Proiect component TADARAV

- Raport ştiinţific şi tehnic în extenso -

Alexandru-Lucian Georgescu, Cristian Manolache, Gheorghe Pop, Dan Oneaţă,
Horia Cucu, Dragoş Burileanu, Corneliu Burileanu

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

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

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

Data: *29.11.2018*

Etapa: *1 / 2018*

Activitatea / activitățile:

- *Activitatea 1.10 - Studiul metodelor din literatură privind utilizarea sistemelor de RAV complementare pentru generarea automată de adnotări*
- *Activitatea 1.11 - Studiul metodelor din literatură pentru alinierea transcrierilor aproximative cu semnalul de vorbire*
- *Activitatea 1.12 - Studiul metodelor din literatură pentru generarea scorurilor de încredere (ŞI) pentru recunoaşterea automată a vorbirii (RAV)*
- *Activitatea 1.13 - Proiectarea şi implementarea unei soluții de bază de adnotare automată a semnalului de vorbire utilizând sisteme de RAV complementar*
- *Activitatea 1.14 - 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 Finanțarea Învățământului 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

Prima etapă a proiectului TADARAV a avut două obiective principale: (i) actualizarea cunoştinţelor consorţiului privind ultimele descoperiri şi invenţii în domeniul adnotării automate a datelor audio şi (îi) evaluarea posibilităţii de utilizare a unor sisteme de recunoaştere automată a vorbirii (RAV) complementare în vederea adnotării automate a semnalului de vorbire. Cele două obiective au fost realizate în proporţie de 100%, în urma activităţilor întreprinse rezultând toate livrabilele asumate de consorţiu la începutul acestei etape.

Concret, etapa 1 / 2018 a proiectului TADARAV a debutat cu cele trei activităţi de studiu al metodelor din literatură privind adnotarea automată a datelor audio. Activităţile 1.10, 1.11 şi 1.12 au fost finalizate în luna iunie cu cele trei rapoarte ştiinţifice asumate, după cum urmează:

- Studiu privind starea artei: Sisteme complementare de recunoaştere automată a vorbirii
- Studiu privind starea artei: Alinierea transcrierilor aproximative cu semnalul de vorbire
- Studiu privind starea artei: Estimarea scorurilor de încredere pentru sistemele de recunoaştere automată a vorbirii

Cele trei livrabile au 21, 25, respectiv 24 de pagini, mai mult decât minimul de 10 pagini asumat la începutul proiectului.

Etapa a continuat cu activitatea 1.13 şi anume proiectarea şi implementarea unei soluţii de bază de adnotare automată a semnalului de vorbire utilizând sisteme de RAV complementar. În cadrul acestei activităţi echipa de implementare a achiziţionat din mediul *online* câteva sute de ore de documente audio ce conţin vorbire şi le-a adnotat în mod automat folosind metodologia de adnotare proiectată şi implementată. În acest moment, soluţia se află la nivelul tehnologic asumat prin propunerea de proiect (TRL3), fapt certificat de rezultatele experimentale de transcriere a vorbirii obţinute folosind noile date audio adnotate automat (a se vedea secţiunea 2 pentru detalii).

Diseminarea rezultatelor proiectului a fost realizată în cadrul consorţiului în cele două workshop-uri organizate pe parcursul acestei etape (Bucureşti: 7 – 8 iunie 2018, şi Iaşi: 22 – 23 noiembrie 2018) şi în comunitatea ştiinţifică la *The 41st International Conference on Telecommunications and Signal Processing (TSP)*. Articolul publicat la această conferinţă, ale cărui date complete sunt indicate mai jos, este în prezent indexat *IEEE Xplore* şi în curs de indexare *Thompson Reuters CPCI* (ISI), iar numele finanţatorului este menţionat în secţiunea *Acknowledgement*, conform indicaţiilor din contractul de finanţare. Chiar dacă articolul nu este încă indexat ISI, avem certitudinea că el va fi indexat în lunile următoare, la fel cum au fost indexate ediţiile 2008 – 2017 ale acestei conferinţe de prestigiu.

