Raport științific și tehnic
Etapa a IV-a, an 2021
„Resurse şi tehnologii pentru dezvoltarea interfeţelor om-maşină în limba română”

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

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

Rezumat:
Acest document prezintă o sinteză a realizărilor de natură științifică și tehnică obținute în a IV-a etapă de implementare în cadrul proiectului PN-III-P1-1.2-PCCDI-2017-0818, Contract Nr. 73 PCCDI/2018 ReTeRom. Rapoartele detaliate ale fiecărui proiect component prezintă amănunțit activitățile și realizările etapei.

Obiectivele principale și rezultatele așteptate ale acestei etape au fost:

Pr. 1: Cobiliro: Exploatarea și diseminarea corpusului bimodal și a tehnologiilor de prelucrări textuale și voce. Aplicații de exploatare a corpusului bimodal și a tehnologiilor de prelucrări textuale și voce, create în proiectele P2, P3, P4; articole științifice.
Pr. 2: Teprolin: Diseminarea platformei dockerizate de prelucrare TEPROLIN. Manual de utilizare a platformei dockerizate TEPROLIN, workshop final de proiect, articole științifice.
Pr. 3: Tadarav: Diseminarea sistemelor de transcriere automata a vorbirii. Articole științifice si cerere de brevet
Pr. 4: Sintero: Evaluare si distribuție finala a tehnologiilor de sinteza vorbirii. Distribuție finală tehnologie pentru interfețe de sinteză a vorbirii. Diseminare și exploatare rezultate finale.

Activitățile de cercetare desfășurate în a IV-a etapă de implementare au condus la obținerea rezultatelor așteptate și ele sunt în concordanță cu obiectivele specifice ale etapei. Rezultatele raportate în acest document și descrise detaliat livrabilele aferente perioadei de raportare, finalizează proiectul. 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.

Raport ştiinţific - tehnic proiect component CoBiLiRo

Număr contract:	PN-III-P1-1.2-PCCDI-2017-0818, Nr. 73 PCCDI/2018
Acronim / titlu:	**„COBILIRO:”**
Titlu livrabil:	**Raport științific și tehnic (Etapa a IV-a, 2021)**
Termen:	**Aprilie 2021**
Editor:	**Dan Cristea**
Adresa de eMail editor:	**danu.cristea@ gmail.com**
Autori, în ordine alfabetică:	**Anca Bibiri, Șerban Boghiu, Dan Cristea, Daniela Gîfu, Felix Cristian Pericică, Ionuț Pistol, Mihaela Onofrei, Andrei Scutelnicu**
Ofițer de proiect:	**Cristian Stroe**

Pentru o descriere a tehnologiilor din proiectele componente menționate în aplicațiile de mai jos se poate consulta raportul 3.4 [12] și rapoartele celorlalte proiecte componente. După secțiunea 3.1 în care se discută progresele făcute la două din aplicațiile descrise în raportul 3.4 urmează descrierea a două aplicații noi ce valorifică tehnologiile și resursele dezvoltate în proiectul complex ReTeRom.

3.1 Actualizări asupra stadiului de dezvoltare a proiectelor propuse în raportul activității 3.4

Nu ne-am propus să dezvoltăm în continuare aplicațiile 3, 4, 5 și 6, față de specificațiile descrise în raportul 3.4. În continuare sunt discutate progresele făcute la aplicațiile 1 și 2.

Aplicația 1: **Suport pentru învățarea limbii române - *PD builder***

Acest proiect propune construirea unui dicționar de pronunție pentru cuvintele limbii române și dezvoltarea tehnologiei ce permite construirea automată a unor resurse similare. În raportul 3.4 au fost descrise două etape principale în realizarea acestui proiect: indexarea referințelor textuale din corpusul aliniat disponibil pe platforma proiectului și extragerea fragmentelor audio corespunzătoare. Prima etapă este finalizată, un index conținând peste 500.000 cuvinte unice și 100.000 forme neflexionate (forme lemă) a fost construit. Etapa va continua până la completarea colecției de resurse multimodale. Extragerea fragmentelor audio este încă în stadiul incipient, dar estimăm ca un prim prototip al aplicației va fi finalizat în acest an.

Aplicația 2: **Analiza corpusurilor bimodale**

Această aplicație propune dezvoltarea unui sistem de sprijin pentru evaluarea unui corpus bimodal aliniat text-sunet. În raportul 3.4 sunt descrise câteva erori posibile și soluții pentru descoperirea lor automată, erorile putând fi de tipul:

- diferențe între transcrierea text și conținutul înregistrării;
- calitatea scăzută a înregistrării;
- erori ale aplicației de aliniere.

Un prototip al acestui sistem a fost implementat ca parte a datelor statistice calculate pentru resursele de pe platforma CoBiLiRo, descrise în raportul 3.3. Un exemplu al acestor date poate fi văzut în figura 1.

