Raport științific și tehnic

Date de identificare proiect

Număr contract:	PN-III-P1-1.2-PCCDI-2017-0818, Nr. 73 PCCDI/2018
Acronim / titlu:	„ReTeRom/Resurse și tehnologii pentru dezvoltarea interfețelor om-mașină în limba română”
Titlu livrabil:	**Raport științific și tehnic (Etapa a II-a, 2019)**
Termen:	**Noiembrie 2019**
Editor:	**Ioan Dan Tufiș**
Adresa de eMail editor:	**tufis@racai.ro**
Ofițer de proiect:	**Cristian STROE**

Obiectivele etapei a II-a, an 2019

COBILIRO: Soluții de realizare a unui corpus bimodal (vorbire/text) pentru limba română
TEPROLIN: Implementarea modulelor NLP noi
TADARAV: Soluţii de bază pentru adnotarea automată a corpusurilor de vorbire folosind sisteme de RAV existente
SINTERO: Implementarea componentelor pentru modelarea prozodiei și adaptarea la noi vorbitori a vocilor sintetice;”

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

Acest document prezintă o sinteză a realizărilor de natură științifică și tehnică obținute în a doua etapă de implementare a proiectului ReTeRom. Cercetătorii din cele patru instituții participante în proiect au documentat activitățile depuse în cursul anului 2019 în 15 rapoarte tehnico-stiințifice. Toate aceste rapoarte sunt disponibile pe situl proiectului.

În continuare sunt prezentate rezumatele acestor rapoarte.

COBILIRO

A doua etapă (2019) a proiectului CoBiLiRO prevede realizarea infrastructurii pentru gestionarea și stocarea corpusului bimodal. Activitățile complementare prevăzute includ descrierea și implementarea unor soluții de armonizare a reprezentărilor colecțiilor existente text/vorbire (metadata și adnotări). Ulterior, aceste soluții urmează să fie utilizate pentru a armoniza formatele colecțiilor existente și a le încărca pe platforma CoBiLiRO. Similar celorlalte etape, este prevăzută o activitate de diseminare atât la evenimente științifice cât și în mass-media. De interes special pentru platforma dezvoltată este respectarea drepturilor de autor și, atunci când este cazul, anonimizarea contributorilor de resurse vorbite în platformă.

Raportul 2.1 aferent activității ***Realizarea infrastructurii comune de calcul care va găzdui resursele și instrumentele de prelucrare și acces la corpusul bimodal*** a avut drept obiectiv implementarea specificațiilor de realizare a infrastructurii funcționale de upload și acces la resurse, care urmărește îndeaproape descrierea din livrabilul *A1.3 Proiectarea funcțională și arhitecturală a infrastructurii care va găzdui resursele și instrumentele de prelucrare și acces ale consorțiului și realizarea unui prototip*.

Arhitectura platformei COBILIRO. Componentele vorbire-text sunt încărcate în platformă de contributori, în general sub formă de perechi de fișiere, iar metadatele sunt completate manual utilizând interfața on-line. În funcție de formatul de intrare este apelat apoi unul dintre convertoare, rezultatul fiind un fișier care respectă formatul standard CoBiLiRo. Doar asupra componentei textuale, Portalul apelează apoi lanțul de prelucrări textuale TEPROLIN, care întoarce adnotări complexe. În final, cele patru componente, voce+text+metadate+adnotări, sunt stocate în Portal. În raportul extins sunt incluse ilustrarea funcționării de principiu a portalului COBILIRO și detalierea arhitecturii sale (v. raportul extins A2.1). Sunt trecute în revistă tehnologiile folosite la implementare, serviciile de prelucrare lingvistică oferite de platformă. Sunt prezentate funcționalitățile principale: categorii de utilizatori și permisiunile lor, autentificarea accesului, modalitatea de încărcare de fișiere, conversia de format a datelor încarcate. Sunt descrise și principalele funcționalități avute în vedere pentru versiunile viitoare.

Raportul 2.2 corespunzător activității ***Raport asupra soluțiilor de armonizare a reprezentărilor colecțiilor existente text /vorbire (metadate și adnotări)*** descrie soluții de conversie automată a celor trei formate identificate ca fiind utilizate de contributorii actuali la formatul CoBiLiRO, descrise în raportul activității A1.3 (formatul PHS/LAB, formatul MULTEXT/TEI și formatul TEXTGRID). Conversia are ca scop facilitarea mecanismelor de căutare și comparare a resurselor contribuite pe platforma proiectului, precum și facilitarea utilizării formatului propus ca standard pentru resursele aliniate text-voce. Pe viitor, în cazul în care noi formate vor fi identificate, vor putea fi propuse procese de conversie similare. Sunt prezentate soluțiile tehnice pentru conversia automată a formatelor de intrate spre formatul armonizat dar și validările necesare înainte de procesele de conversie. Este analizată și problema resurselor de mari dimensiuni (mai mari de 1 Gb) și se prezintă soluțiile hardware alese pentru operaționalizarea arhitecturii (server HP ProLiant ML350 Gen9, cu un procesor E5-2650 v3 de 2.30GHz, care conține 40 de core-uri, cu framework-ul ASP.NET Core).

Raportul 2.3 descrie ***Realizarea de convertoare de format pentru armonizarea diferitelor reprezentări ale partenerilor la reprezentarea standard agreată în consorțiu.*** Această activitate a avut ca obiectiv implementarea mecanismelor de conversie la formatul CoBiLiRo (descris în raportul activității A1.3) a celor trei formate utilizate de parteneri (descrise în raportul activității A2.2). Aceste convertoare sunt aplicate, în prezent, automat, în momentul adaugării unei resurse pe platforma CoBiLiRo (descrisă în raportul activității A2.1). Platforma suportă și conversia fișierelor din arhive. Cum procesul de conversie nu este fără pierdere de informații, unele formate fiind mai detaliate decât standardul propus în raportul A1.3, fișierele originale sunt păstrate pe server împreună cu fișierele convertite. În cazul unor modificări ulterioare a standardului propus, mecanismele de conversie pot fi adaptate și fișierele originale reconvertite. Pentru fiecare format va fi descris pe scurt procesul complet de conversie, cu exemple. Sunt menționate și posibile deficiențe (pierdere de informații, viteză scăzută), cu posibile soluții.

Principalul motiv pentru care s-a ales framework-ul ASP.NET Core a fost faptul că tehnologia este *open-source, necesită resurse de calcul rezonabile, este rapidă și modulară*. Acest framework este în plină dezvoltare și ne pune la dispoziție o serie de librării și pachete *NuGet* ce ne ajută la dezvoltarea rapidă a platformei.

Unul din beneficiile majore aduse de .NET Core este portabilitatea. Acest lucru permite găzduirea aplicației pe orice sistem de operare.

Un alt motiv pentru care am folosit această tehnologie a fost faptul că oferă o securitate a datelor (protecție pentru atacuri de tip *SQL Injection* și *Cross-site request forgery)* și un mod de a securiza API-urile REST folosind *JSON Web Token.*

Performanța ne este garantată de faptul ca aplicația poate fi scalată pe anumite servicii (de exemplu, putem aloca un număr mai mare de instanțe pentru serviciile de conversie - care vor fi mai costisitoare din punct de vedere al procesării).

Pentru programarea la nivel de client - *client-side* s-a fost folosit framework-ul jQuery. Acesta permite manipularea DOM-ului și accesul la anumite animații și validări specifice.

S-a ales să se folosească ca suport pentru bazele de date serverul MariaDB, pentru că este unul dintre cele mai populare servere de baze de date din lume. Este realizat de dezvoltatorii originali ai MySQL și garantat să rămână open source (gratuit). Deoarece serverul pe care va rula aplicația este cu sistem de operare Linux, acest server virtual pentru baze de date este printre puținele care poate rula sub Linux. Un alt motiv pentru care s-a ales MariaDB este securitatea informației. Sistemul de securitate al informației este foarte bine pus la punct, atât prin prisma sistemului de operare - Linux, cât și din cea a configurării mașinii virtuale pe care va rula MariaDB. Ca soluție pentru încărcarea fișierelor mari am ales sa stabilim o limită de încărcare a acestora de 3Gb. Încărcarea se va putea realiza din linie de comandă, utilizând comanda *scp* sau prin utilizarea unui program de tip transfer fișiere gen: *WinSCP*, *FileZilla*. Dacă utilizatorul va încerca să încarce fișiere mai mari de 3Gb, se recomandă ca soluție arhivarea fișierului în pachete multiple, folosind utilitarul *WinRar* sau *7Z*. Aceste utilitare pot împărți o arhivă în pachete de maxim 500Mb/arhivă, iar procesul de încărcare este optimizat, având avantajul că dacă se întrerupe conexiunea de internet, procesul poate fi reluat de la ultima arhivă încărcată.

Raportul 2.4 prezintă activitatea referitoare la ***Armonizarea colecțiilor existente.*** În această etapă, este prevăzută încărcarea pe platforma CoBiLiRo[1] a resurselor partenerilor din Consorțiul ReTeRom și realizarea concordanței/standardizării colecțiilor existente de corpusuri bimodale (text/vorbire) în funcție de soluțiile de armonizare a reprezentărilor acestora (metadate și adnotări) stabilite de comun acord în cadrul consorțiului.

Resursele existente până în prezent pe platforma CoBiLIRo sunt: CoRoLa-IIT, CoRoLa-IIT_2, CoRoLa_RASC, SWARA, SoRoEs, MARA, RoDigits, Rador_2. Resursele încărcate pe platforma CoBiLiRo respectă formatele agreate de comun acord de aliniere text-vorbire: tipul *file* pentru fișierele audio, însoțite de textele corespunzătoare, cel de-al doilea tip, *start-stop*, care comportă un singur fișier, și tipul *file-start-stop*.

Aceste formate au fost descrise exhaustiv în rapoartele anterioare, în special în Activitatea A2.2: *Raport asupra soluțiilor de armonizare a reprezentărilor colecțiilor existente text /vorbire (metadate și adnotări)* și în Activitatea A2.3: *Realizarea de convertoare de format pentru armonizarea diferitelor reprezentări ale partenerilor la reprezentarea standard agreată în consorțiu* (proiectate și implementate de membrii proiectului CoBiLiRo).

[1] http://85.122.23.18:81/

În ceea ce privește respectarea Legii drepturilor de autor, pentru realizarea înregistrărilor audio-video s-au încheiat acorduri de colaborare și au fost semnate fișele de consimțământ ale subiecților intervievați pentru respectarea confidențialității ca datele înregistrate să fie utilizate strict în scopul cercetării științifice.

De exemplu, în cazul înregistrărilor SoRoEs, fiecărui subiect i s-a atribuit un cod de identificare (pentru respectarea legii dreptului de proprietate), iar etichetarea unui enunț cuprinde 4 simboluri, la care am adăugat și simbolul pentru domeniul limbii române folosit în alte proiecte (AMPER și AMPRom), reprezentat de cifra 9: pentru punctul de anchetă respectiv, codul subiectului în funcție de gen, vârstă și nivelul de studii, codul enunțului – cu un simbol alcătuit din cifre și litere și numărul enunțului înregistrat (1, 2, 3, ...). De exemplu, 9I5c_86a reprezintă enunțul 86, rostit de subiectul de gen feminin, cu vârsta peste 50 de ani, având studii superioare, din localitatea Iași, domeniul limbii române. Aceste informații, astfel codificate pentru necesitățile proiectelor de origine (AMPER și AMPRom), au fost însă explicitate și în metadatele asociate resurselor.

Raportul 2.5 aferent activității Diseminare și participare la manifestări tehnico-științifice, inclusiv în mass-media, prezintă acțiunile de diseminare desfășurate în anul 2 al proiectului complex, după cum urmează: *diseminarea corpusului bimodal, valorizare și utilizare; cel puţin 5 articole la conferinţe indexate ISI și alte 2 la reviste de prestigiu, stagii de practică, actualizarea paginii web a proiectului; promovarea întâlnirilor de proiect în mass-media.*

1. Diseminarea corpusului bimodal, valorizare și utilizare

Pentru asigurarea promovării, vizibilității proiectului și, totodată, a diseminării rezultatelor obținute în cadrul acestuia, s-au actualizat continuu următoarele platforme online:

1.1. Proiectare paginii web ReTeROM

Pagina web pentru un nou corpus folosit în sinteza text-vorbire cu expresivitate: https://speech.utcluj.ro/corpora/mara.html

1.2. Proiectare paginii web de control automat al prozodiei

Pagina web / demonstrator conține prima versiune a modulului de control automat al prozodiei: https://speech.utcluj.ro/sintero/prosody_examples/

1.3. Proiectare paginii web demonstrator de sinteză text vorbire

Pagina web pentru demonstrarea sintezei folosește rețele DNN de tip Tacotron-GST: https://speech.utcluj.ro/sintero/dnn-samples/

1.4. Proiectare paginii web demonstrator de modelare a prozodiei și adaptare la noi vorbitori

Pagina web conține mostre audio privind adaptarea sistemelor de sinteză text-vorbire la vorbitor și la stilul de vorbire, folosind Tacotron – GST: https://speech.utcluj.ro/sintero/spkadapt_prosody_examples/

2. Evenimente organizate

(a) 18 noiembrie 2019 - *ReTeRom Workshop 3* - în cadrul celei de-a 14-a ediții a Conferinței Internaționale "Linguistic resources and tools for processing natural language" - https://profs.info.uaic.ro/~consilr/. Workshopul a fost găzduit de Universitatea Tehnică Cluj-Napoca și a fost promovat inclusiv pe site-ul Universității Babeș-Bolyai din Cluj-Napoca, https://news.ubbcluj.ro/cea-de-a-xiv-a-editie-a-conferintei-internationale-resurse-lingvistice-si-instrumente-pentru-prelucrarea-limbii-romane-organizata-la-ubb/.