- Alexandru-Lucian Georgescu, Horia Cucu, "Automatic annotation of speech corpora using complementary GMM and DNN acoustic models," în *the Proceedings of the 41st International Conference on Telecommunications and Signal Processing (TSP)*, 2018, Athens, Greece.

2 Descrierea ştiinţifică şi tehnică a activităţilor

2.1 Activitatea 1.10 – Studiul metodelor din literatură privind utilizarea sistemelor de RAV complementare pentru generarea automată de adnotări

În cadrul acestei activităţi, echipa de cercetare a UPB a studiat peste 30 de lucrări ştiinţifice relevante, dintre care a selectat aproximativ 20 de lucrări care vizează direct subiectul utilizării sistemelor de RAV complementare în vederea generării automate de adnotări. Aceste lucrări au fost analizate în detaliu şi, în urma analizei bibliografice, a rezultat un raport de 21 de pagini despre starea cunoaşterii în acest domeniu. Concluziile raportului sunt prezentate în cele ce urmează.

2.1.1 Activitatea 1.10 – Concluzii

Ideea utilizării sistemelor RAV complementare este una fiabilă în contextul adnotării automate a corpusurilor de vorbire. Diversitatea sistemelor poate fi dată de mai mulți factori, cei mai importanți fiind compoziția seturilor de antrenare acustică/lingvistică, tipul trăsăturilor extrase, tipul modelelor antrenate, tipul algoritmilor utilizați la antrenare/decodare. Nu se poate afirma cu certitudine că vreuna dintre metode oferă complementaritate mai bună, fiecare putând fi eficientă într-o situație particulară.

În ceea ce privește metodele de combinare și selecție a transcrierilor obținute de RAV, se observă faptul că metoda *Recognizer Output Voting Error Reduction* (ROVER) este o procedură clasică, ce a fost aplicată cu succes în multe situații. O altă metodă clasică este cea a utilizării scorurilor de încredere, deși nu întotdeauna sistemele RAV scot la ieșire scoruri corelate cu corectitudinea transcrierii. Totuși, alternativele ce se bazează pe tehnici de învățare automată (clasificatori de tip *Conditional Random Field* – CRF, rețele neurale etc.) au început să fie utilizate în ultima vreme, iar rezultatele sunt mai bune în comparație cu metodele clasice. Metoda stabilității acustice (A-stabil) este derivată din ROVER și a fost propusă pentru situația în care se folosesc sisteme RAV pentru mai multe limbi sursă diferite față de limba țintă.

2.2 Activitatea 1.11 – Studiul metodelor din literatură pentru alinierea transcrierilor aproximative cu semnalul de vorbire

În cadrul acestei activități, echipa de cercetare a UPB a studiat peste 50 de lucrări științifice relevante, dintre care a selectat aproximativ 30 de lucrări care vizează direct subiectul alinierii transcrierilor aproximative cu semnalul de vorbire. Aceste lucrări au fost analizate în detaliu și, în urma analizei bibliografice, a rezultat un raport de 25 de pagini despre starea cunoașterii în acest domeniu. Concluziile raportului sunt prezentate în cele ce urmează.

2.2.1 Activitatea 1.11 – Concluzii

Sarcina de aliniere transcrieri aproximative – vorbire este abordată cu două clase de metode: cele care aliniază transcrierea aproximativă direct la semnalul vocal (aliniere text-audio) și cele care aliniază transcrierea aproximativă la reprezentarea textuală a semnalului vocal, obținută cu un sistem de RAV (aliniere text-text). Aceste clase nu sunt exclusive în sensul că există și metode care combină ambele idei. Cele mai multe metode constau într-o serie de pași de aliniere succesivă și etape de segmentare a semnalului vocal prin împărțirea în unități mai mici. Alinierea este adesea bazată pe algoritmi de aliniere generici care sunt optimizați cu algoritmi de programare dinamică. Din păcate, este dificil de concluzionat care dintre metode funcționează cel mai bine pentru că seturile de date și metodologia de lucru variază – aceste motive fac imposibilă o comparație echitabilă. Din punctul de vedere al proiectului, cele mai promițătoare lucrări sunt cele care folosesc transcrierile foarte imprecise, în principal pentru că genul acesta de transcrieri sunt cele mai uzuale în practică: date încărcate de utilizatori și date descărcate de pe internet.

