Raport ştiinţific şi tehnic

Etapa a III-a, an 2020

„Resurse şi tehnologii pentru dezvoltarea interfeţelor om-maşină în limba română”

Aceste rezultate au fost obţinute prin finanţare în cadrul Programului PN-III Proiecte complexe realizate în consorţii CDI, derulat cu sprijinul MEN – UEFISCDI,
Cod: PN-III-P1-1.2-PCCDI-2017-0818, Contract Nr. 73 PCCDI/2018

Denumirea organizaţiei participante în proiect	Acronim organizaţie	Tip organizaţie	Rolul organizaţiei în proiect (Coordonator/partener)
Institutul de Cercetări Pentru Inteligenţă Artificială “Mihai Drăgănescu”	ICIA	UNI	CO
Universitatea Tehnică din Cluj-Napoca	UTCN	UNI	P1
Universitatea Politehnica din Bucureşti	UPB	UNI	P2
Universitatea "Alexandru Ioan Cuza" din Iaşi	UAIC	UNI	P3

MINISTERUL EDUCAŢIEI NAŢIONALE

UEFISCDI

Raport ştiinţific - tehnic proiect component CoBiLiRo

Date de identificare proiect

Număr contract:	PN-III-P1-1.2-PCCDI-2017-0818, Nr. 73 PCCDI/2018
Acronim / titlu:	**„COBILIRO:"**
Titlu livrabil:	**Raport ştiinţific şi tehnic (Etapa a III-a, 2020)**
Termen:	**Noiembrie 2020**
Editor:	**Dan Cristea**
Adresa de eMail editor:	**danu.cristea@ gmail.com**
Autori, în ordine alfabetică:	**Anca Bibiri, Serban Boghiu, Dan Cristea, Daniela Gîfu, Felix Cristian Pericică, Ionuț Pistol, Mihaela Onofrei, Andrei Scutelnicu**
Ofiţer de proiect:	**Cristian Stroe**

1. Rezumatul etapei III

În această etapă a proiectului complex ReTeRom consorţiul şi-a propus să consolideze şi apoi să exploateze rezultatele acumulate în primii doi ani, cu obiectivele: existenţa unui Portal pregătit a primi şi prelucra resurse bilingve româneşti, dezvoltarea în continuare a unei colecţii de resurse care să corespundă formatului agreat de consorţiu, perfecţionarea lanţurilor de prelucrări lingvistice şi sonore, atât asupra componentelor textuale cât şi vocale ale resurselor bimodale, care să permită alinieri între componentele vocale şi textuale, recunoaşterea cu minimum de erori a vocii, generarea expresivă a vocii şi antamarea de aplicaţii bazate pe aceste tehnologii.

Conţinutul repozitorului de resurse, acumulate pe portal în primii doi ani ai proiectului, a fost completat cu noi colecţii bimodale, iar acestea au fost valorificate atât în cadrul platformei (prin realizarea de instrumente integrate generatoare de statistici) precum şi în afara platformei, prin propunerea unei serii de proiecte ce utilizează tehnologiile dezvoltate în cadrul proiectelor partenere. Similar celorlalte etape, a fost realizată şi o activitate de diseminare, în principal prin publicaţii. De interes special pentru platforma dezvoltată a fost respectarea drepturilor de autor şi anonimizarea datelor (cu precădere cele vocale) contribuite de diverşi furnizori.

2. Descrierea rezultatelor activităţilor etapei III

Activitatea A3.1 este intitulată „**Augmentarea corpusului vocal prin noi înregistrări vocale care dublează texte existente în corpusul CoRoLa**".

Preocuparea membrilor echipei a fost orientată în principal în direcţia achiziţionării din diverse surse ori chiar a creării prin mijloace proprii de înregistrări sonore însoţite de transcrierile lor textuale şi completarea metadatelor corespunzătoare. Ca urmare, **Raportul A3.1** prezintă rezultatele activităţii, sintetizând noile colecţii de documente bimodale ce au fost urcate pe platforma CoBiLiRo în decursul acestui an, după ce au trecut prin etapele de obţinere a drepturilor de autor, unele dintre ele fiind aliniate, parţial automat - parţial manual, cu înregistrările sonore. Resursele nou urcate în Portal în acest an al proiectului includ 125 de piese, aparţinând următoarelor colecţii (descrise în atributul @**colection** din header-ul metadatei):

- *Ghici Cine Vine La Cină?* (cu 13 interviuri, voci de bărbaţi);
- *Alma Mater Iassiensis* (cu 13 interviuri, voci de bărbaţi);
- *SoRoEs* (cu 91 de înregistrări pe teren, voci de bărbaţi şi de femei din diferite regiuni ale României);
- *Poveşti* (cu 8 înregistrări de poveşti realizate în studio).

Toate resursele au atributul @**distribution** setat la valoarea: https://creativecommons.org/licenses/by-nc/3.0/. Colecţiile încărcate pe platforma CoBiLiRo respectă formatul standardizat al fişierelor vorbire-text relativ la metadate (care conţin informaţii header şi alinieri). Formatul, agreat de toţi partenerii proiectului complex ReTeRom, este difuzat în proiect în livrabilul A1.3 şi descrie tipurile:

- *file*: care separă fiecare componentă <speech> de format WAV/MP3, cu includerea atributului @**speechFile**, care identifică numele fişierului conţinând înregistrarea vocală;
- *start-stop*: care marchează bornele temporale de început-sfârşit ale componentelor <speech> în fişiere unice asociate unui obiect <tei>, cu includerea atributelor @**start** şi @**stop**, cu valori reale,

reprezentând momentele de început şi de sfârşit ale înregistrării vocale; în acest caz, numele fişierului în care sunt referite bornele @**start** şi @**stop** ale elementelor <speech> sunt date într-un atribut @**speechFile** încorporat elementului <teiHeader>.

Identificarea în fişierele colecţiei corpusului Cobiliro a primei ori a celei de a doua opţiuni de marcare a segmentării se face în elementele <teiHeader>, care includ un atribut @**speechSegmentation**, cu una din valorile: **"file"**, respectiv **"start-stop"**, tipul de segmentare fiind unitar în lungimea fiecărui obiect. Menţionăm că majoritatea obiectelor înregistrate în Portal sunt de tip *file*.

Înregistrările sonore sunt neuniforme în privinţa apropierii de text, unele fiind prelucrate deja de editori (pentru îndepărtarea unor repetiţii, pauze, bâlbâieli, inerente vorbirii libere) în vederea tipăririi textelor (ca de exemplu colecţiile de interviuri), altele urmărind destul de aproape idiosincraziile din vorbirea liberă.

Calitatea înregistrărilor sonore este, de asemenea, diversă: de la înregistrări realizate în condiţii profesionale (microfon performant, cameră anecoidă), până la înregistrări în condiţii naturale (pe stradă, cu zgomot sau muzică în surdină etc.). Părerea noastră este că fiecare dintre aceste tipuri de resurse bimodale este utilă în experimentele de antrenare a tehnologiei de recunoaştere a vocii, cât şi în cea de sinteză vocală, pentru că, în condiţii reale de transpunere a vocii în text ori invers, toate aceste condiţii pot apărea, ca urmare tehnologiile trebuie să poată fi antrenate pe resurse care să reproducă condiţii de calitate cât mai diverse.

A fost, de asemenea, agreată de către membrii colectivului CoBiLiRo o procedură de dublare în voce (*read speech*) a unei colecţii de fişiere textuale aflate deja în CoRoLa (*Corpusul Reprezentativ al Limbii Române Contemporane*), corpus care a fost construit între anii 2014-2017, prin colaborarea dintre două institute de cercetare ale Academiei Române: Institutul de Inteligenţă Artificială "M.Drăgănescu" din Bucureşti şi Institutul de Informatică Teoretică din Filiala Iaşi a Academiei Române, cu ajutor (în cadrul proiectului DRuKoLA finanţat de Fundaţia Humboldt), din partea Universităţii Bucureşti şi a Institutului Limbii Germane din Mannheim. Această activitate se va continua în perioada de prelungire a proiectului, până în martie 2021. Rezultatul acestei activităţi are o dublă semnificaţie, pentru că astfel se măreşte atât colecţia de înregistrări din Portalul CoBiLiRo, cât şi cea din CoRoLa.

Consorţiul a acordat o grijă deosebită respectării cadrului legal de procurare şi utilizare a resurselor în scopuri de cercetare, pentru respectarea drepturilor de autor şi a securităţii informaţilor private. Pentru fiecare dintre resursele încărcate pe platforma CoBiLiRo, membrii Consorţiului care le-au achiziţionat au semnat protocoale de colaborare cu autorii ori proprietarii lor. Alte discuţii s-au purtat pentru stabilirea unui punct de vedere relativ la posibilitatea de a ceda drepturi de utilizare unor terţe entităţi. Cum părerile au fost divergente în această privinţă, s-a convenit ca fiecare partener să stabilească propriii termeni de licenţiere asupra resurselor procurate ori construite, iar în cazul în care resursele ce sunt solicitate de terţi au intrat în posesia consorţiului prin donaţii din partea altor autori sau furnizori să se revină asupra protocoalelor semnate cu aceştia iniţial pentru a obţine din partea lor acceptul de extensie a dreptului de utilizare către terţi. Opinia membrilor Consorţiului asupra drepturilor de exploatare a resurselor din Portalul Cobiliro este către o utilizare liberă pentru scopuri de cercetare. Atunci când rezultatele cercetării sunt însă explicit enunţate ca urmărind obţinerea de profit, opiniile sunt divergente, pentru că, pe de o parte nu se poate pune problema participării la profit a unora dintre parteneri (spre exemplu, obţinerea de profit nu este admisă în Statutul Academiei Române), iar pe de altă parte, e discutabil aspectul moral al obţinerii de profit din resurse cedate gratuit de proprietari, chiar dacă acestor resurse le-a fost adăugată o plus-valoare rezultată în urma cercetării.

Toate obiectivele incluse în plan la această activitate au fost realizate.

Activitatea A3.2 este intitulată „**Augmentarea corpusului voce-text: completare de metadate, alinieri prin rularea algoritmilor dezvoltaţi în proiectele P2, P3 şi P4 şi adnotări manuale şi semiautomate la corpusul voce/text**".

În **Raportul A3.2** este descris modul în care resurselor bimodale aflate în Platforma CoBiLiRo li s-au asociat alinieri între secţiunile de voce şi de text, majoritar automat, prin rularea tehnologiilor-servicii create în proiectele P2, P3 şi P4. Pentru că unele transcrieri textuale nu redau cu acurateţe echivalentele lor sonore (cum se întâmplă, de exemplu, în colecţia de interviuri, înregistrate în direct, care apoi au fost transcrise în text şi voit „înfrumuseţate" în vederea tipăririi), alinierile au pus în evidenţă doar zonele de coincidenţă.

Folosind aceste servicii, au rezultat două tipuri de procesări:

- Un prim proces de aliniere dintre fişierul audio şi text (Figura 1), în care sunt marcate timpii de început şi sfârşit al fiecărui cuvânt.
- Cel de-al doilea tip de procesare în care au fost marcate *timestamp*-uri aferente secvenţelor de cuvinte. Formatul este după cum urmează: câte un fişier text pentru fiecare resursă audio. Pe fiecare linie este câte o propoziţie (cu *timestamps* la nivel de cuvânt într-o variantă; fără *timestamps* în cealaltă variantă - după cum se poate observa în Figura 2). În cadrul lunilor de implementare rămase aferente proiectului se va

realiza prelucrarea acestui rezultat oferit de serviciul TADARAV, pentru a aduce acest rezultat la formatul CoBiLiRo convenit (livrabilul A3.1).

Datorită unor probleme tehnice întâmpinate pe serverul care găzduieşte Portalul CoBiLiRo, o serie de înregistrări ale vechilor resurse au fost pierdute din baza de date. Pentru ca ele să redevină vizibile în Portal, s-a realizat un soft care a parcurs lista tuturor directoarelor hard-ului care găzduia back-up-urile resurselor şi le-a reintrodus în baza de date a Portalului.

În momentul de faţă pe server există aproximativ 70 de mii de fişiere audio cu perechile textuale respective ce conţin transcrieri ale înregistrărilor sonore, precum şi fişierele corespunzătoare cu metadate.

Toate obiectivele incluse în plan la această activitate au fost realizate.

Activitatea A3.3 este intitulată „**Realizarea de statistici privind corpusul bimodal voce/tex**t". Raportul prezintă toate tipurile de statistici implementate în decursul acestui an şi care pot fi lansate la cerere asupra resurselor voce-text existente în Portalul CoBiLiRo. Statisticile prezintă, la cererea utilizatorului, parametri calculaţi din exploatarea informaţiilor conţinute în fişierele text şi audio, în alinierile dintre cele două, cât şi în metadatele aferente.

Componenta de procesare lingvistică necesară obţinerii de valori statistice, este realizată sub forma unui script Python, încadrat într-un web API, care poate fi accesat folosind principiile arhitecturii Restful, apelat doar la nivel intern (pe server); aici s-a folosit tehnologia **Flask Restful**.

Pentru realizarea prelucrării şi procesării la nivel de limbaj natural a textelor a fost utilizată librăria open-source **Spacy**, mai exact modelul **ro_core_news_sm**; acest model atribuie vectori de tokeni specifici contextului şi etichete ale părţilor de vorbire având o acurateţe de 95.62%. Ca alternativă poate fi folosită componenta de analiză a textului produsă de TEPROLIN; motivul pentru care nu a fost inclusă în versiunea iniţială este că nu toate fişierele de pe platformă au fost procesate de acea componentă. Pentru finalul proiectului avem în plan tranziţia spre tehnologia TEPROLIN.

Statisticile implementate se referă la: liste de obiecte obţinute pe baza informaţiilor din header-ul metadatelor (ca de exemplu: obiecte în care componentele vocale au formatul wav/mp3, obiecte cu componenta vocală segmentată la nivel de fişier/cuvânt, obiecte voce-text aliniate la nivel de frază/cuvânt/fonem etc.), statistici de natură lexicală (densitatea lexicală, numărul de token-uri, de leme diferite, de forme flexionare distincte, raportul tip-token etc.), precum şi statistici la nivelul întregii colecţii de resurse (numărul total de ore de vorbire, de token-uri, densitatea lexicală medie, raportul mediu al numărului de token-uri pe categorii etc.).

Oferirea unor statistici relevante pentru o colecţie de resurse contribuie atât la mărirea încrederii unor potenţiali noi contributori cât şi la potenţialul resurselor din colecţie de a-şi găsi utilitatea în diverse contexte. Statisticile deja incluse pe platformă precum şi cele care vor fi incluse până în aprilie 2021 promit să asigure acest avantaj pentru colecţia de resurse bimodale, principalul rezultat al proiectului CoBiLiRO.

Până la terminarea proiectului avem în vedere ca procesele de determinare a lungimii fişierelor să fie lansate doar asupra fişierelor nou introduse sau updatate. Serverul e capabil să urmărească fişierele asupra cărora s-a umblat şi să calculeze/recalculeze lungimile doar pentru acestea. Odată aflată o lungime, ea se va înscrie în metadate, ceea ce înseamnă că nu va trebui recalculată la fiecare cerere din partea utilizatorului.

Toate obiectivele incluse în plan la această activitate au fost realizate.

Activitatea A3.4 este intitulată „**Proiectare de aplicaţii de exploatare a corpusului bimodal şi a tehnologiilor de prelucrări textuale şi voce, create în proiectele P2, P3, P4**".

Raportul activităţii prezintă un număr de posibile aplicaţii care să demonstraze potenţialul tehnologiilor dezvoltate în celelalte proiecte componente, în scopul măririi vizibilităţii proiectului, în special după finalizarea acestuia. Tehnologiile vizate în special sunt cele descrise în rapoartele 3.7 (Definitivarea, testarea, validarea şi împachetarea într-o soluţie „*ready-to-use*" a platformei integrate şi configurabile de prelucrare a textelor în limba română.), 3.10 (Îmbunătăţirea soluţiei de filtrare şi aliniere a transcrierilor aproximative cu semnalul de vorbire), 3.15 (Dezvoltarea unei noi tehnologii pentru adaptarea vocii sintetice la stilul şi expresivitatea unui nou vorbitor) şi 3.17 (Integrare tehnologie nouă şi demonstrare în realizarea interfeţelor om-maşină pentru sinteza text-vorbire). O parte din proiectele descrise au depăşit deja faza de proiect, fiind în stadii relativ avansate de implementare (în activităţi cu studenţii de la UAIC-FII).

Cele 6 proiecte de aplicaţii descrise sunt prezentate succint mai jos:

Aplicaţia 1: Suport pentru învăţarea limbii române - *PD builder*. Un dicţionar de pronunţie pentru cuvintele unei limbi poate constitui un suport semnificativ pentru cei care doresc să înveţe acea limbă. Colecţia de resurse colectate pe platforma CoBiLiRO are potenţialul să atingă o dimensiune suficientă pentru a include pronunţia marii majorităţi a cuvintelor din limba română; de aici a plecat ideea de a dezvolta o

tehnologie ce caută în resursele din Portal şi reproduce pronunţii ale unor cuvinte, la cererea utilizatorului. Sunt utilizate adnotările produse de TEPROLIN şi alinierile produse de TADARAV, marcaje ce însoţesc majoritatea resurselor existente pe platforma CoBiLiRo.

