarea cuvintelor pe baza unui model de limbaj de tip unigram [Kudo, 2018]. Modelul este antrenat pe 960 de ore din baza de date LibriSpeech [Panayotov, 2015], iar datele sunt augmentate utilizând tehnicile SpecAugment (modificarea vitezei vorbirii, mascare în frecvenţă, mascare în timp) [Park, 2019]. Pentru decodare folosim un model de limbă, care este implementat tot ca Transformer şi este antrenat pe 14,500 de cărţi din domeniu public [Panayotov, 2015]. Vocabularul modelului de limbă constă din aceleaşi 5,000 de token-uri ca şi modelul RAV.

Pentru experimentele cu ansambluri de modele, reantrenăm sistemul de RAV folosind aceeaşi arhitectură şi date, dar iniţializări ale ponderilor diferite. Repetăm procesul de patru ori obţinând patru modele independente. Datorită constrângerilor de calcul, aceste modele au fost antrenate pentru un număr mai scurt de epoci decât sistemul principal (10 versus 120), dar am observat că curba funcţiei de pierdere a validării a început să se aplatizeze şi că performanţa sistemului de RAV este rezonabilă (5.5%±0.4 WER pe Libri `clean` faţă de 2.7% obţinut de modelul preantrenat).

Metrici de evaluare

În mod ideal, ne dorim ca scorurile de încredere să fie corelate cu corectitudinea transcrierii, adică cuvintele corecte ar trebui să aibă un scor de încredere mare, în timp ce cuvintele incorecte, un scor scăzut. Pe baza lucrărilor anterioare [Hendrycks, 2016; Corbière, 2019; Malinin, 2020], folosim metrici care sunt utilizate în general pentru evaluarea clasificatorilor binari, dar variază pragul de discriminare între cele două clase. Mai precis, măsurăm aria de sub curba de precision-recall (area under precision-recall curve; AUPR) şi aria de sub curba receiver operator curve (area under receiver operator curve; AUROC). Cu toate acestea, în funcţie de ceea ce dorim să ne concentrăm (erori sau predicţii corecte) obţinem două variante: dacă suntem interesaţi de clasificări greşite, vom trata erorile ca pe o clasă pozitivă; pe de altă parte, dacă suntem interesaţi de clasificarea corectă, vom trata detecţiile reuşite ca fiind clasa pozitivă. Prin urmare, pentru AUPR folosim două variante AUPR*e* (când erorile sunt tratate ca pozitive) şi AUPR*s* (când succesele sunt tratate ca pozitive). Pentru AUROC se obţine aceeaşi valoare pentru ambele opţiuni, deci nu este necesar să se facă această distincţie.

2.4.4 Rezultate experimentale pentru limba engleză

Această secţiune prezintă rezultatele experimentale pe bazele de date în limba engleză. Începem cu o evaluare a caracteristicilor şi a agregărilor acestora; după care raportăm rezultatele pentru variantele îmbunătăţite: mai întâi, pentru tehnicile de scalare a temperaturii şi dropout, apoi pentru ansamblurile de modele. Încheiem secţiunea cu o discuţie privind alte legături cu starea artei şi posibile direcţii.

Caracteristici şi metode de agregare

Evaluăm caracteristicile pentru reprezentarea incertitudinii din token-uri şi tehnicile de agregare propuse pe cele patru seturi de date descrise în secţiunea precedentă. Folosim modelul preantrenat pentru a obţine predicţii de text pentru toate fişierele audio din partea de testare al fiecărui set de date, şi apoi estimăm

încrederea pe baza metodologiei propuse. Tabelul 2.4.b prezintă rezultatele pentru toate combinaţiile de caracteristici şi agregări.

Comparaţie a caracteristicilor. Observăm că prin folosirea caracteristicile de tip probabilităţi log se obţin rezultate mai bune faţă de situaţia utilizării caracteristicilor pe bază de entropie în toate cazurile (indiferent de agregare sau setul de date). Singura excepţie notabilă este setul de date CommonVoice în care rezultatele sunt comparabile.

Comparaţie a metodelor de agregare. În general, agregarea folosind suma funcţionează mai bine cu caracteristicile de tip log-proba, în timp ce agregarea folosind minimul funcţionează mai bine pentru caracteristicile de entropie. Suma s-ar putea să nu fie potrivită pentru caracteristicile de entropie, deoarece magnitudinea lor este mai mare decât în cazul caracteristicilor de tip log-proba, iar cuvintele lungi sunt penalizate prea mult prin lungime; dar, aşa cum vom vedea mai departe, acest comportament poate fi atenuat prin metoda de scalare a temperaturii. Media este, în general, slabă pentru ambele caracteristici, sugerând că măsurile invariante în lungime sunt dăunătoare. Într-adevăr, o privire mai atentă la frecvenţa erorilor cu dimensiunea lungimii indică următorul fapt: cu cât un cuvânt este format din mai multe token-uri, cu atât este mai probabil că este incorect, vezi figura 2.4.b.

Comparaţie între seturile de date. Aşa cum era de aşteptat, modelul preantrenat are cea mai bună performanţă în ceea ce priveşte datele din domeniu (2.7% WER pe Libri `clean` şi 6.0% pe Libri `other`), performanţa scăzând apoi brusc, pe măsură ce evaluăm datele din afara domeniului (13.3% pe TED şi 28.6% pe CommonVoice). În fiecare dintre aceste situaţii, numărul de cuvinte care sunt clasificate corect se schimbă, trecând de la mai multe pe baza de date Libri la mai puţine pe seturile de date TED şi CommonVoice. Această observaţie explică de ce performanţa pentru AUPRs scade în funcţie de domeniul datelor, şi, dimpotrivă, de ce se îmbunătăţeşte performanţa AUPRe. Din păcate, din exact acest motiv—performanţa diferită a sistemului RAV de bază pe cele patru seturi de date—este imposibil de comparat metodele de încredere între seturile de date, deoarece acestea utilizează o altă adnotare [Ashukha, 2020].

Tabelul 2.4.b *Rezultate pentru evaluarea scorurilor de încredere pentru toate combinaţiile de caracteristici (caract.) si metode de agregare (agreg.) pe cele patru baze de date în engleză. Pentru toate cele trei metrici de evaluare utilizate (AUPRe, AUPRs, AUROC) valorile mai mari indică performanţă mai bună. Raportăm eroarea la nivel de cuvânt pentru sistemul de RAV preantrenat pe fiecare din seturile de date; aceasta este listată în dreapta numelului bazei de date.*

		Libri clean / 2.7%			**Libri other / 6.0%**		
caract.	**agg.**	**AU-PR*e***	**AU-PR*s***	**AU-ROC**	**AU-PR*e***	**AU-PR*s***	**AU-ROC**
log-proba	sum	21.55	99.21	82.41	29.99	98.10	81.75
log-proba	min	21.85	99.19	82.47	28.64	98.06	81.66
log-proba	avg	20.12	99.10	80.90	26.72	97.93	80.47
neg-entropy	sum	17.31	99.10	79.97	26.37	97.86	79.58
neg-entropy	min	19.94	99.09	80.55	26.75	97.82	79.64
neg-entropy	avg	17.55	98.95	77.72	24.26	97.59	77.46
		TED / 13.3%			**CommonVoice / 28.6%**		
caract.	**agg.**	**AU-PR*e***	**AU-PR*s***	**AU-ROC**	**AU-PR*e***	**AU-PR*s***	**AU-ROC**
log-proba	sum	39.97	95.88	79.95	48.98	77.71	64.84
log-proba	min	39.74	95.94	80.58	46.79	76.74	62.67
log-proba	avg	38.74	95.88	80.29	44.51	75.82	60.87
neg-entropy	sum	34.96	95.41	77.57	47.71	77.10	63.74
neg-entropy	min	37.55	95.56	79.01	45.51	76.00	61.21
neg-entropy	avg	36.28	95.42	78.29	42.64	74.83	58.75

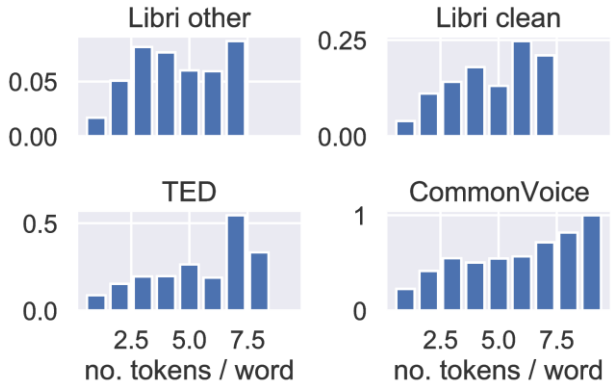

Figura 2.4.b *Raportul de erori ca o funcţie de lungimea cuvântului. Raportul de erori este calculat ca numărul de cuvinte eronate împărţit la numărul total de cuvinte, iar lungimea cuvintelor este măsurată ca număr de token-uri.*