2.3 Activitatea 1.12 – Studiul metodelor din literatură pentru generarea scorurilor de încredere (SI) pentru recunoașterea automată a vorbirii (RAV)

În cadrul acestei activități, echipa de cercetare a UPB a studiat peste 60 de lucrări științifice relevante, dintre care a selectat aproximativ 45 de lucrări care vizează direct subiectul generării scorurilor de încredere pentru RAV. Aceste lucrări au fost analizate în detaliu și, în urma analizei bibliografice, a rezultat un raport de 24 de pagini despre starea cunoașterii în acest domeniu. Concluziile raportului sunt prezentate în cele ce urmează.

2.3.1 Activitatea 1.12 – Concluzii

Estimarea încrederii (EI) prezintă o importanță majoră în îmbunătățirea sistemelor de RAV, contribuind la procesul de detecție și corectare a erorilor. De asemenea metodele de estimare a încrederii care generează scoruri de încredere pentru transcrierile RAV sunt utilizate pe scară largă pentru auto-antrenarea modelelor acustice pentru RAV. Metodele de generare a scorurilor de încredere se încadrează în trei mari categorii: abordări bazate pe clasificare, pe probabilități posteriori și pe verificarea enunțurilor. Pentru evaluarea scorurilor de încredere se folosesc curbele ROC (*Receiver Operating Characteristic*) și DET (*Detection Error Tradeoff*) precum și indicele *Normalized Cross-Entropy* (NCE).

Dintre toate metodele de EI studiate, cele mai slabe rezultate au fost înregistrate de abordările bazate pe probabilitățile aposteriori, întrucât metodele actuale nu reușesc să estimeze corect distribuția aposteriori.

Cele mai bune performanțe au fost obținute în contextul utilizării unor trăsături predictive, ce pot fi extrase din laticea de recunoaştere, din modelul de limbă, sau din informația acustică propriu-zisă. Aceste trăsături sunt ulterior introduse la intrarea unui clasificator. În literatura de specialitate au fost propuşi o serie de clasificatori, dintre care s-au remarcat sistemele bazate pe CRF şi cele bazate pe rețele neurale recurente. Cea mai promițătoare direcție de dezvoltare în prezent o reprezintă utilizarea de rețele neurale recurente bidirecționale (BiDRNN), ce folosesc în procesul de decizie un context stânga-dreapta. Sistemele BiDRNN au depăşit performanțele celor bazate pe CRF, demonstrând o putere mai mare de generalizare.

2.4 Activitatea 1.13 – Proiectarea şi implementarea unei soluții de bază de adnotare automată a semnalului de vorbire utilizând sisteme de RAV complementar

Modelele acustice bazate pe rețele neurale profunde (*Deep Neural Network* – DNN) obțin performanțe direct proporționale cu cantitatea de date folosite la antrenarea rețelei. Prin urmare, dat fiind faptul că adnotarea manuală a resurselor audio presupune o investiție consistentă de efort şi timp, interesul față de tehnicile de adnotare automată a vorbirii a crescut semnificativ. Adnotarea automată a vorbirii presupune colectarea de vorbire în format brut şi folosirea unei metode automate pentru a produce transcrieri cât mai precise pentru cel puțin o parte din corpusul inițial.

Una dintre metodele de adnotare automată presupune folosirea unui număr de sisteme de RAV pentru generarea mai multor transcrieri aproximative, urmând ca apoi transcrierile obținute să fie aliniate iar părțile identice să fie considerate corecte. Aceasta este şi metoda folosită în etapa curentă. Două sisteme de RAV complementare au fost antrenate cu acelaşi corpus de vorbire în limba română, de două tipuri: vorbire citită şi spontană. Diferența dintre sisteme constă însă în modelul acustic folosit: HMM-GMM (*Hidden Markov Model – Gaussian Mixture Model*), respectiv HMM-DNN. În acest fel, s-a plecat de la presupunerea că erorile de transcriere vor fi necorelate, probabilitatea ca ambele sisteme să greşească identic fiind scăzută. Această ipoteză experimentală a fost validată ulterior cu rezultate concrete.