Aplicaţia 2: Analiza corpusurilor bimodale. Proiectul propune dezvoltarea unui sistem capabil să evalueze calitatea unui corpus bimodal voce-text din perspectiva unei alinieri automate sau manuale. O parte din posibilele erori semnalate ar putea fi corectate, lucru care ar mări semnificativ calitatea alinierii automate dar şi a resursei bimodale în general. Prin corelarea statistică a mai multor parametri, pot fi detectate posibile diferenţe între conţinutul înregistrării şi transcrierea ei în text, diferenţe care pot avea cauze diverse: segmente adiţionale prezente în text faţă de înregistrare sau invers, transcrieri aproximative, particularităţi ale pronunţiei şi prozodiei vorbitorului etc.

Aplicaţia 3: ***I listen to my speaking agent reading book fragments as I walk by***. Aplicaţia are la bază o colecţie de texte care abundă în entităţi geografice, marcate XML explicit, textele fiind însoţite de metadate care descriu minimum: autorul şi titlul cărţii, anul de apariţie şi editura. Instalată pe un dispozitiv mobil, ea va semnala proximitatea telefonului faţă de locaţiile menţionate în texte şi va citi acele fragmente care includ menţiunile respective. În felul acesta, o plimbare printr-un mare oraş se poate transforma într-o călătorie literară. Aplicaţia se adresează persoanelor care ar dori să primească sugestii de lecturi şi doresc să afle lucruri noi despre locurile pe care le vizitează. Ea îşi propune să îmbine plăcut literatura şi tehnologia.

Aplicaţia 4: Sistem de sinteză text-vorbire (TTS) şi clonarea vocii în limba română cu metoda învăţării prin transfer. Aplicaţia poate fi utilă persoanelor care şi-au pierdut capacitatea de a vorbi, dar există înregistrări precedente pierderii vocii. Alte utilităţi poy fi imaginate în învăţământul la distanţă, în muzee, în colecţiile de moşteniri culturale etc., unde tehnologia poate fi folosită pentru reproducerea vocii personalităţilor istorice. Clonarea vocii reprezintă un caz particular de TTS, în care un text este sintetizat în vocea unui vorbitor necunoscut folosindu-se un număr limitat de înregistrări audio ale acestuia acompaniate de transcrierile lor.

Aplicaţia 5: Asistent inteligent al şedinţelor online. Pandemia Covid a obligat ca o parte neneglijabilă a activităţilor colectivităţilor umane să migreze în mediul online. Propunem realizarea unui instrument sau a unei colecţii de API-uri care să se poată integra aplicaţiilor de teleconferinţe, în scopul extragerii de informaţii şi generării de rapoarte din discuţiile purtate şi din mesajele schimbate. Astfel, se poate imagina: transcrierea conversaţiilor în text, realizarea de rezumate ale întâlnirilor, extragerea de informaţii punctuale, pe anumite segmente din conferinţă ori din intervenţiile unor anumiţi vorbitori, generarea automată de procese verbale, interogarea minutei ori a procesului verbal generat, pe bază de cuvinte cheie, căutări de secvenţe sonore în înregistrarea sonoră etc.

Aplicaţia 6: ***Tracking assistent.*** O aplicaţie Android care îşi propune să asiste persoanele care suferă de boala Alzheimer în a-şi aminti traseele pe care le-au efectuat în timpul zilei printr-o interfaţă uşor de folosit, în limbaj natural. Pe baza datelor receptate din mai multe canale ale unui telefon mobil, se poate reface traseul şi tipul de activitate al persoanei. La sfârşitul zilei pacientul ar purta un dialog cu sistemul inteligent, care l-ar ajuta să-şi rememoreze activităţile de peste zi.

Toate obiectivele incluse în plan la această activitate au fost realizate.

Activitatea A3.5 este dedicată „**Diseminării, drepturilor de autor şi protejării intereselor vorbitorilor (anonimizare, unde este cazul)**”.

Diseminarea rezultatelor cercetărilor obţinute pe parcursul acestui an s-a realizat, în principal, prin publicarea de lucrări în jurnale, conferinţe şi workshopuri naţionale şi internaţionale. Publicaţiile se adresează comunităţii cercetătorilor din domeniul prelucrării limbajului natural şi al vorbirii, lingviştilor, cu care comunitatea noastră ştiinţifică are legături foarte puternice, cât şi întreprinzătorilor din industria IT, interesaţi să exploateze în practică perspectivele pe care proiectul ReTeRom le deschide în viitor (de ex. realizarea de interfeţe om-maşină care să utilizeze recunoaşterea şi sinteza vorbirii).

Totodată, în cadrul acestei activităţi au fost întreprinse măsuri de respectare a drepturilor de autor ale furnizorilor resurselor achiziţionate şi de protejare a intereselor contributorilor (în esenţă, persoanele cărora li s-a înregistrat vocea). Instituţia implicată în realizarea acestei etape a fost UAIC, realizatoarea componentei CoBiLiRo, sprijinită de toţi membrii consorţiului ReTeRom.

Raportul prezintă lista completă a lucrărilor publicate de parteneri în anul curent: 4 de partenerul UAIC, 5 de partenerul UPB, 5 de partenerul UTCN şi 4 de partenerul ICIA.

Drepturile de autor au fost protejate prin protocoale de colaborare, semnate de furnizor şi de o instituţie parteneră în proiect. La achiziţionarea componentelor vocale din colecţia de obiecte CoBiLiRo au fost respectate prevederile legale de protecţie a intereselor personale ale vorbitorilor.

Raport ştiinţific - tehnic proiect component TEPROLIN

Date de identificare proiect

Număr contract:	PN-III-P1-1.2-PCCDI-2017-0818, Nr. 73 PCCDI/2018
Acronim / titlu:	**„TEPROLIN”**
Titlu livrabil:	**Raport ştiinţific şi tehnic (Etapa a III-a, 2020)**
Termen:	**Noiembrie 2020**
Editor:	**Dan Tufiş**
Adresa de eMail editor:	**tufis@racai.ro**
Autori, în ordine alfabetică:	**Verginica Mititelu, Elena Irimia, Radu Ion, Vasile Păiș, Horia Cucu, Lucian Georgescu, Cristian Manolache, Adriana Stan, Beáta Lőrincz**
Ofiţer de proiect:	Cristian Stroe

1. Rezumatul etapei III

Această activitate are rolul de a evalua erorile facute de sistemul RAV al UPB, şi reantrenarea cu date corectate. De asemenea au fost evidenţiate îmbunătăţirile pe care lexiconul construit de noi le aduce aplicaţiilor de TTS dezvoltate la UTCN, prin comparare cu rezultatele obţinute cu resurse folosite anterior, şi apoi prin analiza a diferite scenarii de evaluare care folosesc informaţia din lexicon integral sau parţial (atât ca număr de intrări, selectate aleatoriu sau sistematic, cât şi ca tipuri de informaţie utilizată). Totodată în această etapă a fost realizată definitivarea, testarea, validarea şi împachetarea într-o soluţie „ready-to-use” a platformei integrate şi configurabile de prelucrare a textelor în limba română.

2. Descrierea rezultatelor activităţilor etapei III

Activitatea 3.6: Analiza erorilor sistemelor ASR şi TTS antrenate în proiectele 3 şi 4 pe corpusul bimodal agregat în proiectul 1, adnotat şi corectat

Recunoaşterea automată a vorbirii în limba română - ASR

În faza actuală a proiectului, echipa UPB a îmbunătăţit performanţele sistemului de RAV pe de o parte prin antrenarea pe noi seturi de date create şi accesibile prin proiectul CoBiLiRo, iar pe de altă parte prin recrearea modelului de limbă prin adăugarea de texte noi. (Pentru date despre cantitatea de texte noi folosite, despre tipurile de informaţii existente în aceste corpusuri folosite la antrenare, despre noile performanţe ale sistemului de RAV, a se vedea raportul fazei întocmit de echipa UPB.)

Evaluarea la nivel de cuvânt arată, aşa cum era de aşteptat, o îmbunătăţire a tuturor criteriilor de analiză: cuvinte recunoscute greşit, cuvinte inserate eronat, cuvinte netranscrise.

Evaluarea rezultatelor sistemului de RAV asupra corpusului din platforma CoBiLiRo a arătat că, în afară de cuvintele obişnuite, se produc o serie de erori în cazul a două tipuri de cuvinte: nume de entităţi şi cuvinte cu cratimă. Toate acestea sunt cuvinte inexistente in lexiconul folosit de sistemul de RAV (engl. „out of vocabulary word”). Pentru fiecare categorie am adoptat o metodă de lucru considerată adecvată.

- Nume de entităţi: am extras din diverse surse liste separate de nume de persoane, nume de locuri, nume de organizaţii/firme/etc. cu scopul de a îmbogăţi lexiconul sistemului de ASR. Listele create sunt disponibile la adresa https://relate.racai.ro/resources/ro_namelists_20201013.zip.
- Cuvinte cu cratimă: folosind un lexicon intern ICIA, am validat o parte dintre cuvintele cu cratimă, adică acelea corecte, dar absente din lexiconul ASR.
- Cuvinte obişnuite: s-a folosit aceeaşi procedură ca la cuvintele cu cratimă.

N.B. Nici una dintre aceste abordări nu rezolvă complet problema evaluării rezultatelor sistemului de RAV, atâta timp cât textele din corpusul CoBiLiRo conţin greşeli de scriere, iar rezultatele sistemului sunt comparate cu forma scrisă a corpusului. Pentru aceasta, am folosit o măsură de similaritate a tuturor tipurilor de cuvinte (Levenshtein distance) pentru a indica forma cea mai apropiată din listele de nume şi lexiconul folosite. Aceste aproximări au fost validate manual şi s-au creat liste de corecturi propuse pentru îmbunătăţirea transcrierilor existente în corpusul din platforma CoBiLiRo.

O altă observaţie rezultată din lucrul cu datele este nevoia de uniformizarea textelor din corpus în ceea ce

priveşte tipul de litere cu diacritice folosit, i.e. utilizarea literelor ş şi ţ conform standardului actual.

În ceea ce priveşte alte tipuri de erori pe care le-am identificat printre rezultatele sistemului de RAV, unele au caracter general şi considerăm că ţin de procedura de evaluare, în sensul că o relaxare a restricţiilor impuse în evaluare ar arăta o îmbunătăţire a performanţelor. Este vorba aici despre posibilitatea de a scrie unele pronume clitice cu sau fără cratimă, distincţia dintre cele două ţinând uneori doar de durata mai scurtă, respectiv mai lungă a unui sunet: ex.: *te-aştept* versus *te aştept.* Întrucât nu se poate stabili o limită de durată dincolo de care să se scrie fără cratimă, considerăm că recunoaşterea oricăreia dintre variante trebuie acceptată ca fiind corectă. Similar, doar cu un sunet în plus, sunt alte clitice, care pot pierde o vocală la o rostire mai rapidă: ex.: *mă aştept* versus *m-aştept.*
În aceeaşi situaţie se află iniţiala în vocala *î* a cuvintelor precedate de anumite cuvinte funcţionale: ex: *la început* versus *la-nceput*, *a început* versus *a-nceput*, *să înceapă* versus *să-nceapă* etc.

Sinteza automată a vorbirii în limba română - TTS

În rapoartele anterioare aferente activităţilor 1.8 şi 2.7 descriam în detaliu lexiconul cu informaţie extinsă, dezvoltat şi validat în cadrul proiectului 2, TEPROLIN; pentru fiecare intrare din lexicon, informaţia asociată reprezintă: lema (forma de dicţionar a cuvântului), eticheta morfo-sintactică în format MSD (Erjavec, 2004), împărţirea în silabe, marcarea accentului (printr-un apostrof) şi transcrierea fonetică a formei ocurenţă. Am subliniat de asemenea importanţa acestui lexicon în aplicaţiile pentru recunoaştere automată şi pentru sinteza vorbirii, susţinând că este esenţial ca o astfel de resursă sa fie de bună calitate. În raportul aferent activităţii 2.7 am descris în detaliu procesul complex de validare prin care lexiconul ReTeRom a trecut până în acest moment. Activitatea de validare a constat atât în corectare manuală intrare cu intrare, cât şi în corectare automată, prin implementarea ca expresii regulate a diverse tipuri de reguli, bazate pe cunoaştere lingvistică.

În această etapă, împreună cu echipa parteneră UTCN, am putut testa şi dovedi utilitatea lexiconului în cadrul unei aplicaţii de sinteză a vorbirii. Pentru a simplifca procesul de evaluare, am restrâns contextul evaluat la nivel textual, deoarece calitatea vorbirii sintetizate este greu de evaluat în mod obiectiv, cuantificabil. Astfel, reţeaua neuronală folosită pentru evaluarea lexiconului are ca scop predicţia (în format text) concurentă a silabificării, accentului şi transcrierii fonetice pornind doar de la forma (ortografică a) cuvântului, în lipsa unui context de utilizare.

În Tabelul 1 puteţi vizualiza câteva perechi de date de intrare/ieşire ale reţelei: punctul marchează limitele silabelor iar accentul este marcat prin apostrof urmând vocala accentuată.

Intrare	Ieşire
abandonarăţi	a . b a n . d o . n a’ . r @ ts i 0
basculantei	b a s . k u . l a’ n . t e j
Ciclopul	tS i . k l o’ . p u l

Tabelul 1. Exemple de date de intrare şi ieşire pentru reţeaua neuronală

În familia arhitecturilor neuronale utilizate pentru a învăţa funcţii secvenţă-la-secvenţă (en. Sequence-to-sequence), reţelele neuronale convoluţionale (CNN, Gehring et al. 2017) şi reţelele bazate pe atenţie (Vaswani et al., 2017) sunt cele care au demonstrat cea mai bună acurateţe în probleme de procesare a limbajului natural. De aceea, într-o primă etapă a acestui studiu, s-a realizat o evaluare CNN vs transformer (vezi (Stan, 2020) pentru detalii despre implementarea acestora), în cadrul căreia rezultatele reţelei transformer au fost evident mai bune. Arhitectura reţelei are o structură de tip encoder-decoder şi este prezentată în figura 1.

Pentru selectarea parametrilor reţelei transformer ne-am bazat pe rezultatele din (Stan, 2020), rezultând o reţea cu 3 unităţi encoder, 4 unităţi decoder, 4 centre de atenţie, dimensiunea stratului ascuns de 1.024 şi dimensiunea vectorului embedding[1] de 128. Ponderile vectorului embedding sunt iniţializate aleatoriu înainte de antrenare. Dimensiunea lotului (eng. „batch”) a fost stabilită la 512 iar pentru actualizarea ponderilor s-a folosit optimizatorul Adam cu o rată iniţială de învăţare de 0,0002. După 50 epoci, rata de învăţare a fost redusă cu un factor de 0,2 iar procesul de antrenare a fost limitat de un criteriu de oprire bazat pe valoarea funcţiei de cost după 5 epoci. În procesul de evaluare, setul de date a fost divizat în loturi de de 70%-10%-20%, dedicate proceselor de antrenare, validare şi, respectiv, testare. Seturile de antrenare şi

[1] reprezentare codificată numeric a șirului sursă/țintă

validare variază în funcţie de natura experimentelor, dar setul de testare este fix (aproximativ 80.000 de intrări).

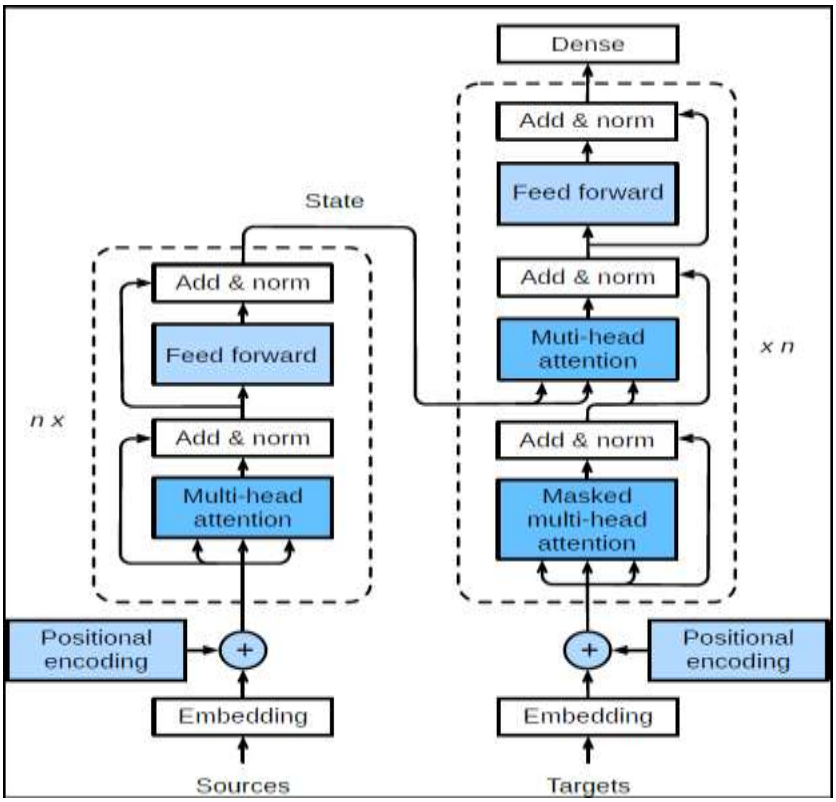


Figura 1. Arhitectura reţelei transformer folosite în experimentele noastre

În prealabil, ne-a interesat să investigăm utilitatea folosirii lexiconului ReTeRom în comparaţie cu folosirea concurentă a resurselor MaRePhor, RoSyllabiDict şi a accentului din DEX (pe care şi lexiconul nostru se bazează, extinzându-le), resurse utilizate până în acest moment pentru predicţie concurentă de informaţie lexicală în aplicaţii de TTS la UTCN. Ulterior, am conceput scenarii de evaluare mai complexe, care să ne ajute să găsim modalităţi mai eficiente de folosire a resursei dezvoltate, după cum se va vedea în continuare.