Scalarea temperaturii şi tehnica de dropout

Evaluăm metodele de estimare a scorurilor de încredere după îmbunătăţirea probabilităţilor token-urilor prin două din tehnicile descrise: scalarea temperaturii şi tehnica de dropout. Folosim sistemul RAV preantrenat şi raportăm rezultatele pe setul de testare TED. Parametrii pentru metoda de scalare a temperaturii sunt învăţaţi pe partea dev a setului de date TED pentru fiecare setare de caracteristică şi metodă de agregare. Când scalarea temperaturii este combinată cu tehnica de dropout, aplicăm mai întâi scalarea temperaturii (folosind aceeaşi temperatură) şi apoi mediem probabilităţile obţinute prin dropout. Pentru dropout mediem 64 de predicţii independente. Tabelul 2.4.c prezintă rezultatele pentru toate combinaţiile de caracteristici, agregări şi tehnici de îmbunătăţire. Rezultatele indică faptul că ambele metode propuse îmbunătăţesc rezultatele la fel ca şi combinaţia lor, ceea ce oferă în general cel mai bun rezultat. Observăm că tehnica de dropout aduce îmbunătăţiri mai mari pentru caracteristicile log-proba, în timp ce caracteristica neg-entropie produce rezultate mai bune atunci când se utilizează scalarea temperaturii. Interesant, cele mai bune rezultate sunt acum obţinute pentru neg-entropie cu agregare sumă (rândul 16). Figura 2.4.c arată că performanţa dropout-ului se îmbunătăţeşte cu numărul de rulări şi se plafonează în jurul valorii alese de 64.

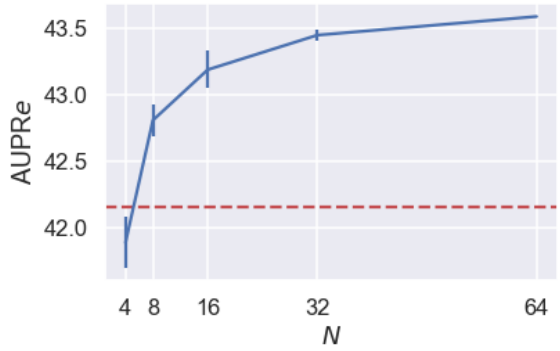

Figura 2.4.c *Performanţa AUPRe ca funcţie de numărul de rulări de dropout N pe baza de date TED. Linia roşie orizontală indică rezultatul metodei fără dropout. Modelul foloseşte caracteristici de tip neg-entropy, agregarea pe bază de sumă şi scalare a temperaturii.*

Ansambluri de modele

În cele ce urmează, prezentăm rezultate pentru estimarea scorurilor de încredere folosind ansambluri de modele şi combinaţiile acestora cu celelalte versiuni îmbunătăţite (scalarea temperaturii şi tehnica de dropout). Pentru fiecare dintre modelele reantrenate, care fac parte din ansamblu, folosim predicţiile modelului preantrenat pentru a selecta transcrierea pentru care dorim să estimăm scoruri de încredere; deci modelul reantrenat este folosit doar pentru scorul de încredere, prin extragerea caracteristicilor de încredere descrise anterior. Rezultatele sunt prezentate în tabelul 2.4.d. Pentru rândurile care nu folosesc ansamblu (rândurile 1, 2, 3 şi 5 din tabel) evaluăm fiecare dintre cele patru modele individuale în mod independent şi raportăm performanţa medie. Modelul preantrenat (tabelul 2.4.c, rândul 13) are, în general, o performanţă mai bună decât cele reantrenate (tabelul 2.4.d, rândul 1), sugerând că performanţa predictivă a unui model se

Tabelul 2.4.c. *Evaluarea scorurilor de încredere pe baza de date TED pentru combinaţii de caracteristici, metode de agregare şi tehnici de îmbunătăţire a probabilităţilor: scalarea temperaturii (ST) şi tehnica de dropout (D). Punctul negru "•" indică dacă e folosită o anumită tehnică, iar linia "–" indică opusul.*

nr.	caract.	agreg.	ST	D	AUPRe	AUPRs	AUROC
1	log-proba	sum	–	–	39.97	95.88	79.95
2			–	•	41.41	96.81	82.78
3			•	–	40.92	96.19	81.11
4			•	•	42.99	97.14	84.10
5	log-proba	min	–	–	39.74	95.94	80.58
6			–	•	42.08	96.94	83.76
7			•	–	39.84	95.98	80.74
8			•	•	42.17	97.00	83.93
9	log-proba	avg	–	–	38.74	95.88	80.29
10			–	•	41.19	96.95	83.73
11			•	–	38.97	95.99	80.66
12			•	•	41.32	97.06	84.08
13	neg-entropy	sum	–	–	34.96	95.41	77.57
14			–	•	33.14	96.22	79.45
15			•	–	42.16	96.91	83.50
16			•	•	43.59	97.62	85.51
17	neg-entropy	min	–	–	37.55	95.56	79.01
18			–	•	38.75	96.53	81.98
19			•	–	41.23	96.87	83.50
20			•	•	42.23	97.60	85.51
21	neg-entropy	avg	–	–	36.28	95.42	78.29
22			–	•	38.01	96.51	81.85
23			•	–	40.22	96.53	82.48
24			•	•	41.15	97.43	85.18

corelează cu performanţa sa de estimare a încrederii. Printre cele trei metode de îmbunătăţire propuse, observăm că scalarea temperaturii oferă cea mai mare creştere a performanţei pentru toate cele trei valori (rândul 2). În mod surprinzător, metoda dropout-ului îmbunătăţeşte numai performanţa AUPR faţă de linia de bază (rândul 3). În combinaţiile de două metode, scalarea temperaturii şi ansamblul se completează reciproc şi obţin performanţe mai bune.

Tabelul 2.4.d *Evaluarea scorurilor de încredere pe baza de date TED pentru combinaţii ale metodelor îmbunătăţite: scalarea temperaturii (ST), dropout (D) şi ansambluri (A). Am folosit caracteristica neg-entropy şi suma ca agregare. Punctul negru "•" indică dacă e folosită o anumită tehnică, iar linia "–" indică opusul.*

nr.	ST	D	A	AUPRe	AUPRs	AUROC
1	–	–	–	28.58	95.30	75.79
2	•	–	–	32.00	96.32	79.47
3	–	•	–	27.49	95.51	75.67
4	–	–	•	30.89	96.26	78.89
5	•	•	–	31.10	96.40	79.06
6	•	–	•	34.57	96.95	81.64
7	–	•	•	28.94	96.26	77.93
8	•	•	•	33.00	96.84	80.82

În continuare, discutăm pe scurt diferite perspective legate de metodologia propusă.

Mărirea setului de caracteristici. Caracteristicile pe care le-am investigat au avantajul de a fi generale, pentru că valorifică probabilităţile la nivel de token, care sunt disponibile în majoritatea, dacă nu în toate, utilitarele pentru RAV-uri de tip end-to-end existente. Cu toate acestea, setul de caracteristici ar putea fi extins cu probabilităţi pentru sunetul de intrare sau textul generat, sau cu informaţii despre durata cuvântului, care pot fi extrase din ponderile de atenţie.

Învăţarea caracteristicilor. Inspiraţi de studiile precedente [Corbière, 2019; Chen, 2019], am experimentat, de asemenea, cu învăţarea unei reţele de estimare a încrederii pe baza caracteristicilor extrase din modelul end-to-end (mai exact, activări logit şi pre-logit). Cu toate acestea, experimentele noastre nu au reuşit să arate îmbunătăţiri faţă de rezultatele prezentate.

Tratarea cuvintelor pierdute. Pentru a genera etichete pentru scorurile de încredere, aliniem textul de referinţă la textul prezis şi marcăm cuvintele corecte din textul prezis ca pozitive şi substituţiile şi inserţiile ca negative. Această abordare este tipică în literatura de evaluare a scorurilor de încredere, dar ratează erorile făcute prin ştergerea (pierderea) cuvintelor din referinţă. Mai multe lucrări au abordat această problemă [Seigel, 2014; Ragni, 2018]; lăsăm pentru viitor să extindem abordarea curentă pentru această sarcină.

2.4.5 Setup-ul experimental pentru experimente pe limba română

Pentru experimentele pe limba română folosim seturile de date de evaluare prezentate în secţiunea 2.1.1: RSC-eval (5.5 ore de vorbire citită) şi SSC-eval1+2 (3.5+1.5 ore de vorbire spontană).

