UNIVERSITATEA POLITEHNICA DIN BUCUREȘTI

Școala Doctorală de Electronică, Telecomunicații și Tehnologia Informației

Decizie nr. 963 **din** 16-11-2022

REZUMAT TEZĂ DE DOCTORAT

Ing. Alexandru-Lucian GEORGESCU

METODE ȘI TEHNOLOGII DE INTELIGENȚĂ ARTIFICIALĂ APLICATE ÎN TEHNOLOGIA VORBIRII

METHODS AND TECHNOLOGIES OF ARTIFICIAL INTELLIGENCE APPLIED IN SPEECH TECHNOLOGY

COMISIA DE DOCTORAT

Prof. Dr. Ing. Gheorghe Brezeanu Univ. Politehnica din București	Președinte
Prof. Dr. Ing. Corneliu BURILEANU Univ. Politehnica din București	Conducător de doctorat
Prof. Dr. Ing. Daniela TĂRNICERIU Univ. Tehnică "Gh. Asachi" din Iași	Referent
Prof. Dr. Ing. Corneliu RUSU Univ. Tehnică din Cluj-Napoca	Referent
Conf. Dr. Ing. Horia CUCU Univ. Politehnica din București	Referent

BUCUREȘTI 2022

Mulţumiri

În primul rând, aș dori să-mi exprim aprecierea specială și cea mai profundă recunoștință conducătorului tezei mele de doctorat, Prof. Dr. Ing. Corneliu Burileanu. Vreau să-i mulțumesc pentru îndrumarea de-a lungul timpului, pentru șansa de a face parte din echipa de cercetare a cărei coordonator este, pentru modelul oferit atât ca inginer, cât și ca om.

Mulțumesc de asemenea comisiei de îndrumare, Prof. Dr. Ing. Dragoș Burileanu, Conf. Dr. Ing. Horia Cucu și Dr. Ing. Dan Oneață, pentru șansa de a evolua și de a învăța din experiența și cunoștințele lor. Sunt foarte recunoscător și am o apreciere deosebită față de Conf. Dr. Ing. Horia Cucu care mi-a oferit foarte multă încredere, mai multă decât am avut decât am avut chiar eu uneori. Tot timpul investit în mine a avut un rol crucial în progresul meu, iar această teză i se datorează în mare parte.

Doresc să mulțumesc tuturor colegilor din Laboratorul Speech and Dialogue pentru colaborare și ideile schimbate.

Multe mulțumiri Centrului de Cercetări Avansate pentru Materiale, Produse și Procese Inovative (CAMPUS) din cadrul Universității Politehnica din București pentru oportunitatea de a lucra într-un veritabil mediu de cercetare.

Vreau să-mi exprim recunoștința membrilor Xilinx Research Labs din Dublin, Irlanda, în special față de Michaela Blott, Alessandro Pappalardo și Lucian Petrică pentru oportunitatea de a efectua un stagiu de cercetare grozav ce a avut un impact major asupra evoluției mele.

Desigur, vreau să mulțumesc membrilor comisiei de doctorat pentru consimțământul de a fi referenți și pentru timpul acordat evaluării acestei lucrări.

Nu în ultimul rând, vreau să-mi exprim profunda recunoștință familiei mele, prietenilor, tuturor celor din jurul meu care m-au susținut și încurajat tot timpul.

Cuprins

Capitolul 1

Introducere

1.1 Motivația tezei

Laboratorul de cercetare SpeeD (Speech and Dialogue) [1] din cadrul Facultății de Electronică, Telecomunicații și Tehnologia Informației a Universității Politehnica București are o experiență îndelungată în domeniul procesării semnalului vocal. În cadrul laboratorului au fost dezvoltate de-a lungul timpului proiecte de diplomă și doctorat de succes.

In anul 3 de studii de licență am intrat pentru prima dată în contact cu laboratorul SpeeD. Am avut ocazia să fac un stagiu de vară, urmat de proiectele de diplomă și disertație. În cadrul acestora, am obținut cunoștințe teoretice solide despre procesarea semnalului vocal, în special recunoașterea automată a vorbirii, recunoașterea vorbitorului, detectarea cuvintelor cheie, lucrând cu algoritmi și utilitare specifice. Am construit proiecte mai simple, precum verificarea vorbitorului, recunoașterea automată a vorbirii pentru cuvinte izolate, apoi sisteme cu vocabular limitat, ajungând la sisteme de vorbire continuă mai complexe.

Astfel, am avut ocazia să încep un program de doctorat în acest domeniu. Rolul meu a fost să duc mai departe abordarea care la acea vreme reprezenta starea artei, mai exact sistemul în limba română de recunoaștere automata a vorbirii (RAV) continue cu vocabular extins, prezentat în [8]. Existau deja suficiente idei și abordări care puteau fi testate în încercarea de a îmbunătăți acel sistem. Dintre acestea, cele mai importante au fost: trecerea la modelarea vorbirii bazată pe rețele neuronale, extinderea corpusurilor de vorbire utilizate pentru antrenarea , extinderea vocabularului și utilizarea unor modele de limbaj mai complexe. Chiar dacă aceste lucruri erau deja abordate în literatura de specialitate pentru limba engleză, în română a existat suficient spațiu pentru a le explora. La vremea respectivă, nu existau multe sisteme RAV pentru limba română cu o precizie și robustețe atât de crescute ca cele pentru limba engleză.

Un prim avantaj a fost sprijinul oferit de colegi și, bineînțeles, de supervizorii din grupul de cercetare. Am avut un mediu competitiv, unde mi-am putut dezvolta

cunoștințele și abilitățile. Am beneficiat de know-how-ul existent în procesarea vorbirii și de experiența vastă a membrilor grupului. Am beneficiat și de resursele deja existente în cadrul grupului și de sistemele dezvoltate anterior.

Un al doilea avantaj a fost oportunitatea de a fi angajat cu normă întreagă în proiecte naționale de cercetare, două dintre ele fiind cele mai importante pentru dezvoltarea mea în domeniul tehnologiei vorbirii: SPIA-VA [18] și ReTeRom-TADARAV [3]. În cadrul acestora, m-am ocupat de recunoașterea vorbirii și a vorbitorului, dar și de adnotarea automată a datelor audio.

Astfel, această teză a luat naștere într-o manieră firească, înglobând preocupările autorului din ultimii ani și continuând eforturile comune ale membrilor grupului, fiind doar o altă piesă adăugată peste cele deja existente.

1.2 Obiective

Obiectivul principal al acestei teze a fost de a profita de paradigmele de inteligență artificială mai performante pentru a obține îmbunătățiri în diverse aplicații legate de tehnologia vorbirii. În primul rând, s-a dorit îmbunătățirea unui sistem RAV pe limba română deja existent, creat și îmbunătățit de-a lungul multor ani în cadrul grupului nostru de cercetare. Acest sistem a ajuns să fie complet înlocuit cu unul nou folosind noi tehnologii bazate pe rețele neuronale. O provocare a fost adaptarea sistemului astfel încât să fie cât mai robust posibil în fața unor seturi de evaluare cu complexitate crescută, care conțin vorbire spontană în diverse scenarii și medii acustice. Pe lângă tehnologia modernă folosită, robustețea sistemului a fost dată de antrenarea acestuia cu cantități din ce în ce mai mari de date audio și text.

În contextul prezentat mai sus, principalele obiective ale tezei sunt:

a) Prezentarea generală a stării artei în procesarea vorbirii, mai exact în ceea ce privește recunoașterea automată a vorbirii, recunoașterea vorbitorului și adnotarea automată a corpusurilor de vorbire.

b) Proiectarea și construirea unor sisteme de recunoaștere a vorbitorilor pe seturi mari de date în limba română, aceste sisteme servind drept bază pentru cercetări ulterioare.

c) Îmbunătățirea unei proceduri de adnotare automată deja existentă folosind sisteme RAV complementare. Obținerea de noi corpusuri de vorbire adnotată în limba română folosind procedura respectivă, dar și experimentarea cu noi procedee.

d) Proiectarea și construirea unor sisteme automate de recunoaștere a vorbirii pentru limba română, folosind algoritmi de ultimă generație și valorificând seturile de date de antrenare obținute în pasul anterior.

1.3 Organizarea tezei

Această teză este organizată în șapte capitole, după cum urmează:

Capitolul 1 introduce conceptele generale de machine learning, deep learning, rețele neuronale. Sunt precizate sarcinile de procesare a vorbirii și provocările acestora. Este prezentată o scurtă istorie a abordărilor de prelucrare a vorbirii la nivel mondial, dar și a celor referitoare la limba română. Capitolul continuă cu motivația, obiectivele și organizarea tezei.

Capitolul 2 prezintă starea artei pe 3 direcții principale: recunoașterea automată a vorbirii, recunoașterea automată a vorbitorului și corpusurile de vorbire existente în limba română. Capitolul prezintă principiile acestor tehnologii și sisteme, punând accent pe arhitecturile bazate pe rețele neuronale. În ceea ce privește corpusurile în limba română, acest capitol rezumă informații despre caracteristicile fiecărui set de date existent.

Capitolul 3 descrie întregul proces de creare a seturilor de date de vorbire, exemplificând modul în care au fost obținute două astfel de seturi de date în limba română. Sunt prezentați toți pașii necesari, începând de la înregistrarea și strângerea datelor audio, până la validarea datelor, împărțirea datelor în subseturi corespunzătoare, rezumarea statisticilor legate de corpus și până la publicarea lor într-un format standard, util în special pentru sarcinile de recunoaștere de vorbire și vorbitor.

Capitolul 4 se ocupă de crearea sistemelor automate de recunoaștere a vorbitorului, antrenate pe seturi de date de vorbire în limba română. Experimentele vizează sarcinile de verificare și identificare a vorbitorului, în scenarii cu set închis și deschis. Sistemele se bazează pe două paradigme diferite: GMM-UBM și UBM-ivectors.

Capitolul 5 descrie metodologia și activitățile privind sarcina de adnotare automată a corpusurilor de vorbire. Am experimentat crearea acestor corpusuri prin filtrarea datelor pe baza metodei ipotezelor multiple, care presupune alinierea transcrierilor furnizate de două sisteme RAV complementare și considerarea părților comune ca fiind corecte. Am folosit diverse sisteme complementare și am comparat această metodă cu alte două metode de filtrare a datelor: metoda transcrierilor aproximative și metoda scorului de încredere. Noile seturi de date obținute au fost utilizate în continuare în Capitolul 6.