În cadrul acestui atelier de lucru au fost prezentate rezultatele pentru cele 4 subproiecte componente (CoBiLiRO, TEPROLIN, TADARAV și SINTERO). Discuțiile s-au axat pe 6 direcții tematice, în care membrii proiectului au interacționat prezentând probleme de mare interes curent, discutând chestiuni care privesc evoluția proiectului și propunând soluții: (1) Integrare unelte software;

(2) Copyright și GDPR date audio și text; (3) Sinteza cu expresivitate; (4) Procesare text (lexicoane, noi algoritmi); (5) Alinieri audio și text (metode, demonstratoare); (6) Noi tendințe în domeniu (deep learning, text, audio, supervised/unsupervised).

Lista care urmează conține conferințe prezentate de membri ai consorțiului ReTeRom pe subiecte apropiate de tematicile proiectului:

(a) Dan Tufiș: *Resources and Technologies for Speech and Language Processing of Romanian*, conferință invitată la cea de a 10-a Conferință *Speech Technology and Human-Computer Dialogue*, 10-12 Oct 2019, Timișoara.
(b) 2 participări (Beata Lorincz și Maria Nuțu) la ***Eastern European Machine Learning Summer School* (EEML)**, 1-6 iulie 2019, București, România - https://www.eeml.eu/.
(c) Dan Cristea: *The COBILIRO Project: Building and Distributing a Bimodal Corpus for Romanian Language*, în cadrul celei de a 14-a Conferințe Internaționale *Linguistic Resources and Tools for Natural Language Processing*, ConsILR-2019, 19 Nov. 2019, Cluj-Napoca.

3. Publicații care menționează proiectul ReTeRom

(1) Pistol, Ionuţ Cristian, and Andrei Arusoaie. "AIM: Designing a language for AI models". In: Procedia Computer Science 159 (2019): 202-211 https://www.sciencedirect.com/science/article/pii/S1877050919313547
(2) Dan Cristea, Cristian Pădurariu, Șerban Boghiu, Daniela Gîfu, Mihaela Onofrei, Diana Trandabăț, Ionuț Cristian Pistol, Anca-Diana Anca Bibiri, Andrei Scutelnicu. "The COBILIRO Project: Building and Distributing a Bimodal Corpus for Romanian Language". In: Proceedings of the 14th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2019, 18-20 Nov. 2019, Cluj-Napoca, Romania, Mihaela Onofrei et al.(eds.), "Alexandru Ioan Cuza" University Publishing House, Iaşi, 2019, pp. 13-23.
(3) Vasile Păiș, Dan Tufiș, Radu Ion. "Integration of Romanian NLP tools into the RELATE platform". In Proceedings of the 14th International Conference on Linguistic Resources and Tools for Natural Language Processing (ConsILR 2019), Mihaela Onofrei *et al.* (eds.), "Alexandru Ioan Cuza" University Publishing House, Iaşi, 2019, pp. 181-192.
(4) Mihaela Onofrei, Anca-Diana Bibiri, Constantin Dragoș Nicolae, Dan Tufiș, Dan Cristea. Proceedings of the 14th International Conference "Linguistic Resources and Tools for processing of the Romanian language", Cluj-Napoca, 2019, ISSN: 1843-911X.
(5) Daniela Gîfu, Alex Moruz, Cecilia Bolea, Anca Bibiri, Maria Mitrofan. "The Methodology of Building CoRoLa". In: Revue Roumaine de Linquistique (Romanian Review of Linguistics), Vol. 64, 2019, Publishing House of the Romanian Academy, ISSN: 0035-3957, indexed by ISI Web of Science, (impact factor = 0.189/2018) - http://www.lingv.ro/index.php?option=com_content&view=article&id=342%3Arrl-arhiva-2019&catid=36%3Areviste-ilb&Itemid=1
(6) Alexandru-Lucian Georgescu, Horia Cucu, Corneliu Burileanu, "Progress On Automatic Annotation Of Speech Corpora Using Complementary ASR Systems", in the Proceedings of the 42nd International Conference on Telecommunications and Signal Processing (TSP), 2019, Budapest, Hungary.
(7) Gheorghe Pop, Șerban Mihalache, Dragoș Burileanu, "Forensic Recognition of Narrowband AMR Signals", in Proceedings of the 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timișoara, Romania, 2019.
(8) Alexandru-Lucian Georgescu, Horia Cucu, Corneliu Burileanu, "Kaldi-based DNN architectures for speech recognition in Romanian", in the Proceedings of the 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timișoara, Romania, 2019.
(9) Gheorghe Pop and Dragoș Burileanu. "Speech Enhancement for Forensic Purposes", in UPB Scientific Bulletin, Series C, Vol. 81, Issue 3, 2019, pp. 41-52.

(10) Florin Iordache, Alexandru-Lucian Iordache, Dan Oneaţă, Horia Cucu. "Romanian Automatic Diacritics Restoration Challenge", in the Proceedings of the 14th International Conference on Linguistics Resources and Tools for Natural Language Processing, Cluj-Napoca, Romania, 2019.
(11) Adriana Stan. "Input Encoding for Sequence-to-Sequence Learning of Romanian Grapheme-to-Phoneme Conversion", in Proceedings of the 10th IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timișoara, Romania, 2019.
(12) Beáta Lőrincz, Maria Nuțu, Adriana Stan. "Romanian Part of Speech Tagging using LSTM Networks", in Proceedings of the IEEE 15th International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, 2019.
(13) Maria Nuțu, Beáta Lőrincz, Adriana Stan. "Deep Learning for Automatic Diacritics Restoration in Romanian", in Proceedings of the IEEE 15th International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, 2019.
(14) David A. Braude, Matthew P. Aylett, Caoimhin Laoide-Kemp, Simone Ashby, Kristen M. Scott, Brian O Raghallaigh, Anna Braudo, Alex Brouwer, Adriana Stan. "All Together Now: The Living Audio Dataset", in Proceedings of Interspeech, Graz, Austria, 2019.
(15) Verginica Barbu Mititelu, Mihaela Cristescu, Mihaela Onofrei. "The Romanian Corpus Annotated with Verbal Multiword Expression". In: Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), August 2019, Florence, Italy, ACL, pp. 13-21.

4. Întreținerea paginii web a proiectului complex ReTeROM

Pentru asigurarea promovării, vizibilității proiectului și diseminarea rezultatelor obținute în cadrul acestuia s-a actualizat continuu platforma online (http://dev.racai.ro/ReTeRom/), compusă dintr-un "landing zone" și patru secțiuni: "DESCRIERE", "RAPOARTE", "ECHIPA" și "CONTACT".

Meniul este unic și poate fi accesat din orice locație a paginii, întregul conținut fiind cuprins în aceste legături. Regăsim aici cele 4 subproiecte componente: CoBiLiRo (Corpus bimodal pentru limba română adnotat pe multiple niveluri); TEPROLIN (Tehnologii pentru procesarea limbajului natural – text); TADARAV (Tehnologii pentru adnotarea automată a datelor audio și pentru realizarea interfețelor de recunoaștere automată a vorbirii); SINTERO (Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate).

5. Comunicate de presă

Confluențe literare, ediția nr. 3244, Anul IX, 18 noiembrie 2019.
https://confluente.org/daniela_gifu_1574084975.html.

6. Concluzii

Livrabilele proiectului component COBILIRO demonstrează că obiectivele asumate prin contract au fost realizate integral.

TEPROLIN

Etapa a doua a proiectului component TEPROLIN a avut ca obiective:

a) Raport asupra implementării modulelor NLP noi;

b) Raport asupra lexiconului îmbogățit cu transcriere fonetică;

c) Raport asupra implementării prototipului de platformă integrată și configurabilă, asupra testării, evaluării și validării prototipului;

d) Raport asupra prelucrării părții textuale a corpusului bimodal colectat în proiectul 1;

În continuare sunt prezentate rezumatele prezentării realizării acestor obiective.

Platforma de prelucrare a textelor românești TEPROLIN, dezvoltată în proiectul ReTeRom, a fost îmbogățită cu două noi module:

1. Un modul de recunoaștere a entităților denumite (engl. „Named Entity Recognition" sau, pe scurt, NER) care recunoaște denumiri de locuri (localități de diverse mărimi, țări, etc.), nume de persoane (femei și bărbați, prefixate sau nu de formule de adresare cum ar fi „d-na", „dr. ing.", etc.) și nume de organizații (instituții cu ar fi de exemplu „Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu"");
2. Un modul de recunoaștere a terminologiei medicale (domeniul bio-medical), bazat pe o versiune anterioară a aplicației NLP-Cube, antrenat pe corpusul românesc MoNERo, care recunoaște părți anatomice („artera iliacă"), substanțe chimice, boli („diabet") și proceduri medicale („abordul arterei iliace").

În **raportul 2.6** este descrisă integrarea celor două module, inclusiv codul auxiliar pentru asigurarea interoperabilității cu celelalte module din fluxul TEPROLIN.
TEPROLIN se poate acum autoconfigura să detecteze ce algoritmi să ruleze pentru a satisface cererea utilizatorului. De exemplu, modulul NER are nevoie de un anumit algoritm pentru segmentare lexicală și adnotarea cu etichete morfosintactice (TTL) și nu funcționează cu un alt algoritm care îndeplinește aceleași funcții. Modulul NER așteaptă fraza de procesat într-un anumit format, lucru realizat de metoda `def_prepareSentences(self, dto)`. Astfel, fiecare frază este reprezentată pe un format cu 5 coloane, începând cu simbolul de start de frază `<s>` , sfârșind cu simbolul de sfârșit de frază `</s>` și, pe fiecare linie, separate de caracterul TAB, avem forma ocurentă, lema, eticheta morfo-sintactică MSD, primele 2 caractere din această etichetă și eticheta morfosintactică redusă. Fraza astfel procesată este trimisă serviciului web NER care adaugă etichetele de entități pentru fiecare unitate lexicală a frazei, de asemenea în format tabular cu 2 coloane: prima conține forma ocurentă a cuvântului iar cea de-a doua conține eticheta entității sau „O" dacă cuvântul nu face parte dintr-un nume de entitate.

Modulul de recunoaștere a terminologiei bio-medicale se bazează pe o versiune anterioară a aplicației NLP-Cube, versiune care a fost încorporată în TEPROLIN, în directorul „bioner" din rădăcina arborelui de surse. Cu ajutorul metodei def createApp(self), încărcăm toate resursele necesare (word embeddings, modele de NER, etc.) o singură dată la pornirea platformei. Similar cu modulul de NER, avem o metodă care formatează fraza pentru NLP-Cube, def_prepareSentences(self, dto) și, cu metoda def _runApp(self, dto, opNotDone) , rulăm modulul de recunoaștere a terminologiei bio-medicale și adăugăm adnotările obiectului de tip DTO (engl. „Data Transfer Object") care conține toate prelucrările platformei și care este primit la intrare și modificat de toate metodele platformei.
De asemenea, în această fază au fost realizate o serie de optimizări și extensii ale modulelor realizate în etapa anterioară (de ex. Inserția de diacritice se face doar dacă nu există deloc, s-a eliminat din TTL execuția fork() care putea eșua, NLP-cube care rula pe un singur fir de execuție acum poate rula in regim multi-thread, etc).

Platforma de prelucrare a textelor românești TEPROLIN îmbogățită cu cele două noi module de recunoaștere a diferitelor tipuri de entități este publică. Codul sursă este stocat pe GitLab, la adresa https://gitlab.com/raduion/teprolin . Accesul este moderat, cererea se poate trimite la adresa de email radu@racai.ro.

Raportul în extenso cu nr. **2.7** aferent activității ”Transcrierea fonetică a cuvintelor din lexiconul validat” descrie îmbogățirea lexiconului (creat în etapa anterioară) cu intrări suplimentare, cu transcrieri fonetice precum și validarea transcrierilor fonetice și a silabației pentru toate intrările. Aceste operații, în mare parte implicând un expert uman, au fost mari consumatoare te timp.

Metodologia pentru orectarea erorilor de transcriere fonetică. Pornind de la premisa că resursele utilizate (RoSyllabiDict și MaRePhor) au fost, așa cum susțin autorii lor, validate manual parțial înainte de lansare, ne-am concentrat pe validarea și corectarea acelor intrări din lexiconul ReTeRom care conțin informație generată automat cu Romanian TTS. Au fost implementate[2] reguli de corectură automată a transcrierii fonetice bazate pe silabificare și accent, care presupun că aceste informații sunt corecte (în practică, a existat o etapă anterioară de validare manuală a lor). S-au efectuat corecturi/validări manuale în cazuri de transcriere fonetică pe care le-am identificat ca ambigue/problematice.