Sistemul de transcriere de vorbire folosit pentru aceste experimente este cel introdus în secţiunea 2.2.1. Sistemul RAV este implementat folosind librăria ESPnet şi se bazează pe o arhitectură de tip Transformer. Vocabularul sistemului cuprinde 1,000 de subcuvinte şi a fost extras din transcrierile materialelor audio cu care a fost antrenat întregul sistem: cele 517 ore de vorbire din seturile de date RSC-train, SSC-train1+2 şi SSC-train3+4-trans-v4. Pentru decodare se foloseşte un model de limbă adiţional implementat tot ca Transformer şi antrenat pe cele ~352M de cuvinte din seturile de date news2017 şi talkshows. Performanţa sistemului de RAV folosit, exprimată sub forma erorii la nivel de cuvânt (WER), este de 3.4% WER pe RSC-eval, 15.3% WER pe SSC-eval1, respectiv 23.5% WER pe SSC-eval2.

Din cauza complexităţii modelului CTC-Transformer şi timpului mare de inferenţă pe configuraţia hardware disponibilă în acest proiect, nu s-au mai putut antrena modele adiţionale cu iniţializari diferite ale ponderilor, pentru ansamblurile de modele.

Metricile de evaluare introduse în secţiunea anterioară (*i.e.* AUROC, AUPRs, AUPRs) sunt universale, independente de limbă şi, în consecinţă, vor fi folosite şi pentru experimentele pe limba română.

2.4.6 Rezultate experimentale pentru limba română

Pentru limba română am reluat următoarele experimente realizate şi pentru sistemul de estimare a încrederii pentru limba engleză:

- evaluarea diverselor caracteristici ce pot fi utilizate pentru estimarea încrederii (Tabelul 2.4.e);
- evaluarea diverselor metode de agregare a caracteristicilor la nivel de subcuvânt pentru a obţine scoruri de încredere la nivel de cuvânt (Tabelul 2.4.e);
- evaluarea tehnicii dropout pentru îmbunătăţirea scorurilor de încredere de mai sus Tabelul 2.4.f).

Caracteristici şi metode de agregare

În această secţiune evaluăm caracteristicile pentru reprezentarea incertitudinii din subcuvinte şi tehnicile de agregare propuse pe cele trei seturi de date prezentate în secţiunea 2.1.1: RSC-eval (5.5 ore de vorbire citită) şi SSC-eval1+2 (3.5+1.5 ore de vorbire spontană). Tabelul 2.4.e prezintă rezultatele pentru toate combinaţiile de caracteristici şi agregări.

Tabelul 2.4.e *Rezultate pentru evaluarea scorurilor de încredere pentru toate combinaţiile de caracteristici (caract.) si metode de agregare (agg.), pe cele trei seturi de evaluare în limba română. Pentru toate cele trei metrici de evaluare utilizate (AU-PRe, AU-PRs, AU-ROC) valorile mai mari indică performanţă mai bună.*

		RSC-eval / 1.8%			**SSC-eval1 / 11.0%**			**SSC-eval2 / 14.0%**		
caract.	**agg.**	**AU-PR*e***	**AU-PR*s***	**AU-ROC**	**AU-PR*e***	**AU-PR*s***	**AU-ROC**	**AU-PR*e***	**AU-PR*s***	**AU-ROC**
log-proba	sum	15.81	98.77	74.50	27.19	96.14	74.57	15.79	91.90	61.52
	min	18.30	98.91	78.66	28.32	96.34	77.24	13.86	91.99	60.46
	avg	16.22	98.74	75.75	26.69	96.04	75.38	14.04	92.09	60.60
neg -entropy	sum	15.82	98.73	75.65	26.44	96.03	75.31	14.30	92.17	61.09
	min	16.28	98.86	77.79	27.10	96.13	76.14	12.75	91.47	57.94
	avg	13.55	98.65	74.02	24.97	95.75	73.62	13.37	91.68	58.88

Comparaţie a caracteristicilor. În ceea ce priveşte tipul de caracteristici (log-proba vs. neg-entropy), rezultatele pe seturile de date în limba română indică exact aceleaşi concluzii ca şi în cazul limbii engleze:

- pe seturile de date pe care sistemul RAV are o acurateţe relativ bună (în cazul nostru RSC-eval şi SSC-eval1) caracteristicile de tip probabilităţi log obţin rezultate mai bune decât caracteristicile pe bază de entropie indiferent de agregare
- pe seturile de date mai dificile (SSC-eval2 pentru română, respectiv CommonVoice pentru engleză) caracteristicile bazate pe entropie agregate cu metoda *sum* prezintă rezultate similare cu, uneori chiar mai bune decât caracteristicile de tip probabilităţi log.

Comparaţie a metodelor de agregare. În ceea ce priveşte metodele de agregare, concluziile sunt relativ diferite pentru limba română:

- pe limba română agregarea minimul funcţionează mai bine indiferent de caracteristici, în timp ce pe limba engleză caracteristicile de tip probabilităţi log erau agregate mai bine folosind suma;
- agregarea prin medie este, în general, slabă pentru ambele caracteristici, sugerând că măsurile invariante în lungime sunt dăunătoare;
- excepţii de la concluziile de mai sus apar din nou pentru seturile de date mai dificile (SSC-eval2 pentru română, respectiv CommonVoice pentru engleză) pentru care agregarea prin sumă produce cele mai bune rezultate, indiferent de caracteristicile utilizate.

Comparaţie între seturile de date. Aşa cum era de aşteptat, modelul preantrenat are cea mai bună performanţă în ceea ce priveşte datele din domeniu (1.8% WER pe RSC-eval şi 11.0% pe SSC-eval1) şi performanţă mai slabă pe date din afara domeniului (14.0% pe SSC-eval2). În fiecare dintre aceste setări, numărul de cuvinte care sunt clasificate corect se schimbă, trecând de la mai multe pe setul RSC la mai

puţine pe seturile de date SSC. Această observaţie explică de ce performanţa pentru AUPRs este mare pe RSC şi scade pe SSC-eval1 şi 2, şi, dimpotrivă, de ce performanţa pentru AUPRe este mai slabă pentru RSC şi mai mică pentru SSC-eval1. Din păcate, din exact acest motiv—performanţa diferită a sistemului RAV de bază pe cele patru seturi de date—este imposibil de comparat metodele de încredere între seturile de date. Performanţa pentru AUPR*e* pe setul de date SSC-eval2 face din nou notă discordantă: dat fiind faptul că sistemul RAV are performanţe slabe de transcriere pe acest set de date ne-am fi aşteptat ca acele cuvinte transcrise greşit să poată fi foarte uşor de identificat (un AURP*e* de valoare mare), însă acest lucru nu se întâmplă.

Tehnica de dropout

În această secţiune evaluăm metodele de estimare a scorurilor de încredere după îmbunătăţirea probabilităţilor subcuvintelor prin tehnica de dropout. Această tehnică presupune transcrierea datelor de intrare de N=4, 8, 16, 32, respectiv 64 de ori folosind dropout. Mediem apoi cele N predicţii independente. Tabelul 2.4.f prezintă rezultatele pentru toate combinaţiile de caracteristici şi agregări şi dropout cu N=64.

Tabelul 2.4.f *Rezultate pentru evaluarea tehnicii dropout (N=64) pentru îmbunătăţirea scorurilor de încredere.*

		RSC-eval / 1.8%			SSC-eval1 / 11.0%			SSC-eval2 / 14.0%		
caract.	**agg.**	**AU-PR*e***	**AU-PR*s***	**AU-ROC**	**AU-PR*e***	**AU-PR*s***	**AU-ROC**	**AU-PR*e***	**AU-PR*s***	**AU-ROC**
log-proba	sum	17.76	98.97	77.14	28.96	96.85	77.31	16.03	90.08	58.47
	min	18.90	99.17	82.02	29.01	97.10	79.90	13.58	90.73	56.77
	avg	15.73	99.13	81.02	27.75	96.95	79.22	13.02	90.78	56.22
neg -entropy	sum	9.53	98.75	71.19	20.40	96.20	71.68	15.18	90.45	56.80
	min	14.43	99.11	80.63	27.01	96.90	78.86	11.85	90.22	53.74
	avg	11.83	99.02	78.71	25.50	96.70	77.84	11.94	90.36	53.85

Comparând rezultatele din Tabelul 2.4.e cu rezultatele din Tabelul 2.4.f observăm îmbunătăţiri pentru seturile de date RSC-eval şi SSC-eval1, însă o scădere de performanţă pentru setul de date SSC-eval2. Nici îmbunătăţirile obţinute pe RSC-eval şi SSC-eval1 nu sunt foarte mari (AU-ROC creşte de la 78.66 la 82.02 pe RSC-eval şi de la 77.24 la 79.90 pe SSC-eval1).

Toate metricile de performanţă se îmbunătăţesc pe măsură ce sunt mediate mai multe rezultate de transcriere (pe măsură ce N creşte) şi, la fel ca şi pentru limba engleză, creşterile de performanţă se plafonează pentru N=64.