Capitolul 6 reprezintă partea cea mai consistentă a tezei din punct de vedere al efortului. Acest capitol este dedicat îmbunătățirii RAV în limba română. Abordarea a fost împărțită în 3 etape majore, fiecare fiind caracterizată de îmbunătățiri aduse la nivelul unei anumite componente a sistemului. Pe scurt, prima etapă a marcat tranziția de la modele acustice probabilistice la rețele neuronale bazate pe convoluție în domeniul timpului, precum și extinderea vocabularului și crearea unor modele de limbă mai complexe. Cea de-a doua etapă a evaluat mai multe tipuri de arhitecturi pentru modelarea acustică și a introdus tehnica reevaluării lingvistice cu modele de limbă bazate pe rețele

neuronale recurente. A treia etapă a reantrenat sistemul RAV utilizând diverse combinații de seturi de date audio, inclusiv cele obținute în Capitolul 5.

Capitolul 7 este rezervat concluziilor. În acest capitol sunt prezentate rezultatele obținute, contribuțiile autorului, lista lucrărilor în care au fost publicate aceste contribuții, precum și idei și direcții care pot face obiectul unor viitoare abordări de cercetare în domeniul tehnologiei vorbirii.

Capitolul 2

Starea artei

Acest capitol cuprinde conceptele fundamentale care stau la baza principalelor direcții ale acestei teze: recunoașterea automată a vorbirii, recunoașterea automată a vorbitorului și corpusuri de vorbire pentru limba română. Este oferită o trecere în revistă a principalelor abordări, pornind de la abordările clasice și punând accent pe cele considerate de ultimă generație în prezent. Acestea din urmă s-au dovedit a fi superioare din punct de vedere al performanței, datorită implementărilor bazate pe rețele neuronale, precum și avansului hardware, în special al computației folosind GPU, care a permis rularea acestor procese de calcul foarte intense.

Secțiunea 2.1 este dedicată recunoașterii automate a vorbirii și explică trecerea de la sistemele tradiționale la cele end-to-end, prezentând principalele abordări ale modelării acustice și lingvistice, parametrizării semnalului vocal și cele mai comune arhitecturi ale acestor sisteme, folosind rețele neuronale. Secțiunea 2.2 tratează subiectul recunoașterii automate a vorbitorului. Începe cu o scurtă istorie a acestor abordări și apoi sunt prezentate atât abordările probabilistice, cât și cele bazate pe neuronale. Secțiunea 2.3 rezumă toate corpusurile de vorbire adnotate în limba română. Sunt prezentate caracteristicile acestora, cum ar fi tipul de vorbire, dimensiunea corpusurilor sau informații despre vorbitori.

2.1 Recunoașterea automată a vorbirii

Această secțiune prezintă conceptele de bază în recunoașterea automată a vorbirii. Ea prezintă drumul de la RAV tradițional la RAV end-to-end. În continuare, aceasta descrie cele mai comune trăsături din vorbire care sunt utilizate în implementările actuale de ultimă generație. Introducem principiile de bază din RAV tradițional și prezentăm caracteristicile diferitelor abordări end-to-end. Sunt rezumate modelele de limbă comune și tehnicile lor de integrare în sistemele RAV. În final, vă prezentăm în detaliu un număr de 8 implementări RAV de ultimă generație, oferind detalii cu privire la caracteristicile lor arhitecturale.

2.2 Recunoașterea automată a vorbitorului

2.2.1 Definirea sarcinii

Biometria, știința care se ocupă cu studiul statistic și cu tehnicile de măsurare aplicate organismelor vii, a devenit un subiect de studiu foarte intens în ultimii ani, mai ales în contextul securității datelor, un domeniu crucial și foarte sensibil. Poate fi integrat în orice sistem de securitate care presupune verificarea sau identificarea utilizatorilor. În fața unui sistem clasic, bazat pe parolă sau jetoane de acces, care pot fi ușor pierdute sau furate, datele biometrice au proprietatea de a se baza pe trăsăturile anatomice.

Recunoașterea vorbitorului este una dintre cele mai la modă tehnologii biometrice, dezvoltată ca o ramură importantă a procesării semnalelor digitale, împreună cu recunoașterea vorbirii. Deoarece vorbirea este una dintre cele mai naturale forme de comunicare umană, iar vocea conține o multitudine de parametri specifici vorbitorului care pot fi extrași destul de simplu din semnalul vocal, evitând interacțiunea directă cu vorbitorul, recunoașterea vorbitorului poate fi regăsită în diverse domenii. Vorbitorii diferă foarte mult prin caracteristicile lor fizice, cum ar fi formele și dimensiunile tractului vocal. În plus, vorbirea unei persoane poate fi caracterizată din punct de vedere comportamental. Mai exact, fiecare individ are propriul mod de a vorbi, în funcție de accent, ritm, intonație, pronunție și multe altele [21]. În funcție de intențiile vorbitorilor, aceștia pot fi cooperanți, dorind ca identitatea lor să fie recunoscută, sau necooperanți. Prima situație se regăsește în unele aplicații, precum controlul accesului, autentificarea tranzacțiilor (tranzacții bancare telefonice, comerț electronic) sau personalizarea dispozitivelor pe baza identității vorbitorului [26]. În a doua situație, vorbitorul nu dorește să-i fie determinată identitatea, acesta fiind un caz des întâlnit în criminalistică.

Când ne referim la recunoașterea vorbitorului, de obicei luăm în considerare două sarcini diferite: verificarea vorbitorului și identificarea vorbitorului. Verificarea vorbitorului ar trebui să verifice dacă identitatea reală a vorbitorului se potrivește cu identitatea revendicată. Identificarea vorbitorului constă în determinarea identității vorbitorului, fără a furniza nicio informație prealabilă despre posibila sa identitate. Dacă vocea lui provine cu siguranță de la unul dintre vorbitorii cunoscuți ai sistemului, se poate spune că aceasta este o sarcină de identificare a vorbitorului cu set închis. În caz contrar, dacă vocea persoanei care urmează să fie identificată nu provine de la niciunul dintre utilizatorii cunoscuți, aceasta este o sarcină de identificare cu set deschis.

Sistemele de recunoaștere automată a vorbitorilor pot fi, de asemenea, clasificate în sisteme dependente de text sau independente de text [20]. Prima categorie include sisteme în care vorbitorul trebuie să spună un cuvânt sau o propoziție predefinită. Acest lucru este obișnuit mai ales în sistemele de control al accesului unde, pentru a fi verificat, vorbitorul trebuie să rostească o parolă. O problemă care poate apărea în această situație este dată de posibilitatea de a păcăli sistemul folosind o înregistrare audio cu vorbitorul legitim în timp ce acesta rostește parola. În a doua categorie, sistemele independente de

text, nu se impune nicio constrângere asupra a ceea ce trebuie pronunțat. Aceste sisteme sunt foarte utile în criminalistică. Cu toate acestea, din punct de vedere al acurateții, sistemele dependente de text sunt mai eficiente.

2.3 Corpusuri de vorbire pentru limba română

Această secțiune trece în revistă corpusurile de vorbire adnotate existente pentru limba română. Sunt oferite detalii despre acestea, cum ar fi tipul discursului, dimensiunea seturilor, atât în ore, cât și ca număr de vorbitori sau număr de propoziții, precum și disponibilitatea seturilor: publice sau private. Sunt prezentate seturi de date aparținând grupului nostru de cercetare, precum și seturi de date externe.

Un rezumat al celor mai importante corpusuri de vorbire românești este prezentat în Tabelul 3.1. După cum arată tabelul, cele mai mari corpusuri sunt cele create în cadrul grupului nostru de cercetare de-a lungul timpului, prezentate în [16], [13], [14] și grupate în lucrarea mult mai recentă [15]. Există, de asemenea, câteva corpusuri mici pentru care sunt date detalii.

2.4 Concluziile capitolului

Acest capitol teoretic a oferit o privire de ansamblu asupra fundamentelor principalelor direcții ale acestei teze: recunoașterea automată a vorbirii, recunoașterea automată a vorbitorului și rezumarea corpusurilor de vorbire românești.

Din punctul de vedere al recunoașterii automate a vorbirii, observăm că sistemele pipeline au fost înlocuite cu sisteme end-to-end, unde parametrizarea semnalului, modelarea acustică și fonetică pot avea loc în cadrul aceleiași rețele neuronale. Astfel de rețele preiau semnal audio brut la intrare și furnizează text la ieșire, fără a fi nevoie de alinieri preexistente între semnalul audio și transcrierile corespunzătoare. În plus, pot fi utilizate modele de limbă implementate cu rețele neuronale recurente, dar ele servesc mai mult ca un pas de reevaluare față de transcrierea inițială. Au fost explorate și analizate în detaliu diferite tipuri de arhitecturi neuronale, evidențiind caracteristicile acestora.

Recunoașterea automată a vorbitorului a beneficiat de dezvoltarea largă a rețelelor neuronale, înlocuind abordările probabilistice care au reprezentat mult timp starea artei.

Din punctul de vedere al corpusurilor de vorbire adnotate în limba română, corpusuri care pot fi folosite pentru antrenarea sistemelor de recunoaștere a vorbirii și a vorbitorului, s-a considerat și se poate considera în continuare că limba română este o limbă cu resurse limitate. Deși există unele seturi de date, nu toate sunt publice și dimensiunea lor relativ mică. Față de engleză, unde există corpusuri de mii sau zeci de mii de ore, cele în limba română sunt de ordinul zecilor sau sutelor de ore de vorbire. De menționat, însă, că în ultimii ani s-au făcut eforturi de îmbogățire a resurselor audio românești.