Alfabetul fonetic SAMPA. Alfabetul fonetic utilizat în MaRePhor și preluat în lexiconul nostru este SAMPA. Legat de sunetele ce/ci/ge/gi/che/chi/ghe/ghi, se poate observa ilustrată în tabel următoarea convenție de transcriere: Atunci când în aceeași silabă cu aceste grupuri există o vocală, aceste grupuri trebuie transcrise ca reprezentând un singur fonem (vezi transcrierile pentru “ceas”, “ciută”, „chiag”, „chiar”, “geană”, „cafegioaică”, „lighean”, „ghiotură”). Dacă în silabă mai există o semivocală, atunci aceste grupuri se transcriu ca reprezentând două foneme.

Ca rezultat al acestei activități a fost creată o resursă lingvistică extrem de valoroasă cu 346.074 intrări cuprinzând informație standardizată asociată unei forme ocurență: lema (forma de dicționar a cuvântului), eticheta morfo-sintactică în format MSD, împărțirea în silabe a formei ocurență, marcarea accentului (printr-un apostrof) în fața vocalei accentuate și transcrierea fonetică a formei ocurență, între paranteze drepte.

Raportul în extenso **2.8** aferent activității ”Implementarea prototipului de platformă integrată și configurabilă, testarea, evaluarea și validarea prototipului” descrie dezvoltările realizate pentru a permite platformei TEPROLIN să prelucreze volume mari de texte prin paralelizarea lanțurilor de prelucrare Acest lucru a condus la implementarea unui mecanism de control al procesărilor sub formă de ”job”-uri care pot fi apoi distribuite la nivelul unui server pe mai multe procese sau la nivelul unei rețele de calcul pe mai multe noduri. Noua platformă, mai complexă (numită RELATE) a incorporat complet funcționalitatea anterioară a TEPROLIN, inclusiv noile module prezentate în raportul 2.6, dar a adăugat mecanisme suplimentare de control al volumelor mari de date și ale prelucrărilor acestora:

- Încărcare fișiere text: direct sau sub formă de arhive
- Posibilitatea configurării resurselor de calcul disponibile în vederea paralelizării fluxurilor de prelucrare, cu definirea adreselor (”endpoint”-urilor) de conectare la diferitele servicii de adnotare
- Procesare paralelă cu utilizarea resurselor disponibile la nivelul unui server sau la nivelul unei rețele de calcul
- Vizualizare rezultate procesare
- Descărcare fișiere originale și prelucrate: atât direct, fișier cu fișier, cât și sub forma unei arhive

Noua platformă are dublă funcționalitate:

- prima permite accesul nerestricționat la adnotări pe texte scurte și interacțiune cu diferitele unelte disponibile în acest context – interfața web publică

[2] scripturile de implementare a regulilor au fost scrise în C#, sub platforma Microsoft Visual Studio 2019

- cea de a doua, necesită autentificare pe bază de nume de utilizator și parolă și permite acces la configurarea și controlul fluxului paralel de adnotare, importarea unui corpus de mari dimensiuni, exportarea rezultatelor – interfața web privată.

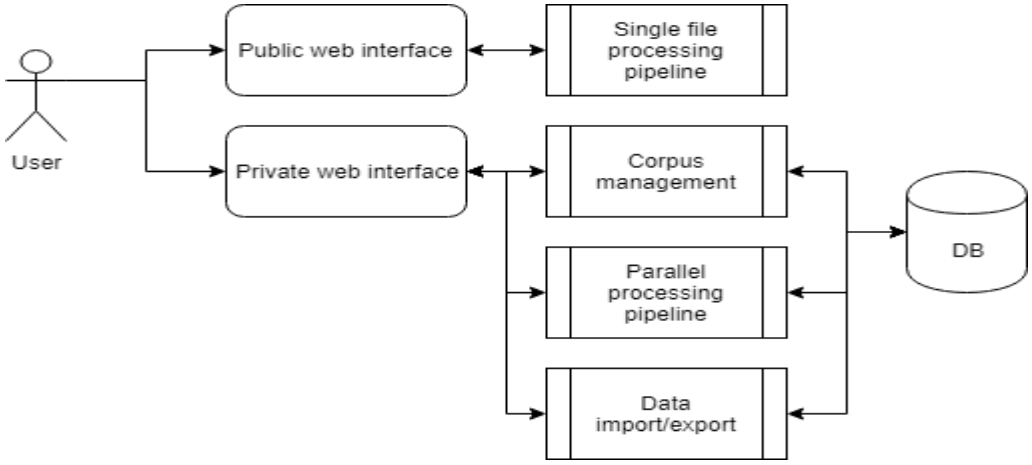

Figura 1. Diagrama generală a platformei

Partea privată a platformei, dedicată procesării unui corpus de mari dimensiuni, utilizează un mecanism pe bază de ”task”-uri pentru a realiza prelucrările. Această structură este prezentată în Figura 2.

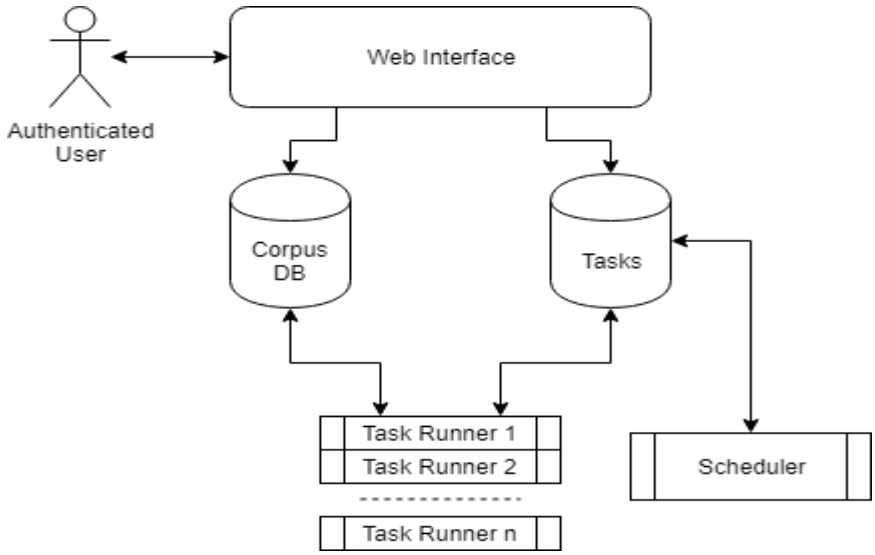

Figura 2. Mecanismul de paralelizare la nivelul platformei

Prin intermediul interfeței web, utilizatorul începe prin încărcarea fișierelor aferente unui corpus. Ulterior, are acces la interfața de management a task-urilor prin care poate lansa în execuție un task de adnotare. Lansarea în execuție este gestionată de un proces de tip ”Scheduler” care permite alocarea diferitelor fișiere către diferite procese de execuție (”Task Runner”) disponibile pe serverul curent sau în rețea, în funcție de resursele configurate la nivelul platformei.

Pe timpul execuției unui task, utilizatorul poate urmări starea acestuia, iar apoi poate accesa fișierele adnotate pe măsură ce acestea devin disponibile. Un task poate avea următoarele stări:

- NEW: task-ul a fost creat și așteaptă să fie preluat de procesul de alocare (”scheduling”)
- SCHEDULING: task-ul este în curs de alocare
- SCHEDULED: task-ul a fost alocat către procesele de execuție disponibile
- RUNNING: task-ul este în curs de execuție de către unul sau mai multe procese

- DONE: task-ul a fost finalizat cu succes
- ERROR: a intervenit o eroare in timpul execuției care a provocat oprirea permanentă a task-ului.

Pentru a permite lucrul cu volume mari de date, toate operațiunile în platformă, inclusiv dezarhivarea fișierelor la upload, arhivarea pentru download, crearea de statistici, sunt executate pe sistemul de task-uri.

Pentru descărcarea unui volum mare de date, platforma oferă posibilitatea realizării unei arhive, fie cu datele neadnotate fie cu cele rezultate în urma procesului de adnotare. Arhivele generate pot fi apoi vizualizare și descărcate.

Platforma a fost testată și validată prin încărcarea de diverse volume de date de la corpusuri mici conținute într-un singur fișier, la un corpus mare distribuit în peste 100,000 de fișiere. S-au putut efectua astfel prelucrări de volume masive, folosind 13 procese distribuite pe infrastructura ICIA. De asemenea, a fost realizată o prezentare a platformei în cadrul Workshop-ului RETEROM organizat la Cluj pe 18 Noiembrie 2019. În cadrul acestei prezentări au fost expuse funcționalitățile disponibile către toți membri proiectului, fiind adunate observații în vederea unor îmbunătățiri ulterioare. Acestea nu au fost orientate spre probleme de bază, confirmând încă o dată variantele alese pentru implementare.

În **raportul 2.9**, aferent activității ”Prelucrarea părții textuale a corpusului bimodal colectat în proiectul 1. Validarea și corectarea (dacă este necesară) a erorilor de prelucrare” se prezintă rezultatele obținute. Această activitate a proiectului TEPROLIN, are rolul de a asigura prelucrarea transcrierilor corespunzătoare înregistrărilor audio colectate în cadrul proiectului COBILIRO. Prelucrarea se face cu un lanț de procesare care presupune 7 tipuri de operații, dezvoltat în cadrul proiectului TEPROLIN.

Activitatea de prelucrare a transcrierilor presupune, în contextul proiectului TEPROLIN, mai multe etape de procesare succesivă:

- segmentare la nivel de propoziție (atunci când este necesar);
- segmentare la nivel de cuvânt: separarea cuvintelor și punctuației din text ca unități de segmentare distincte (sau tokeni);
- introducerea diacriticelor, atunci când lipsesc;
- adnotare morfo-sintactică a cuvintelor;
- lematizarea sau identificarea formei de dicționar a cuvintelor;
- adnotarea sintactică a cuvintelor;
- detectarea entităților denumite.

Aceste procesări au fost efectuate cu lanțul de prelucrare TEPROLIN, parte din platforma RELATE prezentată în raportul asupra Activității 2.8. Corpusurile au fost încărcate în platforma COBILIRO, care integrează lanțul TEPROLIN și îl execută de fiecare dată când o resursă nouă este introdusă și prelucrarea cu TEPROLIN este bifată în interfață. În continuare raportului 2.9 sunt prezentate informații cu privire la numărul de fișiere, propoziții, leme și forme unice, verbe, adjective, adverbe și substantive, corespunzător fiecărui corpus adnotat în cadrul acestei activități.

Corpus	**Nr. fișiere/prop**	**Nr. cuvinte**	**Verbe**	**Adverbe**	**Adjective**	**Substantive**	**Nr. leme unice**	**Nr. forme unice**
IIT	7001/7001	146633	23590	10866	6743	30710	13853	23603
RACAI	3827/3827	87433	14103	3655	4989	27015	17452	26474
RADOR	12551/12551	396804	62286	24554	21085	103692	21642	39394
RASC	2177/2177	30520	4909	1375	2839	9044	6186	9816
SWARA	17/19457	257684	43690	11984	10234	55115	4023	5529
Adevărul	1871/30391	694681	109689	33229	32875	196299	34061	51033

Mara	1/5573	116387	23028	10878	3772	17573	4592	7486
RSC	1/139526	589407	105589	35195	42317	174245	12267	19084
SSC-eval	1/3035	35947	5029	2438	2139	9096	4603	7290
SSC-train	1/53898	241006	37289	18163	13307	64814	8452	16829
SSC-train2	1/170292	983844	122525	63291	56526	256677	12851	26743
SOROES	1/166	896	187	67	13	169	175	223
TOTAL	27450/447894	3581242	551914	215695	196839	944449	127306	233504

Rezultatele acestei faze au fost diseminate prin comunicări la manifestări tehnico-științifice:

(1) Dan Tufiș. 2019. Language Technology and Digital Culture. invited talk at European Conference on Exposing Online the European Cultural Heritage: the Impact of Cultural Heritage on the Transformation of the Society, Iași, 17-18 aprilie, 2019
(2) Dan Tufiș. 2019. Language technologies and the challenges for Digital Single Market. invited pannelist at IMPACT 2019, 21-22 May, Kracow, Poland
(3) Dan Tufiș, Resources and Technologies for Speech and Language Processing of Romanian, invited talk at the 10th Conference on Speech Technology and Human-Computer Dialogue, 10-12 Oct 2019, Timișoara.
(4) Dan Tufiș. 2019. "Inteligența artificială și provocările ingineriei lingvistice". Plenary talk to the international workshop BACStud 2019 (5th edition), 17-19 October, 2019
(5) Verginica Barbu Mititelu, Mihaela Cristescu, Mihaela Onofrei, The Romanian Corpus Annotated with Verbal Multiword Expression. In: Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019, August 2019, Florence, Italy, ACL, pp. 13-21.
(6) Vasile Păiș, Dan Tufiș, Radu Ion, "Integration of Romanian NLP tools into the RELATE platform". In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language – CONSILR 2019, pages 181-192

Pentru asigurarea promovării, vizibilității proiectului și, totodată a diseminării rezultatelor obținute în cadrul acestuia s-a actualizat continuu platforma online - http://dev.racai.ro/ReTeRom/.

Prin activitățile și rezultatele obținute toate obiectivele fazei au fost integral realizate. Instrumentele și resursele create în această etapă sunt publice, iar codul sursa este stocat pe GitLab, la adresa https://gitlab.com/raduion/teprolin.