Un ultim considerent demn de menţionat este acela că pentru a utiliza tehnica dropout în estimarea încrederii este nevoie de un timp de calcul net superior situaţiei iniţiale. Practic, pentru N=64 trebuie realizate 64 de transcrieri, deci timpul de calcul va fi de aproximativ 64 de ori mai mare. În concluzie, avantajele minime obţinute prin utilizarea acestei tehnici pentru îmbunătăţirea probabilităţilor subcuvintelor sunt contrabalansate de un dezavantaj semnificativ ce ţine de timpul de calcul.

2.4.7 Utilizarea scorurilor de încredere propuse pentru generarea de date adnotate

În activitatea 2.12 din etapa 2019 am propus utilizarea scorurilor de încredere la nivel de cuvânt puse la dispoziţie de un sistem de RAV pentru a selecta cuvintele presupus transcrise corect de acesta. Selecţia se face impunând un prag minim pentru scorul de încredere la nivel de cuvânt, prag sub care cuvântul în cauză este considerat ca fiind transcris greşit. Secvenţele de cuvinte presupus transcrise corect, împreună cu datele audio corespunzătoare, ar urma să formeze un nou set de date adnotat. Pentru ca aplicarea acestei metode să genereze seturi de date adnotate suficient de mari, este necesar ca setul de date neadnotat utilizat la intrarea sistemului să aibă o dimensiune considerabilă. De exemplu, în acest proiect am utilizat seturile de date SSC-train3-raw şi SSC-train4-raw care însumează 913 ore de vorbire. Metoda descrisă mai sus se poate aplica numai în cazul în care transcrierea acestui set de date neadnotat cu ajutorul sistemului de RAV se poate

realiza în mod eficient. Din păcate experimentele noastre au arătat că sistemul RAV ESPnet transcrie o oră de vorbire într-un interval de 2 până la 7 ore, în funcţie de cât de similară este acustica respectivei ore de vorbire raportat la setul de date de antrenare. În acest context, transcrierea setului de date de 913 ore ar fi durat între 76 de zile şi 266 de zile. În consecinţă, din considerente ce ţin de infrastructura de calcul, nu am putut utiliza noile metode de estimare a scorurilor de încredere pentru generare de date audio adnotate.

2.5 Activitatea 3.14 - Diseminare

Diseminarea rezultatelor proiectului a fost realizată prin intermediul website-ului proiectului (https://tadarav.speed.pub.ro) şi prin publicarea mai multor articole ştiinţifice după cum urmează:

1. A.-L. Georgescu, H. Cucu, A. Buzo, C. Burileanu, "*RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition*," in the Proceedings of The 12th Language Resources and Evaluation Conference (LREC), pp. 6606-6612, 2020, Marseille, France.
2. C. Manolache, A.-L. Georgescu, A. Caranica, H. Cucu, "*Automatic Annotation of Speech Corpora using Approximate Transcripts*," in the Proceedings of the 43rd International Conference on Telecommunications and Signal Processing (TSP), 2020, Milano, Italy.
3. D. Oneaţă, A.-L. Georgescu, H. Cucu, D. Burileanu, C. Burileanu, "*Revisiting SincNet: An Evaluation of Feature and Network Hyperparameters for Speaker Recognition*," in the Proceedings of the 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 2020.
4. G. Pop, H. Cucu, D. Burileanu, C. Burileanu, "*Cough Sound Recognition in Respiratory Disease Epidemics*," in Romanian Journal of Information Science and Technology, vol. 23, no. S, pp. S77–S89, 2020, ISSN 1453-8245, ISI IF 0.661.
5. A.-L. Georgescu,, C. Manolache, D. Oneaţă, H. Cucu, C. Burileanu, "*Data-filtering methods for self-training of automatic speech recognition systems*," in the Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Virtual, 2021.
6. D. Oneaţă, A. Caranica, A. Stan, H. Cucu, "*An evaluation of word-level confidence estimation for end-to-end automatic speech recognition*," in the Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Virtual, 2021.

Dintre articolele listate mai sus, al patrulea este deja indexat în Web of Science (Thompson Reuters - ISI), al doilea este deja indexat IEEE Xplore şi în curs de indexare în Web of Science, primul şi al treilea au apărut în volumul conferinţelor respective şi sunt în curs de indexare în Web of Science, iar al cincilea şi al şaselea vor apărea în volumul conferinţei respective şi vor fi indexare în Web of Science.

2.6 Crearea unui sistem de RAV îmbunătăţit

2.6.1 Actualizarea modelelor de limbă pentru transcriere de vorbire

În această secţiune sunt prezentate sumar demersurile realizate de colectivul SpeeD în vederea actualizării modelelor de limbă utilizate în sistemele de transcriere de vorbire. Etapele parcurse şi descrise în cele ce urmează sunt: (i) achiziţia unui nou corpus de text pentru actualizarea modelelor de limbă; (ii) actualizarea aplicaţiei de preprocesare a textului de antrenare pentru modelele de limbă; (iii) crearea unor noi modele de limbă; (iv) îmbunătăţiri aduse sistemelor RAV prin utilizarea noilor modele de limbă. Aceste aspecte sunt prezentate în detaliu în [Manolache, 2020].

Un nou corpus de text, format din articole scrise, a fost achiziţionat de pe mai multe website-uri de ştiri româneşti folosind o aplicaţie Java. Această aplicaţie a funcţionat în mod continuu între 26.06.2018 - 22.05.2020, verificând periodic feed-urile RSS a 24 de website-uri de ştiri şi descărcând orice articol întâlnit. Fiecare sursă de ştiri este extrasă pe un fir de execuţie separat, făcând astfel procesul de extracţie mai rapid. Procedura de extracţie constă în descărcarea conţinutului HTML, analiza acestuia în vederea determinării tag-urilor unde ar trebui să se regăsească textul şi apoi preluarea textului propriu-zis.

Acest corpus nou de text, denumit mai departe news2020, constă în text brut, organizat pe propoziţii ce însumează aproximativ 255M de cuvinte. Cele 24 surse de ştiri ce formează corpusul news2020 sunt prezentate în figura 2.6.a.

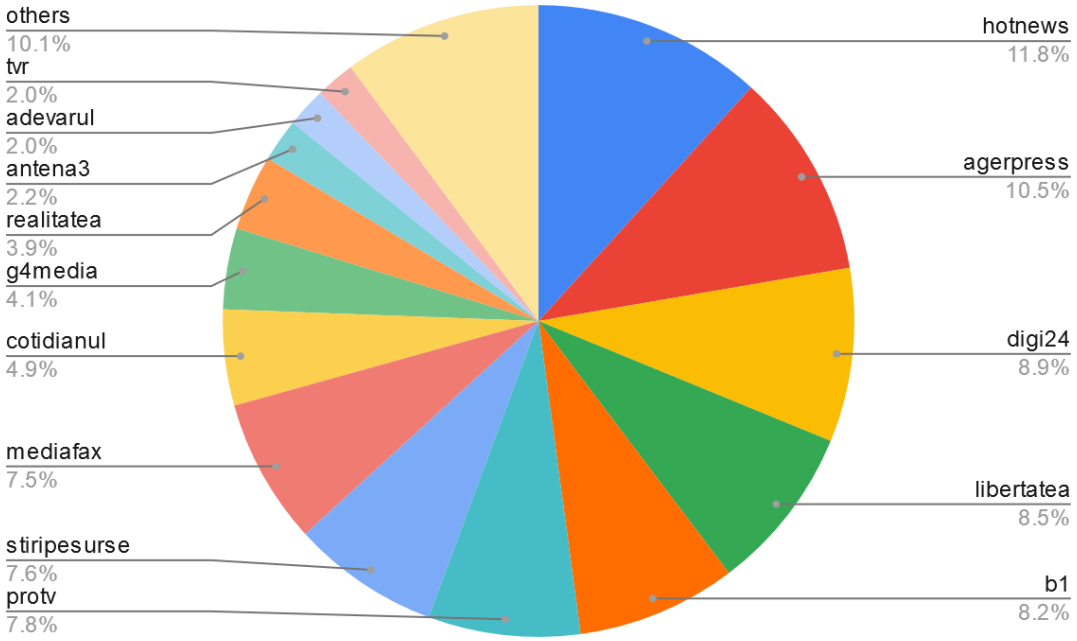

Figura 2.6.a *Sursele de ştiri componente ale corpusului news2020. Procentele sunt relative la dimensiunea întregului corpus (255M cuvinte)*