Astfel, ne-a interesat îmbunătăţirea performanţei sistemului de predicţie atunci cât setul de date creşte gradat şi identificarea momentului de platou al acestei îmbunătăţiri. În acest scop, au fost selectate partiţii aleatoare ale datelor pornind de la un minim de 5.000 de intrări.

Am urmărit de asemenea îmbunătăţirea performanţei în funcţie de cantitatea de date atunci când selecţia unui set redus de date se face asigurându-ne ca dispunem de întreaga acoperire a lexiconului la nivel de leme. Cum limba română este o limbă cu morfologie bogată, dar arareori o resursă de tip lexicon sau corpus cuprinde toate variantele morfologice ale unei leme, am vrut să vedem dacă performanţa scade drastic atunci când doar lema, sau doar o formă (care nu este întotdeauna egală cu lema) sau două forme asociate fiecărei leme se regăsesc în datele de antrenare. Ne-am rezumat la a reduce numărul de forme doar pentru cuvintele conţinut cu morfologie bogată: adjective, verbe si substantive, iar pentru restul tipurilor de cuvinte, numărul formelor asociate s-a păstrat intact. Astfel au fost derivate trei subset-uri, *1-FORM* (30.150 intrări, cu o singură formă pentru adjective, verbe şi substantive) şi *2-FORMS* (55.185 intrări, două forme pentru fiecare adjectiv, verb şi substantiv) şi *LEMMA* (35.890 intrări, cu o singură formă pentru adjective, verbe şi substantive, care este întotdeauna identică cu forma dicţionar, sau lema). În procesul de selectare a formelor din subset-urile *1-FORM* şi *2-FORMS*, ne-am asigurat că reprezentăm în mod balansat fiecare trăsătură morfologică asociată acestor părţi de vorbire (gen, număr, caz, mod, timp, persoană, etc.) pentru a încerca să păstrăm diversitatea terminaţiilor morfologice existentă în limba română. Desigur, într-un scenariu real, nu se poate controla păstrarea integrală a acestei diversităţi (adică nu ne putem asigura că toate terminaţiile sunt reprezentate într-un corpus sau într-un lexicon extras dintr-un corpus), dar o varietate ne-exhaustivă a formelor este naturală, pe când prezenţa exclusivă a formelor de dicţionar este improbabilă. În plus, ne-a interesat să investigăm potenţialul utilizării părţii de vorbire (primul simbol din eticheta morfosintactică) şi a etichetei morfo-sintactice (MSD) ca trăsături suplimentare de intrare. Teoretic, această informaţie ar putea

ajuta sistemul sa diferenţieze între omografele care nu sunt omofone. În practică, în lexiconul nostru, omografele care au rostire diferită în funcţie de POS-ul sau MSD-ul asociat reprezintă aproximativ 2000 de intrări şi nu ne aşteptăm ca rezolvarea acestei probleme de predicţie să aibă un impact mare asupra ratei de erorare. Eventuala rezolvare reprezintă în schimb un plus de acuitate lingvistică adus sistemului. De asemenea, presupunem că reţeaua poate învăţa să asocieze anumite terminaţii morfologice (şi pronunţiile lor, care sunt specifice atunci când sunetele reprezintă terminaţii) cu anumite părţi de vorbire/etichete morfosintactice, fapt ce ar putea compensa pentru lipsa tuturor formelor asociate cuvintelor conţinut. Vom verifica această presupunere în scenariile de evaluare bazate pe subset-urile LEMMA, 1-FORM şi 2-FORMS. În toate aceste scenarii de evaluare, output-ul reţelei transformer a fost, aşa cum am prezentat în Tabelul 1, o combinaţie a celor trei sarcini de predicţie lexicală, iar inputul a fost fie forma ortografică a cuvântului, fie forma împreună cu partea de vorbire, fie forma împreună cu eticheta morfosintactică asociată.

Drept metrici de evaluare, am folosit *word error rate* (WER, ro. "rata de eroare la nivel de cuvânt") şi *character error rate* (CER, ro. „rata de eroare la nivel de caracter"), evaluate pe şiruri de caractere care includ marcajele de silabificare şi accent. CER a fost măsurată folosind distanţa Levenshtein (Levenshtein, 1966) între secvenţa de caractere prezisă şi cea ţintă. În plus, ne-am dorit să evaluăm erorile introduse de fiecare sarcină de predicţie, motiv pentru care am calculat WER şi CER eliminând din predicţia concurentă informaţia furnizată de una sau mai multe dintre sarcini.

Evaluarea prealabilă, în care comparam resursa noastră cu cele folosite anterior, ne arată că, cu lexiconul ReTeRom, putem obţine o rată a erorii WER de 3,08 şi o rată CER de 1,08 pentru predicţia concurenţială, atunci când folosim toată informaţia disponibilă (inclusiv eticheta morfosintactică), ceea ce reprezintă o reducere importantă a ratei erorii de la 10,47 WER şi 3,3 CER, obţinute cu resursele utilizate anterior la UTCN.

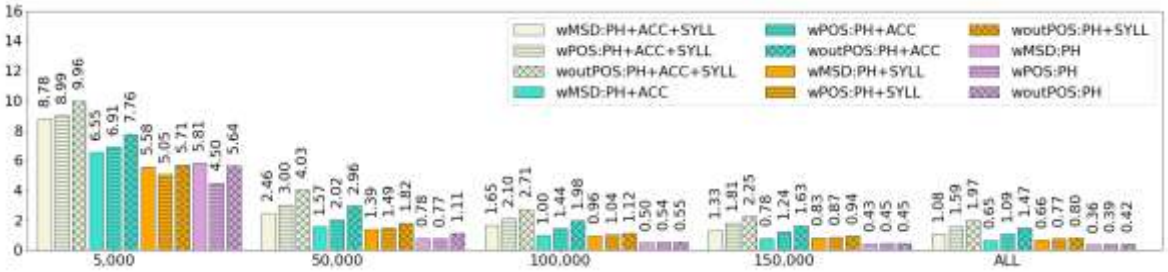


Figura 2. a

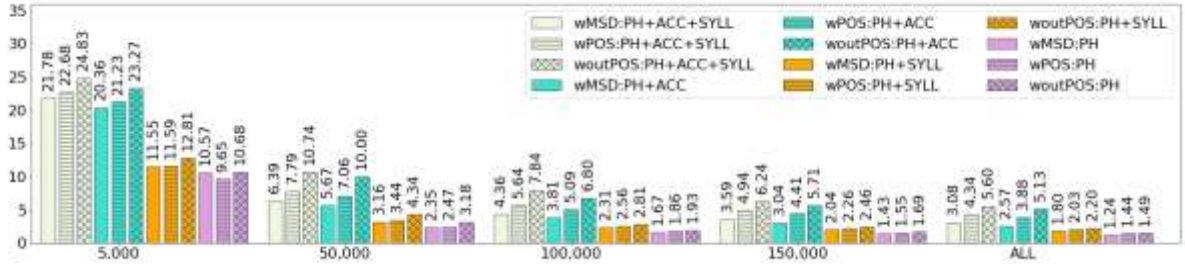


Figura 2. b

Fig 2. (a) CER şi (b) WER) pentru diferite dimensiuni ale datelor de antrenare selectate în mod aleator, evaluate pentru: predicţie lexicală completă (PH+ACC+SYL), eliminând din predicţie silabificare (PH+ACC), eliminând din predicţie silabificarea şi accentul (PH). Rezultatele indică ratele de eroare cu POS (wPOS), cu MSD (wMSD) şi fără nici o informaţie suplimentară în afară de forma cuvântului (woutPOS).

În Figura 2, unde prezentăm rezultatele scenariului de evaluare 1 combinat cu scenariul de evaluare 3, culorile corespund combinaţiei de informaţie menţinută în predicţie iar simbolul de haşurare discriminează între cele două tipuri de input (cu sau fără informaţie morfosintactică). Se poate observa că după 100.000 de intrări, creşterea acurateţii atinge un platou, dar totuşi există o îmbunătăţire relativă de 24% începând de la 100.000 intrări şi până la setul de date complet, în termeni de WER, pentru predicţia întregii secvenţe, atunci când se foloseşte informaţia morfosintactică. Din punct de vedere al contribuţiei sarcinilor de predicţie la rata erorii, accentul are cea mai mare influenţă, lucru de aşteptat pentru limba română unde accentul nu depinde de reguli predefinite (în imagine se pod vedea scăderile bruşte ale ratelor de eroare de la evaluarea PH+ACC+SYLL (culoarea crem) la evaluarea PH+SYLL (culoarea cărămiziu, atunci când nu se evaluează

acentul)). Un rezultat similar a fost prezentat în (Stan şi Giurgiu, 2018). Aşa cum ne aşteptam, în cele mai multe etape de evaluarea, informaţia despre partea de vorbire şi eticheta morfosintactică creşte performanţa modelului.

În Tabelul 2 se poate observa o diferenţă foarte mare între ratele de eroare folosind subsetul de antrenare *LEMMA* şi cele folosind subsetul de antrenare *1-FORM*. De asemenea, E(rror)R(ates)-urile subsetului LEMMA sunt de două ori mai mari decât ER-urile obţinute cu subestul de 5000 de intrări selectate aleator. Ne explicăm scăderea substanţială a ratelor de erorare în cazul subsetului *1-FORM* faţă de subsetul *LEMMA*, deoarece, aşa cum spuneam atunci când am descris procesul de selecţie a subset-urilor am avut grijă să ne asigurăm că păstrăm diversitatea formelor morfologice, chiar dacă am selectat o singură formă pentru fiecare lemă, în subsetul *1-FORM*. În contrast, subset-ul LEMMA conţine doar formele de dicţionar şi nu învaţă nici o terminaţie morfologică. Un subset de intrări selectate aleator de 7 ori mai mic dublează probabilitatea reţelei de a învăţa proprietăţile terminaţiilor morfologice. În continuare, în cazul subsetului 2-FORMS, observăm că rata de eroare continuă să scadă semnificativ. Dacă comparăm performanţele acestor subseturi cu cele ale căror intrări sunt selectate aleatoriu, observăm că: 1-FORM (30.150 intrări), pentru scenariul wMSD, are o CER=4,29% pentru predicţie concurentă, comparabilă cu CER=4,36% a subsetului de 100.000 de intrări în scenariul wMSD; 2-FORM (55.185 intrări), pentru scenariul wMSD, are o CER=3,16%, aproape de CER=3,08 obţinută atunci când se foloseşte întregul set de date.

Aceste performanţe demonstrează că o selecţie strategică a intrărilor reduce foarte mult munca de corectură manuală într-un scenariu de construire incrementală - cu adnotare automată şi corectare manuală pe subset-uri de date – a unui lexicon. Deşi se observă şi aici, în general, o imbunătăţire a performanţei legată de adăugarea informaţiei POS şi MSD, acest lucru nu este valabil pentru subsetul *LEMMA*.

	LEMMA		1-FORM		2-FORMS		All Forms	
	WER	CER	WER	CER	WER	CER	WER	CER
woutPOS								
PH+ACC+SYLL	46.53%	19.03%	13.30%	5.00%	10.20%	3.86%	5.60	1.97
PH+ACC	45.02%	16.01%	12.17%	3.57%	9.28%	2.64%	5.13	1.47
PH+SYLL	22.31%	7.29%	5.47%	2.37%	4.15%	1.83%	2.20	0.80
PH	16.53%	5.30%	3.91%	1.36%	2.86%	0.86%	1.49	0.42
wPOS								
PH+ACC+SYLL	52.55%	20.69%	11.54%	4.49%	8.39%	3.24%	4.34	1.59
PH+ACC	50.87%	18.18%	10.42%	2.93%	7.49%	2.08%	3.88	1.09
PH+SYLL	27.82%	8.92%	5.71%	2.55%	4.21%	1.81%	2.03	0.77
PH	20.87%	6.50%	4.21%	1.50%	3.07%	0.96%	1.44	0.39
wMSD								
PH+ACC+SYLL	47.57%	18.38%	11.07%	4.29%	8.18%	3.16%	3.08	1.08
PH+ACC	46.35%	15.62%	9.86%	2.82%	7.21%	1.95%	2.57	0.65
PH+SYLL	24.01%	7.26%	5.46%	2.42%	4.13%	1.83%	1.80	0.66
PH	20.51%	6.90%	4.04%	1.46%	2.98%	0.95%	1.24	0.36

Tabelul 2. CER şi WER pentru diferite dimensiuni ale datelor de antrenare selectate în modalitatea descrisă pentru scenariul 2, cu subset-urile *1-FORM* (30.150 intrări), *2-FORMS* (55.185 intrări) şi *LEMMA* (35.890 intrări) evaluate pentru: predicţie lexicală completă (PH+ACC+SYL), eliminând din predicţie silabificare (PH+ACC), eliminând din predicţie silabificarea şi accentul (PH). Rezultatele indică ratele de eroare cu (wPOS) sau fără (woutPOS) etichete morfosintactice adăugate la intrarea în transformer; pentru comparaţie, pe ultima coloană reluăm rezultatele pentru lexiconul complet reprezentate în Figura 2.

Aici, informaţia POS şi cea MSD nu poate compensa absenţa variantelor morfologice, pentru că nu are ocazia sa înveţe nici o asociere între terminaţii morfologice şi etichete POS ale cuvintelor conţinut (care reprezintă majoritatea covârşitoare a lexicului unei limbi), pe care să o aplice apoi cuvintelor noi (de exemplu: 1. la transcrierea fonetică a terminaţiilor, pentru care există reguli de transcriere specială a sunetului "i" final, şi a sunetelor "ce/ci/ge/gi/che/chi/ghe/ghi" la sfârşit de cuvânt; 2. la nivel de silabificare, reţeaua nu are ocazia să înveţe despărţirea în silabe a terminaţiilor morfologice). Dimpotrivă, faptul că ratele de eroare wPOS şi wMSD sunt în acest caz mai mari decât ratele de eroare woutPOS ne face să presupunem că restricţia de a prezice informaţie lexicală coerentă cu partea de vorbire a cuvântului (adică, de exemplu, pentru un verb, reţeaua decide în funcţie de ce a învăţat doar despe verbele din datele de antrenare) este de fapt nefericită, iar o predicţie bazată doar pe forma morfologică este de preferat. La nivel de accent, în fişierul de antrenare din scenariul *LEMMA*, 18.981 din 35.890 intrări (mai mult de jumătate) au accent pe ultima silabă (spre deosebire de scenariile *1-FORM* (6.329 din 30.150) şi *2-FORMS* (11.966 din 55.185)). Dintre acestea, verbe care au accent pe ultima silabă sunt 6.421 din 7.158. Asta înseamnă că reţeaua transformer învaţă că probabilitatea ca accentul unui cuvânt să fie pe ultima silabă, în general, este mare, iar pentru verbe, aceasta este foarte mare. Pentru variantele morfologice din datele de test, acest comportament al reţelei este dezavantajos: de exemplu, pentru *"*a mânca*",* accentul cade pe ultima silabă doar pentru 20 din cele 30 de forme morfologice ale conjugării.