Tabelul 3.1 Resurse de vorbire în limba română

Denumire & ref.	Tipul vorbirii	Domeniu	Dimensiune			Disp.
			Utr.	Ore	Vorb.	
RASC [10]	Citită	Articole Wikipedia	3k	4.8	N/A	public
RO-GRID [19]	Citită	General	4.8k	6.6	12	public
IIT [4]	Citită	Literatură	N/A	0.8	3	non-public
N/A [6]	Citită	Eurom-1 traduceri adaptate	4k	10.0	100	non-public
N/A [24]	Spontană	Internet, show-uri TV	N/A	4.0	12	non-public
RSS [29]	Citită	Știri, literatură	4k	4.0	1	public
SWARA [28]	Citită	Presă	19k	21.0	17	public
MaSS [5]	Citită	Biblia	N/A	23.1	N/A	public
N/A [31]	Spontană	Emisiuni știri	N/A	31.0	N/A	non-public
N/A [30]	Spontană	Bancar	N/A	40.0	30	non-public
RoDigits [11]	Citită	Cifre rostite	15k	38	154	public
RSC-train [16]	Citită	Știri, interviuri, literatură	133k	95	157	public
RSC-eval [16]	Citită	Știri, interviuri, literatură	2504	5.5	21	public
SSC-train1 [9]	Spontană	Emisiuni Radio & TV	53k	27	N/A	non-public
SSC-train2 [13]	Spontană	Emisiuni Radio & TV	170k	103	N/A	non-public
SSC-train3 [12]	Spontană	Emisiuni Radio & TV	N/A	42	N/A	non-public
SSC-train4 [14]	Spontană	Emisiuni Radio & TV	277k	250	N/A	non-public
SSC-eval1 [15]	Spontană	Emisiuni Radio & TV	3035	3.5	N/A	public
SSC-eval2 [15]	Spontană	Emisiuni Radio & TV	100	1.5	N/A	public
CoBiLiRo [15]	Spontană	Interviuri	50k	31	N/A	non-public
CoRoLa [15]	Spontană + Citită	Diverse surse: radio, înregistrări studiou, emisiuni știri, vorbitori profesionisti	30k	84	N/A	non-public
CDP-train [15]	Spontană	Parlamentul României (Camera Deputaților)	1.7M	878	2500	non-public
CDP-eval [15]	Spontană	Parlamentul României (Camera Deputaților	300	5	N/A	public

Capitolul 3

Colectarea seturilor de date de vorbire

Acest capitol prezintă activitatea de colectare, curățare și organizare a două corpuri de vorbire în limba română, care s-au desfășurat în timpul studiilor de doctorat și la care autorul a avut contribuții semnificative.

Recunoașterea automată a vorbirii necesită o cantitate mare de date pentru antrenarea modelelor. Pentru unele limbi, inclusiv româna, cea mai mare problemă este încă reprezentată de disponibilitatea resurselor acustice și lingvistice. Astfel de date nu sunt publice sau pur și simplu nu există pentru multe dintre limbile vorbite. În timp ce această problemă nu se pune la engleză, fiind disponibile corpusuri care conțin mii și zeci de mii de ore de vorbire, limba română este considerată o limbă cu resurse reduse, confruntându-se cu o lipsă de resurse care pot fi utilizate în sistemele din tehnologia vorbirii.

Ca rezumat, în timpul studiilor de doctorat am colectat două corpusuri de vorbire în limba română: RoDigits (descris în continuare în secțiunea 3.1) - un set de date de cifre conectate în limba română și Read Speech Corpus (prescurtat RSC, descris în continuare în secțiunea 3.2) - un set de date de vorbire citită colectată în mediul de laborator, fără zgomot de fond.

3.1 RoDigits

În această secțiune vă prezentăm eforturile depuse pentru colectarea, prelucrarea și diseminarea setului de date care conține cifre vorbite în limba română: RoDigits.

3.2 Read Speech Corpus (RSC)

Această secțiune prezintă cel mai mare corpus de vorbire citit în limba română disponibil public, numit RSC. Înregistrările reprezintă enunțuri citite din literatură, știri și interviuri. Acesta este considerat un corpus de vorbire de bază folosit pentru antrenarea sistemelor noastre de recunoaștere a vorbirii.

3.3 Concluziile capitolului

Acest capitol a descris toți pașii implicați în procesul de creare și publicare a seturilor de date de vorbire. Am trecut de la etapele de înregistrare și colectare, parcurgând etapele de corectare și organizare până la expunerea lor într-un format standard pentru a putea fi folosite în continuare în sarcini de recunoaștere automată a vorbirii sau a vorbitorului. Am prezentat acești pași pe câteva sarcini practice, privind realizarea a două astfel de corpuri în limba română. În plus, am furnizat statistici pentru aceste seturi de date, cum ar fi durata, numărul de cuvinte, numărul de propoziții, tipul de vorbitori sau diferite distribuții de date.

Capitolul 4

Recunoașterea vorbitorului

Experimentele de verificare și de identificare a vorbitorului reprezintă subiectul principal al acestui capitol. Acestea au fost realizate folosind corpus RoDigits, un set de date de vorbire care este de câteva ori mai mare decât cele utilizate în alte încercări similare pentru limba română.

Prima secțiune a acestui capitol acționează ca o bază, prezentând un sistem de recunoaștere a vorbitorilor de tip GMM-UBM antrenat pe o versiune mai veche a copusului RoDigits, care conține 31 de ore de cifre conectate rostite de peste 120 de difuzoare.

A doua secțiune a acestui capitol se învârte în jurul comparației dintre sistemul GMM-UBM și sistemul UBM-ivectori cu clasificare PLDA. Pentru acest sistem din urmă a fost utilizată ultima versiune a corpus RoDigits, împreună cu corpus de vorbire RSC-train, SSC-train1 și SSC-train2, însumând un total de peste 200 de ore. În această secțiune, abordarea a fost mai elaborată și pentru o evaluare mai ușoară a ambelor sisteme, metrica EER a fost utilizată pentru sarcina de recunoaștere a vorbitorului și, respectiv, rata de eroare de identificare pentru sarcina de identificare a vorbitorului.

4.1 Sistemul de recunoaștere a vorbitorilor GMM-UBM

Această secțiune prezintă primele experimente și rezultate de recunoaștere a vorbitorului independent de text, utilizând corpus RoDigits. Au fost efectuate experimente în diferite scenarii care implică verificarea și identificarea vorbitorului. Ca și în cazul recunoașterii vorbirii, parametrii principali au fost variați pentru a determina cea mai bună configurație.

4.2 Sistemul de recunoaștere a vorbitorilor UBM-ivector

Experimentele din această secțiune au fost efectuate folosind corpusul RoDigits. Corpusul a fost împărțit într-un set de înrolare și două seturi de testare (unul pentru scenariul cu set deschis și unul pentru scenariul cu set închis). Setul de înrolare cuprinde 11.120

enunțuri: 80 de enunțuri de la fiecare dintre cei 139 de vorbitori înrolati. Setul de test închis este format din 2.780 de enunțuri: 20 de enunțuri de la aceiași 139 de vorbitori de înscriere. Setul de testare deschis este format din 1489 enunțuri: aproximativ 100 de enunțuri de la fiecare dintre ceilalți 15 vorbitori. Pentru aceste experimente au fost folosite și alte corpusuri românești. Primul, RSC-train, este compus din 145.000 de enunțuri citite de la 157 de vorbitori diferiți, însumând un total de 100 de ore de vorbire. Al doilea și al treilea conțin discursuri citite din emisiunile de știri și, de asemenea, vorbire spontană, extrasă din emisiuni radio și TV, uneori afectate de zgomote de fond. Ele cuprind împreună aproximativ 224.000 de enunțuri (aproximativ 130 de ore de vorbire).

4.3 Concluziile capitolului

Acest capitol servește drept bază, oferind experimente de verificare și identificare a vorbitorului folosind vorbire în limba română. În primul rând, prezentăm o abordare simplistă, un sistem GMM-UBM inițial, evaluat în termeni de falsă respingere și falsă acceptare. Apoi, trecem la o abordare mai complexă, efectuând o comparație directă între sistemul GMM-UBM și un sistem UBM-ivectori, evaluându-le în termeni de rata de eroare egală. Un corpus de vorbire citită și un corpus de vorbire spontană au fost folosite pentru a antrena modelele universale de voce. De asemenea, a fost folosit un corpus de cifre conectate în limba română pentru înrolarea vorbitorilor. În timpul experimentelor, numărul de fișiere de înscriere și numărul de densități gaussiene au fost variate. Deoarece acest număr a fost mai mare, performanța a fost mai bună.

Atât pentru verificarea vorbitorului, cât și pentru identificarea vorbitorului în scenariul cu set închis, mai ales când sunt utilizate multe fișiere de înscriere, ambele sisteme funcționează similar, uneori GMM-UBM obținând rezultate mai bune. În schimb, pentru verificarea vorbitorilor cu set deschis, sistemul UBM-ivectors este mult mai bun.

Pentru sarcina de verificare a vorbitorului, EER a fost calculat atât pentru scenariul închis, cât și pentru scenariul deschis. Cele mai bune rezultate s-au obţinut pentru 80 de fișiere de înrolare, valoarea EER fiind egală cu 0,17%. În scenariul cu set închis, sarcina de identificare a vorbitorului obține rezultate competitive, cu erori mai mici de 1%, în timp ce în scenariul cu set deschis, ratele de eroare au fost ridicate.

Capitolul 5

Adnotarea automată a corpusurilor de vorbire

Acest capitol tratează procedura privind adnotarea automată a corpusurilor de vorbire. Se începe cu o introducere teoretică care acționează ca o privire de ansamblu, prezentând aspectele generale ale conceptului, precum și principalele direcții și abordări găsite în literatura de specialitate. Apoi, va fi descrisă o abordare practică folosind una dintre metode, explicând exact modul în care a fost aplicată pentru obținerea de noi date adnotate în limba română în mod automat.

După cum am prezentat anterior, resursele de date audio adnotate în limba română sunt limitate, ceea ce face dificilă antrenarea sistemelor de inteligență artificială, precum sistemele de recunoaștere automată a vorbirii, care necesită cantități foarte mari de date. Pe de altă parte, mass-media online este o sursă continuă de vorbire: emisiuni de radio și televiziune sau înregistrări de la instituții publice, toate fiind ușor accesibile și bogate în discursuri reale, dar neadnotate. În același timp, este evident că adnotarea manuală, deși considerată superioară în ceea ce privește acuratețea, necesită foarte mult timp și efort și dat fiind contextul învățării automate, unde sunt necesare mii de date audio, această soluție devine imposibilă sau mult prea costisitoare. De asemenea, factorul uman are propriile limitări: din diverse motive, o persoană poate greși sau, dacă dorim să transcriem date audio dintr-o limbă rară, disponibilitatea unui vorbitor al acelei limbi devine o problemă. În grupul nostru de cercetare, am abordat de-a lungul timpului mai multe metode de adnotare și apoi de filtrare a datelor adnotate, cum ar fi metoda de scorurilor de încredere, metoda ipotezelor multiple sau metoda transcripției aproximative. Însă, principala preocupare a autorului în această teză este reprezentată de metoda ipotezelor multiple, care face și subiectul acestui capitol.