2.4.1 Descrierea metodei

Metoda utilizată în această etapă are ca scop obținerea într-un mod automat, nesupervizat, a unei adnotări cât mai precise pentru un corpus de vorbire. Corpusul nou obținut s-a dorit a fi utilizat pentru antrenarea sistemelor de RAV existente, crescând astfel variabilitatea acustică a modelelor, îmbunătățind implicit şi acuratețea transcrierilor. Paşii corespunzători metodei vor fi descrişi în continuare, aceştia fiind totodată ilustrați în Figura 1.

Ideea principală a acestei metode de adnotare automată constă în utilizarea a două sisteme RAV pentru a produce transcrieri pentru un corpus neadnotat, urmând ca apoi transcrierile să fie aliniate, iar părțile identice să fie selectate ca fiind corecte. În final, transcrierile selectate şi segmentele de vorbire corespunzătoare sunt folosite pentru a forma un nou corpus adnotat de vorbire.

Pentru ca această metodă să funcționeze este esențial ca cele două sisteme RAV să fie complementare. Mai exact, erorile celor două sisteme RAV trebuie să fie necorelate. Există câteva opțiuni care fac ca acest lucru să fie posibil: modelele acustice sau modelele lingvistice să fie antrenate pe date diferite, algoritmii de decodare să fie diferiți etc.

În abordarea curentă, complementaritatea sistemelor constă în:

- Tipul modelului acustic (HMM-GMM vs. HMM-DNN);
- Dimensiunea vocabularului (64k cuvinte vs. 200k cuvinte);
- Modelul de limbă folosit la decodare (3-gram vs. 2-gram);
- Utilizarea tehnicii de reevaluare lingvistică (fără reevaluare vs. reevaluare folosind model de limbă 4-gram).

În acest context, primul pas a presupus crearea sistemelor RAV complementare, aşa cum a fost menționat anterior. Al doilea pas a fost colectarea de vorbire neadnotată. Materialele produse zilnic de *mass-media* (emisiuni, ştiri, interviuri, reportaje) sunt o foarte bogată sursă de vorbire. Publicarea lor în mediul *online* permite o implementare automată a procesului; strângerea unei cantități semnificative de date devine doar o chestiune de timp, fără vreun efort suplimentar. Al treilea pas a constat în obținerea transcrierilor folosind cele două sisteme RAV. Transcrierile ipotetice au fost aliniate folosind un algoritm bazat pe *Dynamic Time Warping* (DTW), fiind astfel selectate părțile identice. Aşa cum experimentele arată, probabilitatea ca ambele

sisteme să producă erori identice este foarte mică. Secvențe consecutive de cuvinte, ce conțin un număr de cuvinte mai mare decât un prag determinat experimental, sunt considerate a fi corect transcrise. Un alt criteriu utilizat la selecția transcrierilor este durata secvențelor audio, fiind necesar ca aceasta să depășească un anumit prag. De asemenea, distanța în timp între două cuvinte este luată în calcul pentru a asigura faptul că nu există cuvinte intermediare ne-transcrise. La final, după ce secvențele de cuvinte corecte au fost selectate, pe baza ştampilelor de timp oferite de cel mai bun dintre cele două sisteme RAV, s-a recurs la tăierea din datele audio inițiale a părților corespunzătoare secvențelor aliniate.