Locurile de muncă susținute prin program, inclusiv resursa umană nou angajată sunt următoarele: cercetători din personalul permanent: Dan Tufiș, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan (Carp), Radu Ion și Eric Curea. Doctorandul, nou angajat, Vasile Păiș, s-a integrat perfect colectivului implicat in proiect (6 cercetători), contribuind substanțial la atingerea și chiar depășirea obiectivelor institutului.

Prezentarea structurii ofertei de servicii de cercetare și tehnologice este difuzată prin intermediul https://erris.gov.ro/RACAI-ICIA, a sitului nostru de tehnologii și resurse lingvistice pentru limba română http://relate.racai.ro și a GitLab-ului https://gitlab.com/raduion/teprolin.

Valorificarea resurselor existente (cecuri) nu s-a realizat: sumele alocate prin CEC-uri nu au putut fi cheltuite în primul rând pentru ca nu s-au identificat situații concrete în care se putea utiliza suma respectivă.

TADARAV

A doua etapă a proiectului TADARAV a avut trei obiective principale:

1. evaluarea posibilităţii utilizării transcrierilor aproximative ale materialelor ce conţin vorbire, împreună cu un sistem de recunoaştere automată a vorbirii (RAV) inițial, pentru a produce în mod automat transcrieri precise pentru o parte a semnalului de vorbire;
2. evaluarea posibilităţii utilizării scorurilor de încredere generate de un sistem de RAV inițial pentru a produce în mod automat transcrieri precise pentru o parte a semnalului de vorbire;
3. îmbunătățirea soluţiei de adnotare automată a semnalului de vorbire utilizând sisteme de RAV complementare.

Toate metodele de adnotare automată au fost evaluate și în contextul reantrenării sistemului de RAV inițial cu datele nou generate. Cele trei obiective au fost realizate în proporţie de 100%, în urma activităţilor întreprinse rezultând toate livrabilele asumate de consorţiu la începutul acestei etape.

Concret, în urma activităților A2.11, A2.12 și A2.13 din etapa 2/2019 a proiectului TADARAV, au rezultat următoarele livrabile:

- Soluţie de bază de filtrare şi aliniere a transcrierilor aproximative cu semnalul de vorbire (TRL3), funcțională
- Soluţie de bază pentru generarea de scoruri de încredere pentru RAV (TRL3), funcțională
- Soluţie îmbunătăţită de adnotare automată a semnalului de vorbire utilizând sisteme de RAV complementare (TRL4), funcțională

Diseminarea rezultatelor proiectului a fost realizată: în cadrul consorţiului în cele workshopul organizat la Cluj-Napoca pe 18 noiembrie 2019 şi în comunitatea ştiinţifică la trei conferințe internaționale de prestigiu. Suplimentar, unele dintre rezultate au fost publicate într-un articol în Buletinul Științific al Universității Politehnica din București. De asemenea, progresul înregistrat în această etapă a fost diseminat prin intermediul website-ului proiectului: https://tadarav.speed.pub.ro.

1 Descrierea ştiinţifică şi tehnică a activităţilor

1.1 Introducere

Modelele acustice bazate pe reţele neurale profunde (*Deep Neural Network* – DNN) obţin performanţe direct proporţionale cu cantitatea de date folosite la antrenarea reţelei. Prin urmare, dat fiind faptul că adnotarea manuală a resurselor audio presupune o investiţie consistentă de efort şi timp, interesul faţă de tehnicile de adnotare automată a vorbirii a crescut semnificativ. Adnotarea automată a vorbirii presupune colectarea de vorbire în format brut şi folosirea unei metode automate pentru a produce transcrieri cât mai precise pentru cel puţin o parte din corpusul iniţial.

1.1.1 Seturi de date

Activităţile A2.11, A2.12 şi A2.13 presupun (i) utilizarea unor seturi de date de vorbire deja existente pentru antrenarea și evaluarea unor sisteme de RAV necesare în aplicarea metodelor de adnotare automată și (ii) utilizarea unor seturi de date brute, neadnotate sau adnotate parţial ca date de intrare pentru cele trei metode de adnotare automată. Aceste seturi de date sunt rezumate în tabelele 2.1.a și 2.1.b.

Seturi de date de vorbire adnotată

Pentru antrenarea şi evaluarea sistemelor de RAV, au fost folosite două seturi de date de vorbire în limba română: *Read Speech Corpus* (RSC), ce conţine vorbire citită, colectată în condiţii de laborator, fără zgomot de fundal şi *Spontaneous Speech Corpus* (SSC), ce conţine vorbire continuă, spontană, preluată de la posturi de radio şi TV, uneori afectată de zgomot. Ambele corpusuri cuprind fişiere audio

şi transcrieri corespunzătoare şi sunt divizate în seturi de antrenare şi seturi de evaluare. RSC-train este setul de antrenare din RSC, ce conţine 100 ore de vorbire citită, cuvinte izolate sau fraze de la 157 de vorbitori diferiţi. RSC-eval este setul de evaluare din RSC; acesta conţine vorbire de la 22 de vorbitori diferiţi, însumând 5.5 ore de vorbire. SSC-train este setul de antrenare din SSC şi conţine 130 ore de vorbire spontană, majoritatea din emisiuni de ştiri şi *talkshow*-uri. SSC-eval este setul de evaluare din SSC şi însumează 3.5 ore de vorbire.

În etapa anterioară a proiectului, ca parte a activităţii A1.13 au fost obţinute seturile de date de vorbire adnotată SSC-train3-compl și SSC-train4-compl. Ele sunt prezentate, alături de seturile de vorbire adnotată RSC și SSC în Tabelul 2.1.a.

Tabelul 2.1.a Seturile de vorbire adnotată folosite pentru antrenarea și evaluarea sistemelor de RAV și seturile de vorbire adnotată obținute în etapa anterioară (1/2018)

Corpus	Set	Durată	
Antrenare	RSC-train	94h, 46m	225h, 30m
	SSC-train	130h, 44m	
Evaluare	RSC-eval	5h, 29m	8h, 58m
	SSC-eval	3h, 29m	
SSC-train3-compl-2018	radio #1	6h, 20m	49h, 13m
	TV #1	10h, 00m	
	TV #2	32h, 53m	
SSC-train4-compl-2018	radio #1	25h, 16m	280h, 00m
	TV #1	66h, 02m	
	TV #2	188h, 42m	

Seturi de date brute

Seturile de date brute, neadnotate sau adnotate parţial, utilizate ca date de intrare pentru cele trei metode de adnotare automată sunt denumite SSC-train3-raw și SSC-train4-raw și sunt prezentate în Tabelul 2.1.b. Primul set de date neadnotat, SSC-train3-raw, a fost achiziţionat din mass-media românească, mai exact de pe 2 website-uri de ştiri şi un post de radio, de-a lungul unei perioade de o lună calendaristică. Al doilea set de date de vorbire neadnotată, SSC-train4-raw, a fost achiziţionat de asemenea din cele 3 surse din mass-media românească, de-a lungul unei perioade de nouă luni calendaristice.

Seturile de date au fost achiziţionate cu o aplicaţie creată în cadrul proiectului, aplicaţie ce parcurge *feed*-urile RSS al acestor *website*-uri, identifică ştirile noi și descarcă fişierele audio (eşantionate la 16 kHz, 16 biţi pe eşantion) și textele corespunzătoare ştirilor respective.

Tabelul 2.1.b Seturi de date de vorbire neadnotată (+ transcrieri aproximative) utilizate ca date de intrare pentru cele trei metode de adnotare automată. Numărul de cuvinte se referă la textul brut descărcat de pe fiecare website în parte

	SSC-train3-raw	SSC-train4-raw

	# cuvinte	# ore	Nr. de cuvinte	Nr. de ore
radio #1	30.049	19,3	120.121	78,2
TV #1	357.926	51,5	2.241.389	331,6
TV #2	825.722	65,9	4.111.690	367,3
Total	1.213.697	136,7	6.473.200	777,2

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

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

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

Corpus	Sursa	Durată [# ore]		Eficienţă aliniere [% ore]	
SSC-train3-compl-2019	radio #1	12h, 10m	96h, 38m	63.1%	70.6%
	TV #1	20h, 05m		39.2%	
	TV #2	64h, 23m		98,4%	
SSC-train4-compl-2019	radio #1	50h, 20m	535h, 53m	64.1%	68.8%
	TV #1	125h, 12m		37.8%	
	TV #2	360h, 21m		98.1%	
SSC-train3-trans-v3	radio #1	1,0	37,5	5,0%	27,4%
	TV #1	12,8		25,0%	
	TV #2	23,6		35,9%	
SSC-train4-trans-v3	radio #1	2,7	228,8	3,5%	29,4%
	TV #1	87,9		26,5%	
	TV #2	138,1		37,6%	
SSC-train3-conf	radio #1	5h, 30m	55h, 51m	28,5%	44,1%
	TV #1	19h, 07m		37,1%	
	TV #2	31h, 13m		47,4%	
SSC-train4-conf	radio #1	22h 03m	315h, 34m	28,2%	40,6%
	TV #1	124h 29m		37,5%	
	TV #2	169h 01m		46,0%	

1.2 Activitatea 2.11 - Proiectarea şi implementarea unei soluţii de bază de filtrare și aliniere a transcrierilor aproximative cu semnalul de vorbire

Activitatea A2.11 a avut ca scop evaluarea unei metode de generare de seturi de date de vorbire adnotată folosind materiale audio disponibile pe diverse website-uri de mass-media împreună cu ştirile text de pe paginile respective. Folosind un singur sistem RAV s-a generat un set de transcrieri aproximative, urmând ca apoi acestea să fie aliniate cu transcrierile de pe site. Părțile identice dintre cele 2 seturi de transcrieri au fost considerate ca fiind corecte. Motivul pentru care nu s-au folosit doar transcrierile de pe site este că acestea nu conțin întotdeauna textul vorbit din fișierul audio; există cazuri în care lipsesc părți din vorbire precum și cazuri în care apar informații adiționale în transcriere. Astfel, prin efectuarea alinierii dintre transcrierile de pe site și cele obținute cu sistemul RAV sperăm să obținem secvențe audio-text cât mai precise.

1.2.1 Descrierea metodei

Metoda utilizată în această etapă are ca scop obţinerea într-un mod automat, nesupervizat, a unei adnotări cât mai precise pentru un corpus de vorbire. Corpusul nou obţinut s-a dorit a fi utilizat pentru antrenarea sistemelor de RAV existente, crescând astfel variabilitatea acustică a modelelor, îmbunătăţind implicit şi acurateţea transcrierilor. Paşii corespunzători metodei vor fi descrişi în continuare.

Ideea principală a acestei metode de adnotare constă în utilizarea unui singur sistem RAV pentru a produce transcrieri pentru un corpus neadnotat, dar pentru care există transcrieri aproximative. În urma alinierii transcrierilor RAV cu transcrierile aproximative, vor fi selectate ca fiind corecte părțile identice dintre cele 2 seturi de transcrieri. În final, transcrierile selectate şi segmentele de vorbire corespunzătoare sunt folosite pentru a forma un nou corpus adnotat de vorbire.

Resursele brute. Resursele brute utilizate în această metodă au fost prezentate în tabelul 2.1.b. Materialele brute conţin (i) vorbire neadnotată (audio) și (ii) transcrieri aproximative (text).

Transcrierea vorbirii neadnotate. Sistemul RAV folosit pentru transcrierea materialelor audio este sistemul HMM-DNN creat în activitatea A1.13 din etapa 1/2018. Mai multe informaţii tehnice despre acest sistem pot fi obţinute consultând raportul etapei 1/2018. Transcrierile obținute în urma folosirii acestui sistem RAV conţin doar din litere mici, nu conţin semne de punctuaţie sau cifre, iar cuvintele sunt însoțite de ştampile de timp (timpul de început al rostirii cuvântului și timpul de încheiere al rostirii cuvântului). Iată, cu titlul de exemplu, o astfel de transcriere RAV: bărbatul(3.71,4.14) de(4.14,4.25) treizeci(4.25,4.55) şi(4.55,4.65) şase(4.65,4.93) de(4.93,5.05) ani(5.05,5.19) povesteşte(5.19,5.66) că(5.66,5.75) muncise(5.75,6.22) toată(6.22,6.62) noaptea(6.62,6.99).

Preprocesarea transcrierilor brute. Materialele de pe site (vorbire și transcrieri aproximative) provin din *mass-media* (emisiuni, ştiri, interviuri, reportaje) și reprezintă o foarte bogată sursă de vorbire și text. Însă, transcrierile brute au o formă diferită față de transcrierile RAV, în sensul că acestea conțin litere mari, numere scrise cu cifre, abrevieri etc. Pentru a efectua procesul de aliniere, acestea trebuie aduse la o formă cât mai apropiată de transcrierile RAV. Astfel, s-au efectuat următoarele operații de preprocesare asupra transcrierilor brute: restaurarea de diacritice, înlocuirea URL-urilor cu forma lor vorbită, înlocuirea numerelor cu text, înlocuirea abrevierilor cu forma lor neabreviată, înlocuirea adreselor de email cu forma lor vorbită, mutarea textelor din paranteză pe linii separate şi înlăturarea parantezelor, ștergerea liniilor din alte limbi, înlocuirea literelor mari cu litere mici.