Prin urmare, secțiunea 5.1 reprezintă un rezumat teoretic al metodelor de adnotare automată care au fost utilizate de-a lungul timpului în diferite grupuri de cercetare. Secțiunea 5.2 descrie în detaliu particularitățile metodei de adnotare automată folosind mai multe ipoteze. Secțiunea 5.3 prezintă experimente de adnotare automată folosind această metodă pe niște corpuri de vorbire brute, culese din mass-media din România.

Secțiunea 5.4 prezintă experimente comparative între cele 3 metode de filtrare a datelor pentru adnotarea automată: metoda ipotezelor multiple, metoda transcrierilor alternative și metoda scorurilor de încredere.

5.1 Adnotarea automată a corpusurilor de vorbire. Abordări și metodologii

Cantități mari de vorbire sunt disponibile cu ușurință în multe limbi diferite. Cu toate acestea, doar o mică parte din vorbire este transcrisă, în timp ce marea majoritate este neetichetată. Având în vedere că pentru limba română există puține seturi de date audio însoțite de transcrierea corespunzătoare, această secțiune se ocupă de sarcina de a valorifica cantități mari de vorbire neetichetată pentru a îmbunătăți sistemele de recunoaștere a vorbirii existente.

În mod tradițional, învățarea atât din datele etichetate, cât și din cele neetichetate este cunoscută sub denumirea de învățare semi-supravegheată [7] și, fără îndoială, cea mai comună clasă de metode semi-supravegheate este *self-training* [32]: antrenarea un sistem inițial pe datele etichetate existente, apoi utilizarea sistemului pentru a adnota automat datele neetichetate și folosirea acelor noi mostre pentru a reinstrui sistemul. Acest proces poate fi apoi repetat pentru mai multe iterații.

În mod crucial, predicțiile sistemului RAV inițial asupra datelor neetichetate ar putea fi eronate și ar putea produce transcrieri incorecte. Reantrenarea cu transcrieri greșite poate afecta performanța următorului model, deci trebuie să filtram predicțiile și să selectăm doar acele părți ale transcrierilor care sunt considerate de încredere. În comunitate au fost propuse mai multe abordări, dar majoritatea lucrărilor abordează doar o singură metodă. În schimb, investigăm trei clase de metode, care se bazează pe scorul de încredere, mai multe ipoteze RAV și transcrieri aproximative. În continuare, discutăm despre lucrările anterioare relevante pentru fiecare dintre aceste metode.

5.2 Metoda de adnotare a ipotezelor multiple. O abordare practică

Schema de principiu a metodei ipotezelor multiple este prezentată în Figura 2.1. Corpusul de vorbire brut, neetichetat, este transcris folosind mai multe sisteme RAV existente. În cazul acestei lucrări, am folosit un număr de două sisteme. Deoarece sistemele RAV nu sunt perfecte și transcrierile obținute conțin erori, este necesară o procesare ulterioară a transcrierilor inițiale. Prin urmare, stenogramele trec apoi printr-un proces de filtrare și selecție, cu condiția ca sistemele RAV să fie complementare, cu alte cuvinte, sistemele trebuie să greșească diferit. Pe baza acestei presupuneri, părțile identice transcrise sunt considerate a fi corecte. Complementaritatea sistemelor RAV poate fi obținută în mai

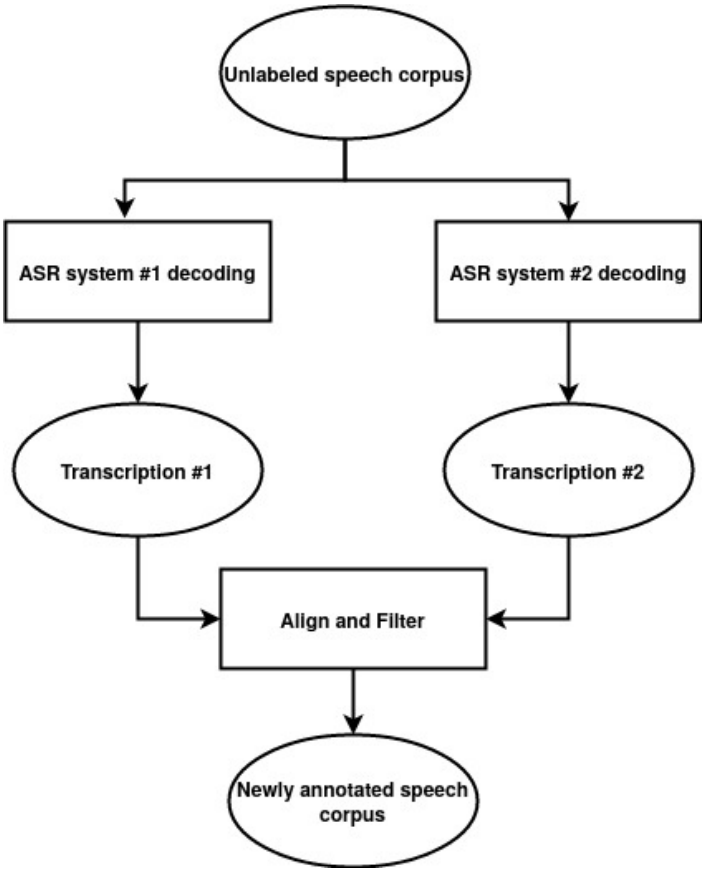

Figura 2.1 Metoda ipotezelor multiple. Sursa: [Georgescu, 2018b]

multe moduri, precum: diferite date de antrenare, diferite tipuri de caracteristici extrase din semnalul vocal, diferite arhitecturi de modele acustice sau lingvistice sau diferiți algoritmi de decodare.

5.3 Experimente de adnotare automată pentru vorbirea în limba română folosind metoda ipotezelor multiple

Această secțiune prezintă abordarea practică a adnotării automate a seturilor de date de vorbire în limba română. Sunt date detalii despre seturile de date implicate, și anume seturile de date brute, care sunt adnotate automat. Sunt descrise pe scurt sistemele RAV inițiale, evidențiind elementele care le deosebesc și le fac complementare, pentru a produce diferite erori. În primul rând, complementaritatea lor este evaluată pe un set de date adnotat manual de evaluare și apoi aceleași sisteme sunt utilizate pentru a adnota seturile de date brute. În final, sunt prezentate rezultatele procedurii de adnotare și se trag câteva concluzii.

5.4 Experimente comparative între metoda ipotezelor multiple și alte două metode

Până în acest punct, capitolul curent a fost dedicat întregului proces de adnotare automată a seturilor de date de vorbire folosind metoda ipotezelor multiple. Am discutat despre principiile care stau la baza metodei, sistemele complementare utilizate și elementele prin care acestea diferă. Am evaluat metoda din punct de vedere calitativ și cantitativ și am arătat că putem obține date cu o corectitudine de peste 95% pentru mai mult de 50% din datele inițiale, neadnotate. Am reantrenat sistemele inițiale adăugând noile date obținute automat și nu am fost tocmai mulțumiți de îmbunătățirile relative. Când am mărit setul de antrenare cu 50%, îmbunătățirile au fost de 8% și 12%, iar când l-am dublat, am obținut îmbunătățiri relative de 12% și 16%.

Ne-am propus să testăm în continuare dacă am putea fi mai eficienți folosind alte două metode de adnotare automată, pe lângă metoda ipotezelor multiple: metoda transcripțiilor aproximative și metoda scorurilor de încredere. Secțiunea actuală compară aceste trei abordări de filtrare a datelor într-un cadru experimental corect și oferă răspunsuri la întrebări precum: Care dintre abordările de filtrare este cea mai utilă? Este benefic să fie combinate? Care sunt avantajele fiecărei abordări?

5.5 Concluziile capitolului

Acest capitol a fost dedicat sarcinii de adnotare automată a corpusurilor de vorbire. Ne-am concentrat pe experimentele de adnotare automată folosind metoda de filtrare a datelor cu ipoteze multiple. Am folosit mai multe sisteme complementare, pentru care am evidențiat diferențiatorii și am testat complementaritatea acestora pe seturi de date de testare adnotate manual. În acest fel, am arătat că sistemele, diferite în unele aspecte, fac diferite erori de transcriere. În funcție de setul de date de testare, acuratețea datelor selectate este de 95% -99%, iar cantitatea variază de la 20%-30% la 70% în comparație cu corpusul brut. Desigur, există un compromis între cantitatea și calitatea datelor adnotate automat.

Reantrenând sistemele inițiale prin dublarea datelor de antrenament prin adăugarea de date nou adnotate, am demonstrat că îmbunătățirile sunt mici în comparație cu cantitatea de date noi de antrenament. Am efectuat o analiză detaliată a corpusului nou dobândit pentru a determina dacă metoda de adnotare automată produce artefacte nedorite sau nu. Concluzia a fost că noul corpus are caracteristici similare cu seturile de antrenament inițial în ceea ce privește distribuția duratei de timp pe fișiere, numărul de cuvinte și caractere pe fișier și distribuția caracterelor pe întregul set. Numai după ce cantitatea de date nou adăugate a fost de câteva ori mai mare decât datele inițiale, îmbunătățirile în acurateție au fost semnificative.

În cele din urmă, am evaluat trei metode de filtrare a datelor în contextul de sistemelor ASR *self-training* folosind date adnotate automat: ipoteze multiple, transcrieri aproximative și scor de încredere. Aceste metode au fost utilizate pentru a filtra transcrierile brute generate de un sistem RAV inițtial pentru un set de date neetichetat de aproximativ 900 de ore. Toate seturile de date filtrate (confi, 477 h; multi, 555 h; aprox, 292 h), s-au dovedit a fi benefice în reantrenarea RAV, îmbunătățind performanța sistemului RAV inițial cu 8,8%, 21,3% și, respectiv, 26,2%. Deși este cel mai mic, setul de date aprox aduce cele mai diverse date la antrenarea modelului acustic, ajutând la generalizarea vorbirii în condiții degradate. Pe partea opusă, setul de date confi cuprinde numai date care au fost deja transcrise cu încredere de sistemul RAV inițial, aducând puține informații noi. Evaluarea noastră empirică asupra vorbirii în limba română arată o îmbunătățire relativă de peste 25% față de cel mai bun sistem de până acum.