Aplicaţia de procesare a limbajului natural ce are scopul de normalizare a seturilor de date text ce urmează a fi folosite pentru antrenarea modelelor de limbă a fost supusă mai multor actualizări. Cea mai notabilă actualizare este legată de procesarea cuvintelor ce conţin cratimă. Această actualizare a fost necesară deoarece s-au întâmpinat mai multe cazuri în care sistemul RAV genera cuvinte care nu conţineau cratimă precum: ”ratat o”, ”lăsându se”, ”TVA ul”. Cauza principală a acestei probleme a fost aplicaţia de procesare a limbajului natural folosită pentru preprocesarea textului folosit mai departe pentru antrenarea modelelor de limbă folosite în sistemele RAV. Vechea aplicaţie de procesare a limbajului natural căuta cuvintele cu cratimă într-un lexicon dedicat. Dacă cuvântul era găsit în lexicon, acesta rămânea neschimbat, în caz contrar, cratima era înlocuită cu spaţiu (e.g. “lovit o”, “lăsându se”, “FMI ul”). Lexiconul dedicat pentru cuvintele cu cratimă nu conţinea toate cuvintele posibile cu cratimă deoarece acest lucru ar fi însemnat ca vocabularul pentru modelul de limbă să crească exponenţial. De exemplu pentru acronimul SMURD, mai multe forme ar fi trebui inserate în vocabular: SMURD-ul, SMURD-ului. Alt exemplu mai problematic îl reprezintă diferitele forme ale verbelor (e.g. gerunziu, imperativ): “lăsându-mă”, “lăsându-te”, “lăsându-l”, “nelăsându-ne”. Din cauza acestei probleme, lexiconul a fost limitat la cele mai frecvente cuvinte cu cratimă, iar restul cuvintelor întâlnite în text aveau cratima înlocuită cu spaţiu. Acest fapt a dus la crearea unui model de limbă ce conţine cuvinte incomplete: “ul”, “ului”, “lăsându” ce urmau să apară la în textul de la ieşirea sistemului ASR ca şi erori. Soluţia la această problemă a constat în modificarea procedurii de procesare a cuvintelor cu cratimă. Lexiconul a fost păstrat, dar numai pentru cuvinte compuse, precum: Târgu-Jiu. Pe lângă lexicon, au fost create 2 liste adiţionale, una de prefixe cu cratimă (e.g. te-, ne-, istorico-), respectiv una de sufixe cu cratimă (e.g. -se, -vă, -ul). Având în vedere aceste resurse, procesarea cuvintelor cu cratimă întâlnite în text se face conform Algoritmului 2.6.a.

Astfel, după etapa de procesare de text, textul folosit pentru antrenarea modelului de limbă conţinea cuvinte precum: “să”, “-mi”, “să”, “-i”, “m-”, “a”, “ratat”, “-o”, “FMI”, “-ul”. Noul sistem ASR ce foloseşte modelul de limbă generează acum text de forma: “**TVA -ul** a crescut cu 3 la sută”. În final, un simplu procesor de text va unii cratimele de cuvintele adiacente: “**TVA-ul** a crescut cu 3 la sută”.

Cu privire la modelele de limbă noi, nu au fost create tipuri noi de modele de limbă, ci s-au antrenat modele de limbă folosind corpusuri de text extinse şi preprocesate mai bine. Modelele de limbă de bază au fost antrenate folosind numai corpusul news002 (352M de cuvinte), pe când noile modele de limbă au fost antrenate pe corpusul news002 plus corpusul news2020 (352M + 255M de cuvinte). Textul folosit pentru antrenarea modelelor de limbă noi a fost preprocesat folosind noul procesor de limbaj natural discutat în acest subcapitol. Modelele de limbă îmbunătăţite sunt evaluate în funcţie de perplexitate (PPL) şi cuvinte

if (cuvântul cu cratimă este întâlnit în lexiconul pentru cuvinte cu cratimă)	Exemple
then se păstrează forma iniţială	*Coca-Cola*
else cuvântul este împărţit	
if (una din părţi se regăseşte în lexiconul de cuvinte comune) && (cealaltă parte se regăseşte în lista de sufixe/prefixe)	TVA *-ul*
then cuvântul şi sufixul/prefixul sunt separate cu spaţiu	*TVA-ul => TVA -ul*
else cratima este înlocuită cu un spaţiu, iar forma iniţială este trecută într-o listă de cuvinte nerecunoscute pentru o analiză manuală ulterioară	*două-trei* *bugetari-particulari*

Algoritmul 2.6.a *Procedura actualizată de procesare a cuvintelor cu cratimă*

out-of-vocabulary (OOV). Modelele de limbă mai simple (2-gram şi 3-gram) au fost încorporate în etapa de decodare a sistemului ASR, în timp ce modelele de limbă mai complexe (4-gram şi RNN) au fost folosite în etapa de rescoring. Modelele mai complexe nu au fost folosite în etapa de decodare din cauza constrângerilor arhitecturale şi de memorie.

Pentru etapa de decodare am experimentat cu diferite ordine n-gram (2-gram şi 3-gram), dimensiuni ale vocabularului (200k, 250k and 300k words) şi n-gram pruning (no pruning, 1e-7, 3e-7). Cele mai bune rezultatele obţinute după evaluarea pe seturile de transcrieri pentru evaluare pot fi observate în Tabelul 2.6.a.

Din punct de vedere al PPL, cel mai bun model de limbă, LM-2020-3g-large-200k, prezintă o scădere relativă între 40% şi 45% pe seturile de evaluare, comparativ cu modelul de limbă de bază (Tabel 2.6.a linia 1). Cu privire la rezultatele OOV, se poate observa o îmbunătăţire în cazul seturilor de evaluare de vorbire spontană (SSC-eval1 şi SSC-eval2). În cazul setului de evaluare de vorbire citită, RSC-eval, numai modelele cu un vocabular de 300k cuvinte au generat un rezultat mai bun.

Similar cu modelele de limbă pentru decodare prezentate în Tabelul 2.6.a, în Tabelul 2.6.b sunt evaluate modelele de limbă destinate pentru etapa de reevaluare lingvistică, anume 4-gram şi RNN.

Analizând rezultatele din Tabelul 2.6.b, putem observa pentru modelul LM-2020-4g-large-200k o scădere relativă de 7% şi 40% PPL pe seturile RSC-eval şi SSC-eval2 comparativ cu modelul de bază cu rescoring (Tabelul 2.6.b, linia 1). Pentru setul de evaluare SSC-eval1, modelul LM-2020-5g-RNN a obţinut rezultate mai bune, cu o scădere relativă de 14% PPL. Din punct de vedere al rezultatelor OOV, modelul LM-2020-4g-large-300k a generat cele mai bune rezultate pentru seturile de evaluare de vorbire spontană, pe când pentru setul RSC-eval cel mai bun rezultat a fost obţinut la modelul de bază şi LM-2020-5g-RNN. Procentajul de 0.01% OOV pentru modelul de bază şi LM-2020-5g-RNN a fost obţinut datorită faptului că aceste modele nu am avut o dimensiune limită pentru vocabular, faţă de modelele 4-gram care au avut limită.

Tabelul 2.6.a *Modele de limbă pentru decodare îmbunătăţite (LM-2020-*) vs model de limbă pentru decodare de bază (LM-2017). Numele modelelor de limbă indică ordinul n-gram (2g, 3g), dacă s-a aplicat sau nu n-gram prunning (small = 3e-7, large = no pruning) şi dimensiunea vocabularului (200k, 300k words)*

	PPL [%]			**OOV [%]**		
Model de limbă	**RSC-eval**	**SSC-eval1**	**SSC-eval2**	**RSC-eval**	**SSC-eval1**	**SSC-eval2**
LM-2017-2g-large-200k	316	219	356	0.13	1.96	0.91
LM-2020-2g-small-300k	393	280	411	**0.09**	**0.37**	**0.28**
LM-2020-3g-large-200k	**178**	**131**	**210**	0.19	0.51	0.37

Tabelul 2.6.b *Modele de limbă pentru rescoring îmbunătăţite (LM-2020-*) vs model de limbă pentru rescoring de bază (LM-2019). Numele modelelor de limbă indică ordinul n-gram (4g, RNN), n-gram pruning (large = no pruning) şi dimensiunea vocabularului (200k, 250k, 300k words)*

	PPL [%]			**OOV [%]**		
Model de limbă	**RSC-eval**	**SSC-eval1**	**SSC-eval2**	**RSC-eval**	**SSC-eval1**	**SSC-eval2**
LM-2019-RNN	176	132	325	0.01	0.82	0.66
LM-2020-RNN	246	**113**	206	**0.01**	0.43	0.35
LM-2020-4g-large-200k	**162**	118	**196**	0.19	0.51	0.37
LM-2020-4g-large-250k	163	119	197	0.14	0.44	0.34
LM-2020-4g-large-300k	165	120	198	0.09	**0.37**	**0.28**

Noile sisteme RAV au fost obţinute folosind modelele de limbă prezentate anterior. Pentru etapa de decodare am optat pentru modelele 3-gram 200k, iar pentru rescoring am ales LM-2020-4g-large-200k şi LM-2020-5g-RNN. Cele mai bune rezultate s-au obţinut pentru sistemul RAV ce foloseşte LM-2020-RNN pentru rescoring. În Tabelul 2.6.c sunt prezentate rezultatele WER ale noilor sisteme RAV cu şi fără rescoring.