Alinierea și filtrarea transcrierilor. Alinierea transcrierilor RAV cu transcrierile brute WEB s-a făcut folosind distanța Levenstein. Această metrică compară 2 secvențe de cuvinte ținând cont de numărul de substituții, inserții și ștergeri dintre cele 2 secvențe. După alinierea celor două transcrieri, selecția părților identice ce urmează să facă parte din corpusul nou de vorbire adnotată s-a făcut be baza mai multor criterii, după cum urmează. Secvenţe consecutive de cuvinte, ce conţin un număr de caractere

mai mare decât un prag determinat experimental (8 caractere), sunt considerate a fi corect transcrise. Un alt criteriu utilizat la selecţia transcrierilor este durata secvenţelor audio, fiind necesar ca aceasta să depăşească un anumit prag ales tot empiric (1 secundă). De asemenea, distanţa în timp între două cuvinte consecutive este limitată superior la 2 secunde pentru a asigura faptul că nu există cuvinte intermediare netranscrise. În urma efectuării alinierii și filtrării rezultă setul de transcrieri aliniate cu ștampile de timp, ștampile ce vor fi folosite pentru selecția segmentelor de vorbire corespunzătoare transcrierilor.

Selecția segmentelor de vorbire. La final, după ce secvenţele de cuvinte corecte au fost selectate, ştampilele de timp asociate acestor cuvinte au fost folosite pentru tăierea secvenţelor audio corespunzătoare din datele audio brute.

Corpus nou de vorbire adnotată. Corpusul nou de vorbire adnotată este format din transcrierile aliniate și segmentele de vorbire corespunzătoare. Corpusul poate fi folosit la reantrenarea sistemului RAV. Detalii privind corpusul obținut la finalul acestei activități au fost prezentate în tabelul 2.1.c.

1.2.2 Utilizarea transcrierilor aproximative pentru generare de date

În urma aplicării procedurii de aliniere și filtrare prezentată în secțiunea anterioară pe seturile de date brute SSC-train3-raw și SSC-train4-raw, și a trei iterații de identificare de probleme și soluții corespunzătoare, au fost obținute seturile de date denumite SSC-train3-trans-v3 și SSC-train4-trans-v3. Dimensiunile acestora, exprimate în număr de cuvinte, respectiv număr de ore de vorbire și eficiența procesului de adnotare automată, exprimată sub forma procentului de date brute ce au putut fi adnotate, raportat la dimensiunea datelor brute sunt prezentate în Tabelul 2.2.a. Comparativ cu numărul de ore aliniate precedent în activitatea 1.13 (compl-2018), în această etapă (trans-v3) s-au aliniat mai puține ore per total (Tabelul 2.1.c).

Tabelul 2.2.a Statistici pentru seturile de date SSC-train3-trans-v3 și SSC-train4-trans-v3, obținute în urma aplicării metodei alinierii transcrierilor aproximative cu transcrierile RAV

Corpus	Sursa	Durată [# ore]		Eficienţă aliniere [% ore]		Dimensiune [# cuvinte]	Eficienţă aliniere [% cuvinte]
SSC-train3-trans-v3	radio #1	1,0	37,5	5,0%	27,4%	8.833	29,4%
	TV #1	12,8		25,0%		135.874	38,0%
	TV #2	23,6		35,9%		249.271	30,2%
SSC-train4-trans-v3	radio #1	2,7	228,8	3,5%	29,4%	24.345	20,3%
	TV #1	87,9		26,5%		920.785	41,1%
	TV #2	138,1		37,6%		1.426.334	34,7%

1.2.3 RAV utilizând corpusul nou creat

Sistemul RAV bazat pe HMM-DNN folosit anterior în activitatea A1.13 a fost antrenat folosind seturile de date RSC-train și SSC-train, obținând un WER de 2.87% pe setul de evaluare RSC-eval, respectiv 15.87% pe setul de evaluare SSC-eval. Sistemul a fost reantrenat ulterior folosind corpusurile SSC-train3-compl-2018 și SSC-train4-compl-2018 rezultate din activitatea A1.13 împreună cu cele inițiale. Sistemul reantrenat a obținut rezultate puțin mai bune după cum se poate observa în Tabelul 2.2.e.

Acesta a avut o îmbunătățire relativă a WER de 8.36% pe setul RSC-eval, respectiv 12.03% pe setul SSC-eval.

Tabelul 2.2.e Performanţa sistemelor RAV după reantrenare

Corpus antrenare	Model acustic	WER [%]		Îmbunătăţire relativă a WER [%]	
		RSC-eval	SSC-eval	RSC-eval	SSC-eval
RSC-train + SSC-train	HMM-DNN	2.87	15.87	n/a	n/a
+ SSC-train3-compl + SSC-train4-compl	HMM-DNN	2.63	13.96	8.36	12.03
+ SSC-train3-trans + SSC-train4-trans	HMM-DNN	2.41	12.97	16.03	18.27

Același sistem RAV a fost reantrenat folosind corpusurile obținute în această activitate împreună cu cele deja existente. Noul sistem a obținut un WER de 2.41% pe setul RSC-eval, respectiv 12.97% pe setul SSC-eval. Comparativ, sistemul RAV rezultat din această activitate are o îmbunătățire relativă a WER față de sistemul inițial de 16.02%, pe când sistemul obținut anterior în activitatea 1.13 are o îmbunătățire relativă a WER de 8.36%. În cazul setului SSC-eval, noul sistem a obținut o îmbunătățire relativă a WER de 18.27% față de doar 12.03% a sistemului anterior din activitatea A1.13.

1.3 Activitatea 2.12 - Proiectarea şi implementarea unei soluţii de bază pentru generarea de scoruri de încredere pentru RAV

Majoritatea sistemelor de recunoaștere automată a vorbirii (RAV) oferă pe lângă transcrierea fișierului audio și o secvență de scoruri de încredere. Fiecare scor corespunde unui cuvânt și reprezintă gradul de încredere al sistemului de RAV în transcrierea cuvântului respectiv. Aceste scoruri sunt de obicei între 0 și 1 și pot fi interpretate ca probabilități – cu cât scorul este mai mare cu atât este mai probabil ca transcrierea furnizată să fie corectă.

Pentru a construi baze de date de vorbire într-un mod automat folosim scorurile de încredere astfel: *(i)* pornim de la un set de vorbire neadnotat pe care îl trecem prin sistemul de RAV pentru a produce o transcriere și secvența aferentă de scoruri de încredere; *(ii)* transcrierea este filtrată pe baza unui prag τ aplicat scorurilor de încredere: dacă un cuvânt are scorul asociat mai mare sau egal cu pragul atunci este păstrat, altfel este ignorat. Repetând acest procedeu pentru fiecare fișier audio din setul de date, construim o nouă bază de date adnotată într-un mod total automat. Pragul τ controlează compromisul dintre cantitatea și corectitudinea datelor generate: un prag mic rezultă în multe date, dar incerte din punct de vedere al transcrierilor; invers, un prag mare rezultă în puține date, dar corecte.

În această secțiune prezentăm rezultate experimentale pentru această metodă de generare de date. Arătăm rezultate pentru utilizarea metodei pentru sarcina de interes, și anume, generarea de baze de date în mod automat și apoi pentru reantrenarea sistemului de RAV cu datele nou generate.

1.3.1 Utilizarea scorurilor de încredere în generarea de date

Aplicând procedura descrisă anterior pe seturi de date pentru care nu avem transcrieri manuale, SSC-train3 și SSC-train4, și utilizând diferite praguri de filtrare $\tau \in \{0.9, 0.95, 1.0\}$obținem noi seturi de date; de asemenea, am exclus cuvintele mai scurte de 200 ms. Cantitatea de date rezultată pentru fiecare dintre aceste configurații este descrisă în tabelul 2.3.d. Aceste date sunt apoi utilizate pentru augmenta setul de date standard și pentru a reantrena sistemul de RAV.

Tabelul 2.3.d Cantitatea de date obținute după filtrarea folosind scorurile de încredere. Prezentăm atât valori absolute (în ore *h* și minute *m*), cât și valori relative (în procente *%*) raportate la cantitatea totală de date. Filtrarea s-a realizat pe baza a diferite praguri $\tau \in \{0.9, 0.95, 1.0\}$– cu cât pragul este mai mare cu atât se obțin mai puține date, dar mai corecte din punctul de vedere al transcrierilor.

Sursa	SSC-train3-conf			SSC-train4-conf		
	τ = 0.9	τ = 0.95	τ = 1	τ = 0.9	τ = 0.95	τ = 1
radio #1	8h 12m (42.5%)	7h 21m (38.1%)	5h 30m (28.5%)	33h 28m (42.8%)	29h 39m (37.9%)	22h 3m (28.2%)
TV #1	27h 58m (54.3%)	25h 15m (49.0%)	19h 7m (37.1%)	181h 40m (54.8%)	164h 24m (49.6%)	124h 29m (37.5%)
TV #2	42h 7m (63.9%)	39h 1m (59.2%)	31h 13m (47.4%)	229h 13m (62.4%)	212h 0m (57.7%)	169h 1m (46.0%)
Total	78h 17m (61.8%)	71h 38m (56.5%)	55h 51m (44.1%)	444h 22m (57.2%)	406h 4m (52.2%)	315h 34m (40.6%)

Corpus antrenare	Model acustic	WER [%]		Îmbunătăţire relativă a WER [%]	
		RSC-eval	SSC-eval	RSC-eval	SSC-eval
RSC-train + SSC-train	HMM-DNN	2.87	15.87	n/a	n/a
+ SSC-train3-compl-2018 + SSC-train4-compl-2018	HMM-DNN	2.63	13.96	8.36	12.03
+ SSC-train3-conf-090 + SSC-train4-conf-090	HMM-DNN	2.67	14.88	6.97	6.24
+ SSC-train3-conf-095 + SSC-train4-conf-095	HMM-DNN	2.59	15.01	9.76	5.42
+ SSC-train3-conf-100 + SSC-train4-conf-100	HMM-DNN	2.76	14.93	3.83	5.92

Rezultatele experimentale indică mai multe aspecte:

1. metoda prezentată și evaluată mai sus poate fi utilizată pentru generare de date pentru antrenarea RAV, sistemele rezultate obținând rezultate mai bune decât sistemul RAV inițial;
2. metoda sistemelor RAV complementare, evaluată în activitatea A1.13 de anul trecut produce sisteme RAV mai performante decât metoda prezentată și evaluată în această secțiune;
3. e nevoie de metode mai precise de estimare a scorurilor de încredere pentru a produce seturi de date mai corecte, dar și pentru a putea selecta date cu o incertitudine mai mică în vederea reantrenării RAV.

1.4 Activitatea 2.13 - Îmbunătăţirea soluţiei de adnotare automată a semnalului de vorbire utilizând sisteme de RAV complementare

1.4.1 Introducere

Proiectarea și implementarea inițială a metodei de adnotare automată a semnalului de vorbire utilizând sisteme de RAV complementare a făcut obiectul activității A1.13 din etapa 1/2018. Activitatea curentă a vizat dezvoltarea suplimentară și îmbunătățirea metodei folosind ca punct de pornire concluziile activității de anul trecut.

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

Pentru ca această metodă să funcţioneze este esenţial ca cele două sisteme RAV să fie complementare. Mai exact, erorile celor două sisteme RAV trebuie să fie necorelate. Există câteva opţiuni care fac ca

acest lucru să fie posibil: tipurile de modele acustice sau lingvistice să fie diferite, modelele să fie antrenate pe date diferite, algoritmii de decodare să fie diferiţi etc.

În cadrul activității A1.13 din etapa 1/2018 au fost utilizate două sisteme de RAV inițiale care difereau prin următoarele caracteristici:

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

Am arătat atunci că cele două sisteme fac erori diferite, necorelate: practic numai 1.0% - 1.3% din datele adnotate în mod automat cu această metodă sunt adnotate greșit. Restul transcrierilor sunt realizate corect, iar datele nou create pot fi utilizate pentru reantrenarea sistemului de RAV.

Cu toate acestea, experimentele au arătat că datele nou generate ajută foarte puțin la creșterea performanțelor celui mai bun sistem de RAV inițial: eroarea la nivel de cuvânt (WER) a scăzut:

- de la 4.50% la 4.33% pentru vorbire citită și
- de la 20.20% la 18.41% pentru vorbire spontană

Rezultatele sumarizate ale sistemelor de RAV inițiale și ale sistemului de RAV îmbunătățit obținut în A1.13 din etapa anterioară sunt prezentate în tabelul 2.4.a.

Tabelul 2.4.a Performanţa sistemelor RAV inițiale și a sistemului RAV îmbunătățit din A1.13, etapa 1.