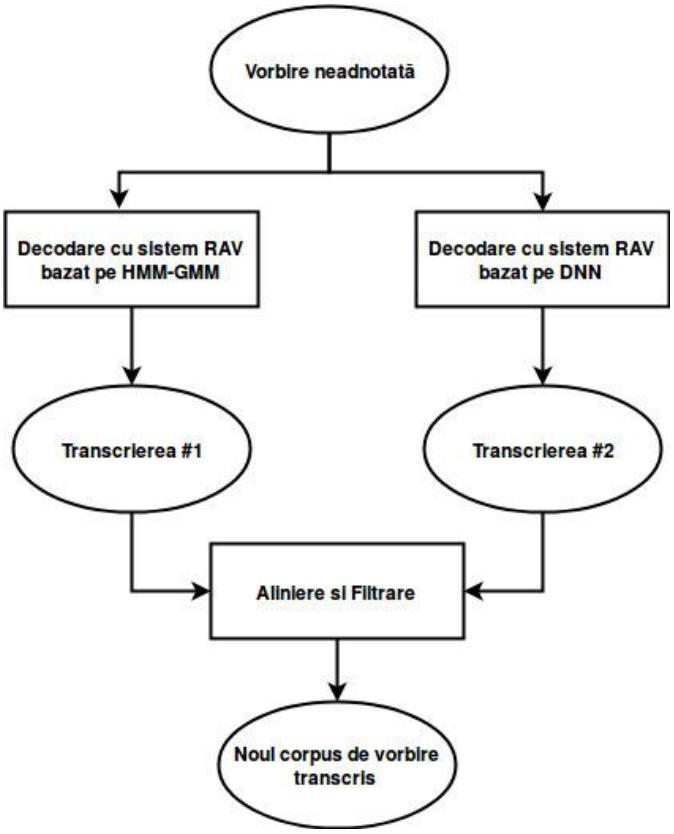

Figura 1. Diagrama metodei de adnotare automată

2.4.2 Evaluarea metodei

Utilitatea metodelor de adnotare automată poate fi măsurată folosind două criterii:

- Cantitatea de vorbire obținută în urma alinierii, raportat la dimensiunea totală a corpusului inițial;
- Calitatea adnotării, măsurabilă în eroarea la nivel de cuvânt (WER) şi/sau caracter (ChER).

Aceste metrici de performanță pot fi calculate folosind un corpus de vorbire pentru care există deja o adnotare de referință, împreună cu ştampile de timp la nivel de cuvânt. Această transcriere de referință poate fi obținută în două feluri: prin crearea ei în mod manual sau prin folosirea unui sistem RAV care să realizeze alinierea forțată a corpusului de evaluare. Deşi a doua metodă pare să fie predispusă la a genera erori, dacă sistemul RAV este destul de performant, alinierea forțată are o acuratețe suficient de bună. Astfel, corpusul este considerat la început a fi un corpus neadnotat şi va fi transcris cu ambele sisteme RAV, fiind obținute transcrieri ipotetice ce vor fi aliniate. Pe baza ştampilelor de timp din părțile aliniate, sunt selectate părțile corespunzătoare din transcrierea de referință. În acest mod, WER şi ChER pot fi calculate între cele două seturi de text. De asemenea, tot pe baza ştampilelor de timp pentru părțile aliniate, se poate calcula durata segmentelor de vorbire selectate.

2.4.3 Pregătirea experimentelor

Această secțiune prezintă resursele principale utilizate la implementarea metodei de adnotare automată propusă. Sunt oferite detalii despre corpusurile de vorbire folosite, atât pentru antrenare cât şi pentru evaluarea sistemelor RAV, dar şi despre corpusul de vorbire neadnotat. Cele două sisteme RAV sunt prezentate din punct de vedere al performanțelor şi al modelelor acustice şi lingvistice.

A. Seturile de date de vorbire

Pentru antrenarea şi evaluarea sistemelor RAV, au fost folosite două mari corpusuri de vorbire în limba română: *Read Speech Corpus* (RSC), ce conține vorbire citită, colectată în condiții de laborator, fără zgomot de fundal şi *Spontaneous Speech Corpus* (SSC), ce conține 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 conține 100 ore de vorbire citită, cuvinte izolate sau fraze de la 157 de vorbitori diferiți. RSC-eval este setul de evaluare din RSC; acesta conține vorbire de la 22 de vorbitori diferiți, însumând 5.5 ore de vorbire. SSC-train este setul de antrenare din SSC şi conține 130 ore de vorbire spontană, majoritatea din emisiuni de ştiri şi *talkshow*-uri. SSC-eval este setul de evaluare din SSC şi însumează 3.5 ore de vorbire.

Primul corpus de vorbire neadnotat a fost achiziționat din mass-media românească, mai exact de la 3 *website*-uri de ştiri şi un post de radio într-o perioadă de o lună calendaristică. Prin parcurgerea *feed*-urilor RSS al acestor *website*-uri, au fost extrase fişierele audio, eşantionate la 16 kHz, 16 biți pe eşantion.