În Tabelul 2.6.c se poate observa că cel mai bun sistem RAV este cel ce foloseşte LM-2020-3g-large-200k cu LM-2020-RNN pentu rescoring. Acest model are o scădere relativă de 9% şi 7% WER pe seturile de evaluare de vorbire spontană, anume SSC-eval1 şi respectiv SSC-eval2, comparativ cu sistemul RAV de bază cu rescoring (Tabel 2.6.c linia 1). Însă, pe setul de evaluare RSC-eval sistemul RAV a înregistrat o creştere de 16% WER.

2.6.2 Utilizarea seturilor de date rezultate din proiect pentru antrenarea RAV

În această secţiune analizăm impactul utilizării unor seturi de date de antrenare mai mari asupra performanţei sistemelor de RAV SpeeD bazate pe utilitarul Kaldi. Pentru a putea compara direct îmbunătăţirile rezultate strict din creşterea numărului de ore de antrenare, toţi ceilalţi hiperparametri ai sistemelor au fost păstraţi constanţi. Printre altele, toate sistemele prezentate în această secţiune folosesc pentru decodare de vorbire modelul LM-2017-2g-large-200k, iar pentru reevaluare lingvistică modelul LM-2017-RNN.

Tabelul 2.6.d prezintă evoluţia sistemului RAV SpeeD în perioada 2018-2020. Fiecare linie reprezintă un model acustic diferit, îmbunătăţit faţă de cel precedent, pentru antrenarea căruia au fost adăugate seturi de date suplimentare. Fiecare sistem este evaluat pe cele 3 seturi de evaluare existente: RSC-eval, SSC-eval1, SSC-eval2. Primul model acustic, denumit în tabel *train-base,* ese cel din cadrul sistemului iniţial RAV #3, prezentat în secţiunea 2.2.2. Acesta a fost antrenat folosind setul de antrenare *baseline*: RSC-train1 + SSC-train1 + SSC-train2.

Tabelul 2.6.c *Rezultatele WER ale sistemelor RAV după decodare şi rescoring*

Model de limbă folosit pentru rescoring	**Model de limbă folosit pentru decodare**	**WER[%] fără rescoring**			**WER[%] cu rescoring**		
		RSC-eval	**SSC-eval1**	**SSC-eval2**	**RSC-eval**	**SSC-eval1**	**SSC-eval2**
LM-2017-RNN	LM-2017-2g-large-200k	2.8	13.3	16.4	**1.8**	11.0	14.0
LM-2020-RNN	LM-2020-3g-large-200k	**2.7**	**11.0**	**14.4**	2.1	**10.0**	**12.9**
	LM-2020-3g-medium-200k	3.1	11.9	15.1	2.2	10.4	13.3
	LM-2020-3g-small-200k	3.3	12.6	15.8	2.2	10.5	13.7

Cu ajutorul sistemului iniţial RAV #3, au fost decodate seturile SSC-train3-raw şi SSC-train4-raw, iar transcrierile obţinute, împreună cu transcrierile aproximative deja existente, au fost aliniate în etapa A2.11/2019, rezultând seturile SSC-train3-trans-v4 şi SSC-train4-trans-v4. Aceste seturi au fost adăugate la setul de antrenare *baseline,* adunând 292 ore de vorbire la cele 225 ore anterioare. Prin reantrenarea sistemului, s-a obţinut astfel cel de-al doilea model acustic, denumit *train3-v4*. Acest fapt a dus la o scădere relativă a ratei de eroare la nivel de cuvânt de 5% pe RSC-eval, respectiv 26%-30% pe SSC-eval1 şi SSC-eval2. Faptul că pe vorbire spontană s-a obţinut o îmbunătăţire mai mare, se datorează caracteristicilor noilor seturi de antrenare acestea cuprinzând preponderent vorbire spontană.

Sistemul RAV #3, corespunzător modelului acustic *train-base*, a fost utilizat pentru a transcrie setul de date CoBiLiRo, prezentat în secţiunea 2.1.2. În secţiunea 2.3.2 este prezentat modul în care a fost aliniată transcrierea obţinută cu transcrierile aproximative deja existente, rezultând în acest fel setul de date CoBiLiRo-trans-v4. Acesta a fost adaugat la seturile de antrenare anterioare, fiind obţinut un nou model acustic, denumit *train14*. În comparaţie cu modelul *train3-v4*, modelul *train14* obţine rezultate mai bune pe vorbire citită (11% scădere relativă a WER pe RSC-eval), dar rezultate mai slabe pe vorbire spontană. Acest lucru se datorează similitudinii dintre CoBiLiRo-trans-v4 şi setul de date RSC. CoBiLiRo-trans-v4 conţine interviuri şi dialoguri într-un mediu fără zgomot. Totodată, trebuie ţinut cont şi de cantitatea mică a datelor nou adăugate, 31 ore, faţă de cele 517 deja existente.

Ultimul model acustic, *train11*, a presupus reantrenarea sistemului RAV după ce a fost adăugat setul de date CoRoLa, introdus în secţiunea 2.1.1. Îmbunătăţirile relative faţă de sistemul anterior sunt cuprinse între 8%-12% pe vorbire spontană, în timp ce în cazul vorbirii citite, nu se înregistrează schimbări. Acest lucru se explică prin conţinutul setului CoRoLa: vorbire spontană, de la vorbitori profesionişti, în cadrul unor emisiuni TV. În total, faţă de modelul *train-base,* acest ultim model obţine o scădere relativă a erorii de 15% pe vorbire citită, respectiv 31%-39% pe vorbire spontană.

În continuare ne-am pus problema contribuţia fiecărui nou set de date de vorbire la îmbunătăţirea totală a sistemului de RAV. Astfel, am antrenat sisteme de RAV pe seturi de date formate din setul baseline şi un singur alt set de date de vorbire. Tabelul 2.6.e prezintă rezultatele. Rezultatele obţinute pe setul de date RSC-eval diferă foarte puţin, dacă nu chiar deloc, între diversele experimente. Tragem concluzia că modelul din adăugarea de date noi numai pentru creşterea performanţei pe vorbire spontană. Deoarece seturile nou adăugate conţin în mare măsură vorbire spontană, îmbunătăţirile substanţiale au rezultat pe seturile de evaluare SSC-eval1 şi SSC-eval2 (ce conţin astfel de vorbire). Este foarte probabil ca în urma reantrenărilor, modificări ale parametrilor reţelei neuronale să fi apărut numai în acele zone din reţea care se ocupă de prelucrarea vorbirii spontane.

O alta concluzie se poate trage cu privire la dimensiunea seturilor noi de date. De exemplu, chiar dacă modelul *train3-v4* conţine mai mult decât dublul datelor folosite la antrenarea lui *train-base*, îmbunătăţirile obţinute nu sunt de două ori mai mari. Nu există o dependenţă direct proporţională între cantitatea de date folosite la antrenare şi performanţa sistemului. Cu cât acurateţea creşte, este nevoie de mult mai multe date pentru a avea parte de îmbunătăţiri mici.

***Tabelul 2.6.d**. Evolutia RAV SpeeD în perioada 2018 - 2020. Îmbunătăţirile rezultate prin adăugarea la setul de date de antrenare BAS (RSC-train + SSC-train1 + SSC-train2) a seturilor de date rezultate din proiect SSC (SSC-train3-trans-v4 + SSC-train4-trans-v4), COB (CoBiLiRo-trans-v4) şi COR (CoRoLa)*

Cod model acustic	**Set de antrenare**					**Set de evaluare (WER[%])**		
	BAS 225h	**SSC 292h**	**COB 31h**	**COR 84h**	**CDP 879h**	**RSC-eval**	**SSC-eval1**	**SSC-eval2**
train-base	x					1.9	15.0	20.0
train3-v4	x	x				1.8	11.0	14.0
train14	x	x	x			1.6	11.3	14.4
train11	x	x	x	x		1.6	10.3	12.2

***Tabelul 2.6.e**. Contribuţia fiecărui set de date nou la îmbunătăţirea RAV SpeeD 2018 - 2020. Seturile de date de antrenare folosite sunt: BAS (RSC-train + SSC-train1 + SSC-train2), SSC (SSC-train3-trans-v4 + SSC-train4-trans-v4), COB (CoBiLiRo-trans-v4) şi COR (CoRoLa)*

Cod model acustic	Set de antrenare					Set de evaluare (WER[%])		
	BAS 225h	SSC 292h	COB 31h	COR 84h	CDP 879h	RSC-eval	SSC-eval1	SSC-eval2
train-base	x					1.9	15.0	20.0
train3-v4	x	x				1.8	11.0	14.0
train15	x		x			1.8	14.0	21.1
train17	x			x		1.8	11.9	15.4

Modelul acustic *train15* prezintă influenţa adăugării datelor din setul CoBiLiRo-trans-v4. Dată fiind cantitatea redusă a datelor din acest set, în comparaţie cu datele folosite la antrenarea modelului *train-base*, performanţele sistemului nou nu sunt cu mult diferite faţă de cel iniţial.