Procentul de cuvinte cu accent pe ultima silabă din lista de erori aferentă scenariului wPOS (24.125 intrări din 42.510, dintre care pentru 10.835 eroarea provine doar de la acest accent pe ultima silabă) este chiar mai mare decât cel din lista de erori aferentă scenariului wout POS (19.366 intrări din 37.806, dintre care pentru 10.183 eroarea provine doar de la accentul pe ultima silabă). Analizând în detaliu erorile specifice apărute cu subset-ul LEMMA wPOS vs LEMMA woutPOS, am observat că, din cele 12.085 de erori obţinute cu LEMMA wPOS care nu se găsesc în lista de erori obţinute cu LEMMA woutPOS, 7.365 de erori sunt asociate verbelor. În timp ce numărul erorilor distincte asociat celorlalte părţi de vorbire rămâne relativ acelaşi, numărul de erori asociate verbelor creşte foarte mult (de mai mult de 5 ori), când trecem de la scenariul woutPOS la cel wPOS. Aceste rezultate demonstrează că preponderenţa cuvintelor cu accent pe ultima silabă din setul LEMMA produce un procent mare de erori în scenariile wPOS şi wMSD. Aşa cum spuneam anterior, restricţia cuvântului la clasa părţii de vorbire căruia îi aparţine - în cazul acesta faptul că forţăm reţeaua să trateze verbele din datele de test asemănător verbelor din datele de antrenare - conduce la erori care ar fi putut fi evitate dacă reţeaua se uita doar la forma cuvântului.

Activitatea 3.7: Definitivarea, testarea, validarea şi împachetarea într-o soluţie „ready-to-use" a platformei integrate şi configurabile de prelucrare a textelor în limba română

1. Platforma de prelucrare a textelor TEPROLIN a fost îmbunătăţită după cum urmează:
   - *Am introdus dependenţe de tip graf între operaţiile de prelucrare a textelor*. Acest tip de a preciza ce operaţii trebuie rulate mai întâi pentru a se putea rula operaţia de prelucrare dorită e mult mai eficient decât tipul de rulare în secvenţă pe care se baza platforma. De exemplu, pentru a putea rula operaţia de adnotare cu etichete morfo-sintactice, nu mai sunt necesare operaţii precum silabificarea sau detecţia accentului. Modulul Python 3 în care sunt precizate aceste operaţii este TeproAlgo.py iar metoda se numeşte _assignAlgorithmsToOperations().
   - *Am adăugat modulul de prelucrare a textelor UDPipe* (http://ufal.mff.cuni.cz/udpipe/1) ca alternativă la TTL şi NLPCube. Este foarte rapid (cea mai rapidă componentă de prelucrare din cele trei) şi are performanţe bune. A fost configurat ca modul implicit dacă se solicită operaţii precum adnotare cu etichete morfo-sintactice sau analiză cu relaţii de dependenţă sintactică.
   - *Modulul de inserare a diacriticelor* DiacRestore.py *a fost îmbunătăţit* pentru a detecta mai bine când un text este scris fără diacritice sau cu puţine diacritice şi a rula, astfel, automat pentru a insera diacriticele lipsă.
   - Am eliminat numărarea caracterelor din modulul de statistici pentru că frecvenţa caracterelor prelucrate de TEPROLIN putea creşte foarte mult când se prelucrau texte de zeci de milioane de cuvinte. Au rămas statisticile despre numărul de accesări la serviciu şi despre numărul de cuvinte prelucrate pe zi.

Păiş et al., (2020) descriu un studiu de caz în care TEPROLIN rulează pe mai multe fire de execuţie în RELATE şi adnotează corpusul legislativ din proiectul MARCELL. În total, au fost adnotaţi aprox. 456 de milioane de tokeni, *ceea ce demonstrează că testarea şi validarea platformei s-au încheiat cu succes*.

2. Soluţia „ready-to-use” a platformei TEPROLIN

TEPROLIN se poate utiliza într-unul din următoarele patru moduri:

- Pentru testare cu fraze scurte (pentru evaluarea performanţelor) cu efectuarea tuturor operaţiilor disponibile, se poate accesa link-ul https://relate.racai.ro/index.php?path=teprolin/complete şi se pot vizualiza adnotările făcute;
- Pentru rularea unor operaţii la alegere pe fraze scurse, folosind algoritmii preferaţi, se poate accesa link-ul https://relate.racai.ro/index.php?path=teprolin/custom;
- Pentru adnotarea corpusurilor cu mai mult de 1000 de cuvinte, se poate solicita acces la platforma RELATE care rulează TEPROLIN pe mai multe fire de execuţie;
- Ca modul Python 3, clonând repository-ul https://gitlab.com/raduion/teprolin şi urmând indicaţiile din fişierul README.md. Recomandăm ca toate pachetele necesare să fie instalate într-un mediu dedicat Python 3 (eng. „virtual environment”), executând comenzile:
    a. python3 -m venv /calea/către/mediul/dedicat/teprolin
    b. pip3 install -r requirements.txt

Concluzii

Am arătat că folosirea resursei noi dezvoltate în ReTeRom aduce, în mod evident, îmbunătăţiri unui sistem de TTS, în comparaţie cu resusele utilizate anterior. Am presupus şi demonstrat că, atunci când ne interesează să evaluăm influenţa pe care cantitatea de date o are asupra ratelor de eroare, selectarea datelor în mod strategic şi nu aleatoriu este esenţială. De exemplu, în procesul de corectare manuală a unui lexicon de tipul celui pe care l-am construit noi, se poate reduce foarte mult timpul de corectare dacă selectăm un set esenţial de intrări pe care să le corectăm şi pe care să antrenăm instrumentul de adontare automată pentru a adnota intrările rămase, care ne aşteptăm să fie mai corecte dacă am ales un set de anternare potrivit. În cazul nostru, vezi diferenţa între setul de antrenare *1-FORM*, setul de antrenare *LEMMA* şi setul de antrenare selectat aleatoriu, de dimensiuni similare, dar producând efecte foarte diferite asupra reţelei la antrenare. Recomandăm deci o selecţie a datelor conform scenariului 2 de evaluare atunci când este importantă reducerea set-ului de date de antrenare. În plus, după cum s-a observat, prin acest tip de selecţie putem observa şi o descreştere a ratelor de eroare atunci când sunt adăugate informaţii morfo-sintactice.

Platforma de prelucrare a textelor TEPROLIN se află la https://gitlab.com/raduion/teprolin. Pentru a avea acces, trebuie să aveţi cont pe GitLab şi să solicitaţi accesul autorului platformei.

Toate obiectivele incluse în plan la aceste activități au fost realizate.

Bibliografie

Erjavec, Tomaz. "MULTEXT-East Morphosyntactic Specifications: Version 3.0." Supported By EU Projects Multext-East, Concede And TELRI (2004)

Gehring, J., M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, “Convolutional sequence to sequence learning,” CoRR, vol. abs/1705.03122, 2017.

V. Levenshtein, “Binary Codes Capable of Correcting Deletions, Insertions and Reversals,” Soviet Physics Doklady, vol. 10, p. 707, 1966.

Stan, A. and M. Giurgiu, “A Comparison Between Traditional Machine Learning Approaches And Deep Neural Networks For Text Processing In Romanian,” in Proceedings of the 13th International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR), 2018.

Stan, Adriana. “RECOApy: Data recording, pre-processing and phonetic transcription for end-to-end speech-based applications,” in Proceedings of Interspeech, Shanghai, China, 2020.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008.

Păiş, V., Tufiş, D. şi Ion, R. (2020) A Processing Platform Relating Data and Tools for Romanian Language. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC), European Language Resources Association, Marseille, France, pages 81—88.

Raport ştiinţific - tehnic proiect component TADARAV

Date de identificare proiect

Număr contract:	PN-III-P1-1.2-PCCDI-2017-0818, Nr. 73 PCCDI/2018
Acronim / titlu:	**„TADARAV:" Tehnologii pentru adnotarea automată a datelor audio şi pentru realizarea interfeţelor de recunoaştere automată a vorbirii**
Titlu livrabil:	**Raport ştiinţific şi tehnic (Etapa a III-a, 2020)**
Termen:	**Noiembrie 2020**
Editor:	**Horia Cucu (Universitatea Politehnica din Bucureşti)**
Adresa de eMail editor:	**horia.cucu@upb.ro**
Autori, în ordine alfabetică:	**Alexandru-Lucian Georgescu, Cristian Manolache, Dan Oneaţă, Gheorghe Pop, Horia Cucu, Corneliu Burileanu, Dragoş Burileanu**
Ofiţer de proiect:	**Cristian Stroe**

1. Activităţile etapei de raportare în contextul general al proiectului

Activităţile realizate în etapa 3/2020 au fost următoarele:

- **Activitatea 3.9 - Analiza impactului utilizării de RAV complementare pentru generarea de adnotări în contextul îmbunătăţirii sistemelor de RAV.** Această activitate a presupus un studiu comparativ asupra performanţelor sistemului de RAV iniţial şi a sistemelor de RAV rezultate în cadrul activităţilor A1.13/2018 şi A2.13/2019. De asemenea, în cadrul acestei activităţi s-a încercat aplicarea metodei proiectate în activităţile anterioare cu un sistem de RAV complet nou, bazat pe platforma ESPnet. Sistemul de RAV rezultat s-a dovedit a fi extrem de lent în transcriere, astfel că nu a putut fi utilizat corespunzător. Cu toate acestea, per ansamblu, sistemul de RAV iniţial (disponibil la începutul proiectului) a putut fi îmbunătăţit folosind metoda proiectată în activităţile A1.13/2018 şi A2.13/2019 cu aproximativ 50% atât pe vorbire citită, cât şi pe vorbire spontană.
- **Activităţile 3.10 Îmbunătăţirea soluţiei de filtrare şi aliniere a transcrierilor aproximative cu semnalul de vorbire şi 3.12 Analiza impactului utilizării transcrierilor aproximative în vederea reantrenării sistemelor de RAV.** Activităţile A3.10 şi A3.12 au avut ca scop îmbunătăţirea şi evaluarea metodei de generare de seturi de date de vorbire adnotată folosind materiale audio brute împreună cu transcrieri aproximative. Ideea principală a acestei metode este următoarea: un sistem de RAV iniţial este folosit pentru a genera transcrieri pentru materialul audio brut, iar ulterior aceste transcrieri sunt aliniate cu textele aproximative deja existente. Părţile aliniate sunt considerate corecte şi sunt folosite pentru reantrenarea sistemului iniţial de RAV. În urma realizării activităţii A2.11 din etapa anterioară am tras concluzia că între două secvenţe de text aliniate corect există transcrieri aproximative care sunt în mare proporţie corecte. Îmbunătăţirea propusă şi evaluată în cadrul activităţilor din anul 2020 constă în utilizarea acestor transcrieri aproximative aflate între două secvenţe de text aliniate corect. Rezultatele experimentale au arătat că îmbunătăţirea RAV rezultată este nesemnificativă (mai mică de 1%).
- **Activităţile 3.11 Îmbunătăţirea soluţiei pentru generarea de scoruri de încredere pentru RAV şi 3.13 Analiza impactului utilizării scorurilor de încredere pentru filtrarea transcrierilor RAV în vederea reantrenării sistemelor RAV.** În cadrul acestor activităţi au fost propuse o serie de noi metode de generare de scoruri de încredere pentru RAV folosind o platformă modernă de tip end-to-end: ESPnet. Metodele propuse au fost evaluate şi comparate, iar metoda cea mai performantă a fost selectată pentru utilizarea ulterioară în vederea producerii de transcrieri precise pentru o parte a semnalului de vorbire. Sistemul de RAV ESPnet s-a dovedit însă a fi prea lent pentru a putea fi utilizat într-o astfel de metodă de adnotare automată de date: cu un factor de timp real de aproximativ 3, transcrierea celor 913 de ore de vorbire brută ar fi durat aproximativ 100 de zile. În consecinţă, noile metode de estimare a scorurilor de încredere nu au putut fi evaluate în contextul reantrenării sistemului de RAV iniţial, folosind datele adnotate rezultate.

2. Gradul de realizare a obiectivelor specifice pentru Etapa a III-a, 2020

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

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

3. Rezultatele etapei şi descrierea lor ştiinţifică şi tehnică

Rezultatele etapei 3/2020 a proiectului TADARAV sunt listate mai jos. Primele două rezultate din lista de mai sus sunt prezentate pe larg în livrabilele anexate: [Manolache, 2020a], [Oneaţă, 2020]. Al treilea rezultat, sistemul RAV actualizat este descris în continuare.

- Soluţie îmbunătăţită de filtrare şi aliniere a transcrierilor aproximative cu semnalul de vorbire (TRL4);
- Soluţie îmbunătăţită pentru generarea de scoruri de încredere pentru RAV (TRL4);
- Sistem de RAV actualizat (TRL5).

3.1 Actualizarea modelelor de limbă pentru transcriere de vorbire

În această secţiune sunt prezentate sumar demersurile realizate de colectivul SpeeD în vederea actualizării modelelor de limbă utilizate în sistemele de transcriere de vorbire. Etapele parcurse penmtru această actualizare sunt descrise în detaliu în [Manolache, 2020b].

Primul pas întreprins a fost colectarea unui nou corpus de text. Acest corpus nou de text, denumit mai departe news2020, constă în text brut, organizat pe propoziţii ce însumează aproximativ 255M de cuvinte. Cele 24 surse de ştiri ce formează corpusul news2020 sunt prezentate în Figura T1.

Al doilea pas întreprins a vizat actualizarea aplicaţiei de preprocesare de text în vederea îmbunătăţirii modului de tratare a cuvintelor cu cratimă [Manolache, 2020b].

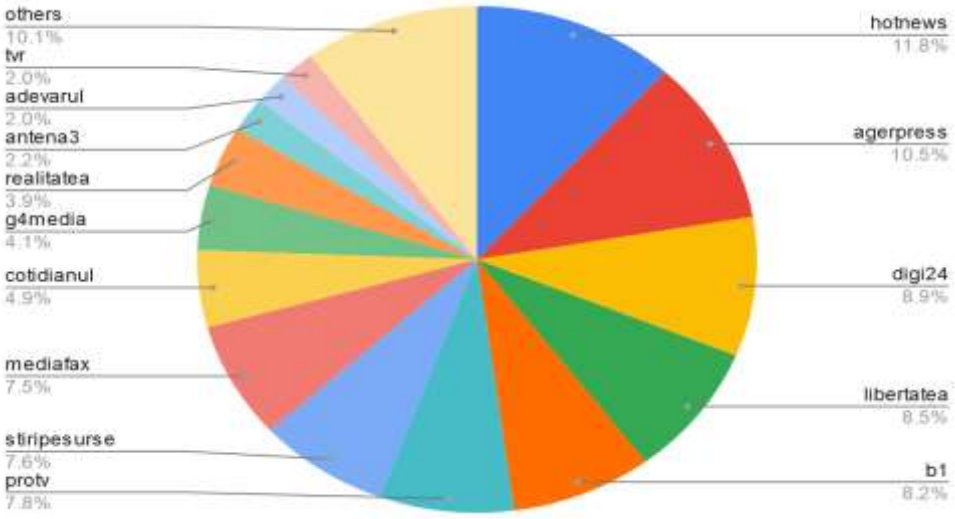


Figura T1 *Sursele de ştiri componente ale corpusului news2020. Procentele sunt relative la dimensiunea întregului corpus (255M cuvinte)*

Al treilea pas întreprins a vizat actualizarea modelelor de limbă. Nu au fost create tipuri noi de modele de limbă, ci au fost antrenat modele de limbă folosind corpusuri de text extinse şi preprocesate mai bine.

Modelele de limbă de bază au fost antrenate folosind numai corpusul news002 (361M de cuvinte), pe când noile modele de limbă au fost antrenate pe corpusul news002 plus corpusul news2020 (361M + 255M de cuvinte). Textul folosit pentru antrenarea modelelor de limbă noi a fost preprocesat folosind noul procesor de limbaj natural discutat în acest subcapitol. Modelele de limbă îmbunătăţite sunt evaluate în funcţie de perplexitate (PPL) şi cuvinte out-of-vocabulary (OOV). Modelele de limbă mai simple (2-gram şi 3-gram) au fost încorporate în etapa de decodare a sistemului ASR, în timp ce modelele de limbă mai complexe (4-gram şi RNN) au fost folosite în etapa de rescoring. Modelele mai complexe nu au fost folosite în etapa de decodare din cauza constrângerilor arhitecturale şi de memorie.

În final, în al patrulea pas, modelele de limbă nou create au fost încorporate în sistemul de RAV. Pentru etapa de decodare am optat pentru modelele 3-gram 200k, iar pentru rescoring am ales LM-2020-4g-large-200k şi LM-2020-5g-RNN. Cele mai bune rezultate s-au obţinut pentru sistemul RAV ce foloseşte LM-2020-RNN pentru rescoring. În Tabelul T1 sunt prezentate rezultatele WER ale noilor sisteme RAV cu şi fără rescoring.

Tabelul T1 *Rezultatele WER ale sistemelor RAV după decodare şi rescoring*

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

În Tabelul T1 se poate observa că cel mai bun sistem RAV este cel ce foloseşte LM-2020-3g-large-200k cu LM-2020-RNN pentu rescoring. Acest model are o scădere relativă de 9% şi 7% WER pe seturile de evaluare de vorbire spontană, anume SSC-eval1 şi respectiv SSC-eval2, comparativ cu sistemul RAV de bază cu rescoring (Tabel T1 linia 1). Însă, pe setul de evaluare RSC-eval sistemul RAV a înregistrat o creştere de 16% WER.