Nu în ultimul rând, adnotarea automată a corpusurilor audio poate fi considerată un subiect care poate fi încă explorat mai mult. Compromisul dintre acuratețe și cantitatea de date adnotate automat prezintă în continuare loc de îmbunătățire. Deși există diverse abordări de filtrare a datelor adnotate automat, o eventuală prelucrare a datelor care nu pot fi adnotate, datele care se pierd acum, va reprezenta cheia acestei sarcini în viitor. Cu siguranță, acele date dificile pot fi cele mai valoroase în antrenarea sistemelor bazate pe învățarea automată.

Capitolul 6

Recunoașterea automată a vorbirii pentru limba română

Capitolul de față își propune să prezinte evoluția primului sistem de recunoaștere continuă a vorbirii, cu vocabular mare, pentru limba română, bazat pe rețele neuronale, aceasta fiind principala contribuție a autorului în timpul studiilor sale de doctorat. Această lucrare a venit ca o continuare a eforturilor grupului de cercetare SpeeD de-a lungul timpului.

Îmbunătățirile aduse sistemului RAV pe limba română au fost aduse treptat la diferite niveluri și pot fi grupate în trei etape majore, după cum urmează. Prima etapă majoră de îmbunătățire (secțiunea 6.1) constă în înlocuirea utilitarului CMU Sphinx [22] cu Kaldi [25], începând a fi utilizate modele acustice bazate pe rețele neuronale, precum și tehnica reevaluării lingvistice. A doua etapă majoră de îmbunătățire (secțiunea 6.2) este dedicată explorării de noi arhitecturi de rețele neuronale pentru modelarea acustică, precum și introducerea modelelor de limbă bazate pe rețele neuronale recurente. A treia și ultima etapă majoră de îmbunătățire (secțiunea 6.3) a implicat antrenarea modelelor acustice și lingvistice folosind corpusuri extinse masiv, precum şi unele schimbări conceptuale în modelarea limbajului.

6.1 Prima abordare bazată pe DNN pentru RAV cu vocabular extins în limba română

Această secțiune prezintă prima etapă majoră în ceea ce privește îmbunătățirea sistemului nostru RAV cu vocabular extins în limba română. Conținutul acestei secțiuni poate fi considerat începutul, din punct de vedere al contribuțiilor autorului, al unui drum lung, început în 2017, care își propune să actualizeze, folosind tehnici de ultimă generație, sistemul de inițial din 2014. Prin urmare, punctul de plecare este un sistem dezvoltat în laboratorul nostru anterior acestei lucrări și descris în [9]. Ceea ce s-a folosit mai departe de la vechiul sistem la noul sistem sunt datele de antrenare și evaluare, adică corpusul

vocal, text și dicționarele fonetice. Această primă etapă de îmbunătățiri a adus câteva schimbări semnificative:

- un nou utilitar pentru antrenarea RAV: de la CMU Sphinx la Kaldi
- trecerea de la modele acustice probabilistice, de la HMM-GMM, la modele acustice bazate pe rețele neuronale: TDNN din implementarea Kaldi NNET2 și apoi TDNN din implementarea Kaldi NNET3
- noi tehnici de îmbunătățire a modelării acustice: de exemplu: speaker Adaptive Training - SAT
- algoritmi suplimentari pentru procesarea trăsăturilor vocale: am adăugat i-vectori în plus față de MFCC care sunt caracteristici standard
- progres în ceea ce privește modelarea limbajului: am păstrat modelele probabilistice de tip n-gram, dar am extins vocabularul de la 64k cuvinte la 200k cuvinte și am folosit modele de ordin 4 și 5, în timp ce în trecut ordinea maximă folosită era de 3-gram
- tehnica de reevaluare lingvistică folosind n-gram a fost folosită pentru prima dată în contextul RAV pentru limba română.

6.2 Noi arhitecturi de modelare acustică și lingvistică neuronală

Această secțiune își propune să prezinte a doua etapă majoră în ceea ce privește îmbunătățirile sistemului RAV pe limba română. Ne-am concentrat pe crearea de modele acustice de ultimă generație, bazate pe TDNN, pentru limba română și pe recalarea rezultatelor RAV cu rețele neuronale recurente profunde (RNNs) folosind implementările disponibile în biblioteca Kaldi NNET3, care a atins de multă vreme stadiul rezultate ce se incadrează la starea artei pe limba engleză, pe sarcini cunoscute, cum ar fi LibriSpeech sau TED-LIUM.

Au fost propuse, implementate și evaluate mai multe versiuni ale arhitecturii TDNN pentru modelarea acustică cu Kaldi: TDNN simplu, CNN-TDNN, TDNN-LSTM și TDNN-LSTM cu atenție. Diferitele modele acustice au fost evaluate împreună cu modele n-gram și modele de limbă recurente. Raportăm rezultate semnificativ mai bune față de sistemele RAV anterioare pentru limba română.

6.3 Modele îmbunătățite prin utilizarea unor resurse mai mari de vorbire și limbă. Actualizări ale modelului de limbă

Această secțiune prezintă a treia și ultima etapă majoră în ceea ce privește îmbunătățirile aduse sistemului nostru RAV pe limba română. Am colectat și folosit mai multe date text și audio pentru antrenarea modelelor de limbă și acustice. Mai ales în cazul modelelor de limbă, acestea necesită actualizări periodice, deoarece în limbă apar constant cuvinte noi, precum substantive proprii (nume de persoane sau entități), expresii sau cuvinte care nu există în vocabularul modelului de limbă.

O altă îmbunătățire substanțială este reprezentată de găsirea unei soluții la problema legată de cratime în modelele de limbă. Din cauza constrângerii dimensiunii vocabularului, numai cele mai frecvente cuvinte cu cratime pot fi transcrise prin sistemul RAV. În consecință, sistemul RAV încalcă normele ortografice în cazul cuvintelor cu cratime care nu se află în vocabularul LM. Soluția constă în utilizarea unei proceduri mai complexe de procesare a cratimelor în aplicația Natural Language Processing (NLP) care prelucrează textul brut, înainte de modelarea limbajului.

RAV pentru limba română este pe un trend ascendent de interes pentru comunitatea științifică. În ultimii doi ani, mai multe grupuri de cercetare au raportat rezultate valoroase privind recunoașterea vorbirii și sarcinile de dialog pentru limba română. Pentru a facilita compararea directă cu alte sisteme RAV, oferim rezultate de acuratețe pe toate seturile de date de evaluare pe care le avem și pe care le-am pus la dispoziția publicului, însumând aproximativ 15 ore de discurs adnotat manual. În comparație cu cel mai bun sistem anterior, am obținut rezultate de ultimă generație pentru vorbirea citită (WER de 1,6%) și rezultate semnificativ mai bune la vorbirea spontană (îmbunătățire relativă în jur de 40%).

6.4 Concluziile capitolului

Îmbunătățirea sistemului RAV pe limba română a fost un proces îndelungat, desfășurat treptat pe trei etape majore, care a presupus varierea arhitecturii sistemelor, varierea datelor de antrenare și o mulțime de experimente de reglare fină la nivelul parametrilor sistemului. Toate aceste etape de îmbunătățire sunt ilustrate în Figura 3.1, punând accent pe elementele inovatoare în fiecare moment, pe fiecare dintre principalele axe de dezvoltare ale sistemului RAV: modelare acustică, modelare lingvistică, vocabular, caracteristici de vorbire, corpusuri de vorbire, corpusuri de text. Figura oferă informații cu privire la progresul acurateții sistemului în vorbirea citită, în timp ce pentru vorbirea spontană rezultatele diferă în funcție de natura și dificultatea sarcinii. Descrierea detaliată a tuturor acestor abordări a constituit capitolul curent.

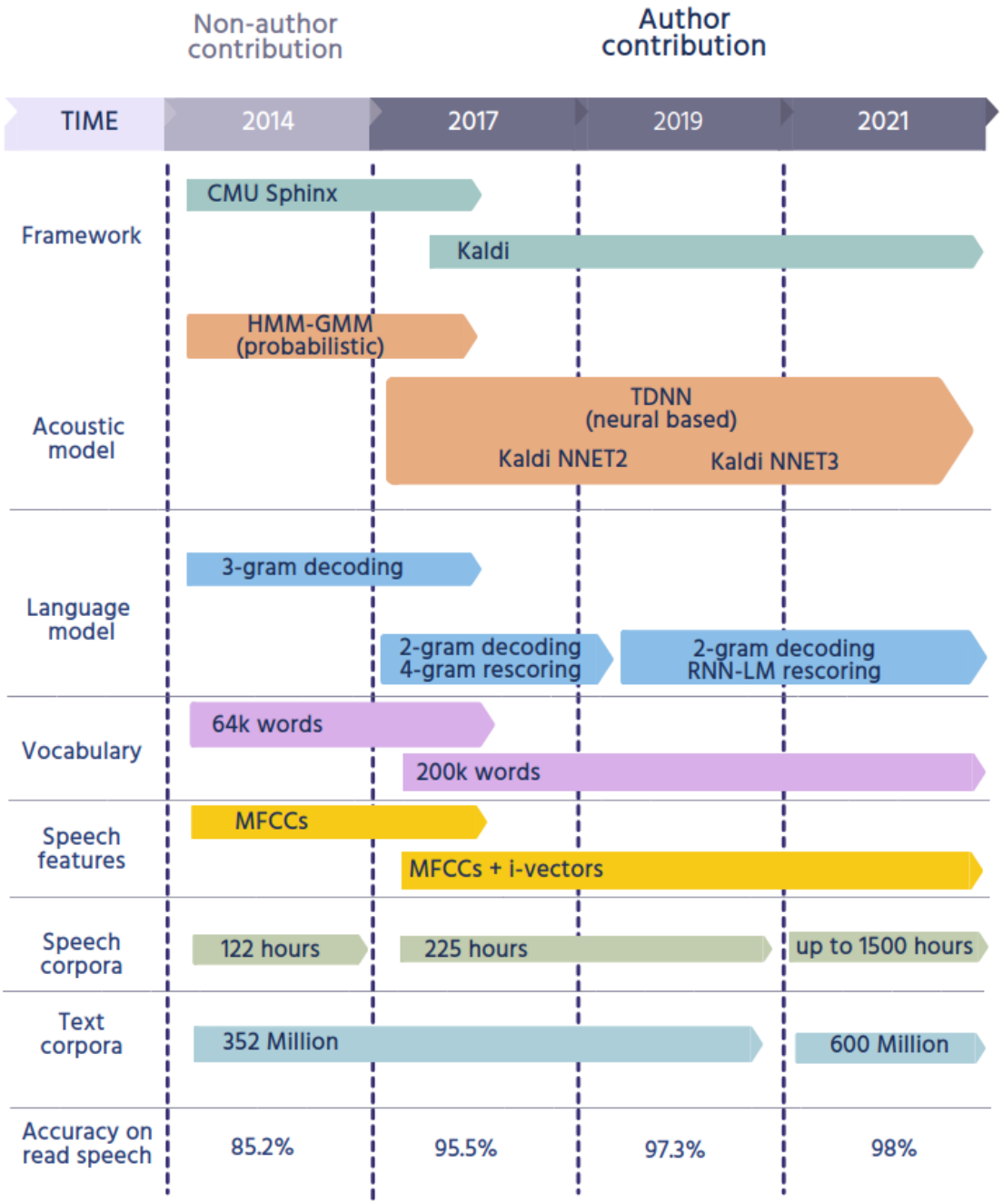

Figura 3.1 Evoluția sistemului RAV pe limba română. Diferențe la nivelul fiecărei componente de-a lungul celor 3 etape majore de imbunătățiri. Comparație cu modelul de bază din 2014.