Modelul acustic *train17*, antrenat prin adăugarea setului de date CoRoLa, însumând 84 de ore, ajunge să obţină performanţe asemănătoare cu modelul *train3-v4*, ce totalizează 292 ore. Deşi ambele seturi, atât CoRoLa, cât şi SSC-train3-trans-v4 + SSC-train4-trans-v4, conţin vorbire spontană, diferenţa este dată de lungimea rostirilot din cele două seturi. CoRoLa a fost adnotat manual, iar rostirile au dimensiuni mai mari decât în cazul lui SSC-train3-trans-v4 + SSC-train4-trans-v4, care a fost adnotat automat, prin metoda alinierii transcrierilor aproximative, unde au fost selectate doar segmente comune de lungimea a câtorva cuvinte. Reţeaua neuronală folosită la antrenarea modelului acustic beneficiază în prima situaţie de un context temporal mult mai larg, în timp ce în a doua situaţie, reţeaua are un context temporal insuficient pentru a învăţa interdependenţele fonemelor.

Tabelul 2.6.f prezintă contribuţia noilor seturi de date introduse în 2020 în mod comparativ. Unul din cele 3 seturi noi din 2020, pe lângă CoBiLiRo-trans-v4 şi CoRoLa, este cdep-trans-v4, prezentat în secţiunea 2.3.3. Setul de date cdep-raw a fost transcris folosind modelul acustic *train3-v4*, iar apoi, prin alinierea cu transcrierile aproximative deja existente, a rezultat setul cdep-trans-v4. Surprinzător este faptul că deşi setul de date cdep-trans-v4 însumează 879 ore, fiind de 28 de ori mai mare decât CoBiLiRo-trans-v4, respectiv de 10 ori mai mare decât CoRoLa, adăugarea acestui set la antrenare nu produce rezultate mai bune decât adăugarea celorlalte două seturi amintite. O pistă de investigat în acest sens este lungimea rostirilor din setul cdep-trans-v4, fiind posibil ca acestea să fie foarte scurte (e posibil ca transcrierile aproximative şi transcrierile date de sistemul RAV să se suprapună într-o mică măsură).

***Tabelul 2.6.f**. Contribuţia noilor seturi de date introduse in 2020. Sistemul iniţial a fost antrenat pe seturile de date BAS (RSC-train + SSC-train1 + SSC-train2) şi SSC (SSC-train3-trans-v4 + SSC-train4-trans-v4). În etapa 3/2020 au fost introduse seturile de date COB (CoBiLiRo-trans-v4), COR (CoRoLa) şi CDP (cdep-trans-v4)*

Cod model acustic	Set de antrenare					Set de evaluare (WER[%])		
	BAS 225h	SSC 292h	COB 31h	COR 84h	CDP 879h	RSC-eval	SSC-eval1	SSC-eval2
train-base	x					1.9	15.0	20.0
train3-v4	x	x				1.8	11.0	14.0
train14	x	x	x			1.6	11.3	14.4
train18	x	x		x		1.8	10.5	12.5
train13	x	x			x	1.7	12.3	15.4

2.6.3 Sistemul de RAV SpeeD îmbunătăţit

În concluzie, selectând cea mai performantă configuraţie de model de limbă pentru decodare, model de limbă pentru reevaluare lingvistică (conform secţiunii 2.6.1) şi model acustic (conform secţiunii 2.6.2), cel mai performant sistem de RAV SpeeD are actualmente performanţele listate în Tabelul 2.6.g.

***Tabelul 2.6.g**. Sistemul RAV SpeeD 2020*

Model acustic		**Model lingvistic**		**WER [%]**			
Set de antrenare	**Tip model**	**Corpus antrenare**	**Tip model**	**RSC-eval**	**SSC-eval1**	**SSC-eval2**	**CDep-eval**
RSC-train + SSC-train +SSC-train3-trans-v4 +SSC-train4-trans-v4 +CoBiLiRo-trans-v4 +CoRoLa	HMM-DNN (TDNN3) Kaldi toolkit	news2017 +news2020	Decod.: LM-2020-3g-large-200k Reev. lingv.: LM-2020-RNN	1.9	9.4	11.4	7.0

2.7 Bibliografie

[Ardila, 2020] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor We- ber, “Common Voice: A massively-multilingual speech corpus,” in International Conference on Language Resources and Evaluation, 2020.

[Ashukha, 2020] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov, “Pitfalls of in-domain uncertainty estimation and ensembling in deep learning,” in International Conference on Learning Representations, 2020.

[Barbu, 2018] V. Barbu-Mititelu, D. Tufiş, E. Irimia, “*The Reference Corpus of the Contemporary Romanian Language (CoRoLa)*,” in Proceedings of LREC 2018, Japan, p.1178-1185.

[Chen, 2019] Tongfei Chen, Jirí Navrátil, Vijay Iyengar, and Karthikeyan Shanmugam, “Confidence scoring using whitebox meta-models with linear classifier probes,” in International Conference on Artificial Intelligence and Statistics, 2019, pp. 1467–1475.

[Corbière, 2019] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez, “Addressing failure prediction by learning model confidence,” in Advances in Neural Information Processing Systems, 2019, pp. 2902– 2913.

[Cortina, 2016] Isaías Sánchez Cortina, Jesús Andrés-Ferrer, Alberto Sanchis, and Alfons Juan, “Speaker-adapted confidence measures for speech recognition of video lectures,” Computer Speech & Language, vol. 37, pp. 11–23, 2016.

[Del-Agua, 2018] M. A. Del-Agua, A. Gimenez, A. Sanchis, J. Civera, and A. Juan, “Speaker-adapted confidence measures for ASR using deep bidirectional recurrent neural networks,” Transactions on Audio, Speech, and Language Processing, vol. 26, no. 7, pp. 1198–1206, 2018.

[Errattahi, 2018] Rahhal Errattahi, Salil Deena, Asmaa El Hannani, Hassan Ouahmane, and Thomas Hain, “Improving ASR error detection with RNNLM adaptation,” in IEEE Spoken Language Technology Workshop, 2018, pp. 190–196.

[Gal, 2016] Yarin Gal and Zoubin Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in International Conference on Machine Learning, 2016, pp. 1050–1059.

[Georgescu, 2017] A.-L. Georgescu, H. Cucu, C. Burileanu, “*SpeeD’s DNN Approach to Romanian Speech Recognition*,” in Proc. 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD), 8p, 2017.

[Georgescu, 2018] A.-L. Georgescu, H. Cucu, “Automatic annotation of speech corpora using complementary GMM and DNN acoustic models,” In Proc. TSP 2018, pp. 1-4.

[Georgescu, 2019a] A.-L. Georgescu, C. Manolache, G. Pop, D. Oneaţă, H. Cucu, D. Burileanu, C. Burileanu, *“Proiect component TADARAV: Raport ştiinţific şi tehnic în extenso 2019”*.

[Georgescu, 2019b] A.-L. Georgescu, H. Cucu, C. Burileanu, *"Kaldi-based DNN architectures for speech recognition in Romanian,"* in the Proceedings of the 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2019, Timişoara, Romania.

[Georgescu, 2020] A.-L. Georgescu, H. Cucu, A. Buzo, C. Burileanu, "*RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition*," in the Proceedings of The 12th Language Resources and Evaluation Conference (LREC), pp. 6606-6612, 2020, Marseille, France.

[Georgescu, 2021] A.-L. Georgescu, C. Manolache, D. Oneaţă, H. Cucu, C. Burileanu, "*Data-filtering methods for self-training of automatic speech recognition systems*," in the Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Virtual, 2021.

[Guo, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Wein- berger, "On calibration of modern neural networks," in International Conference on Machine Learning, 2017, pp. 1321–1330.

[Hadian, 2018] Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur, "End-to-end speech recognition using lattice-free MMI," in Interspeech, 2018, pp. 12–16.

[Hazen, 2002] Timothy J Hazen,Stephanie Seneff, and Joseph Polifroni, "Recognition confidence scoring and its use in speech understanding systems," Computer Speech & Language, vol. 16, no. 1, pp. 49–67, 2002.

[Hendrycks, 2016] Dan Hendrycks and Kevin Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in International Conference on Learning Representations, 2016.

[Hinton, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[Hori, 2019] Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Shinji Watanabe, *"Advanced Methods for Neural End-to-End Speech Processing – Unification, Integration, and Implementation,"* Interspeech 2019.