Al doilea corpus de vorbire neadnotat a fost achiziționat de asemenea din cele 3 surse din mass-media românească, într-o perioadă de nouă luni calendaristice. Detalii despre cele 4 corpusuri de vorbire se găsesc în Tabelul 1.

Tabelul 1. Corpusurile de vorbire

Corpus	Set	Durată	
Antrenare	RSC-train	94 h 46 m	225 h 30 m
	SSC-train	130 h 44 m	
Evaluare	RSC-eval	5 h 29 m	8 h 58 m
	SSC-eval	3 h 29 m	
Corpus nou neadnotat #1	Sursa #1	65 h 55 m	136 h 41 m
	Sursa #2	51 h 27 m	
	Sursa #3	19 h 19 m	
Corpus nou neadnotat #2	Sursa #1	367 h 57 m	777 h 53 m
	Sursa #2	331 h 44 m	
	Sursa #3	78 h 12 m	

B. Modelele acustice

Sistemele RAV utilizate în această abordare sunt bazate pe modele acustice ce aparțin de două paradigme diferite: HMM-GMM şi HMM-DNN. Ambele sisteme folosesc trăsături acustice de tipul coeficienților cepstrali, mai exact vectori de 13 elemente, împreună cu derivatele lor de ordinul 1 şi 2. Coeficienții au fost extraşi cu o fereastră Hamming de 25 ms, cu suprapuneri 50%. Modelul acustic bazat pe DNN este de asemenea antrenat pe bazat unor i-vectori de dimensiune 100. Ambele sisteme modelează 36 de foneme dependente de context din limba română.

Modelul acustic bazat pe HMM-GMM cuprinde 4.000 de stări acustice (senone), fiecare dintre acestea fiind modelată de un număr de 128 densități Gaussiene. Modelul acustic bazat pe DNN este construit folosind alinierile obținute cu un model HMM-GMM. Arhitectura DNN folosită este de tipul *Time Delay Neural Network*, modelând dinamica temporală a semnalului. Stratul de intrare al rețelei constă în 3.500 neuroni, ce procesează câte 9 cadre de semnal la un moment de timp. Rețeaua are 6 straturi ascunse şi 1200 neuroni pe fiecare strat. Stratul de ieşire cuprinde 350 neuroni. Modelul a fost antrenat pe parcursul a 5 epoci, cu o rată de învățare inițială de 0.015 şi a rata finală de învățare de 0.00015. Setul de antrenare a fost divizat în mini-loturi de mărime 512.

C. Modelele lingvistice

Modelele lingvistice folosite în ambele sisteme RAV sunt de tip probabilistic – modele de limbă *n*-gram. Textele folosite la crearea acestora provin din ştiri *online* şi transcrieri ale unor conversații. Acestea au fost interpolate cu o pondere de 0.5. Primul corpus conține 315M cuvinte, în timp ce al doilea conține 40M cuvinte. Sistemul RAV bazat pe HMM-GMM foloseşte un model de limbă 3-gram ce conține 64k cuvinte, în

timp ce sistemul RAV bazat pe DNN foloseşte un model de limbă 2-gram cu 200k cuvinte pentru decodare, respectiv un model de limbă 4-gram cu 200k cuvinte pentru reevaluarea lingvistică.

D. Sistemele RAV inițiale

Tabelul 2 prezintă performanțele celor două sisteme RAV folosite în procesul de adnotare automată. Ambele au fost antrenate şi evaluate folosind corpusurile de vorbire prezentate anterior în Tabelul 1.