Prima etapă a implicat schimbări destul de drastice în comparație cu sistemul existent la acea vreme. În primul rând, am schimbat algoritmii furnizați de setul de utilitarul CMU Sphinx cu alții mai moderni, disponibili în utilitarul Kaldi. Utilizarea modelelor acustice bazate pe DNN în locul modelelor HMM-GMM este una dintre cele mai importante schimbări în sistemul nostru RAV. Această modificare a condus la obținerea

unei îmbunătățiri de 20,7% până la 30,8% în funcție de tipul de vorbire (conversațional sau citit).

Creșterea dimensiunii vocabularului modelului de limbă, împreună cu utilizarea reevaluării lingvistice, a declanșat și noi îmbunătățiri ale vorbirii citite (WER cu 27,4% mai mică), dar nu au adus îmbunătățiri în vorbirea conversațională. Faptul că WER a scăzut doar cu 3% la vorbirea conversațională atunci când modelul de limbă mai mare (200k cuvinte) a fost folosit pentru decodare, este foarte intrigant. Același model de limbă a fost evaluat foarte bine în ceea ce privește rata OOV pe vorbirea conversațională (rată OOV mai mică cu 83%). Din punct de vedere lingvistic, am ajuns la concluzia că, având în vedere dimensiunea și variabilitatea corpusului nostru de text, cele mai bune rezultate se obțin folosind un LM 2-gram pentru decodarea RAV și un LM de 4-gram pentru reevaluare lingvistică, ambele bazate pe un vocabular de 200.000 de cuvinte. În general, am obținut o îmbunătățire relativă a WER de 69,6% la vorbirea citită și 48,3% la vorbirea conversațională, comparativ cu sistemul anterior.

A doua etapă a îmbunătățirilor majore a implicat testarea mai multor adaptări ale TDNN pentru modelarea acustică. Am analizat și utilizat următoarele arhitecturi: TDNN, CNN-TDNN, TDNN-LSTM și TDNN-LSTM-Attention. Pentru modelarea limbajului, am început să aplicăm reevaluarea lingvistică bazată pe modele recurente, în combinație cu decodificarea n-grame. Am comparat reevaluarea lingvistică folosind n-gram cu reevaluarea lingvistică folosind RNN-LM, dovedind că aceasta din urmă este net superioară.

S-a demonstrat că TDNN pur obține în general rezultate mai bune decât celelalte rețele. Reevaluarea lingvistică cu n-gram oferă transcrieri mai precise față de cele obținute inițial la deocodare.Folosirea unui RNN-LM are de obicei cele mai bune rezultate. Îmbunătățirile WER relative globale ale sistemului nostru RAV românesc, comparativ cu sistemul după prima etapă majoră de îmbunătățiri, sunt de 38% la citirea vorbirii și 17% la vorbirea spontană.

A treia etapă de îmbunătățiri majore a adus o creștere semnificativă a corpusurilor de vorbire și text utilizate pentru antrenarea sistemului nostru RAV pe limba română. Astfel, am folosit până la 1500 de ore de vorbire pentru a antrena modelul acustic, față de 225 de ore în trecut, respectiv am dublat numărul de cuvinte folosite pentru modelarea limbajului, ajungând la 600 M. Am efectuat diverse experimente pentru a evidenția contribuția fiecărui set de date audio din punct de vedere al preciziei. Am tras concluzii pe baza originii și tipului de vorbire corespunzător fiecărui set. Performanțele sistemului RAV pe limba română de ultimă generație este în jur de 98%-99% pentru vorbirea citită, respectiv peste 90% pentru vorbirea spontană.

Capitolul 7

Concluzii

Obiectivul principal al acestei teze a fost utilizarea metodelor și tehnologiilor de inteligență artificială pentru a aduce îmbunătățiri în 3 sarcini din domeniul tehnologiei vorbirii: recunoașterea automată a vorbirii, adnotarea automată a corpusurilor audio și recunoașterea automată a vorbitorului. Primele două sarcini au fost explorate cu succes de mult timp, în timp ce a treia sarcină are încă mult spațiu pentru explorare. Cele mai mari eforturi și cele mai multe contribuții au fost făcute în domeniile adnotării automate a vorbirii și antrenării sistemelor RAV pe limba română, aceste două sarcini fiind interdependente: evoluția uneia dintre direcții a atras și evoluția celeilalte.

Această teză a prezentat etapele succesive întreprinse în pregătirea unui sistem RAV performant pentru limba română, în paralel cu crearea automată a noilor corpuri de vorbire. După cunoștințele noastre, sistemul final RAV obține rezultate de ultimă generație.

7.1 Rezultate obținute

Capitolul 1 a descris principalele concepte de învățare automată, învățare profundă, rețele neuronale. Au fost precizate sarcinile de procesare a vorbirii și provocările acestora. A fost prezentată o scurtă istorie a abordărilor de prelucrare a vorbirii la nivel mondial, dar și a celor referitoare la limba română. Capitolul a continuat cu motivația, obiectivele și organizarea tezei.

Capitolul 2 a prezentat stadiul tehnicii în ceea ce privește principalele direcții ale tezei, și anume recunoașterea vorbirii și a vorbitorului, precum și un rezumat al seturilor de date existente privind vorbirea în limba română. Rezultatul acestui capitol a fost o imagine de ansamblu destul de complexă a celor mai utilizate sisteme pentru aceste sarcini. Foarte valoroasă a fost analiza detaliată la nivel scăzut a fiecărei rețele neuronale, descriind rolul fiecărei componente, precum și prezentarea lor prin intermediul unor diagrame conceptuale.

Capitolul 3 a prezentat o abordare concretă a colectării vorbirii în cazul a două corpusuri de vorbire în limba română, pentru a fi utilizate în sarcinile automate de recunoaștere a vorbirii și a vorbitorului. Deși la prima vedere poate părea banal, colectarea setului de date implică mult mai mult decât înregistrarea sau achiziționarea efectivă a fișierelor audio. Audio-ul a trecut printr-un proces de validare semi-automat, apoi a fost adus la un format standard. Setul de date a fost împărțit în subseturi de antrenare, dezvoltare și evaluare. Pentru a oferi un set de date pe deplin util pentru sarcinile de învățare automată, au fost efectuate o analiză detaliată a seturilor de date, cu privire la numărul de fișiere, numărul de vorbitori, precum și distribuția lor de vârstă și sex, tipul de vorbire, numărul mediu de cuvinte pe propoziție, durata medie a fișierelor. Rezultatul acestui capitol constă în publicarea a două seturi de date privind vorbirea românească, care sunt disponibile și pe site-ul SpeeD [2].

Capitolul 4 a prezentat două sisteme de recunoaștere a vorbitorului, antrenate pe seturi de date de vorbire românească, dar bazate pe paradigme diferite. Sistemele au fost comparate direct din punct de vedere al performanței, rezultatul capitolului fiind rezultatele evaluării acestora. Acest capitol poate servi drept bază pentru sarcina de recunoaștere a vorbitorilor de limbă română și poate invita alți cercetători să-și compare propriile sisteme, având în vedere că seturile de date utilizate sunt publice, conform capitolului anterior.

Capitolul 5 a prezentat adnotarea automată bazată pe metoda ipotezelor multiple, aplicată pe două corpusuri de vorbire brute, SSC-train3-raw și SSC-train4-raw, cuprinzând 136 de ore, respectiv 777 ore. Seturile de date rezultate au fost analizate, din punct de vedere calitativ și cantitativ, precum și din punct de vedere al contribuției aduse de reantrenarea sistemelor RAV. Am ajuns la concluzia că dintre metodele de adnotare automată studiate, metoda transcrierilor aproximative este cea mai utilă. Rezultatul acestui capitol au fost experimentele menționate mai sus, împreună cu noile seturi de date de antrenare obținute, SSC-train3 și SSC-train4, care conțin 42 și 250 de ore de vorbire adnotată.

Capitolul 6 a prezentat un efort îndelungat de îmbunătățire a sistemului RAV pe limba română. Prima etapă a adus cele mai multe schimbări, atât la nivel de cadru, cât și la nivel de caracteristici de vorbire, modelare acustică, modelare de limbă sau vocabular. Această primă etapă a fost marcată în special de trecerea de la sistemele probabilistice la rețelele neuronale. A doua etapă a implicat explorarea mai multor variații pentru modelarea acustică, respectiv utilizarea tehnicii de reevaluare lingvistică bazată pe RNN-LM. Ultima etapă a constat într-o creștere semnificativă a datelor de antrenare audio și text. Rezultatul acestui capitol este reprezentat de rezultatele de ultimă generație ale sistemului de recunoaștere automată a vorbirii, cu o acuratețe de 99% la vorbirea citită și de aproximativ 90%-95% la vorbirea spontană, în funcție de dificultatea sarcinii de transcriere. De asemenea, datorită faptului că seturile de evaluare au fost lansate public [2], rezultatele acestui capitol pot servi drept bază pentru alți cercetători.