[Huggins-Daines, 2006] David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W. Black, Mosur Ravishankar, and Alexander I. Rudnicky. *"Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices."* in Proc. 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1, pp. I-I. 2006.

[Kalgaonkar, 2015] Kaustubh Kalgaonkar, Chaojun Liu, Yifan Gong, and Kaisheng Yao, "Estimating confidence scores on ASR results using recurrent neural networks," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2015, pp. 4999–5003.

[Karita, 2019] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang, "A comparative study on transformer vs RNN in speech applications," in Workshop on Automatic Speech Recognition and Understanding, 2019, pp. 449–456.

[Kemp, 1997] Thomas Kemp and Thomas Schaaf, "Estimating confidence using word lattices," in Eurospeech, 1997.

[Kim, 2017] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in Proc. ICASSP, pp. 4835-4839, 2017.

[Kudo, 2018] Taku Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in Association for Computational Linguistics, 2018, pp. 66–75.

[Lakshminarayanan, 2017] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Advances in Neural Information Processing Systems, 2017, pp. 6402–6413.

[Li, 2019] Qiujia Li, PM Ness, Anton Ragni, and Mark JF Gales, "Bi-directional lattice recurrent neural networks for confidence estimation," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 6755–6759.

[Lüscher, 2019] Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitza, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney, "RWTH ASR systems for LibriSpeech: Hybrid vs attention," in Interspeech, 2019, pp. 231–235.

[Malinin, 2020] Andrey Malinin and Mark Gales, “Uncertainty in structured prediction,” arXiv preprint arXiv:2002.07650, 2020.

[Manolache, 2019] C. Manolache, “*Natural language processing using artificial intelligence: keyword spotting and speech transcription intelligibility*“, MSc. Thesis, June 2019 (scientific coordinator: Conf. H. Cucu).

[Manolache, 2020] C. Manolache, A.-L. Georgescu, H. Cucu, V. B. Mititelu, C. Burileanu, *“Improved text normalization and language models for SpeeD’s Automatic Speech Recognition System”*, in Proc. 15th International Conference “Linguistic Resources and Tools for Processing the Romanian Language” (ConsILR) 2020.

[Ogawa, 2017] Atsunori Ogawa and Takaaki Hori, “Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks,” Speech Communication, vol. 89, pp. 70–83, 2017.

[Ovadia, 2019] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek, “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift,” in Advances in Neural Infor- mation Processing Systems, 2019, pp. 13991–14002.

[Panayotov, 2015] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, *"Librispeech: An ASR corpus based on public domain audio books,"* in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206-5210, 2015.

[Park, 2019] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” in Interspeech, 2019, pp. 2613–2617.

[Paszke, 2019] Adam Paszke et al., *“PyTorch: An Imperative Style, High-Performance Deep Learning,”* Advances in Neural Information Processing Systems 32, pp. 8024-8035, 2019.

[Povey, 2011] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hanne- mann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, “The Kaldi speech recognition toolkit,” in Workshop on Automatic Speech Recognition and Understanding, 2011.

[Ragni, 2018] Anton Ragni, Qiujia Li, Mark JF Gales, and Yongqiang Wang, “Confidence estimation and deletion prediction using bidirectional recurrent neural networks,” in IEEE Spoken Language Technology Workshop, 2018, pp. 204– 211.

[Rousseau, 2014] Anthony Rousseau, Paul Deléglise, and Yannick Estève, “Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks,” in International Conference on Language Resources and Evalua- tion, 2014, pp. 3935–3939.

[Seigel, 2013] Mathew Stephen Seigel, Confidence estimation for automatic speech recognition hypotheses, Ph.D. thesis, University of Cambridge, 2013.

[Seigel, 2014] Matthew Stephen Seigel and Philip C Woodland, “Detecting deletions in ASR output,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2014, pp. 2302–2306.

[Sennrich, 2016] Rico Sennrich, Barry Haddow, Alexandra Birch, “Neural Machine Translation of Rare Words with Subword Units,” In Proc. 54th Annual Meeting of the Association for Computational Linguistics, pp 1715-1725, 2016.

[Sperber, 2017] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel, “Neural lattice-to-sequence models for uncertain inputs,” in Empirical Methods in Natural Language Processing, 2017, pp. 1380–1389.

[Srivastava, 2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 1 2014.

[Swarup, 2019] Prakhar Swarup, Roland Maas, Sri Garimella, Sri Harish Mallidi, and Björn Hoffmeister, “Improving ASR confidence scores for Alexa using acoustic and hypothesis embeddings,” in Interspeech, 2019, pp. 2175–2179.

[Tokui, 2019] Seiya Tokui, Ryosuke Okuta, Takuya Akiba, Yusuke Niitani, Toru Ogawa, Shunta Saito, Shuji Suzuki, Kota Uenishi, Brian Vogel, and Hiroyuki Yamazaki Vincent, *"Chainer: A deep learning framework for accelerating the research cycle,"* In Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2002-2011, 2019.

[Tüske, 2019] Zoltán Tüske, Kartik Audhkhasi, and George Saon, “Advancing sequence-to-sequence based speech recognition,” in Interspeech, 2019, pp. 3780–3784.

[Vaswani, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[Vesely, 2013] Karel Vesely, Mirko Hannemann, and Lukas Burget, “Semi-supervised training of deep neural networks,” in Workshop on Automatic Speech Recognition and Understanding, 2013, pp. 267–272.

[Vyas, 2019] Apoorv Vyas, Pranay Dighe, Sibo Tong, and Hervé Bourlard, “Analyzing uncertainties in speech recognition using dropout,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 6730–6734.

[Watanabe, 2017] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey and Tomoki Hayashi, *"Hybrid CTC/Attention Architecture for End-to-End Speech Recognition,"* IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240-1253, 2017.

[Watanabe, 2018] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in Proc. Interspeech, pp. 2207-2211, 2018.

[Weintraub, 1997] Mitch Weintraub, Francoise Beaufays, Zeév Rivlin, Yochai Konig, and Andreas Stolcke, “Neural-network based measures of confidence for word recognition,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 1997, vol. 2, pp. 887–890.

[Yu, 2010] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero, “Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion,” Computer Speech & Language, vol. 24, no. 3, pp. 433–444, 2010.

3 Structura ofertei de servicii de cercetare şi tehnologice

Laboratorul de cercetare *Speech and Dialogue* (SpeeD) din cadrul Universităţii Politehnica din Bucureşti (UPB), reprezentantul UPB în proiectul TADARAV, oferă pe platforma ERRIS serviciile de cercetare şi tehnologice enumerate în Tabelul 3.

Tabelul 3. *Servicii de cercetare şi tehnologice oferite de Laboratorul de cercetare Speech and Dialogue*

Serviciu	**Detalii**
Serviciu şi aplicaţie web de transcriere de documente ce conţin vorbire în limba română	https://transcriptions.speed.pub.ro
Serviciu şi aplicaţie web de identificare de cuvinte cheie în documente ce conţin vorbire în limba română	https://keywords.speed.pub.ro
Serviciu şi aplicaţie web de restaurare de diacritice în limba română	https://diacritics.speed.pub.ro
Proiectarea şi implementarea de aplicaţii personalizate de transcriere a vorbirii continue	La cerere
Proiectarea şi implementarea de aplicaţii personalizate de identificare de cuvinte şi termeni de interes	La cerere
Proiectarea şi implementarea de aplicaţii personalizate de sinteză de vorbire pornind de la text	La cerere
Proiectarea şi implementarea de sisteme de recunoaştere de pattern-uri folosind inteligenţă artificială	La cerere

Laboratorul de cercetare *Speech and Dialogue* (SpeeD) este prezent pe platforma ERRIS la adresa https://erris.gov.ro/SpeeD---UPB.

4 Locuri de muncă susţinute prin program

Echipa de cercetare a Universităţii Politehnica din Bucureşti pentru proiectul component TADARAV este prezentată în Tabelul 4.

Tabelul 4. *Echipa de cercetare UPB*

Nr.	**Nume**	**Calitatea**	**Poziţia**	**Normă**
1	Horia CUCU	Conf. Univ.	Responsabil proiect component	Parţială
2	Corneliu BURILEANU	Prof. Univ.	Membru cercetător	Parţială
3	Dragoş BURILEANU	Prof. Univ.	Membru cercetător	Parţială
4	Alexandru-Lucian GEORGESCU	ACS	Membru cercetător	Parţială
5	Dan Theodor ONEAŢĂ	CS	**Membru cercetător nou**	**Întreagă**
6	Gheorghe POP	ACS	**Membru cercetător nou**	**Întreagă**
7	Cristian MANOLACHE	ACS	**Membru cercetător nou**	**Întreagă**

5 Valorificarea şi îmbunătăţirea competenţelor şi resurselor existente la nivelul consorţiului

În această etapă proiectul TADARAV nu a avut fonduri la capitolul bugetar CEC-uri.