Tabelul 2. Sistemele RAV inițiale

Model acustic	Model lingvistic	WER[%]		OOV[%]	
		RSC-eval	SSC-eval	RSC-eval	SSC-eval
HMM-GMM	Decodare RAV: 64k cuvinte, 3-gram	12.6	32.3	1.41	2.64
HMM-DNN	Decodare RAV: 200k cuvinte, 2-gram Reevaluare lingvistică: 200k cuvinte, 4-gram	4.5	20.2	0.13	1.96

2.4.4 Rezultate experimentale

A. Evaluarea metodei de adnotare automată

Evaluarea metodei de adnotare automată a fost realizată prin compararea transcrierii ipotetice rezultată în urma aplicării metodei pe corpusul de evaluare, cu transcrierea de referință. Transcrierile generate automat nu acoperă întreg corpusul de evaluare. În consecință, comparația propusă anterior trebuie efectuată pe o selecție a corpusului de referință.

Pentru a realiza acest lucru, corpusul de evaluare a fost aliniat forțat, iar ştampilele de timp rezultate au fost utilizate (împreună cu ştampilele de timp ale transcrierii ipotetice) pentru a selecta părțile corespunzătoare din transcrierea de referință.

Tabelul 3 prezintă detalii despre părțile aliniate: erori la nivel de cuvânt şi caracter, durata datelor aliniate şi numărul de cuvinte aliniate. Rezultatele arată că metoda de adnotare automată produce corpus de calitate înaltă: 99% din cuvintele selectate sunt corecte. Acest fapt confirmă eficiența celor două sisteme RAV: erorile făcute de acestea sunt identice numai într-o mică măsură, iar părțile transcrise identic sunt aproape (99%) corecte. În concluzie, utilizarea celor două sisteme RAV care diferă prin tipul modelului acustic (HMM-GMM vs. HMM-DNN) este de departe mult mai eficientă decât utilizarea de sisteme RAV ce diferă prin alte caracteristici (seturile de date folosite sau metoda de decodare).

Tabelul 3. Evaluarea metodei

Corpus	WER [%]	ChER [%]	# ore aliniate	% ore aliniate
RSC-eval	1.0	0.3	2.62	48%
SSC-eval	1.3	0.4	0.69	20%

B. Adnotarea seturilor de vorbire nou achiziționate

Tabelul 4 prezintă cantitatea de date obținută după ce metoda de adnotare automată a fost aplicată pentru seturile de date nou achiziționate.

Pentru primul set, aproape 50% din cantitatea totală de date din sursa #1 a fost adnotată. Pentru sursa #2 au putut fi adnotate doar 20% din date, în timp ce pentru sursa #3, aproape 31% din date au fost aliniate şi adnotate automat. Aceleaşi procente de adnotare s-au obținut şi pentru cel de-al doilea set de date.

Tabelul 4. Rezultate de adnotare pe noile seturi de date achiziționate; setul #1 - stânga, setul #2 - dreapta

Corpus	# ore aliniate	% ore aliniate
Sursa #1	32 h 53 m	50%
Sursa #2	10 h 00 m	31%
Sursa #3	6 h 20 m	20%

Corpus	# ore aliniate	% ore aliniate
Sursa #1	188 h 42 m	51%
Sursa #2	66 h 02 m	30%
Sursa #3	25 h 16 m	20%

C. Recunoaşterea automată a vorbirii

Noile seturi de date de vorbire adnotată (49 ore, respectiv 280 de ore, după cum a fost prezentat anterior) au fost adăugate la seturile de antrenare deja existente. Ambele sisteme RAV au fost reantrenate şi evaluate (Tabelul 5).

Tabelul 5. Performanța sistemelor RAV după reantrenare

Corpus antrenare	Model acustic	WER[%]		Îmbunătățire relativă a WER [%]	
		RSC-eval	SSC-eval	RSC-eval	SSC-eval
RSC-train + SSC-train + set nou #1	HMM-GMM	11.70	31.30	7.14	3.10
	HMM-DNN	5.00	20.90	–10.67	–3.47
RSC-train + SSC-train + set nou #1 + set nou #2	HMM-GMM	–	–	–	–
	HMM-DNN	4.33	18.41	3.78	8.86

Surprinzător, adăugarea primului corpus adnotat la setul de antrenare, a fost avantajoasă numai pentru sistemul bazat pe HMM-GMM: pe ambele corpusuri de evaluare, noul sistem bazat pe HMM-GMM a obținut îmbunătățiri relative minore. Situația diferă însă în cazul sistemului bazat pe HMM-DNN, care a obținut rezultate mai slabe decât sistemul inițial.