3.2 Utilizarea seturilor de date rezultate din proiect pentru antrenarea RAV

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

Tabelul T2 prezintă evoluţia sistemului RAV SpeeD în perioada 2018-2020. Fiecare linie reprezintă un model acustic diferit, îmbunătăţit faţă de cel precedent, pentru antrenarea căruia au fost adăugate seturi de date suplimentare. Fiecare sistem este evaluat pe cele 3 seturi de evaluare existente: RSC-eval, SSC-eval1, SSC-eval2. Primul model acustic, denumit în tabel *train-base,* este cel din cadrul sistemului iniţial RAV #3, prezentat în secţiunea 2.2.2 a raportului [Georgescu, 2020b]. Acesta a fost antrenat folosind setul de antrenare *baseline*: RSC-train1 + SSC-train1 + SSC-train2.

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

Sistemul RAV #3, corespunzător modelului acustic *train-base*, a fost utilizat pentru a transcrie setul de date CoBiLiRo, prezentat în secţiunea 2.1.2 a raportului [Georgescu, 2020b]. În secţiunea 2.3.2 a raportului

[Georgescu, 2020b]este prezentat modul în care a fost aliniată transcrierea obţinută cu transcrierile aproximative deja existente, rezultând în acest fel setul de date CoBiLiRo-trans-v4. Acesta a fost adaugat la seturile de antrenare anterioare, fiind obţinut un nou model acustic, denumit *train14*. În comparaţie cu modelul *train3-v4*, modelul *train14* obţine rezultate mai bune pe vorbire citită (11% scădere relativă a WER pe RSC-eval), dar rezultate mai slabe pe vorbire spontană. Acest lucru se datorează similitudinii dintre CoBiLiRo-trans-v4 şi setul de date RSC. CoBiLiRo-trans-v4 conţine interviuri şi dialoguri într-un mediu fără zgomot. Totodată, trebuie ţinut cont şi de cantitatea mică a datelor nou adăugate, 31 ore, faţă de cele 517 deja existente.

Ultimul model acustic, *train11*, a presupus reantrenarea sistemului RAV după ce a fost adăugat setul de date CoRoLa, introdus în secţiunea 2.1.1 a raportului [Georgescu, 2020b]. Îmbunătăţirile relative faţă de sistemul anterior sunt cuprinse între 8%-12% pe vorbire spontană, în timp ce în cazul vorbirii citite, nu se înregistrează schimbări. Acest lucru se explică prin conţinutul setului CoRoLa: vorbire spontană, de la vorbitori profesionişti, în cadrul unor emisiuni TV. În total, faţă de modelul *train-base,* acest ultim model obţine o scădere relativă a erorii de 15% pe vorbire citită, respectiv 31%-39% pe vorbire spontană.

În continuare ne-am pus problema contribuţia fiecărui nou set de date de vorbire la îmbunătăţirea totală a sistemului de RAV. Astfel, am antrenat sisteme de RAV pe seturi de date formate din setul baseline şi un singur alt set de date de vorbire. Tabelul T3 prezintă rezultatele. Rezultatele obţinute pe setul de date RSC-eval diferă foarte puţin, dacă nu chiar deloc, între diversele experimente. Tragem concluzia că modelul din adăugarea de date noi numai pentru creşterea performanţei pe vorbire spontană. Deoarece seturile nou adăugate conţin în mare măsură vorbire spontană, îmbunătăţirile substanţiale au rezultat pe seturile de evaluare SSC-eval1 şi SSC-eval2 (ce conţin astfel de vorbire). Este foarte probabil ca în urma reantrenărilor, modificări ale parametrilor reţelei neuronale să fi apărut numai în acele zone din reţea care se ocupă de prelucrarea vorbirii spontane.

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

***Tabelul T2**. Evolutia RAV SpeeD în perioada 2018 - 2020. Îmbunătăţirile rezultate prin adăugarea la setul de date de antrenare BAS (RSC-train + SSC-train1 + SSC-train2) a seturilor de date rezultate din proiect SSC (SSC-train3-trans-v4 + SSC-train4-trans-v4), COB (CoBiLiRo-trans-v4) şi COR (CoRoLa)*

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

***Tabelul T3**. Contribuţia fiecărui set de date nou la îmbunătăţirea RAV SpeeD 2018 - 2020. Seturile de date de antrenare folosite sunt: BAS (RSC-train + SSC-train1 + SSC-train2), SSC (SSC-train3-trans-v4 + SSC-train4-trans-v4), COB (CoBiLiRo-trans-v4) şi COR (CoRoLa)*

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

train15	x		x			1.8	14.0	21.1
train17	x			x		1.8	11.9	15.4

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

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

Tabelul T4 prezintă contribuţia noilor seturi de date introduse în 2020 în mod comparativ. Unul din cele 3 seturi noi din 2020, pe lângă CoBiLiRo-trans-v4 şi CoRoLa, este cdep-trans-v4, prezentat în secţiunea 2.3.3 a raportului [Georgescu, 2020b]. Setul de date cdep-raw a fost transcris folosind modelul acustic *train3-v4*, iar apoi, prin alinierea cu transcrierile aproximative deja existente, a rezultat setul cdep-trans-v4. Surprinzător este faptul că deşi setul de date cdep-trans-v4 însumează 879 ore, fiind de 28 de ori mai mare decât CoBiLiRo-trans-v4, respectiv de 10 ori mai mare decât CoRoLa, adăugarea acestui set la antrenare nu produce rezultate mai bune decât adăugarea celorlalte două seturi amintite. O pistă de investigat în acest sens este lungimea rostirilor din setul cdep-trans-v4, fiind posibil ca acestea să fie foarte scurte (e posibil ca transcrierile aproximative şi transcrierile date de sistemul RAV să se suprapună într-o mică măsură).

***Tabelul T4**. Contribuţia noilor seturi de date introduse in 2020. Sistemul iniţial a fost antrenat pe seturile de date BAS (RSC-train + SSC-train1 + SSC-train2) şi SSC (SSC-train3-trans-v4 + SSC-train4-trans-v4). În etapa 3/2020 au fost introduse seturile de date COB (CoBiLiRo-trans-v4), COR (CoRoLa) şi CDP (cdep-trans-v4)*

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

3.3 Sistemul de RAV SpeeD îmbunătăţit

În concluzie, selectând cea mai performantă configuraţie de model de limbă pentru decodare, model de limbă pentru reevaluare lingvistică (conform secţiunii 3.1) şi model acustic (conform secţiunii 3.2), cel mai performant sistem de RAV SpeeD are actualmente performanţele listate în Tabelul T5.

***Tabelul T5**. Sistemul RAV SpeeD 2020*

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

4. Oferta de servicii de cercetare, locuri de muncă susţinute şi valorificarea resurselor

Laboratorul de cercetare *Speech and Dialogue* (SpeeD) din cadrul Universităţii Politehnica din Bucureşti (UPB), reprezentantul UPB în proiectul TADARAV, oferă pe platforma ERRIS (https://erris.gov.ro/SpeeD---UPB) serviciile de cercetare şi tehnologice enumerate în Tabelul T6.

***Tabelul T6.** Servicii de cercetare şi tehnologice oferite de Laboratorul de cercetare Speech and Dialogue*

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

Echipa de cercetare a Universităţii Politehnica din Bucureşti pentru proiectul component TADARAV este prezentată în Tabelul T7.

***Tabelul T7.** Echipa de cercetare UPB*

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

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

5. Concluzii

Activităţile de cercetare desfăşurate în etapa a III-a de implementare a proiectului (2020) au condus la obţinerea rezultatelor aşteptate şi ele sunt în concordanţă cu obiectivele specifice ale etapei. Astfel, rezultatele raportate în acest document şi descrise detaliat în cele 3 livrabile aferente perioadei de raportare, asigură testarea şi evaluarea finală a tehnologiei dezvoltate în etapa finală a proiectului (2021).

6. Referinţe

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

[Georgescu, 2020b]A.-L. Georgescu, A. Caranica, C. Manolache, G. Pop, D. Oneaţă, H. Cucu, C. Burileanu, D. Burileanu, *„Proiect component TADARAV: Raport ştiinţific şi tehnic în extenso 2020"*.

[Georgescu, 2020c]A.-L. Georgescu, A. Caranica, H. Cucu, D. Burileanu, C. Burileanu, *„Raport de analiză a impactului utilizării de RAV complementare pentru generarea de adnotări în contextul îmbunătăţirii sistemelor de RAV,"* Livrabil proiect ReTeRom, 2020.

[Oneaţă, 2020] D. Oneaţă, A. Caranica, H. Cucu, C. Burileanu, D. Burileanu, *„Raport de analiză a impactului utilizării scorurilor de încredere pentru filtrarea transcrierilor RAV în vederea reantrenării sistemelor RAV,"* Livrabil proiect ReTeRom, 2020.

[Manolache, 2020a] C. Manolache, A.-L. Georgescu, H. Cucu, C. Burileanu, D. Burileanu, *„Raport de analiză a impactului utilizării scorurilor de încredere pentru filtrarea transcrierilor RAV în vederea reantrenării sistemelor RAV,"* Livrabil proiect ReTeRom, 2020.

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

Raport ştiinţific - tehnic proiect component SINTERO

Date de identificare proiect

Număr contract:	PN-III-P1-1.2-PCCDI-2017-0818, Nr. 73 PCCDI/2018
Acronim / titlu:	**„SINTERO: Tehnologii de realizare a interfeţelor om-maşină pentru sinteza text-vorbire cu expresivitate”**
Titlu livrabil:	**Raport ştiinţific şi tehnic (Etapa a III-a, 2020)**
Termen:	**Noiembrie 2020**
Editor:	**Mircea Giurgiu (Universitatea Tehnică din Cluj-Napoca)**
Adresa de eMail editor:	**Mircea.Giurgiu@com.utcluj.ro**
Autori, în ordine alfabetică:	**Mircea Giurgiu, Beata Lorincz, Maria Nuţu, Adriana Stan**
Ofiţer de proiect:	**Cristian Stroe**

1. Activităţile etapei de raportare în contextul general al proiectului

În a treia etapă (2020) a proiectului SINTERO *„Dezvoltarea unei noi tehnologii de realizare a interfeţelor om – maşină pentru sinteza text – vorbire cu expresivitate”*, s-au desfăşurat activităţile: **a)** dezvoltarea unei noi tehnologii de adaptare a vocii sintetizate la stilul şi expresivitatea unui nou vorbitor (Livrabil D3.15)**, b)** dezvoltarea unei metode de adaptare rapidă a vocii sintetizate folosind date audio atipice (Livrabil D3.16), **c)** integrarea diferitelor tehnologii într-un flux de prelucrări automatizat (Livrabil D3.17), respectiv activităţi de testare, validare, diseminare şi demonstrare online (Livrabil D3.18).

2. Gradul de realizare a obiectivelor specifice pentru Etapa a III-a, 2020

Ob. Pr4.3.15: *Dezvoltarea unei tehnologii pentru adaptarea vocii sintetizate la stilul şi expresivitatea unui vorbitor* (Obiectiv realizat integral). *Rezultate:* ● 1 nou corpus cu 29 de vorbitori si aproximativ 65 de ore de vorbire ● 1 aplicaţie software (RecoApy) pentru automatizarea înregistrărilor ● 1 vocoder de tip WaveGlow antrenat cu vocile din noul corpus ● 1 sistem de sinteză de tip DC-TTS pentru adaptare la noi vorbitori folosind o nouă metrică pentru calculul erorii de antrenare ● 1 metodă de adaptare de la date ne-expresive la date expresive ● 2 articole publicate la 2 conferinţe internaţionale ● un livrabil (D3.15) cu titlul „Tehnologie de adaptare a vocilor sintetice la noi vorbitori”.

Ob. Pr4.3.16: *Dezvoltarea unei metode de adaptare rapidă a vocii sintetizate folosind date audio atipice* (Obiectiv realizat integral). *Rezultate:* ● 1 metodă de predicţie simultană cu reţele neuronale a informaţiei lexicale ● 1 lexicon RoLEX utilizat pentru antrenarea modelelor de text ● câte 5 sisteme de sinteză text-vorbire folosind informaţii lexical prezise, evaluate atât pentru limba română, cât şi engleză ● 1 demonstrator pe bază de Flowtron pentru transferul stilului de vorbire, respectiv pentru antrenarea cu date atipice ● pagini web cu demonstrarea calităţii semnalului audio sintetizat ● 2 articole la conferinţe internaţionale ● 1 livrabil (D3.16) cu titlul „Metodă de adaptare rapidă a sistemului de sinteză cu date atipice”.

Ob. Pr4.3.17: *Integrarea tehnologiilor şi demonstratoare online* (Obiectiv realizat integral). *Rezultate:* ● integrarea tehnologiilor dezvoltate ● 20 de sisteme de sinteză text vorbire disponibile pentru demonstrare online in interfaţa //speech.utcluj.ro/ronna ● 1 livrabil (D3.17) cu titlul „Tehnologie de realizare a interfeţelor om – maşină pentru sinteza text - vorbire”.

Ob. Pr4.3.18: *Diseminarea rezultatelor* (Obiectiv realizat integral). *Rezultate:* ● realizarea şi actualizarea web site-ului ● pagini web cu demonstratoare cu vocile sintetizate ● 1 livrabil referitor la activităţile de diseminare (D3.18).

3. Rezultatele etapei şi descrierea lor ştiinţifică şi tehnică

3.1. Tehnologie de adaptare a vocilor sintetice la noi vorbitori

Adaptarea vorbirii sintetice la stilul şi expresivitatea unui nou vorbitor a fost realizată prin intermediul a trei noi metode: 1) adaptarea de la un sistem cu vorbitori multipli la un sistem cu un singur vorbitor folosind un set redus de date, 2) adaptarea de la un sistem cu voce ne-expresivă la un sistem cu voce

expresivă, 3) adaptarea sistemului de sinteză la un nou vorbitor printr-o procedură de adaptare bazată pe post-filtrare. Descriere în extenso în D3.15.

(1) Crearea unui nou corpus (SWARA 2.0) pentru adaptarea sistemului de sinteză la noi vorbitori. Acest corpus a fost înregistrat de către studenţi voluntari în condiţii ambientale pentru o mai mare variabilitate a datelor audio şi folosind aplicaţia RecoApy[2]. A rezultat un set de înregistrări ce conţine 51.839 de segmente audio de la 29 de vorbitori: 14 masculini şi 15 feminini şi o durata totală a înregistrărilor de aproximativ 65 de ore (vezi D3.15). Noul corpus a fost utilizat pentru a antrena un vocoder de tip *WaveGlow* utilizat împreună cu Tacotron2, precum şi o serie de noi voci folosind metoda *Tranfer Learning* pentru a valida tehnologia de adaptare a sistemului de sinteză la noi vorbitori. Mostre audio sunt disponibile la adresa: https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home.

(2) Aplicaţia RecoApy[3] permite înregistrarea secvenţială a propoziţiilor din corpusul SWARA 2.0 cu salvarea unei copii în caz de eroare, respectiv afişarea transcrierii fonetice a textului pe bază de reţele neuronale convoluţionale (CNN - Convolutional Neural Networks) şi de tip Transformer antrenate cu date din Wiktionary[4] în opt limbi: română, engleză, spaniolă, franceză, germană, italiană, cehă şi poloneză. Performanţa transcrierii fonetice cu această aplicaţie este le nivelul metodelor de dată recentă din domeniu.

(3) Adaptarea unui sistem de sinteză cu vorbitori multipli pe baza unei funcţii de cost suplimentare. Pentru îmbunătăţirea învăţării identităţii vorbitorului şi adaptarea unui sistem de vorbitori multipli la un nou vorbitor, s-a propus folosirea unui sistem de identificare a vorbitorului, combinat cu un sistem de sinteză text-vorbire (sistemul adoptat este DC-TTS), prin adăugarea unei funcţii de cost suplimentare bazată pe reprezentările interne ale sistemului de identificare a vorbitorului. Pornind de la dezvoltările software realizate de grupul de cercetare CSTR[5], s-a implementat reţeaua DC-TTS cu posibilitatea de concatenare a unei reprezentări vectoriale a identităţii vorbitorului (embedding) la cele 3 module ale componentei Text2Mel, respectiv prin utilizarea reprezentării vectoriale într-o strategie de învăţare a contribuţiei reprezentării vectoriale la canalele de informaţie din reţea. Astfel, este posibilă inserarea identităţii vorbitorului în mai multe puncte ale reţelei. Sistemul cu vorbitori multipli a fost antrenat în 3 scenarii:

Scenariul 1 **(B)**: sistem cu vorbitor unic sau vorbitori multipli considerat sistemul de bază.

Scenariul 2 **(B+CS)**: sistem cu vorbitor unic sau vorbitori multipli şi adăugarea unei funcţii de cost suplimentare obţinută prin calculul similarităţii spectrale.