Chapter 7 este rezervat concluziilor.

7.2 Contributii originale

Contribuțiile personale ale autorului acestei lucrări se regăsesc parțial în capitolul 2 și capitolul 3, dar mai ales în capitolele 4, 5 și 6. De menționat că aceste contribuții au fost posibile datorită membrilor grupului nostru de cercetare, sugestiile acestora fiind fructuoase pe parcurs. De asemenea, unele metodologii și abordări au fost dezvoltate pe lângă ceea ce exista deja în cadrul grupului. Această secțiune rezumă aceste contribuții, indicând secțiunea tezei în care apar, precum și numărul lucrării la care au fost publicate, conform celor din secțiunea 7.3, astfel:

a) Expunerea sistematizata a celor 8 tipuri de implementari RAV bazate pe retele neuronale, care a constat in mod specific in:

    a.1) descrierea amănunțită a fiecărei rețele, oferind detalii privind blocurile componente și straturile specifice și dimensionalitatea acestora
    a.2) reprezentare grafică detaliată a fiecărei rețele

   Din câte știm, aceasta este prima încercare de a oferi o astfel de analiză aprofundată pentru aceste rețele. Mai multe detalii sunt în [Georgescu, 2021c] unde am prezentat compromisul dintre performanța RAV și cerințele hardware ale acestor rețele neuronale.

b) Abordarea din Capitolul 3 de creare și publicare a două seturi de date de vorbire în limba română, RoDigits și Read Speech Corpus (RSC), într-un format util pentru sarcinile automate de recunoaștere a vorbirii și a vorbitorului. Contribuțiile au constat în:

    b.1) validarea semiautomată a corpusurilor
    b.2) organizarea subseturilor
    b.3) statistici privind caracteristicile corpusurilor

   Aceste corpusuri au fost de asemenea prezentate în [Georgescu, 2018d] și [Georgescu, 2020]. Ambele corpuri au fost utilizate în sistemele de recunoaștere a vorbitorului din capitolul 4, precum și în sistemele de recunoaștere a vorbirii din capitolul 6 și capitolul 5.

c) Proiectarea și implementarea sistemelor de recunoaștere a vorbitorului din capitolul 4, care a constat în mod specific în:

    c.1) sarcini de verificare și identificare a vorbitorilor în scenariul cu set deschis și închis folosind seturi de date în limba română

c.2) reglarea fină a parametrilor sistemului

c.3) experimente comparative între sistemele de recunoaștere a vorbitorilor bazate pe două paradigme diferite: GMM-UBM și UBM-iVectors

Toți acești pași au fost de asemenea prezentați în [Georgescu, 2018d], [Georgescu, 2018a] și [Georgescu, 2018c].

d) Îndeplinirea sarcinii de adnotare automată a corpusurilor de vorbire din Capitolul 5, care a dus la obținerea de noi date audio adnotate românești de ordinul a câteva sute de ore:

d.1) adaptarea metodei de filtrare a datelor cu ipoteze multiple folosind noi sisteme complementare RAV

d.2) experimente folosind diverse sisteme RAV complementare, evaluându-se gradul de complementaritate al acestora

d.3) proiectarea și implementarea setului de instrumente utilizat pentru colectarea datelor audio brute de pe Internet

d.4) analiza aprofundata a metodologiei de adnotare prin inspectarea caracteristicilor datelor adnotate automat, care fac obiectul sectiunii 5.3

d.5) experimente comparative între 3 metode automate de adnotare, care fac obiectul secțiunii 5.4, în vederea efectuării unei comparaţii cantitative şi calitative a datelor adnotate rezultate

Aceste experimente și rezultatele aferente lor au fost publicate în [Georgescu, 2018b], [Georgescu, 2019a], [Manolache, 2020a] și [Georgescu, 2021a].

e) Îndeplinirea sarcinii de îmbunătățire a sistemului RAV pe limba română, ceea ce a condus la obținerea unui nou sistem RAV de ultimă generație în cadrul grupului de cercetare SpeeD. Această abordare este descrisă în capitolul 6, care reprezintă contribuţia cea mai consistentă a autorului:

e.1) crearea unui numar destul de mare de sisteme RAV pentru limba romana - antrenarea si reevaluarea modele acustice, modele de limbă, modele de reevaluare, in diferite configuratii, cu date de antrenare diferite

e.2) experimente extinse de reglare fină a parametrilor sistemului

e.3) evaluarea tuturor sistemelor create și interpretarea rezultatelor, pentru a anticipa direcția următoarei serii de experimente care pot aduce mai multe îmbunătățiri ale preciziei

Toate aceste etape succesive de implementare-îmbunătățire-implementare a sistemelor RAV românești au fost publicate în [Georgescu, 2017], [Georgescu, 2018b], [Georgescu, 2018d], [Georgescu, 2019a], [Georgescu, 2019b], [Georgescu, 2020],

[Georgescu, 2021a], [Georgescu, 2021b]. De asemenea, sarcinile care implicau antrenarea noilor sisteme RAV din [Manolache, 2020a] și [Manolache, 2020b] sunt contribuția personală a autorului acestei teze.

7.3 Lista publicațiilor originale

7.3.1 Articole de jurnal

[Georgescu, 2018d] **Georgescu A-L.**, Caranica A., Cucu H., Burileanu C., "RoDigits – a Romanian connected-digits speech corpus for automatic speech and speaker recognition," *in University Politehnica of Bucharest Scientific Bulletin*, Series C, vol. 80, issue 3, pp. 45-62, Bucharest, 2018, ISSN: 2286-3540, WOS: 000440896700004, Impact Factor: 0.25, **Q4**.

[Georgescu, 2021c] **Georgescu A-L.**, Pappalardo A., Cucu H., Blott M., "Performance vs. hardware requirements in state-of-the-art automatic speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing 2021*, no. 1 (2021): 28, ISSN: 1687-4722, DOI: 10.1186/s13636-021-00217-4, WOS: 000675403700001, Impact Factor: 2.66, **Q2**.

7.3.2 Articole de conferință

[Georgescu, 2017] **Georgescu A-L.**, Cucu H., Burileanu C., "SpeeD's DNN Approach to Romanian Speech Recognition," *in the Proceedings of the 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, Romania, 2017, pp. 1-8, DOI: 10.1109/SPED.2017.7990443, WOS: 000425849600018.

[Georgescu, 2018a] **Georgescu A-L.**, Cucu H., "GMM-UBM modeling for speaker recognition on a Romanian large speech corpora," *in the Proceedings of the 12th Romanian International Conference on Communications (COMM)*, 2018, Bucharest, Romania, pp. 547-551, DOI: 10.1109/ICComm.2018.8453633, WOS: 000449526000104.

[Georgescu, 2018b] **Georgescu A-L.**, Cucu H., "Automatic annotation of speech corpora using complementary GMM and DNN acoustic models," *in the Proceedings of the 41st International Conference on Telecommunications and Signal Processing (TSP)*, 2018, Athens, Greece, pp. 794-797, DOI: 10.1109/TSP.2018.8441374, WOS: 000454845100178.

[Georgescu, 2018c] **Georgescu A-L.**, Cucu H., Burileanu C., "Comparison of i-vector and GMM-UBM speaker recognition on a Romanian large speech corpus", *in the Proceedings of the 13th Edition of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR-2018)*, Iași, Romania, pp. 25-32, WOS: 000610358400003.

[Georgescu, 2019a] **Georgescu A-L.**, Cucu H., Burileanu C., "Progress on automatic annotation of speech corpora using complementary ASR systems," *in the Proceedings of the 42nd International Conference on Telecommunications and Signal Processing (TSP)*, 2019, Budapest, Hungary, pp. 571-574, DOI: 10.1109/TSP.2019.8769087, WOS: 000493442800124.

[Georgescu, 2019b] **Georgescu A-L.**, Cucu H., Burileanu C., "Kaldi-based DNN architectures for speech recognition in Romanian," *in the Proceedings of the 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Timișoara, Romania, 2019, pp. 1-6, DOI: 10.1109/SPED.2019.8906555, WOS: 000571718700012.

[Oneaţă, 2019] Oneaţă D., **Georgescu A-L.**, Cucu H., Burileanu D., Burileanu C., "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, pp. 361-365, DOI: 10.23919/Eusipco47968.2020.9287794, WOS: 000632622300073.

[Manolache, 2020a] Manolache C., **Georgescu A-L.**, Caranica A., Cucu H., "Automatic Annotation of Speech Corpora using Approximate Transcripts," *in the Proceedings of the 43rd International Conference on Telecommunications and Signal Processing (TSP)*, 2020, Milan, Italy, pp. 386-391, DOI: 10.1109/TSP49548.2020.9163405, WOS: 000577106400084.

[Georgescu, 2020] **Georgescu A-L.**, Cucu H., Buzo A., Burileanu C., "RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition", *in the Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, Marseille, France, 2020, pp. 6606-6612, WOS: 000724697202048.

[Manolache, 2020b] Manolache C., **Georgescu A-L.**, Cucu H., Barbu Mititelu V., Burileanu C., "Improved text normalization and language models for SpeeD's Automatic Speech Recognition System", *in the Proceedings of the 13th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, ConsILR 2020, Bucharest, pp. 115-128, Romania, WOS: 000659362800011.

[Georgescu, 2021a] **Georgescu A-L.**, Manolache C., Oneaţă D., Cucu H., Burileanu C., "Data-Filtering Methods for Self-Training of Automatic Speech Recognition Systems.", *in 2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 141-147. IEEE, 2021, DOI: 10.1109/SLT48900.2021.9383577, WOS: 000663633300020.

[Georgescu, 2021b] **Georgescu A-L.**, Cucu H., Burileanu C., "Improvements of SpeeD's Romanian ASR system during ReTeRom project," *in the Proceedings of the 11th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, Romania, 2021, pp. 177-182, DOI: 10.1109/SpeD53181.2021.9587383, WOS: 000786794700032.

7.4 Perspective pentru dezvoltări ulterioare

Autorul își propune să continue munca începută în această teză cu privire la diverse sarcini de procesare a semnalului vocal folosind inteligența artificială.