Model acustic		**Model linvistic**	**WER [%]**		**Îmbunătăţire relativă a WER [%]**	
Corpus antrenare	**Tip model**		**RSC-eval**	**SSC-eval**	**RSC-eval**	**SSC-eval**
RSC-train + SSC-train	HMM-GMM	Decodare RAV: 64k cuvinte, 3-gram	12.60	32.30	-	-
RSC-train + SSC-train	HMM-DNN (TDNN2)	Decodare RAV: 200k cuvinte, 2-gram Reev. lingv.: 200k cuvinte, 4-gram	4.50	20.20	-	-
+ SSC-train3-compl-2018 + SSC-train4-compl-2018	HMM-DNN (TDNN2)	Decodare RAV: 200k cuvinte, 2-gram Reev. lingv.: 200k cuvinte, 4-gram	4.33	18.41	3.78	8.86

Pornind de la rezultatele de mai sus, în activitatea A2.13 din această etapă am decis să abordăm următoarele sarcini:

- crearea unor noi sisteme complementare de RAV, țintind cu performanțe inițiale mai bune, similare cu performanțele celui mai bun sistem inițial de anul trecut și
- combinarea mai multor transcrieri de RAV în procesul de aliniere și selecție a transcrierilor cu scopul obținerii mai multor date adnotate.

1.4.2 Sisteme de RAV inițiale îmbunătățite

Din punctul de vedere al caracteristicilor cheie și al componentelor constitutive, sistemele RAV inițiale utilizate în A1.13 din etapa anterioară pot fi caracterizate astfel:

- Sistemul RAV #1: creat cu toolkitul CMU Sphinx, model acustic HMM-GMM, model de limbă pentru decodare RAV de tip 3-gram cu vocabular de 64k cuvinte, performanțe scăzute;

● Sistemul RAV #2: creat cu toolkitul Kaldi, model acustic HMM-DNN (TDNN2), model de limbă pentru decodare RAV de tip 2-gram cu vocabular de 200k cuvinte, model de limbă pentru reevaluare lingvistică de tip 4-gram cu vocabular de 200k cuvinte.

În cadrul A2.13 din etapa curentă au mai fost dezvoltate alte două sisteme de RAV cu următoarele caracteristici:

● Sistemul RAV #3: creat cu toolkitul Kaldi, model acustic HMM-DNN (TDNN3), model de limbă pentru decodare RAV de tip 2-gram cu vocabular de 200k cuvinte, model de limbă pentru reevaluare lingvistică de tip RNN cu istorie de 5 cuvinte și vocabular de 200k cuvinte;
● Sistemul RAV #4: creat cu toolkitul NVIDIA OpenSeq2Seq, model acustic și model de limbă pentru decodare integrate într-o singură rețea neurală de tip DeepSpeech, model de limbă pentru reevaluare lingvistică de tip 4-gram cu vocabular de 200k cuvinte.

Performanțele acestor două noi sisteme de RAV sunt prezentate în tabelul 2.4.b. Așa cum se observă sistemul RAV #3 are performanțe net superioare sistemului RAV #2 (cel mai performant sistem RAV inițial din etapa anterioară). Concret, eroarea la nivel de cuvânt (WER) a acestui sistem este de 2.87% pentru vorbire citită (față de 4.50% pentru SRAV #2), respectiv de 15.87% pentru vorbire spontană (față de 20.20% pentru SRAV #2). în plus, în cadrul acestei activități am reantrenat SRAV #3 și cu setul de date generat anul trecut (SSC-train{3,4}-compl-2018), iar sistemul astfel rezultat a fost evaluat și mai bine: eroarea la nivel de cuvânt de 2.63% pentru vorbire citită, respectiv de 13.96% pentru vorbire spontană.

Sistemul de RAV #4 s-a dovedit a avea performanțe foarte slabe. Acesta este chiar mai puțin performant decât sistemul RAV #1, utilizat în A1.13 din etapa anterioară, SRAV bazat pe o tehnologie veche (CMU Sphinx cu modele acustice de timp HMM-GMM). Concluzia pe care o putem trage din acest experiment este că tehnologia de RAV de tip end-to-end (model acustic și model de limbă integrate într-o singură rețea neurală profundă de tip sequence-to-sequence) nu este încă suficient de matură pentru a putea fi utilizată în practică. Dat fiind această concluzie, SRAV #4 nu a mai fost utilizat în continuare în această activitate. Nu a fost evaluată nici complementaritatea lui față de celelalte SRAV inițiale și nici nu a fost folosit pentru a genera noi seturi de date adnotate automat.

Tabelul 2.4.b Performanţa sistemelor RAV inițiale din A2.13, etapa 2/2019. Performanța sistemului RAV inițial reantrenat folosind și setul de date generat în cadrul A1.13 din etapa 1/2018.

Model acustic		**Model linvistic**	**WER [%]**		**Îmbunătăţire relativă a WER [%]**	
Corpus antrenare	**Tip model**		**RSC-eval**	**SSC-eval**	**RSC-eval**	**SSC-eval**
RSC-train + SSC-train	HMM-DNN (TDNN3)	Decodare RAV: 200k cuvinte, 2-gram Reev. lingv.: RNN 5-gram	2.87	15.87	-	-
RSC-train + SSC-train	DeepSpeech	Decodare RAV: integrat DeepSpeech Reev. lingv.: 200k cuvinte, 4-gram	15.12	43.61	-	-
+ SSC-train3-compl-2018 + SSC-train4-compl-2018	HMM-DNN (TDNN3)	Decodare RAV: 200k cuvinte, 2-gram Reev. lingv.: RNN 5-gram	2.63	13.96	9.67	21.65

1.4.3 Evaluarea calitativă a metodei: complementaritatea sistemelor de RAV inițiale

În contextul situației descrise mai sus (renunțarea la SRAV #4 din motive de performanță scăzută), sistemele de RAV inițiale au fost considerate ca fiind SRAV #1, SRAV #2 și SRAV #3. Complementaritatea perechii de sisteme (SRAV #1 - SRAV #2) a fost evaluată în etapa anterioară. S-a demonstrat atunci că aplicarea metodei folosind cele două sisteme RAV ca sisteme inițiale conduce la adnotarea automată a 48%, respectiv 20% din seturile de date RSC-eval, respectiv SSC-eval și că eroarea de adnotare se plasează în gama 1.0% - 1.3%. Aceste rezultate au fost reluate în tabelul 2.4.c.

În continuare, în această etapă a fost evaluată complementaritatea perechii de sisteme (SRAV #2 - SRAV #3). Rezultatele evaluării complementarității și implicit a eficienței și calității metodei de adnotare automată sunt rezumate, pentru comparație, tot în tabelul 2.4.c. Se poate observa că sistemele analizate sunt mai asemănătoare: ele generează transcrieri mai similare și, implicit, fac și mai multe greșeli identice. Acest lucru este indicat de eroarea la nivel de cuvânt mai mare (2.6%, respectiv 2.7%) comparativ cu eroarea la nivel de cuvânt obținută pentru perechea de sisteme SRAV #1 - SRAV #2. Pe de altă parte, cantitatea de date selectată prin aplicarea metodei folosind sistemele SRAV #2 și #3 ca sisteme inițiale este semnificativ mai mare (79%, respectiv 73%) comparativ cu situația de anul trecut (48%, respectiv 20%).

Putem concluziona că utilizând perechea de SRAV #2 + #3 reușim să adnotăm automat o cantitate de date 2 ori mai mare, cu o eroare de adnotare de aproximativ 2 ori mai mare. În ce măsură acest lucru este benefic se va vedea în experimentele ulterioare.

Tabelul 2.4.c Calitatea și cantitatea datelor obținute prin aplicarea metodei folosind ca SRAV inițiale perechile (SRAV #1 - SRAV #2), respectiv (SRAV #2 - SRAV #3).

Set evaluare	RSC-eval		SSC-eval	
Pereche SRAV	SRAV #1 - SRAV #2	SRAV #2 - SRAV #3	SRAV #1 - SRAV #2	SRAV #2 - SRAV #3
WER [%]	1.0	2.6	1.30	2.7
ChER [%]	0.3	0.7	0.4	1.0
Durată	2h, 37m (48 %)	4 h, 14 m (79 %)	0h, 41m (20 %)	2 h, 33 m (73%)

1.5 Activitatea 2.14 - Diseminare

Diseminarea rezultatelor proiectului a fost realizată: în cadrul consorţiului în cele workshopul organizat la Cluj-Napoca pe 18 noiembrie 2019 şi în comunitatea ştiinţifică la trei conferințe internaționale de prestigiu: 42nd International Conference on Telecommunications and Signal Processing, 10th Conference on Speech Technology and Human-Computer Dialogue și 14th International Conference on Linguistics Resources and Tools for Natural Language Processing. Suplimentar, unele dintre rezultate au fost publicate într-un articol în Buletinul Științific al Universității Politehnica din București. De asemenea, progresul înregistrat în această etapă a fost diseminat prin intermediul website-ului proiectului: https://tadarav.speed.pub.ro.

Dintre publicațiile menționate mai jos, articolele 1 și 4 sunt deja indexate în Web of Science (Thompson Reuters - ISI), articolele 2 și 3 sunt deja indexate IEEE Xplore şi în curs de indexare în Web of Science (Thompson Reuters - ISI), iar articolul 5 a apărut în volumul conferinței și este în curs de indexare în Web of Science (Thompson Reuters - ISI). În toate aceste articole numele finanţatorului este menţionat în secţiunea Acknowledgement, conform indicaţiilor din contractul de finanţare.

Lista completă a publicațiilor din etapa 2/2019 este următoarea:

1. Alexandru-Lucian Georgescu, Horia Cucu, Corneliu Burileanu, "Progress on automatic annotation of speech corpora using complementary ASR systems," in the Proceedings of the 42nd International Conference on Telecommunications and Signal Processing (TSP), 2019, Budapest, Hungary.
2. Gheorghe Pop, Serban Mihalache, Dragos Burileanu, "Forensic Recognition of Narrowband AMR Signals," in the Proceedings of the 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timișoara, Romania, 2019.
3. Alexandru-Lucian Georgescu, Horia Cucu, Corneliu Burileanu, "Kaldi-based DNN architectures for speech recognition in Romanian," in the Proceedings of the 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timișoara, Romania, 2019.

4. Gheorghe Pop and Dragos Burileanu, "Speech Enhancement for Forensic Purposes," in UPB Scientific Bulletin, Series C, Vol. 81, Issue 3, pp. 41-52, 2019.
5. Florin Iordache, Alexandru-Lucian Iordache, Dan Oneaţă, Horia Cucu, "Romanian Automatic Diacritics Restoration Challenge", in the Proceedings of the 14th International Conference on Linguistics Resources and Tools for Natural Language Processing, Cluj-Napoca, Romania, 2019.

2 Structura ofertei de servicii de cercetare şi tehnologice

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

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

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

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

3 Locuri de muncă susţinute prin program

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

Tabelul 4. Echipa de cercetare UPB

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

1 Valorificarea şi îmbunătăţirea competenţelor şi resurselor existente la nivelul consorţiului

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

SINTERO

Activitățile de cercetare desfășurate în a doua etapă de implementare au condus la obținerea rezultatelor așteptate și ele sunt în concordanță cu obiectivele specifice ale etapei. În plus, s-a reușit să se proiecteze, implementeze și testeze un model integrat bazat pe structura Tacotron și care unifică modelarea acustică, modelarea prozodiei și modelarea vorbitorilor într-un sistem unic. Astfel, rezultatele raportate în acest document și descrise detaliat în cele 5 livrabile aferente perioadei de raportare, pregătesc cadrul etapei viitoare pentru dezvoltarea unei noi tehnologii de realizare a interfețelor de sinteză text vorbire cu expresivitate. De asemenea, acest raport prezintă detalii referitoare la oferta de servicii de cercetare și tehnologice, activitățile de management și comunicare, modul de valorizare a resursei umane și dezvoltarea acesteia prin activități colaborative la nivelul consorțiului.

1. Activitățile etapei de raportare în contextul general al proiectului

În a doua etapă (2019) a proiectului SINTERO, etapă cu denumirea *„Implementarea componentelor pentru modelarea prozodiei și adaptarea la noi vorbitori a vocilor sintetice”*, s-a pornit de la resurse și module software dezvoltate deja în etapa 2018 de către partenerii UTCN (corpusuri de date audio și text, sistemul preliminar de sinteză fără expresivitate) și ICIA (module de adnotare a textului disponibile pe platforma Relate[3]) și au fost desfășurate o serie de activități pentru: **a)** implementarea modulului de identificare și codare a stilului de vorbire prin vectori de stil**, b)** implementarea și validarea unei metode de adaptare la noi vorbitori cu set redus de date, **c)** implementarea și testarea unei metode de modificare a prozodiei pentru voci expresive, **d)** îmbunătățirea componentei de control a prozodiei, respectiv activități de testare, validare, diseminare și demonstrare online.

În etapa următoare a proiectului SINTERO vom realiza *„Dezvoltarea unei noi tehnologii pentru realizarea interfețelor de sinteză text vorbire cu expresivitate”* (2020).