Scenariul 3 **(B+E)**: sistem cu vorbitor unic sau vorbitori multipli cu adăugarea unei funcţii de cost suplimentare obţinută cu rata de eroare egală (EER - Equal Error Rate) folosind sistemul de verificare de vorbitor. Acest sistem a fost antrenat cu corpusul SWARA folosind o implementare[6] de la Clova AI Research. Cele 3 scenarii au fost antrenate cu diferite seturi de date:

- **ALL -** 18 vorbitori, între 1.000 şi 1.500 de propoziţii / vorbitor, 21.302 pronunţii.
- **RND1** – 18 vorbitori, aproximativ 500 de pronunţii / vorbitor, 8.932 de pronunţii.
- **RND1-100** – 18 vorbitori, aproximativ 100 de pronunţii / vorbitor, 1.787 de pronunţii.
- **RND1-SAM** – 1 vorbitor, în total 500 de pronunţii.

Evaluare. Sistemele au fost evaluate obiectiv calculând valoarea EER. Fiecare semnal sintetizat este comparat cu o mostră naturală de la un acelaşi vorbitor şi de la un alt vorbitor ales aleator. Tabelul 1 prezintă rezultatele obţinute. De asemenea, rezultatele obţinute prin aplicarea metodei de codare a vorbitorilor prin embedding pot fi vizualizate cu ajutorul metodei de t-Distributed Stochastic Neighbour Embedding (t-SNE) (Fig 1, 2).

Tabel 1. EER pentru sistemele Baseline, CosSim şi EER, antrenate cu diferite seturi de date

[2] *https://gitlab.utcluj.ro/sadriana/recoapy/*

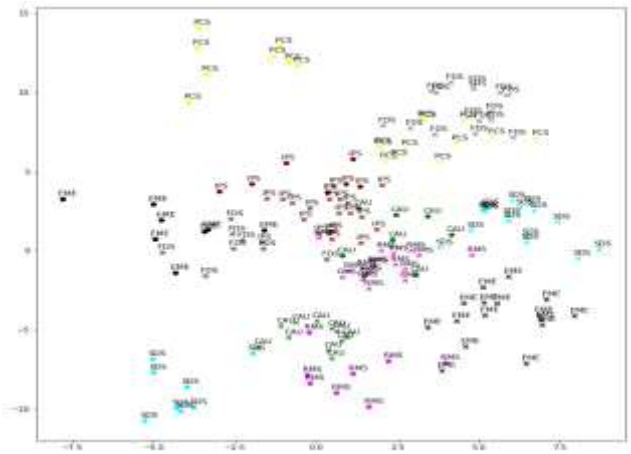
[3] *https://www.isca-speech.org/archive/Interspeech_2020/pdfs/1184.pdf*

[4] *https://www.wiktionary.org/*

[5] http://www.cstr.ed.ac.uk/

[6] https://github.com/clovaai/voxceleb_trainer

Sistem	ALL (EER)	RND1 (EER)	RND1-100 (EER)	RND1-SAM (EER)
B	*6.94*	*4.86*	*8.33*	*2.43*
B+CS	*6.25*	***4.66***	*6.25*	*2.43*
B+E	***4.66***	*8*	***6***	*2.43*



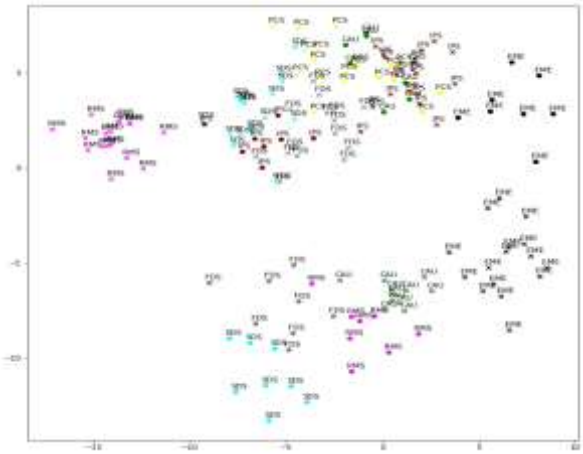


Fig. 1. Embedding pentru sistemul B (Baseline) *Fig. 2. Embedding pentru sistemul B + E*

(4) Adaptarea vorbirii din date audio ne-expresive la date audio cu expresivitate. Antrenarea sistemelor a folosit implementarea[7] CSTR bazată pe Tensorflow 1.13. Am efectuat două tipuri de experimente.

Date audio pentru Experiment tip A. Datele folosite pentru antrenarea primului tip de experimente aparţin vorbitorului SAM din corpusul SWARA şi au fost împărţite în două categorii, în funcţie de evaluarea subiectivă a gradului lor de expresivitate. Cele două categorii selectate sunt bazate pe selectare manuală: date neexpresive: ● rnd1, rnd2, rnd3, diph1, diph2 (2.476 propoziţii) ● date expresive: citirea nuvelelor "Ivan Turbincă", „Stan păţitul" (704 de propoziţii)

Date audio pentru Experiment tip B. Pentru al doilea tip de experimente am adăugat date neexpresive de la vorbitorii BEA, EME şi IPS din corpusul SWARA. De la aceşti vorbitori am utilizat câte 40 sau 100 de pronunţii selectate din rnd1, fiind considerate date neexpresive.

Date text. Textele corespunzător datele audio au fost folosite în trei forme diferite:

(1) Forma ortografică	*Pe de altă parte, conform rezultatelor obţinute*
(2) Forma transcrisă fonetic	*p e <> d e <> a l t @ <> p a r t e <,> k o n f o r m <> r e z u l t a t e l o r <> o p ts i n u t e*
(3) Forma transcrisă fonetic cu accent	*p e0 <> d e0 <> a1 l t @0 <> p a1 r t e0 <,> k o0 n f o1 r m <> r e0 z u0 l t a1 t e0 l o0 r <> o0 p ts i0 n u1 t e0*

Rezultate Experiment A. După 3.000 de epoci antrenarea a fost continuată cu date expresive de la vorbitorul SAM, textul de intrare fiind acelaşi cu cel folosit în modelul antrenat pe date neexpresive. Mostre audio pot fi ascultate aici: https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home. Sistemele antrenate sunt listate în Tabelul 2.

Tabel 2. Descrierea experimentelor din prima categorie (A)

[7] *https://github.com/CSTR-Edinburgh/ophelia*

Forma text	Date audio	Epoci	Număr de propoziţii	Date audio suplimentare	Epoci suplimentare	Număr de propoziţii
Ortografică	*neexpresiv*	*3.000*	*2.476*	*Expresiv*	*3.000*	*704*
Fonetică	*neexpresiv*	*3.000*	*2.476*	*Expresiv*	*7.000*	*704*
Fonetic + accent	*neexpresiv*	*3.000*	*2.476*	*Expresiv*	*6.000*	*704*

Rezulate Experiment B. Textul de intrare a fost furnizat sistemului în forma transcrisă fonetic. Pentru a învăţa mai mulţi vorbitori, ID-ul de vorbitor a fost concatenat textului de intrare şi sistemul a fost antrenat cu vorbitori multipli. Identitatea mai multor vorbitori în sistem a fost învăţată cu ajutorul metodei LCC (Learning Channel Contribution). Aceste categorii de experimente au avut rezultate modeste, identitatea vorbitorului fiind păstrată sau învăţată, dar caracterul de expresivitate al vorbirii a fost pierdut în procesul de învăţare. Mostre audio: https://rb.gy/6pbnny. Sistemele antrenate sunt listate în Tabelul 3.

Tabel 3. Descrierea experimentelor din a doua categorie (B)

Forma text	Date audio	Epoci	Număr de propoziţii	Vorbi-tor	Date suplimentare	Epoci adiţionale	Nr propoziţii	Vorbi tor
Fonetică	*Expresiv*	*2.000*	*704*	*SAM*	*neexpresiv*	*1.500*	*40*	*BEA*
Fonetică	*Expresiv*	*2.000*	*704*	*SAM*	*neexpresiv*	*1.500*	*100*	*BEA*
Fonetică	*Expresiv*	*2.000*	*704*	*SAM*	*neexpresiv*	*1.500*	*40*	*EME*
Fonetică	*Expresiv*	*2.000*	*704*	*SAM*	*neexpresiv*	*1.500*	*100*	*EME*
Fonetică	*Expresiv*	*2000*	*704*	*SAM*	*neexpresiv*	*1.500*	*40*	*IPS*
Fonetică	*Expresiv*	*2000*	*704*	*SAM*	*neexpresiv*	*1.500*	*100*	*IPS*

5) O nouă metodă de adaptare la vorbitor folosind post-filtrarea[8]. Extinzând diferitele experimente realizate pentru sistemele de sinteză text - vorbire bazate pe reţele neuronale, am ajuns la concluzia că există o serie de limitări în adaptarea acestora la un nou vorbitor. În special, atunci când se doreşte adaptarea reţelei cu un set redus de date, problemele sunt şi mai complexe. Astfel, noi am propus o nouă metodă de adaptare a sistemului de sinteză la un nou vorbitor folosind post-filtrarea semnalului vocal sintetizat cu o reţea pre-antrenată cu date puţine, tot cu o structură de reţea neuronală care este antrenată să mapeze semnalul vocal sintetizat în semnal vocal natural.

Sistemele de sinteză implementate prin această metodă în laborator au fost validate atât prin metode obiective (MSD), cât şi subiective (MOS). Vezi şi înregistrarea, https://youtu.be/OLAJGaQmjqA.

3.2. Metodă de adaptare rapidă a sistemului de sinteză folosind date atipice (vezi D3.16)

(1) Augmentarea datelor de intrare de tip text prin predicţia simultană a informaţiei lexicale de nivel înalt. Pornind de la resursele existente la partnerul ICIA: RoSyllabiDict, MaRePhor, şi DEX online, s-au antrenat o serie de reţele neuronale de tip LSTM, BLSTM, CNN, CNN cu atenţie, ce au ca obiectiv predicţia simultană a transcrierii fonetice, silabificării, poziţionării accentului, lematizării, respectiv părţii de vorbire (rezultatele au fost prezentate la conferinta KES 2020[9]). Experimentul s-a realizat comparativ şi pentru limba engleză, iar detaliile aferente sunt prezentate doar în livrabilul D3.16. (vezi Tabel 4).

Tabelul 4. Rezultate ale predicţiei simultane a informaţiei lexicale din textul de intrare

[8] *https://youtu.be/OLAJGaQmjqA (Prezentare la conferință)*

[9] *https://youtu.be/-iLf2ZvTeKY (Prezentare conferință)*

Arhitectură	Limba	Acurateţe [%]			
		3 informaţii lexicale	**fără silabificare**	**fără accent**	**la nivel de caracter**
CNN cu atenţie	*română*	***86.64***	***88.83***	***93.84***	-
LSTM cu atenţie	*română*	*86.26*	*88.68*	*92.87*	-
LSTM	*română*	*84.60*	*86.89*	*91.19*	-
BLSTM	*română*	*86.10*	*88.13*	*92.73*	-
CNN cu atenţie	*engleză*	***58.96***	***59.70***	***64.00***	*85.53*
LSTM cu atenţie	*engleză*	*53.79*	*54.43*	*57.24*	*81.15*
LSTM	*engleză*	*52.81*	*53.41*	*56.33*	*90.30*
BLSTM	*engleză*	*56.02*	*56.72*	*59.71*	***94.94***

Aceste rezultate promiţătoare au determinat atât extinderea corpusului de intrare (corpus de 330.000 cuvinte adnotate, RoLEX – Romanian Lexicon), cât şi a arhitecturii de reţea neuronală la una de tip Transformer antrenată cu subseturi de 5.000, 50.000, 100.000, 150.000 şi 330.000 de cuvinte. Vezi şi livrabilul D3.6 (ICIA) pentru o analiză şi discuţie mai detaliată.

Adiţional, pentru a îmbogăţi descriptorii pentru textul de intrare s-a dezvoltat şi un modul de predicţie automată a **lemei unui cuvânt** folosind reţele neuronale recurente LSTM şi convoluţionale CNN. Sistemele au fost antrenate folosind setul de date DEX online precum şi un subset din setul de date CoRoLa, fie doar prin perechi cuvânt – lemă, fie introducând şi atributul de parte de vorbire (POS – Part of Speech Tagging) la nivel de trigram (doar pentru CoRoLa). Pentru setul de date colectat din DEX, sistemele bazate pe reţele LSTM ating o acurateţe de 99.43% la nivel de caracter şi cu adnotare POS, iar pentru subsetul CoRoLa cea mai mare acurateţe este de 99.69% pentru CNN.

(2) Aplicarea informaţiei lexicale pentru augmentarea reprezentării textului în sisteme de sinteză text – vorbire cu date atipice. S-au antrenat mai mute sisteme de sinteză text-vorbire, atât pentru limba română (doar pe acestea le prezentăm aici), cât şi pentru limba engleză (vezi livrabil D3.16 pentru detalii), toate bazate atât pe arhitectura DC-TTS cât şi Tacotron2 şi antrenate cu date audio, respectiv text adnotat cu informaţiile lexicale prezise conform cu descrierea anterioara. **Evaluare**: în cadrul etapei de evaluare s-a analizat în ce măsură corectitudinea informaţiilor lexicale suplimentare influenţează sinteza sistemelor în următoarele scenarii de testare: a) silabificare şi/sau accent corecte pentru întreaga propoziţie, b) silabificare/accent incorecte pentru un singur cuvânt, c) silabificare/accent absente pentru un singur cuvânt, d) silabificare/accent incorecte pentru toate cuvintele (alocate aleator), e) silabificare/accent absente pentru toate cuvintele (alocate aleator)

Sistemele descrise mai sus au fost validate folosind metrica distanţa cepstrală MCD (*Mel Cepstral Distortion)*. Pentru calcularea valorilor MCD au fost sintetizate câte 20 de propoziţii cu maximum 20 de cuvinte şi 20 de propoziţii scurte cu maxim 5 cuvinte, pentru fiecare sistem, în toate cele 5 scenarii de testare. Rezultatele sunt prezentate în figurile alăturate, iar mostre audio pentru aceste sisteme pot fi accesate şi ascultate aici: https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home .

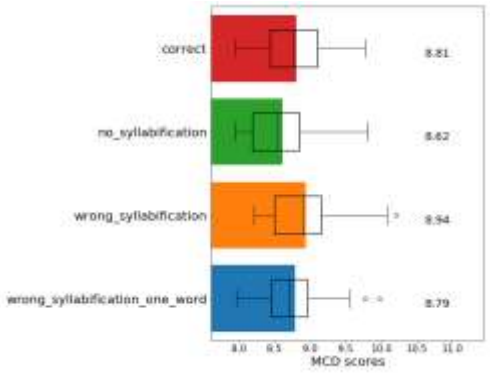


Fig. 3 MCD natural vs sintetic

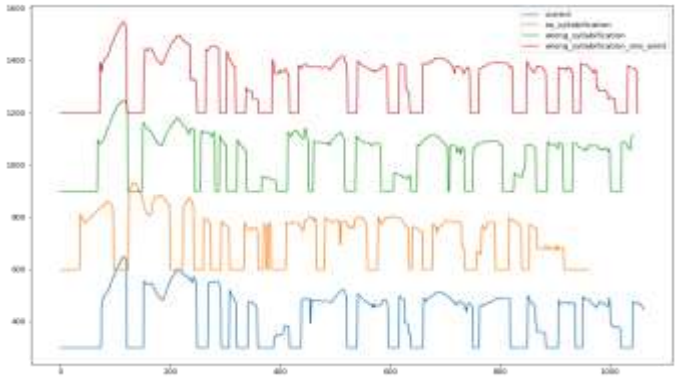


Fig. 4 Variaţia F0, natural vs sintetic, un vorbitor

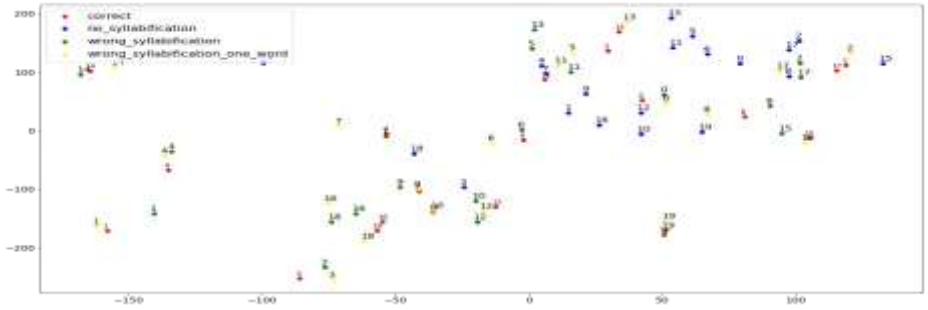


Fig. 5 Proiecţiile t-SNE pentru embedding vorbitor la cele 20 de propoziţii

(3) Adaptarea sistemului de sinteză cu date atipice prin transfer de stil de vorbire şi Flowtron. Cercetările din această etapă au avut în vedere şi transferul stilului de vorbire folosind un set redus de date prin intermediul metodei *Transfer Learning*. Am adoptat soluţia de antrenare *One-Shot Learning*. Această soluţie este posibilă datorită fluxurilor de normalizare implementate la nivelul acestei reţelei neuronale.