Adăugarea celui de-al doilea corpus adnotat s-a utilizat numai pentru reantrenarea sistemului ASR bazat pe HMM-DNN. Pe ambele corpusuri de evaluare au fost obținute uşoare îmbunătățiri.

2.4.5 Concluzii

În această etapă a fost utilizată o metodă de adnotare automată a corpusurilor de vorbire ce presupune utilizarea transcrierii de la două sisteme RAV complementare. Experimentele au arătat că sistemele RAV bazate pe HMM-GMM, respectiv HMM-DNN, fac semnificativ mai puține erori comparativ cu situația când ambele sisteme folosesc modele acustice de tip HMM-GMM. În prima situație, eroarea la nivel de cuvânt (WER) este de aproximativ 1%, în timp ce rezultatele prezentate într-o lucrare anterioară indică o rată de eroare la nivel de cuvânt de aproximativ 10%.

Este de remarcat faptul că sistemele de RAV inițiale au o precizie destul de bună, fiind antrenate cu o foarte mare cantitate de date. Acesta poate fi un motiv pentru care reantrenarea sistemelor adăugând noile corpusuri obținute prin această abordare nesupervizată nu aduce îmbunătățiri semnificative. Din cantitatea totală de date a fiecărui nou corpus achiziționat, un procent de aproximativ 36% au fost adnotate corect. După reantrenarea modelelor acustice adăugând primul corpus achiziționat, sistemul de RAV bazat pe HMM-GMM a obținut o mică îmbunătățire, în timp ce sistemul de RAV bazat pe HMM-DNN a pierdut puțin în ceea ce priveşte performanța. Reantrenarea sistemului HMM-DNN adăugând ambele corpusuri nou achiziționate a condus la o uşoară îmbunătățire a performanțelor.

2.5 Activitatea 1.14 – Diseminare

Diseminarea rezultatelor proiectului a fost realizată în cadrul consorțiului în cele două workshop-uri organizate pe parcursul acestei etape (Bucureşti: 7 – 8 iunie 2018, şi Iaşi: 22 – 23 noiembrie 2018) şi în comunitatea ştiințifică la *The 41st International Conference on Telecommunications and Signal Processing (TSP)*. Articolul publicat la această conferință, ale cărui date complete sunt indicate mai jos, este în prezent indexat *IEEE Xplore* şi în curs de indexare *Thompson Reuters CPCI* (ISI), iar numele finanțatorului este menționat în secțiunea *Acknowledgement*, conform indicațiilor din contractul de finanțare. Chiar dacă articolul nu este încă indexat ISI, avem certitudinea că el va fi indexat în lunile următoare, la fel cum au fost indexate edițiile 2008 – 2017 ale acestei conferințe de prestigiu.

- Alexandru-Lucian Georgescu, Horia Cucu, "Automatic annotation of speech corporă using complementary GMM and DNN acoustic models," în *The Proceedings of the 41st International Conference on Telecommunications and Signal Processing (TSP)*, 2018, Athens, Greece.

3 Structura ofertei de servicii de cercetare şi tehnologice

Laboratorul de cercetare *Speech and Dialogue* (SpeeD) din cadrul Universității Politehnica din Bucureşti (UPB), reprezentantul UPB în proiectul TADARAV, oferă pe platforma ERRIS serviciile de cercetare şi tehnologice enumerate în Tabelul 6.

Tabelul 6. Servicii de cercetare şi tehnologice oferite de Laboratorul de cercetare *Speech and Dialogue*

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

Echipa de cercetare a Universității Politehnica din Bucureşti pentru proiect component TADARAV este prezentată în Tabelul 7.

Tabelul 7. Echipa de cercetare UPB

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

Cei trei noi membri cercetători au fost angajați începând cu data de 25.04.2018 cu funcțiile de cercetător ştiințific (CS), respectiv asistenți cercetare ştiințifică (ACS), cu normă întreagă.

5 Valorificarea şi îmbunătățirea competențelor şi resurselor existente la nivelul consortului

La nivelul proiectului component TADARAV CEC-urile nu au fost valorificate.