2. Gradul de realizare a obiectivelor specifice pentru Etapa a II-a, 2019

Ob. Pr4.2.15: *Identificarea automată a stilului de vorbire și expresivității din analiza textului*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 2 module software pentru identificarea automată a stilului din text[4]
- rezultate privind evaluarea performanțelor celor 2 module
- îmbunătățirea modulelor de identificarea a stilului de vorbire prin corecția automată a diacriticelor și adnotarea părților de vorbire, inclusiv evaluare
- 1 nouă metodă de codare a textului pentru transcrierea fonetică
- 3 articole publicate la 2 conferințe internaționale
- un livrabil (D2.15) cu titlul *„Implementarea modulului de identificare a stilului de vorbire și nivelului de expresivitate din analiza textului”.*

Ob. Pr4.2.16: *Implementarea unui modul de adaptare la un nou vorbitor*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 2 metode de adaptare a sistemului de sinteză la un nou vorbitor: (a) o metodă bazată pe posfiltrarea cu rețele neuronale, (b) o metodă bazată pe arhitectura Tacotron și GST (Global Style Tokens), implementate pe arhitecturi paralele de procesare cu GPU
- 8 sisteme de sinteză text vorbire adaptate la un nou vorbitor prin metoda de postfiltrare
- evaluarea obiectivă și prin teste de ascultare a celor 8 sisteme și pagină web cu mostre audio[5]

[3] *https://relate.racai.ro/*

[4] *https://github.com/speech-utcluj/romanian-text-classification-cnn*

[5] *https://speech.utcluj.ro/pf_lrec2020/*

- 5 sisteme de sinteză text vorbire bazate pe arhitectura Tacotron GST adaptate la stilul și expresivitatea unui nou vorbitor cu implementare pe sisteme cu GPU, evaluarea sistemelor și pagină web cu mostre audio[6]
- 1 articol care descrie metoda bazată pe postfiltrare
- 1 livrabil (D2.16) cu titlul *„Implementarea unui modul de adaptare la un nou vorbitor a sistemului de sinteză”.*

Ob. Pr4.2.17: *Modul de transfer a prozodiei unui vorbitor în sistemul de sinteză*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 1 modul software pentru transferul prozodiei bazat pe Tacotron GST
- 6 sisteme de sinteză text vorbire experimentale pentru transfer prozodie
- 1 livrabil (D2.17) cu titlul *„Implementarea unui modul de transplantare a prozodiei unui vorbitor în sistemul de sinteză”.*

Ob. Pr4.2.18: *Îmbunătățirea componentei de modelare și control a prozodiei*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 1 metodă originală de antrenare a sistemelor de sinteză bazate pe rețele neuronale pentru voci expresive, folosind chiar voce sintetizată cu sisteme de tip parametric
- 5 sisteme de sinteză text vorbire pentru voce expresivă
- web site pentru evaluarea subiectivă a celor 5 sisteme de sinteză[7]
- pagină web cu mostre audio generate cu acest nou model[8]
- 1 livrabil (D2.18.) *„Îmbunătățirea componentei de modelare și control a prozodiei. Activități de testare, validare și demonstrare online a modulelor implementate”.*

Ob. Pr4.2.19: *Diseminarea rezultatelor intermediare*

Grad realizare: Obiectiv realizat integral

Rezultate:

- realizarea și actualizarea web site-ului proiectului[9]
- pagini web cu demonstratoare cu vocile sintetizate
- 1 livrabil referitor la activitățile de diseminare (D2.19).

3. Rezultatele etapei și descrierea lor științifică și tehnică

3.1. Implementarea modulului de identificare a stilului de vorbire

Rezultatele raportate în această secțiune corespund obiectivului Pr4.2.15 și ele sunt descrise în extenso in livrabilul D2.15. Pentru un control mai bun al sistemelor de sinteză și pentru ca acestea să poată reda textul introdus într-o manieră cât mai apropiată de vocea naturală este util ca textul de intrare să poată fi clasificat automat în funcție de stilul și expresivitatea acestuia. Au fost dezvoltate două metode de detecție a stilului textului bazate pe: 1) modele probabiliste de tip LDA (Latent Dirichlet Allocation), 2) rețele neuronale convoluționale multistrat. De asemenea, pentru a îmbunătăți sistemul de detecție a stilului, vom prezenta și două module pentru restaurarea diacriticelor și determinarea părții de vorbire a cuvintelor. Acestea sunt incluse în fluxul de procesare la intrarea sistemului de clasificare a textului.

Metodele au fost aplicate asupra unui set de date text extrase din corpusul CoRoLa al Institutului de Inteligență Artificială al Academiei Române din București. Setul de date conține text în stilurile:

[6] https://speech.utcluj.ro/sintero/dnn-samples

[7] *http://romaniantts.com/lrec/*

[8] *http://speech.utcluj.ro/lrec2020_mara/.*

[9] *http://speech.utcluj.ro/sintero/*

beletristic, științific, publicistic, memorialistic și juridic. Pentru fiecare subset am avut la dispoziție aproximativ 1 milion de cuvinte, organizate în 40.000 de fraze. Media numărului de cuvinte dintr-o frază este de 20.

(1) Implementarea LDA, ca model probabilist, este capabilă să modeleze în mod ierarhic fiecare stil de exprimare ca o combinație finită de probabilități de stiluri de exprimare, din cele disponibile în mod latent în setul de antrenare. Această modelare este potrivită, mai ales având în vedere spectrul larg de posibilități de exprimare în diferitele stiluri de vorbire.

Rezultatele sunt vizibile sub formă grafică, sub forma probabilităților de clasificare a unui text necunoscut către unul dintre stilurile din corpusul de antrenare (mai jos grafic rezultat din 4 stiluri), precum și a scorului de coerență ale modelelor (ca măsură a gradului de separabilitate ale modelelor).

(2) Implementarea cu rețele neuronale convoluționale (CNN – Convolutional Neural Networks) s-a bazat pe un studiu detaliat[10] și pe experimente preliminare de pre-procesare și reprezentare vectorială a textului, care sunt descrise detaliat în livrabilul D2.15. Arhitectura rețelei[11] este inspirată din aceste studii și conține următoarele niveluri de prelucrare (codul este disponibil online[12]): 1) reprezentare vectorială compactă, 2) reducere număr neuroni prin tehnica dropout, 3) strat convoluțional cu filtre de dimensiune 3, 4, sau 5, 4) nivel cu conectare totală de tip dens, 5) reducere neuroni prin dropout și funcție de activare ReLU (Rectified Linear Unit), 6) nivel final de clasificare cu conectare totală a neuronilor și funcție de activare de tip SoftMax.

Au fost evaluate mai multe scenarii (vezi Tabel 1) pentru această rețea prin modificarea volumului de date de antrenare a rețelei, a dimensiunii stratului convoluțional și a numărului de epoci de antrenare. Pentru fiecare dintre aceste combinații s-au reținut din datele de antrenare 20% pentru testare și 10% pentru validare. Se poate observa din rezultatele prezentate în tabel că folosind această arhitectură, algoritmul este capabil să clasifice datele cu o acuratețe relativ mare și poate fi folosit în pașii următori ai sintezei text-vorbire. Chiar și cu date puține (1000 de propoziții/stil) rezultatele algoritmului sunt de aproximativ 91%.

Tabel 1. Rezultatele identificării stilului din text folosind rețele CNN

Nr.	Număr de propoziții pentru antrenare	Dimensiune convoluție	Număr epoci	Acuratețe
1	5*3000	512	25	93.45%
2.			50	93.37%
3.		1024	25	92.77%
4			50	93.46%
5.	5*1000	512	25	91.83%
6.			50	91.81%
7.		1024	25	91.37%
8.			50	91.63%
9.	5*8000	512	25	92.69%
10.			50	92.41%
11.		1024	25	92.72%
12.			50	92.28%
13.	5*38000	512	25	90.10%
14.		1024	25	90.44%

Pentru preprocesarea textului în aceste două metode de clasificare au fost incluse și modulele de restaurare automată a diacriticelor (utilă în special pentru texte fără diacritice colectate din mediul

[10] *http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/*

[11] *https://richliao.github.io/supervised/classification/2016/11/26/textclassifier-convolutional/*

[12] *https://github.com/speech-utcluj/romanian-text-classification-cnn*

online), respectiv de detecție automată a părții de vorbire (pentru realizarea dezambiguizării înțelesului unui cuvânt). Descrierea completă a sistemelor este realizată în cele două articole publicate la conferința ICCP 2019 (vezi livrabil D2.19 – Diseminare). De asemenea, tot în domeniul prelucrării textului au fost experimentate diferite metode de reprezentare a caracteristicilor textuale folosind informații auxiliare de natură lingvistică (de exemplu accent, silabificare), cu publicarea unui articol la conferința SPED 2019. Cercetările demonstrează faptul că prin analiza textului de intrare se poate determina stilul și expresivitatea în vorbire, cu scopul de a controla modul de generare a semnalului vocal.

3.2. Implementarea modulului de adaptare la un nou vorbitor a sistemului de sinteză

Rezultatele raportate în această secțiune corespund obiectivului Pr.4.2.16, iar ele sunt descrise în extenso în livrabilul D2.16 „Implementarea unui modul de adaptare la un nou vorbitor a sistemului de sinteză". Cerința esențială pentru modulul de adaptare este ca pornind de la un model existent să realizeze adaptarea către o nouă voce folosind date audio cât mai puține. Astfel, se deschide perspectiva creării de sisteme de sinteză de voce personalizate. În acest sens, au fost implementate 2 metode de adaptare bazate pe arhitecturi cu rețele neuronale: (1) o metodă de adaptare folosind postfiltarea, și (2) o metodă de adaptare folosind arhitectura Tacotron și GST (Global Style Tokens).

(1)Metoda de adaptare folosind postfiltrarea. Ideea acestei metode este de a crea un sistem de sinteză text vorbire generic, antrenat cu mai multe voci pentru a mări volumul de date necesar antrenării, și bazat pe arhitecturi DNN (Deep Neural Network). Ieșirea acestui sistem este adaptată către noul vorbitor tot prin intermediul unei rețele neuronale, cu rol de postfiltrare. Acest postfiltru are rolul de a condiționa caracteristicile acustice, sintetice, generate la ieșire, către caracteristicile acustice naturale ale noului vorbitor. Antrenarea sistemului de sinteză s-a realizat cu 8 voci feminine din corpusul SWARA pe o arhitectură de tip Merlin[13]. Pentru partea de postfiltrare s-a implementat o rețea neuronală cu nivele total conectate, combinate cu nivele recurente de tip LSTM (Long Short Term Memory), cu număr variabil de neuroni în fiecare strat (256, 512 sau 1024) și cu diferite funcții de activare (tanh, ReLu).

Tabel 2. Sisteme de sinteză antrenate pentru metoda de post-filtrare (detalii în D2.16)

Acronim de sistem de sinteză	Numărul de propoziții folosite pentru antrenarea sistemului de sinteză	Numărul de propoziții folosite pentru post-filtrare
M050_PF050	*50*	*50*
M100_PF100	*100*	*100*
M100Db_PF100Db	*2x100*	*2x100*
M500_PF500	*500*	*500*
M100_PF_MSPK	*100*	*10x100*

Sistemele prezentate mai sus au fost validate cu metode obiective și subiective. Pentru metoda obiectivă s-a folosit măsura de distorsiune cepstrală (en. *Mel Cepstral Distortion (MCD)*). Acuratețea alinierilor pe nivel de stare nu este cunoscută, motiv pentru care valoarea MCD a fost obținută folosind un pas de aliniere a datelor cu ajutorul algoritmului DTW. Pentru calcularea valorilor MCD au fost sintetizate 50 de propoziții cu fiecare sistem. Aceste propoziții nu au fost incluse în datele de antrenare. Sistemele M050, M100 și M500 sunt sisteme de sinteză minimale, ce utilizează 50, 100 și 500 de propoziții de la un singur vorbitor fără post-filtrare. Cea mai bună valoare pentru sistemele antrenate pe 100 de propoziții este obținută cu ajutorul post-filtrării, urmată de adaptarea de vorbitor.

[13] *https://github.com/CSTR-Edinburgh/merlin*

Testul de ascultare a fost completat de 20 de ascultători și arată că metoda de post-filtrare și adaptare la vorbitor conduce la o voce mai naturală și inteligibilă. Identificatorii sistemelor sunt următorii: A - Sistem de *sinteză* antrenat cu 50 de propoziții; B - sistem de sinteză antrenat cu 100 de propoziții; C - sistem de sinteză antrenat cu 100 de propoziții și post-filtrare cu 100 de propoziții; D - sistem de sinteză și post-filtrare antrenate cu 100 de propoziții dublate artificial; E - sistem de sinteză antrenat cu 100 de propoziții și post-filtrare multi-vorbitor; F - sistem de adaptare la vorbitor pornind de la o rețea preantrenată cu date multi-vorbitor; G - sistem de sinteză antrenat cu 500 de propoziții; H - vocea naturală. Un demonstrator online pentru aceste voci se găsește la adresa *https://speech.utcluj.ro/pf_lrec2020/*

(2)Metoda de adaptare folosind arhitectura Tacotron și GST (Global Style Tokens). Această arhitectură[14] permite antrenarea cu mai mulți vorbitori, dar mai ales permite învățarea automată a expresivității din vorbire prin intermediul unor vectori (GST – Gloal Style Tokens) incluși în această arhitectură. În plus, modulul de control al prozodiei poate fi folosit și la învățarea identității vorbitorilor. Cerința principală a unui astfel de sistem este volumul foarte mare de date audio pentru antrenare.