Astfel, din punct de vedere al recunoașterii automate a vorbirii, deși pe limba engleză autorul a lucrat cu sistemele end-to-end studiate și prezentate în [Georgescu, 2021c], pe limba română au fost folosite până acum doar rețelele hibride TDNN-HMM pentru modelare acustică, împreună cu n-gram, respectiv RNN-LM pentru reevaluare lingvistică. Prin urmare, există suficient spațiu pentru a explora arhitecturi end-to-end pentru limba română. Acest lucru a fost abordat doar superficial de către autor până acum, în cadrul proiectului de cercetare ReTeRom [Georgescu et al.], unde unele dintre experimente au fost realizate cu DeepSpeech, dar rezultatele obținute nu au fost foarte satisfăcătoare.

Deși transcrierea vorbirii citite, cu pronunții clare, fără zgomot și dicție bună, este aproape perfectă, vorbirea spontană este totuși o provocare care poate fi explorată în continuare. Precizia vorbirii spontane de la un set de evaluare la altul este destul de diferită. Prin urmare, vorbirea spontană poate prezenta o mulțime de particularități, de la zgomot de fond, pronunții incorecte sau incomplete, accente diferite, emoții manifestate în vorbire, toate acestea fac ca sarcina de transcriere să fie foarte dificilă. Singura soluție pentru aceste scenarii este crearea de modele acustice și modele de limbă dedicate. Ar putea fi explorate în continuare abordări care presupun antrenare cu date poluate, atât natural, cât și artificial, antrenare cu date care conțin vorbire cu accente din diverse regiuni ale țării, modele de antrenament cu date specifice unui anumit domeniu de activitate.

În ceea ce privește adnotarea automată a vorbirii, nu am găsit încă o metodă de prelucrare a datelor care nu au fost păstrate în urma procesului de filtrare. De obicei, acestea corespund unor zone dificile din semnalul audio, zone cu zgomot puternic, accent puternic sau vorbire neinteligibilă uneori. Aceste zone au fost transcrise incorect de sistemele complementare RAV și ipotezele lor nu au putut fi aliniate. În acest caz, nici metoda bazată pe scoring de încredere nu funcționează, deoarece astfel de date sunt transcrise cu un scor de încredere scăzut, fiind ulterior respinse la pasul de filtrare. În momentul de față, singura metodă care poate valorifica aceste date este metoda transcrierilor aproximative deoarece presupune alinierea între o transcriere manuală și una automată, care este obligatoriu să fie corectă, chiar dacă are un scor de încredere scăzut. Folosirea acestor date pentru antrenarea modelelor acustice ar putea produce îmbunătățiri semnificative, având în vedere distribuția diferită a datelor, fiind diferită de ceea ce sistemul RAV poate deja transcrie cu succes. Acesta este și motivul pentru care cele mai utile date pentru reantrenarea sistemului RAV au fost cele obținute cu metoda transcrierilor aproximative.

În ceea ce privește recunoașterea automată a vorbitorului, această direcție nu a fost încă explorată atât de mult de către autor, cel puțin în comparație cu direcția recunoașterii

automate a vorbirii. Sistemele implementate și descrise în capitolul 4 au mai mult rolul de sisteme de bază, fiind printre puținele instruite pe cantități mari de date de vorbire românească. Sistemele folosite, GMM-UBM și UBM-ivectori, chiar dacă obțin rezultate bune la prima vedere, ele sunt, cel puțin din punct de vedere arhitectural, nu tocmai de ultimă generație. O abordare mai profundă a recunoașterii vorbitorului folosind rețele neuronale este cea descrisă în [Oneață, 20219], unde am antrenat și analizat un sistem automat de recunoaștere a vorbitorului folosind SincNet, dar autorul acestei teze are doar o contribuție parțială în lucrarea respectivă. Prin urmare, scopul în această direcție este de a lucra cu mai multe sisteme de ultimă generație. O motivație bună pentru îndeplinirea acestor sarcini ar putea fi participarea la o provocare de recunoaștere a vorbitorilor, cum ar fi provocarea de recunoaștere a vorbitorilor VoxCeleb [23] sau provocarea de recunoaștere a vorbitorilor NIST [27].

În cele din urmă, un alt subiect fierbinte în procesarea vorbirii folosind învățarea automată este cel al reprezentărilor vorbirii. Acest concept își propune să valorifice caracteristicile învățate prin antrenarea unei rețele neuronale care rezolvă o sarcină proxy, pentru a utiliza ulterior aceste caracteristici pentru sarcina principală, cum ar fi recunoașterea automată a vorbirii sau recunoașterea vorbitorului. Unele tipuri de sarcini proxy sunt: predicția cadrului următor, predicția benzilor de frecvență/cadrele mascate, verificarea ordinii cadrelor. Această abordare ar putea fi de interes în viitor, în special pentru îmbunătățirea sistemelor RAV românești.

Bibliografie

[1] (2022a). SpeeD's and Dialogue Research Laboratory. https://speed.pub.ro/. [Online; accessed 22-July-2022].

[2] (2022b). SpeeD's speech datasets. https://speed.pub.ro/downloads/speech-datasets/. [Online; accessed 22-July-2022].

[3] (2022). TADARAV. http://tadarav.speed.pub.ro/ro/. [Online; accessed 22-July-2022].

[4] Bibiri, A.-D., Cristea, D., Pistol, L., Scutelnicu, L.-A., and Turculeț, A. (2013). Romanian corpus for speech-to-text alignment. In *Proc. of the 9th International Conference on Linguistic Resources And Tools For Processing The Romanian Language*, pages 151–162.

[5] Boito, M. Z., Havard, W. N., Garnerin, M., Ferrand, É. L., and Besacier, L. (2019). Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible. *arXiv preprint arXiv:1907.12895*.

[6] Boldea, M., Munteanu, C., and Doroga, A. (1998). Design, collection and annotation of a romanian speech database. In *Proceedings of the First LREC-Workshop on Speech Database Development for Central and Eastern European Languages*. Citeseer.

[7] Chapelle, O., Schlkopf, B., and Zien, A. (2010). *Semi-Supervised Learning*. The MIT Press, 1st edition.

[8] Cucu, H. (2011). Towards a speaker-independent, large-vocabulary continuous speech recognition system for romanian. *Ph. D. dissertation, PhD Thesis*.

[9] Cucu, H., Buzo, A., Petrică, L., Burileanu, D., and Burileanu, C. (2014). Recent improvements of the speed romanian lvcsr system. In *2014 10th International Conference on Communications (COMM)*, pages 1–4. IEEE.

[10] Dumitrescu, S. D., Boroș, T., and Ion, R. (2014). Crowd-sourced, automatic speech-corpora collection–building the romanian anonymous speech corpus. *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, pages 90–94.

[11] Georgescu, A. L., Caranica, A., Cucu, H., and Burileanu, C. (2018). Rodigits-a romanian connected-digits speech corpus for automatic speech and speaker recognition. *University Politehnica of Bucharest Scientific Bulletin (submitted to)*.

[12] Georgescu, A.-L. and Cucu, H. (2018). Automatic annotation of speech corpora using complementary gmm and dnn acoustic models. In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, pages 1–4. IEEE.

[13] Georgescu, A.-L., Cucu, H., and Burileanu, C. (2017). Speed's dnn approach to romanian speech recognition. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–8. IEEE.

[14] Georgescu, A.-L., Cucu, H., and Burileanu, C. (2019). Progress on automatic annotation of speech corpora using complementary asr systems. In *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, pages 571–574. IEEE.

[15] Georgescu, A.-L., Cucu, H., and Burileanu, C. (2021). Improvements of speed's romanian asr system during reterom project. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 177–182. IEEE.

[16] Georgescu, A.-L., Cucu, H., Buzo, A., and Burileanu, C. (2020). Rsc: A romanian read speech corpus for automatic speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6606–6612.

[Georgescu et al.] Georgescu, A.-L., Manolache, C., Pop, G., Oneaţă, D., Cucu, H., Burileanu, D., and Burileanu, C. Proiect component tadarav.

[18] Ionescu, B., Ghenescu, M., Răstoceanu, F., Roman, R., and Buric, M. (2020). Artificial intelligence fights crime and terrorism at a new level. *IEEE MultiMedia*, 27(2):55–61.

[19] Kabir, A. and Giurgiu, M. (2011). A romanian corpus for speech perception and automatic speech recognition. In *The 10th International Conference on Signal Processing, Robotics and Automation*, pages 323–327.

[20] Kaur, K. and Jain, N. (2015). Feature extraction and classification for automatic speaker recognition system-a review. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(1):1–6.

[21] Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1):12–40.

[22] Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., Warmuth, M., and Wolf, P. (2003). The cmu sphinx-4 speech recognition system. In *Ieee intl. conf. on acoustics, speech and signal processing (icassp 2003), hong kong*, volume 1, pages 2–5.

[23] Nagrani, A., Chung, J. S., Huh, J., Brown, A., Coto, E., Xie, W., McLaren, M., Reynolds, D. A., and Zisserman, A. (2020). Voxsrc 2020: The second voxceleb speaker recognition challenge. *arXiv preprint arXiv:2012.06867*.

[24] Popescu, V., Petrea, C., Haneş, D., Buzo, A., and Burileanu, C. (2008). Spontaneous speech database for romanian.

[25] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society.

[26] Reynolds, D. A. (2001). Automatic speaker recognition: Current approaches and future trends. *Speaker Verification: From Research to Reality*, 5:14–15.

[27] Sadjadi, S. O., Greenberg, C., Singer, E., Mason, L., and Reynolds, D. (2022). The 2021 nist speaker recognition evaluation. *arXiv preprint arXiv:2204.10242*.

[28] Stan, A., Dinescu, F., Ţiple, C., Meza, Ş., Orza, B., Chirilă, M., and Giurgiu, M. (2017). The swara speech corpus: A large parallel romanian read speech dataset. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6. IEEE.

[29] Stan, A., Yamagishi, J., King, S., and Aylett, M. (2011). The romanian speech synthesis (rss) corpus: Building a high quality hmm-based speech synthesis system using a high sampling rate. *Speech Communication*, 53(3):442–450.

[30] Suciu, G., Toma, Ş.-A., and Cheveresan, R. (2017). Towards a continuous speech corpus for banking domain automatic speech recognition. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6. IEEE.

[31] Tarján, B., Mozsolics, T., Balog, A., Halmos, D., Fegyó, T., and Mihajlik, P. (2012). Broadcast news transcription in central-east european languages. In *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 59–64. IEEE.

[32] Triguero, I., García, S., and Herrera, F. (2015). Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems*, 42(2):245–284.