Practic, s-au antrenat 2 sisteme: unul ce foloseşte doar corpusul Mara (vorbitor unic) şi unul ce utilizează corpusul SWARA (vorbitori multipli). Pentru transferul stilului au fost folosite date audio provenite din buletine de ştiri, precum şi date preluate de la alţi vorbitori din corpusul SWARA. Deşi Flowtron este capabil să realizeze transferul stilului pe baza modificărilor spaţiului latent[10], este nevoie ca sistemele antrenate să folosească un set de date extins şi de foarte bună calitate. Se poate observa în figura alăturată că alinierea text-audio nu este realizată în mod ideal, ceea ce introduce erori de pronunţie şi artefacte audio. Rezultate audio ale acestor experimente sunt disponibile la adresa: https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home.

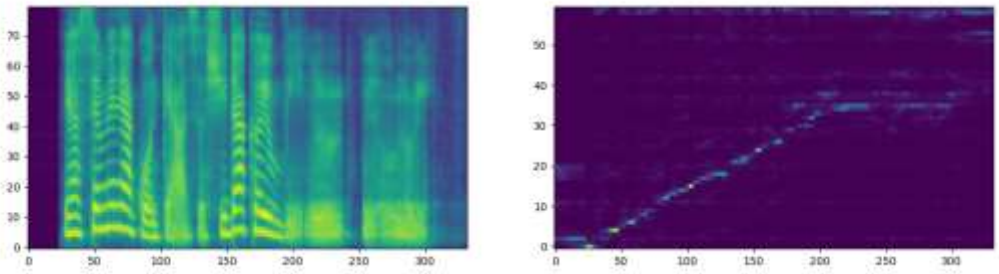


Fig. 6. Exemplu Mel-spectrogramă şi aliniere text-audio în sistemul Flowtron-Mara

(4) Rezultate experimentale ale metodei de adaptare folosind date atipice. Datele atipice se referă la date imperfecte precum: transcrieri inexacte, lipsa segmentărilor la nivel de propoziţie, condiţii de zgomot diferite, etc. Toate aceste probleme se transpun în calitate scăzută a ieşirii sistemului de sinteză. Astfel, este importantă dezvoltarea unor metode ce reduc influenţa acestor date atipice în vocea sintetizată. Pentru acest

[10] https://nv-adlr.github.io/Flowtron

scop, s-au antrenat 2 sisteme ce utilizează fie o voce a unei prezentatoare de ştiri şi pentru care erau disponibile doar 20 de minute de audio, fie o voce masculină extrasă dintr-o resursă colectată în P1 şi transcrisă în P3. Pentru vocea din urmă, transcrierile nu sunt 100% corecte, astfel că se poate analiza influenţa erorilor de transcriere. Mostre audio ale acestor voci sunt disponibile în pagina https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home, sistem FEM şi sistem COB. Rezultatekle indică o foarte bună flexibilitate a arhitecturii sistemului în ceea priveşte adaptarea la date atipice.

3.3. Tehnologie de realizare a interfeţelor om-maşină pentru sinteza text–vorbire[11] (D3.17)

Au fost extinse tehnologiile de sinteză text-vorbire în limba română bazate pe modele Markov şi disponibile în website-ul www.romaniantts.com , la tehnologiile bazate pe reţele neuronale multistrat şi utilizarea informaţiilor lingvistice suplimentare în transcrierea textului de intrare. Deoarece serverele pe baza de GPU care deservesc demonstraţia online sunt folosite în mod curent de grupul de cercetare pentru diferite experimente, utilizatorii au nevoie de o cheie de API ce poate fi obţinută de la coordonatorii P4.

Tabel 5. Sistemele de sinteză disponibile în platforma API:

Tehnologie pentru sistemul de sinteză	**Voci[12] disponibile pentru demonstrare online**
Sistem bazat pe reţele neuronale convoluţionale (DC-TTS)	*Mara, BAS, EME, CAU, DDM, DCS, FDS, IPS, PCS, SAM*
Sistem bazat pe reţele neuronale recurente (Tacotron 2) şi vocoder WaveGlow	*Mara, Mara + infomaţii lexicale, EME, NL, DOL, DLL, ZPL, SRL, FEM*
Sistem bazat pe fluxuri de normalizare (Flowtron)	*Mara*

4. Oferta de servicii de cercetare, locuri de muncă susţinute şi valorificarea resurselor

Tabel 6. Sinteză privind oferta de servicii, locuri de muncă şi valorificarea resurselor în UTCN

Oferta de servicii în UTCN	● oferta unor noi tehnologii de sinteză text-vorbire cu expresivitate, în limba română, bazată pe reţele neuronale şi aliniată la standardele internaţionale (Tacotron GST, DC-TTS, Flowtron) ● servicii de adnotare automată a resurselor de date text pornind de la corpusurile CoRoLa şi RoLEX ● servicii de înregistrare audio prin aplicaţia RecoApy ● servicii de procesare paralelă a datelor folosind tehnici de învăţare automată pe noile echipamente achiziţionate în anul 2019 din proiect ● servicii software pentru dezvoltarea modelelor bazate pe învăţare automată. *ERRIS: https://erris.gov.ro/speech.utcluj.ro*
Locuri de muncă susţinute în UTCN	1 x CS I, 1 x CS III, 1 x Tehnician 2 x ACS nou angajaţi începând cu luna ianuarie 2019
Resursa umană nou angajată în UTCN	Conform acordului de grant au fost angajate 2 ACS, doctoranzi, începând cu 1 ianuarie 2019.
Valorificare resurse în parteneriat	● UTCN a preluat de la ICIA resurse de date text (CoRoLa, RoLEX) pentru predicţia simultană a informaţiilor lexicale de nivel înalt ● UTCN a preluat de la UAIC un set de înregistrări audio / interviuri radio pentru a demonstra antrenarea cu date atipice, respectiv transferul de stil şi expresivitate ● UTCN a furnizat pentru ICIA o arhitectură de tip Transformer pentru a realiza adnotări automate a textului ● UTCN a preluat de la UPB un set de transcrieri text a unor interviuri furnizate de UAIC ● UAIC a furnizat pentru UTCN acces la o platformă online pentru stocarea corpusurilor bimodale.
Cecuri	● Nu au fost planificate cecuri pentru UTCN în această etapă de raportare.

5. Management şi comunicare

Activităţile de management au fost orientate în special către managementul proiectului complex în vederea integrării grupurilor de cercetare şi a schimbului de resurse de date sau software între partenerii din proiect. S-a solicitat în martie 2020 de către Autoritatea Contractantă întocmirea unui act aditional pentru redistribuirea unei părţi din buget pe anul 2021. S-au organizat mai multe conferinţe Skype pentru prezentarea rezultatelor intermediare. Din punct de vedere financiar s-au primit 3 tranşe de avans cu o

[11] http://speech.utcluj.ro/ronna/

[12] *Acronimele indică vorbitori anonimizați*

regularitate adecvată. Resursele financiare alocate UTCN pentru anul 2019 au fost utilizate în majoritate, cu excepţia unor sume în categoria salarii pentru tinerii cercetători (concediu de maternitate pentru o nou angajată), respectiv deplasări (participarea la conferinţe s-a realizat online şi doar taxele de participare au fost plătite).

6. Diseminarea rezultatelor

O preocupare în UTCN şi în această etapă de raportare a fost implementarea şi îndeplinirea cu succes a obiectivelor stabilite în strategia de diseminare a rezultatelor elaborată în cadrul propunerii de proiect. Astfel, adecvat acestei etape iniţiale s-a acţionat pe următoarele direcţii:

a) actualizarea paginii web a proiectului SINTERO *(http://speech.utcluj.ro/sintero/)*,

b) crearea de pagini Wiki interne grupului de cercetare UTCN (mostre audio cu semnal sintetizat - https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home, aplicaţia RecoApy - https://gitlab.utcluj.ro/sadriana/recoapy), pagini web dedicate pentru demonstrarea online a tehnologiilor dezvoltate în această etapă (interfaţa online //speech.utcluj.ro/ronna), prezentări video ale unor articole susţinute online la conferinţe internaţionale (https://youtu.be/OLAJGaQmjqA,http://adrianastan.com/docs/AStan_RecoApy_Interspeech2020.mp4, https://youtu.be/-iLf2ZvTeKY,).

c) publicaţii ştiinţifice cu rezultatele cercetărilor la conferinţe internaţionale în domeniu

d) implicarea studenţilor în stagii de practică, proiecte de diplomă sau disertaţii cu tematică apropiată de cea a proiectului: *5 studenţi în stagii de practică în perioada iulie – august 2020, 5 studenţi la proiect de diplomă susţinut în Iulie 2020, 1 student la lucrarea de disertaţie susţinută în Iulie 2020.*

e) acorduri de partneriat cu mediul economic: 1 acord de partneriat (în lucru) cu firma SC DiktaCom pentru cesionarea reciprocă de date audio în vederea dezvoltării unor tehnologii pe bază de prelucrare automată a semnalului vocal.

7. Concluzii

Activităţile de cercetare desfăşurate în etapa a III-a de implementare a proiectului (2020) au condus la obţinerea rezultatelor aşteptate şi ele sunt în concordanţă cu obiectivele specifice ale etapei. Astfel, rezultatele raportate în acest document şi descrise detaliat în cele 4 livrabile aferente perioadei de raportare (vezi Secţiunea 8 a acestui raport), asigură testarea şi evaluarea finală a tehnologiei dezvoltate în etapa finală a proiectului (2021).

8. Referinţe la livrabilele aferente etapei 2020 (Anexe la raport)

[1] Livrabil D3.15:	*„Dezvoltarea unei noi tehnologii pentru adaptarea vocii sintetice la stilul şi expresivitatea unui nou vorbitor”,* Noiembrie 2020.
[2] Livrabil D3.16:	*„Dezvoltarea unei noi metode de adaptare rapidă a vocii sintetice folosind date audio atipice”,* Noiembrie 2020.
[3] Livrabil D3.17	*„Integrare tehnologie nouă şi demonstrarea în realizarea interfeţelor om-maşină pentru sinteza text – vorbire”,* Noiembrie 2020.
[4] Livrabil D3.18:	*„Diseminare”,* Noiembrie 2020.

Diseminarea rezultatelor proiectului complex ReTeRom

Diseminarea rezultatelor proiectului a fost realizată prin intermediul website-ului proiectului complex http://www.racai.ro/p/reterom/index_en.html care include trimiteri la paginile proiectelor componente: http://www.racai.ro/p/reterom/index_en.html#pr2, http://cobiliro.info.uaic.ro:3083/, https://tadarav.speed.pub.ro, https://speech.utcluj.ro/sintero/#).

De asemenea rezultatele proiectului au fost diseminate prin publicarea mai multor articole ştiinţifice:

1. V. Păiș, R. Ion, D.Tufiș. A Processing Platform Relating Data and Tools for Romanian Language. In: Proceedings of the 1st International Workshop on LanguageTechnology Platforms (IWLTP 2020), European Language Resources Association (ELRA), Georg Rehm et al. (eds.), pp. 81-88 - indexed by DBLP and ISI Web of Science.
2. V. Păiș, D. Tufiș, R. Ion. MWSA Task at GlobaLex 2020: RACAI's Word Sense Alignment System using a Similarity Measurement of Dictionary Definitions. In Proceedings of The 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 69-75, - indexed by DBLP and ISI Web of Science, May 2020
3. V. Păiș, R. Ion. TermEval 2020: RACAI's automatic term extraction system. In Proceedings of the 6th International Workshop on Computational Terminology. European Language Resources Association, Marseille, France, pp. 101-105, indexed by DBLP and ISI Web of Science, May 2020
4. G. Rehm, ... D. Tufiș,..., F. Yvon. The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe. In Proceedings of The 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 3315-3325, May 2020, indexed by DBLP and ISI Web of Science
5. Ș. Boghiu, D. Gîfu. Event Detection from the Spatial-Temporal Perspective at the 3rd Conference of Doctoral Schools from the UNIVERSITARIA Consortium, see, CSDCU-MIF2020, Oct. 22-24, 2020, Iași, Romania.
6. Ș., D. Gîfu. A Spatial-Temporal Model for Event Detection in Social Media. In: Procedia Computer Science, Vol. 176, Matteo Cristani, Carlos Toro, Cecilia Zanni-Merk, Robert J. Howlett, Lakhmi C. Jain (eds.), ELSEVIER, 2020, pp. 541-550, ISSN 1877-0509
7. D. Cristea, I. Pistol, Ș.Boghiu, A. Bibiri, D. Gîfu, A. Scutelnicu, M. Onofrei, D. Trandabăț, G. Bugeag. CoBiLiRo: a Research Platform for Bimodal Corpora. In: Proceedings of the 1st International Workshop on LanguageTechnology Platforms (IWLTP 2020), European Language Resources Association (ELRA), George Rehm et al. (eds.), pp. 22-27 - indexed by DBLP and ISI Web of Science.
8. A.Iftene, D. Gîfu, A. R. Miron, M. Șt. Dudu. A Real-Time System for Credibility on Twitter. In: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), Nicoletta Calzolari et al. (eds.), European Language Resources Association (ELRA), pp. 6168-6175 - indexed by DBLP and ISI Web of Science - Rank C.
9. A.-L. Georgescu, H. Cucu, A. Buzo, C. Burileanu, "RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition," in the Proceedings of The 12th Language Resources and Evaluation Conference (LREC), pp. 6606-6612, 2020, Marseille, France.
10. C. Manolache, A.-L. Georgescu, A. Caranica, H. Cucu, "Automatic Annotation of Speech Corpora using Approximate Transcripts," in the Proceedings of the 43rd International Conference on Telecommunications and Signal Processing (TSP), 2020, Milano, Italy.
11. D. Oneaţă, A.-L. Georgescu, H. Cucu, D. Burileanu, C. Burileanu, "Revisiting SincNet: An Evaluation of Feature and Network Hyperparameters for Speaker Recognition," in the Proceedings of the 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 2020.
12. G. Pop, H. Cucu, D. Burileanu, C. Burileanu, "Cough Sound Recognition in Respiratory Disease Epidemics," in Romanian Journal of Information Science and Technology, vol. 23, no. S, pp. S77–S89, 2020, ISSN 1453-8245, ISI IF 0.661.
13. A.-L. Georgescu,, C. Manolache, D. Oneaţă, H. Cucu, C. Burileanu, "Data-filtering methods for self-training of automatic speech recognition systems," in the Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Virtual, 2021.
14. D. Oneaţă, A. Caranica, A. Stan, H. Cucu, "An evaluation of word-level confidence estimation for end-to-end automatic speech recognition," in the Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Virtual, 2021.
15. B. Lőrincz, M. Nutu, A. Stan, M. Giurgiu. "An Evaluation of Postfiltering for Deep Learning Based Speech Synthesis with Limited Data", IEEE 10th International Conference on Intelligent Systems (IS), Bulgaria, 2020
16. B. Lőrincz. "Concurrent phonetic transcription, lexical stress assignment and syllabification with deep neural networks", Proceedings of the 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems KES2020, 2020.
17. A.Stan, "RECOApy: Data Recording, Pre-Processing and Phonetic Transcription for End-to-End Speech-Based Applications", In Proceedings of the Interspeech, Shanghai, China, 2020
18. K. M Scott, S. Ashby, A. Stan ."Designing a Synthesized Content Feed System for Community Radio", Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, Estonia, 2020

Stadiul elaborării si asumării Programului comun de CDI, cu evidențierea modului de colaborare ulterioară între parteneri și atragerea de noi fonduri naționale/internaționale.

Planul comun este prezentat in anexa acestui raport.

Detalii privind angajarea si menţinerea noilor cercetători

Nr. posturi asumate de noi cercetători	7
Nr. posturi ocupate de noi cercetători	8
Nr. posturi ocupate de noi cercetători (in prezent)	7

Lista noi cercetători								
Nr. crt.	**Instituție**	**Nume**	**Prenume**	**Poziția ocupata in cadrul proiectului**	**Data angajare in proiect**	**Perioada implicare in proiect**	**Costuri salariale alocate**	**Costuri salariale plătite**
1.	ICIA	Păiș	Vasile	Doctorand/ CS	01.09.2018	1,09.2018 - prezent	211.463	193.033
2.	UAIC	Pădurariu	Constantin Cristian	Membru doctorand	03.12.2018	03.12.2018 – 30.11.2019	110.871	110.436
3.	UAIC	Boghiu	Șerban Paul	Membru doctorand	01.03.2020	01.03.2020 - prezent	102.471	106.761
4.	UPB	Oneață	Dan Theodor	CȘ	24.04.2018	24.04.2018 - prezent	206.066	186.926
5.	UPB	Pop	Gheorghe	AȘC	24.04.2018	24.04.2018 - prezent	74.610	67.586
6.	UPB	Manolache	Cristian	AȘC	24.04.2018	24.04.2018 - prezent	103.871	93.686
7	UTCN	Beáta	Lőrincz	AȘC	03.01.2019	03.01.2019 13.10.2020	116.320	77.696
8	UTCN	Maria	Nuțu	AȘC	03.01.2019	03.01.2019 - prezent	116.320	105.188

ICIA va menține cercetătorul angajat pe proiect căruia ii va oferi un contract de muncă în institut pe perioada nedeterminată, activitatea sa fiind foarte apreciată. El va fi finațat atât prin fonduri bugetare (1/2 normă) cât și din proiecte extrabugetare (1/2 normă: proiectele europene MARCELL, CURLICAT).