În consecință, pentru evaluarea metodei s-au folosit corpusurile MARA[15] (1 vorbitor, 11 ore de vorbire preluată din audiobook) și SWARA[16] (17 vorbitori, 21 de ore de vorbire, din care au fost preluate aproximativ 50 de minute în configurație de 10 vorbitori feminine, respectiv 5 feminini și 5 masculini). În livrabilul D2.16 sunt prezentate în detaliu rezultatele experimentale. Concluziile arată că adaptarea la un nou vorbitor se face relativ rapid (aproximativ 10 epoci, 5-10 minute de rulare pe sistem cu o singură placă GPU). Identitatea vorbitorilor este controlată prin intermediul unui strat de reprezentări vectoriale, învățate tot în cadrul antrenării sistemului. Exemple audio pentru aceste sisteme sunt disponibile și pot fi ascultate aici: *https://speech.utcluj.ro/sintero/dnn-samples/*. Pe această direcție, se propune ca în viitor să fie explorate și alte metode de învățare, de exemplu doar cu o singură mostră audio sau cu un set redus de mostre audio.

3.3. Implementarea modulului de transfer a prozodiei

Rezultatele raportate în această secțiune corespund obiectivului Pr.4.2.17 și se referă la posibilitatea de a modifica prozodia generată de un sistem de sinteză text vorbire la prozodia unui anumit vorbitor. Pentru implementarea modulului de transfer a prozodiei s-a ales arhitectura Tacotron GST, deoarece aceasta permite codificarea prozodiei prin intermediul vectorilor GST (Global Style Tokens). Modulul este compus din următoarele componente

- componenta Tacotron, adică sistemul de sinteză text vorbire de bază
- componenta Style Tokens prin intermediul căreia se codează și reprezintă latent prozodia din vorbire. Această componentă e compusă din (a) codorul parametrilor acustici prin intermediul unei rețele neuronale cu nivele recurente, respectiv convoluționale și (b) codorul pentru stil și prozodie, bazat pe un nivel cu neuroni de atenție, capabili să exploreze corelații pe termen lung (prozodie) în parametrii acustici
- componenta de decodare, capabilă să producă - prin intermediul unei rețele neuronale și pe baza codurilor generate în procesul de antrenare – semnalul sintetizat. La intrarea decodorului se prezintă pe de o parte textul de sintetizat, pe de altă parte, fie (a) un semnal referință a cărui prozodie se dorește a fi transferată pe semnalul sintetizat, fie (b) o combinație a vectorilor prin care s-a codat prozodia în procesul de antrenare.

Deoarece sistemele de sinteză bazate pe arhitectura Tacotron necesită un volum foarte mare de date audio, s-a abordat o strategie prin care s-a folosit un sistem deja pre-antrenat cu date audio cu

[14] *https://github.com/mozilla/TTS*

[15] *https://speech.utcluj.ro/marasc/*

[16] *https://speech.utcluj.ro/swarasc/*

expresivitate colectate din audiobook-ul Mara. Modelul a fost antrenat 800 de epoci, suficient pentru ca modulul GST (în configurație cu 10, respectiv 5 tokeni de stil) să surprindă variabilitatea prozodică a datelor audio. În etapa de sinteză s-au aplicat 2 strategii:

- setarea manuală a tokenilor de stil prin intermediul unei ponderi, astfel că s-a putut observa că fiecare token a învățat un stil prozodic diferit
- utilizarea unei referințe audio conținând prozodia care se dorește a fi transferată, ocazie cu care s-a constatat că această prozodie nu influențează în mod semnificativ prozodia semnalului de ieșire și de aici concluzia că transferul de prozodie este eficient prin intermediul ponderilor tokenilor.

Tabel 3. Experimente de transfer a prozodiei cu sistemele de sinteză implementate

Modelul inițial	Date de adaptare	Concluzii
MARA	*2 vorbitori din SWARA, cu adaptare tokeni*	*Sistemul s-a adaptat la cei 2 vorbitori, însă prozodia doar parțial*
MARA	*10 vorbitori din SWARA, cu adaptare GST și ponderi tokeni*	*GST s-au adaptat, dar surprind identitatea vorbitorilor și mai puțin prozodia*
MARA	*10 vorbitori din SWARA, cu ponderi GST fixe*	*GST s-au adaptat și deși ponderile lor sunt fixe a fost învățată foarte rapid identitatea vorbitorilor, ignorând prozodia învățată în modelul inițial*
MARA	*10 vorbitori din SWARA, dar întreg modulul GST e fix*	*Prozodia inițială e uitată și este suprascrisă de prozodia vorbitorilor din corpusul SWARA*
MARA	*10 vorbitori din SWARA + propoziții expresive din MARA, ponderi GST fixe*	*Prozodia inițială e parțial reținută, iar tokenii au învățat identitatea vorbitorilor*

Astfel, Tacotron GST permite modelarea prozodiei unui vorbitor prin manipularea unor reprezentări latente ale stilurilor de vorbire. S-a observat că în arhitectura modulului GST tokenii rețin dimensiunea de variabilitate maximă a datelor de antrenare (de exemplu, prozodia pentru un singur vorbitor, respectiv identitatea vorbitorilor pentru sisteme antrenate cu date de la mai mulți vorbitori). Ca urmare, păstrarea informației anterioare în cadrul modulului GST nu este fezabilă.

În dezvoltările următoare, pentru a îmbunătăți transferul prozodiei, vor fi abordate alte tehnici care folosesc rețelele neuronale: învățarea continuă, învățarea folosind set redus de date. O altă metodă ar fi augmentarea setului de date neutre de antrenare cu date sintetice generate de o voce expresivă. Modulul software dezvoltat este descris mai amplu în livrabilul D2.17 „Implementarea unui modul de transplantare a prozodiei unui vorbitor în sistemul de sinteză", iar mostre audio sintetizate cu sistemele implementate sunt disponibile în pagina https://speech.utcluj.ro/sintero/prosody_examples_2019/.

3.4. Îmbunătățirea componentei de modelare și control a prozodiei

Rezultatele raportate în această secțiune corespund obiectivului Pr.4.2.18 și sunt descrise în detaliu in articolul transmis spre publicare la conferința internațională Language Resources an Evaluation 2020, titlu "*The MARA Corpus: Expressivity in end-to-end TTS systems using synthesised speech data"*. Articolul este prezentat în Anexa la livrabilul D2.18 „Îmbunătățirea componentei de modelare și control a prozodiei". Componenta de modelare și control a prozodiei a fost evaluată conform cu cele descrise în secțiunea anterioară și s-a evidențiat faptul că pentru transferul prozodiei sunt necesare date de antrenare bogate în conținut prozodic.

Astfel, în lipsa acestor date cu caracteristici prozodice variate, s-a propus o idee originală, și anume posibilitatea utilizării datelor audio sintetizate ca date de antrenare ale acestor modele. Datele sintetizate au fost obținute cu ajutorul sistemelor de sinteză parametrice bazate pe modele Markov sau

rețele neuronale cu straturi complet conectate unidirecționale. Datele sintetizate au utilizat conturul frecvenței fundamentale și durata la nivel de fonem extrase din înregistrările audio originale și în care există o mare variabilitate prozodică.

Au fost evaluate 5 sisteme de sinteză complete ce folosesc date expresive provenite fie de la vorbitorul original, fie din sistemele statistic-parametrice anterior dezvoltate (HTS). Rezultatele acestor 5 sisteme au fost evaluate atât din punct de vedere obiectiv, folosind o măsură a distorsiunii spectrogramei pe scală Mel, precum și din punct de vedere subiectiv folosind teste de ascultare. Testele de ascultare au inclus două secțiuni: naturalețe și expresivitate în format MuSHRA (TU-R Recommendation BS.1534-1). În urma evaluării nu au fost determinate diferențe statistice semnificative între sisteme. Pentru testele de ascultare a fost creată o pagină web distincta la adresa http://romaniantts.com/lrec/ unde se pot asculta și mostrele de test[17]. Variația conturului F0 pentru metoda propusă demonstrează faptul că acest contur este foarte apropiat de conturul F0 al vorbirii naturale expresive, iar măsurarea obiectivă a distanței spectrale între semnalul original și semnalul sintetizat indică valori care sunt tipice pentru sisteme de sinteză de înaltă calitate.

4. Oferta de servicii de cercetare, locuri de muncă susținute și valorificarea resurselor

Tabel 4. Sinteză privind oferta de servicii, locuri de muncă și valorificarea resurselor în UTCN

Oferta de servicii în UTCN	● oferta unei noi tehnologii de sinteză text-vorbire cu expresivitate, în limba română, bazată pe rețele neuronale și aliniată la standardele internaționale (Tacotron GST) ● servicii de adnotare automată a resurselor de date audio pe noul corpus MARA ● servicii de înregistrare audio de înaltă fidelitate ● servicii de procesare paralelă a datelor folosind tehnici de învățare automată pe noile echipamente achiziționate în anul 2019 din proiect ● servicii software pentru dezvoltarea modelelor bazate pe învățare automată. *ERRIS: https://erris.gov.ro/speech.utcluj.ro*
Locuri de muncă susținute în UTCN	1 x CS I, 1 x CS II, 1 x CS III, 1 x Tehnician 2 x ACS nou angajați începând cu luna ianuarie 2019
Resursa 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 (4 corpusuri) pentru clasificarea stilurilor de exprimare ● UTCN a folosit serviciile web oferite de ICIA pe platforma online „Relate” pentru adnotarea corpusului MARA ● UTCN a furnizat pentru ICIA și UAIC corpusurile de date audio disponibile și adnotările acestora ● UTCN a furnizat pentru ICIA module software pentru a fi integrate în platforma „Relate” ● UAIC a furnizat pentru UTCN acces la o platformă online pentru stocarea corpusurilor bimodale.
Cecuri	● UTCN a folosit 2 cecuri de tip C pentru formarea resursei umane nou angajate prin participarea la Școala de Vară Eastern European Summer School (1 săptămână) organizată de partenerul UPB.

5. Management și comunicare

Activitățile de management au fost orientate în special către managementul proiectului complex în vederea integrării diferitelor grupuri de cercetare și a resurselor tehnice ale acestora. S-au organizat mai multe conferințe Skype și o reuniune a parteneriatului în 18 Noiembrie 2019 la Cluj-Napoca. Este de notat faptul că s-a asigurat de către ICIA (prin responsabilul de achiziții) o bună comunicare și coordonare pentru realizarea planului de achiziții global, respectiv pentru documentația de raportare

[17] *http://speech.utcluj.ro/lrec2020_mara/.*

etapă. Din punct de vedere administrativ s-au primit 4 tranșe de avans cu o regularitate adecvată. Resursele financiare alocate UTCN pentru anul 2019 au fost utilizate în majoritate, cu excepția unei sume în categoria cecuri, care a trecut la economii.

6. Diseminarea rezultatelor

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

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

b) crearea de pagini web dedicate pentru demonstrarea online a modulelor dezvoltate în această etapă (corpusul Mara cu 11 ore de vorbire expresivă - *https://speech.utcluj.ro/marasc*/, demonstrator privind sinteza pe baza corpusului Mara și adaptarea la noi vorbitori - *https://speech.utcluj.ro/lrec2020_mara*/, evaluarea online a 7 sisteme de sinteză folosind testul Mushra - *http://romaniantts.com/lrec/,* demonstrator privind adaptarea sistemului de sinteză la noi vorbitori - *https://speech.utcluj.ro/sintero/dnn-samples/).*

c) publicații științifice cu rezultatele cercetărilor la conferințe internaționale în domeniu

[1] B. Lorincz, M. Nuțu, A. Stan, „Romanian Part of Speech Tagging using LSTM Networks", In Proc. of the IEEE 15th International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Sept 2019, Cluj-Napoca.

[2] M. Nuțu, B. Lorincz, A. Stan, „Deep Learning for Automatic Diacritics Restoration in Romanian", In Proc. of the IEEE 15th International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Sept 2019, Cluj-Napoca.

[3] A. Stan, „Input Encoding for Sequence-to-Sequence Learning of Romanian Grapheme-to-Phoneme Conversion", In Proc. of the 10th Conference on Speech Technology and Human-Computer Dialogue, 10-12 Oct 2019, Timişoara, Romania.

[4] David A. Braude, Matthew P. Aylett, Caoimhın Laoide-Kemp, Simone Ashby, Kristen M. Scott, Brian O'Raghallaigh, Anna Braudo, Alex Brouwer, Adriana Stan, „All Together Now: The Living Audio Dataset", Proceedings of Interspeech 2019, 16-19 Sept 2019, Graz, Austria.

7. Concluzii

Activitățile de cercetare desfășurate în etapa a II-a de implementare a proiectului (2019) au condus la obținerea rezultatelor așteptate și ele sunt în concordanță cu obiectivele specifice ale etapei. Astfel, rezultatele raportate în acest document și descrise detaliat în cele 5 livrabile aferente perioadei de raportare, asigură modulele software pentru etapa finală a proiectului.

8. Referințe la livrabilele aferente etapei 2019 (Anexe la raport)

[1] Livrabil D2.15:	*„Implementarea modulului de identificare a stilului de vorbire şi nivelului de expresivitate din analiza textului",* Mai 2019.
[2] Livrabil D2.16:	*„Implementarea unui modul de adaptare la un nou vorbitor a sistemului de sinteză",* Noiembrie 2019.
[3]Livrabil D2.17	*„Implementarea unui modul de transplantare a prozodiei unui vorbitor în sistemul de sinteză",* Noiembrie 2019.
[4] Livrabil D2.18:	*„Îmbunătățirea componentei de modelare şi control a prozodiei",* Noiembrie 2019.
[5] Livrabil D2.19:	*„Diseminare",* Noiembrie 2019.

Director Proiect Complex
Ioan Dan Tufiș