Universitatea ”A.I. Cuza” din Iași va menține în cadrul proiectului noul cercetător angajat și îl va plăti din fonduri proprii până în luna februarie 2021, când i se va oferi un post de ASC în cadrul Facultății de Informatică pentru o perioadă de 2 ani.

Universitatea Politehnica din Bucureşti va menține în vigoare posturile de noi cercetători oferind granturi interne instituției. Modalitatea de finantare pentru personalul nou angajat in proiecte PCCDI se va face prin urmatoarele mecanisme:

- Cristian Manolache va fi încadrat cu normă întreagă în proiectul de cercetare “Climatologia aerosolului - de la masuratori de teledetectie la invatare profunda structurata” (proiect CLARA), contract de finanţare nr. 295PED/ 2020; el are contract până pe 16.08.2022.
- Dan Theodor ONEAŢĂ va fi încadrat cu normă întreagă în proiectul de cercetare Recunoaşterea automată a vorbirii din semnale multi-modale (VORBIS), contract de finanţare nr. PD 97/ 2020; el are contract pana pe 31.08.2022
- Gheorghe POP va aplica la granturile interne UPB de tip Proof of Concept (PoC) destinate tinerilor cercetatori care au participat la proiecte de tip PCCDI, in vederea finantarii activitatii lor de cercetare in perioada de sustenabilitate de 2 ani din fondurile UPB.

Dupa aceste date, cei doi vor putea accesa si ei un proiect UPB-POC pentru a-si completa cele 3 luni ramase din perioada de sustenabilitate a proiectului ReTeRom.

Universitatea Tehnică Cluj Napoca va menține posturile de noi cercetători și va oferi granturi interne instituției pe bază de competiție. Modalitatea de finantare pentru personalul nou angajat in proiecte PCCDI se va face prin urmatoarele mecanisme: a) incadrare cu norma intreaga pe un singur proiecte sau cumulat pana la o norma intreaga din mai multe proiecte de cercetare din universitate in care sunt necesare competentele noilor cercetatori; b) incadrare cu norma intreaga pe posturi sustenabile financiar din Fondul de Sustinere a Cercetarii cu un salar care sa maximizeze perioada de sustenabilitate (Cf. HCA 126/17.11.2020).

ANEXĂ

Stadiul elaborării si asumării Programului comun de CDI, cu evidențierea modului de colaborare ulterioară între parteneri și atragerea de noi fonduri naționale/internaționale.

a) Considerăm că este oportun sa formăm (consorțiul ReTeRom) un pol de cercetare la nivel național in domeniul tehnologiei vorbirii. Aceasta implica agregare de resurse începând de la HW, SW și resurse umane (le avem) + un plan de cercetare (de construit împreuna). Consider ca doar așa putem intra in marile proiecte ale CE, dar si la nivel național.
b) Derivat din (a), acțiuni punctuale pentru: formare/educație, colaborare cu mediul economic, atragere de fonduri pentru sustenabilitate.

Acest proiect nu este neapărat pentru noi, cei din lista de proiect, cât mai degrabă pentru cei care vin după noi și în care am investit timp, credință, onestitate.

ICIA are în vedere aprofundarea cercetărilor și continuarea dezvoltării de aplicații de IA

A. Ne vom concentra o parte din eforturi pentru realizarea unor sisteme ASR care să poată fi folosite pentru transcrierea unor înregistrări de lungă durată. Sistemele realizate în cadrul acestui proiect nu permit transcrierea unei înregistrări mai lungi de aproximativ 20 secunde. Această restricție apare din cauza volumului de memorie RAM necesară prelucrării semnalului vocal. Pentru depășirea acestei limitări avem în vedere studiul și dezvoltarea unui flux de prelucrare inteligentă presupunând segmentarea înregistrării audio și prelucrarea succesivă a segmentelor audio cu concatenarea transcrierilor acestor segmente. Pentru că niciun sistem ASR nu este perfect, fluxul de prelucrare menționat trebuie incorporat unui mediu inteligent și prietenos de corectare a erorilor de transcriere. Problema nu este trivială și o astfel de aplicație este extrem de necesară și solicitată intens de parteneri academici dar și din media. Nu avem cunoștință de un astfel de sistem operațional pentru limba română.

B. Valorificând rezultatele obținute de ICIA într-un proiect anterior respectiv dezvoltarea unui sistem de traducere automată EN↔RO a limbajului scris și rezultatele comune cu UPB și UTCN obținute in cadrul acestui proiect în domeniul prelucrării limbajului vorbit, avem în vedere un proiect de traducere a limbajului vorbit EN↔RO. Pentru limba română vom folosi sistemele noastre ASR (ICIA și UPB) și TTS (ICIA si UTCN) iar pentru limba engleză fie vom folosi sisteme open-source existente fie vom dezvolta noi modulele corespunzătoare. Această soluție în cascadă, deși au apărut sisteme neuronale ce realizează traducerea automată direct din semnalul vocal în semnal vocal-fără a mai trece prin fazele de ASR și TTS, este relativ ușor de realizat. Eventual, într-o fază ulterioară am putea să ne propunem realizarea unui sistem de traducere automată EN↔RO direct din semnalul vocal fără a mai trece prin etapa textuală. Rezultatele direcției de cercetare/dezvoltare de la punctul A, vor fi direct relevante pentru crearea de resurse de antrenare pentru sisteme dezvoltate în direcția B de cercetare/dezvoltare.

C. Valorificând rezultatele cercetărilor curente ale ICIA în ce privește prelucrarea textelor românești cu rețele neuronale complexe și cercetări recente în reprezentarea cunoștințelor în grafuri de relații semantice, ne interesează construirea automată a unui graf de cunoștințe cuprinzător din situl enciclopedic Wikipedia, utilizând rețele neuronale de tip BERT pentru a descoperi relațiile semantice care se stabilesc între entități. Cu un astfel de graf disponibil, care este de așteptat să conțină sute de mii de relații semantice, vom construi un sistem de întrebare-răspuns (de ex. „Câte planete sunt în sistemul solar?") cu care se va putea interoga Wikipedia în limba română, utilizând sistemele de ASR și TTS descrise la punctul B.

UAIC are în vedere un număr de posibile aplicații care să demonstreze potențialul tehnologiilor dezvoltate în celelalte proiecte componente, în scopul măririi vizibilității proiectului, în special după finalizarea acestuia. Tehnologiile vizate în special sunt cele descrise în rapoartele 3.7 (Definitivarea, testarea, validarea și împachetarea într-o soluție „*ready-to-use*" a platformei integrate și configurabile de prelucrare a textelor în limba română.), 3.10 (Îmbunătățirea soluției de filtrare și aliniere a transcrierilor aproximative cu semnalul de vorbire), 3.15 (Dezvoltarea unei noi tehnologii pentru adaptarea vocii sintetice la stilul și expresivitatea unui nou vorbitor) și 3.17 (Integrare tehnologie nouă și demonstrare în realizarea interfețelor om-mașină pentru sinteza text-vorbire). O parte din proiectele descrise au depășit deja faza de proiect, fiind în stadii relativ avansate de implementare (în activități cu studenții de la UAIC-FII). Cele 6 proiecte de aplicații descrise sunt prezentate succint mai jos:

Aplicația 1: Suport pentru învățarea limbii române - *PD builder*. Un dicționar de pronunție pentru cuvintele unei limbi poate constitui un suport semnificativ pentru cei care doresc să învețe acea limbă.

Colecția de resurse colectate pe platforma CoBiLiRO are potențialul să atingă o dimensiune suficientă pentru a include pronunția marii majorități a cuvintelor din limba română; de aici a plecat ideea de a dezvolta o tehnologie ce caută în resursele din Portal și reproduce pronunții ale unor cuvinte, la cererea utilizatorului. Sunt utilizate adnotările produse de TEPROLIN și alinierile produse de TADARAV, marcaje ce însoțesc majoritatea resurselor existente pe platforma CoBiLiRo.

Aplicația 2: Analiza corpusurilor bimodale. Proiectul propune dezvoltarea unui sistem capabil să evalueze calitatea unui corpus bimodal voce-text din perspectiva unei alinieri automate sau manuale. O parte din posibilele erori semnalate ar putea fi corectate, lucru care ar mări semnificativ calitatea alinierii automate dar și a resursei bimodale în general. Prin corelarea statistică a mai multor parametri, pot fi detectate posibile diferențe între conținutul înregistrării și transcrierea ei în text, diferențe care pot avea cauze diverse: segmente adiționale prezente în text față de înregistrare sau invers, transcrieri aproximative, particularități ale pronunției și prozodiei vorbitorului etc. Un astfel de sistem ar fi util atât dezvoltatorilor și utilizatorilor tehnologiilor de aliniere automată cât și celor care colectează și utilizează resurse bimodale.

Aplicația 3: *I listen to my speaking agent reading book fragments as I walk by*. Aplicația are la bază o colecție de texte care abundă în entități geografice, marcate XML explicit, textele fiind însoțite de metadate care descriu minimum: autorul și titlul cărții, anul de apariție și editura. Instalată pe un dispozitiv mobil, ea va semnala proximitatea telefonului față de locațiile menționate în texte și va citi acele fragmente care includ mențiunile respective. În felul acesta, o plimbare printr-un mare oraș se poate transforma într-o călătorie literară. Aplicația se adresează persoanelor care ar dori să primească sugestii de lecturi și doresc să afle lucruri noi despre locurile pe care le vizitează. Ea își propune să îmbine într-un mod plăcut literatura și tehnologia.

Aplicația 4: Asistent inteligent al ședințelor online. Pandemia Covid a obligat ca o parte neneglijabilă a activităților colectivităților umane să migreze în mediul online. Propunem realizarea unui instrument sau a unei colecții de API-uri care să se poată integra aplicațiilor de teleconferințe, în scopul extragerii de informații și generării de rapoarte din discuțiile purtate și din mesajele schimbate. Astfel, se poate imagina: transcrierea conversațiilor în text, realizarea de rezumate ale întâlnirilor, extragerea de informații punctuale, pe anumite segmente din conferință ori din intervențiile unor anumiți vorbitori, generarea automată de procese verbale, interogarea minutei ori a procesului verbal generat, pe bază de cuvinte cheie, căutări de secvențe sonore în înregistrarea sonoră etc.

Aplicația 5: *Tracking assistent*. O aplicație Android care își propune să asiste persoanele care suferă de boala Alzheimer în a-și aminti traseele pe care le-au efectuat în timpul zilei printr-o interfață ușor de folosit, în limbaj natural. Pe baza datelor receptate din mai multe canale ale unui telefon mobil, se poate reface traseul și tipul de activitate al persoanei. La sfârșitul zilei pacientul ar purta un dialog cu sistemul inteligent, care l-ar ajuta să-și rememoreze activitățile și traseele de peste zi.

UPB are în vedere dezvoltarea continuă a sistemelor de transcriere automată a vorbirii și integrarea lor cu alte sisteme inteligente de procesare de text, respectiv sinteză de vorbire. Suplimentar față de participarea în proiectele expuse de ceilalți parteneri, UPB intenționează să demareze următoarele proiecte, implicând bineînțeles și partenerii din proiectul ReTeRom, în funcție de ariile lor de expertiză:

Chat-bot vocal. În ultima perioadă aplicațiile de chat-bot text au luat o amploare remarcabilă. Majoritatea site-urilor web ale marilor prestatori de servicii de telefonie, banking, etc. din România pun la dispoziție pe pagina principală un sistem interactiv de transmitere/recepție de mesaje text cu automatizarea răspunsurilor pentru întrebările frecvent apărute (chat-bot text). Tehnologiile dezvoltate sau îmbunătățite în proiectul ReTeRom pot face obiectul extinderii acestor aplicații de tip chat-bot prin adăugarea de funcționalități pentru interacțiunea prin voce. Astfel, **sistemul de recunoaștere automată a vorbirii** dezvoltat de **UPB** poate fi folosit pentru a transcrie mesajele vocale ale utilizatorilor, **tehnologiile de procesare de limbaj natural** dezvoltate de **ICIA** și **UAIC** pot fi folosite pentru înțelegerea mesajelor și generarea de răspunsuri text adecvate, iar **sistemul de sinteză de vorbire** dezvoltat de **UTCN** poate fi utilizat pentru sinteza și redarea răspunsurilor vocale către utilizator.

Sistem de IVR inteligent. Sistemele de IVR (interactive voice response) implementate în România folosesc în acest moment meniuri prin care se poate naviga prin voce, utilizatorul fiind foarte restricționat din punctul de vedere a ceea ce poate să rostească (e.g. *„Vă rugăm să ne spuneți care este problema dumneavoastră rostind: abonamentul de voce, abonamentul de internet, cartela telefonului, telefonie fixă”*). Un sistem IVR inteligent ar trebui să se adreseze utilizatorului în limbaj natural, similar cu un operator (e.g. *„Cu ce vă putem ajuta”*) și să poată să răspundă oricăror solicitări vocale ale utilizatorului. Pentru a implementa un astfel de sistem IVR inteligent pot fi folosite sistemele implementate în cadrul proiectului

ReTeRom. Astfel, **sistemul de recunoaștere automată a vorbirii** dezvoltat de **UPB** poate fi folosit pentru a transcrie solicitările vocale ale utilizatorilor, **tehnologiile de procesare de limbaj natural** dezvoltate de **ICIA** și **UAIC** pot fi folosite pentru înțelegerea solicitărilor și generarea de răspunsuri text adecvate, iar **sistemul de sinteză de vorbire** dezvoltat de **UTCN** poate fi utilizat pentru sinteza și redarea răspunsurilor vocale către utilizator.

UTCN are in vedere aprofundarea cercetărilor realizate în cadrul ReTeRom în domeniul Sintezei din Text a Vorbirii (STV) în limba română și consolidarea colaborării fructuoase pe care am avut-o cu partenerii în acest proiect, în vederea realizării unui pol de cercetare foarte puternic în domeniul tehnologiei vorbirii. Ne propunem o agendă de cercetare comună, care se dezvoltă pe următoarele direcții:

A. Colaborare cu partenerii din proiect și cu mediul economic pentru dezvoltarea resurselor audio și de text.

(A1) Extinderea de către UTCN a corpusului SWARA de la 65 de ore de vorbire, la mai mult de 100 de ore de vorbire, în următorul an, respectiv îmbogățirea lui cu date audio expresive și cu diversificarea modelului de limbaj pentru a putea deservi și crearea de sisteme de Recunoaștere Automată a Vorbirii (RAV).

(A2) Extinderea adnotărilor automate pentru resursele de semnal vocal și de text colectate în ReTeRom prin aplicarea tehnologiilor de prelucrare a textului dezvoltate de ICIA și a datelor audio dezvoltate de către UPB și UTCN.

(A3) Cooperarea cu mediul economic pentru extinderea resurselor de date audio și text, atragerea de fonduri pentru a asigura sustenabilitatea cercetărilor, respectiv pentru transfer tehnologic. UTCN și UAIC au dezvoltat un protocol de colaborare cu o firmă privată în vederea cesionării reciproce de resurse necesare sistemelor de antrenare (RAV, STV)

B. Dezvoltarea în parteneriat a tehnologiilor rezultate din proiect și aducerea lor pe un nivel apropiat de funcționarea în condiții reale.

(B1) Dezvoltarea unui sistem integrat de interacțiune om – mașină care să includă RAV și STV. Tehnologiile de RAV disponibile la UPB și ICIA, respectiv de STV, disponibile la UTCN, creează premisele unui proiect comun privind comunicarea om – mașină în limbaj natural. O provocare în cercetare, rămâne interpretarea semantică a mesajului vocal.

(B2) Dezvoltarea tehnologiei de STV prin creșterea naturaleței semnalului sintetizat cu apropierea de vorbirea spontană, respectiv adaptare la noi vorbitori prin aplicarea noilor tehnologii de conversie de voce.

(B3) Optimizarea sistemului de STV, respectiv creșterea gradului de paralelizare a procesărilor, pentru a răspunde unor aplicații de timp real.

C. Atragerea de noi fonduri pentru cercetare și susținerea noilor cercetători.

(C1) UTCN a aplicat pentru mai multe proiecte de cercetare în anul 2020, iar în prezent este în negociere pentru a participa într-un proiect de tip Horizon 2020 în apelul Green Deal.

(C2) UTCN a aplicat ca și partener într-o propunere de tip COST (Cooperation in Science and Technology) cu titlul „Speech Technology in the Age of Security and Privacy Legislation” și coordonată de EURECOM (Franța).

(C3) UTCN are in strategia instituțională obiective ample privind atragerea de fonduri pentru cercetare, respectiv susținerea noilor cercetători prin realizarea unor competiții interne de proiecte de cercetare.

Director Proiect Complex ReTeRom

Acad. Dan Tufis