File	AutoText	AlignedText	LongestUnmatch	StartsAt	SingleUnmatched
Alma-Mater-Iassiensi-24-Facultatea-de-Drept	5379	3761	44	30.51	193
Alma-Mater-Iassiensis-1-Aniversarile-Universitatii	7078	4400	121	2944.26	248
Alma-Mater-Iassiensis-11-Prof-Univ-Dr-Andrei-Marga	7072	4947	65	2688.87	266
Alma-Mater-Iassiensis-12-Facultatea-de-Fizica	5232	3312	52	3282.42	166
Alma-Mater-Iassiensis-13-Universitatea-din-Iasi-prezent-si-perspective	7006	5220	61	2109.15	281
Alma-Mater-Iassiensis-14-Universitatea-din-Iasi-Dimensiunea-culturala	6301	4558	41	2090.97	265

Fig.1 : Statistici pe alinieri la nivel de cuvânt

În figura 1 semnificaţia coloanelor este:

- *AutoText* - dimensiunea (în cuvinte) a textului extras automat din fișierul audio;
- *AlignedText* - dimensiunea (în cuvinte) a textului aliniat cu transcrierea originală;
- *LongestUnmatch* - cea mai lungă secvență de cuvinte din transcrierea originală pentru care nu există nici o aliniere;
- *StartsAt* - indexul cuvântului de la care începe secvența de mai sus;
- *SingleUnmatched* - numărul de cuvinte din transcrierea originală care nu au fost aliniate, deși cuvântul imediat precedent și cel imediat ulterior au fost aliniate.

O serie de observații pot fi făcute pe aceste statistici, de exemplu se observă că procentul cuvintelor aliniate corect (pe cele 6 fișiere din figura 1) este de aproximativ 70%. Se observă că în fișierul al doilea există o zonă lungă de cuvinte nealiniate (121 cuvinte) ce poate indica prezența în transcrierea manuală a unui fragment care nu se regăsește și în înregistrarea audio. O altă observație ce ar putea fi făcută pe aceste statistici ar fi că în fișierul 5 există un număr relativ mare de cuvinte izolate nealiniate (281) ceea ce ar putea indica o calitate scăzută a semnalului audio sau un accent/prozodie nestandard al/a vorbitorului.
Finalizarea acestui sistem și punerea lui la dispoziție ca o aplicație independentă de platforma CoBiLiRO este planificată pentru începutul anului viitor.

3.2 Aplicația 7 (nouă): **Sistem suport pentru crearea corpusurilor bimodale**

Acest proiect propune o aplicație mobilă care să permită crearea și editarea unor resurse bimodale aliniate text-vorbire. Dispozitivele mobile ce permit prelucrări audio-video complexe sunt deja la îndemâna unui segment important din populație. Un astfel de dispozitiv este capabil să permită atât înregistrarea unui semnal audio de calitate rezonabilă cât și interacțiunea utilizatorului cu o interfață de editare. Aplicația “Sistem de suport pentru crearea corpusurilor bimodale” dorește să pună la dispoziție unui posesor de telefon “inteligent” sau tabletă un mediu în care să poată crea alinieri text-voce.

Utilizatorul poate selecta sau adăuga fie un fișier text, fie o înregistrare audio și poate crea și alinia resursa pereche. De exemplu, dacă are la dispoziție un text, poate citi acel text, aplicația înregistrează acea citire și deschide o interfață în care utilizatorul vede atât textul

(derulat cuvânt cu cuvânt) cât și un fragment din corespondentul fișier audio pe care poate delimita zona de început și cea de sfârșit a pronunției acelui cuvânt.

Versiunea curentă este implementată pentru platforma Android (în Java/Kotlin), urmând ca după finalizarea tuturor funcționalităților propuse ea să fie transferată și pe platforma MacOS. O primă versiune disponibilă public a acestei aplicații este propusă pentru sfârșitul acestui an. Avem în vedere și posibilitatea urcării resurselor bimodale create de această aplicație pe platforma CoBiLiRo, dacă creatorul acestei resurse dorește acest lucru.

3.3 Aplicația 8 (noutate): **Sinteză text-vorbire și clonarea vocii în limba română cu metoda învățării prin transfer**

Acest proiect propune un sistem de sinteză text-vorbire, folosind metoda învățării prin transfer, care poate fi antrenat pe corpusuri în limba română de dimensiuni mici, conținând doar elemente audio și transcrierea acestora, fără a fi necesară strângerea altor date. Este o abordare cu rezultate mai bune decât ceea ce reține literatura de specialitate. Sistemele clasice necesită potrivirea manuală a fonemelor cu formele de undă pentru fiecare înregistrare, ceea ce face strângerea de date costisitoare și dificilă.
În general, sistemele bazate pe rețele neurale profunde necesită cantități mari de date pentru a da rezultate mulțumitoare, ceea ce face incomodă antrenarea de la zero a modelelor.

Totodată, acest proiect propune și un sistem de clonare vocală, folosind învățarea prin transfer de la modelul de sinteză general. Clonarea vocală este o sarcină, de asemenea, dificilă, capacitățile sistemelor de sinteză a vorbirii fiind și limita superioară a acestora, iar, în maniera clasică, ar trebui transcrise și alineate manual fonemele cu fiecare înregistrările audio ale vorbitorului țintă. Momentan, pentru limba română, sistemele de sinteză vocală publicate sunt într-o fază de pionierat. Clonarea vocii[1] rămâne o sarcină abordată relativ recent pentru limba română, cu toate că s-au făcut eforturi în întocmirea de corpusuri de vorbire paralelă [10].
Persoanele care și-au pierdut capacitatea de a vorbi și-o pot recăpăta prin clonarea digitală a vocii, putând să-și recupereze astfel o parte din identitate. De asemenea, în învățământul de la distanță tehnologia poate fi folosită pentru reproducerea vocii personalităților din istoria recentă și de azi, întrucât elevii să poată învăța despre respectivele personalități discutând chiar cu ele. Spre exemplu, aplicația FreudBot[2], un psihanalist robotizat, sugerează faptul că se poate învăța inclusiv învingerea agresivității, îndoielii și a fricii. Recunoaște peste 100 de vibrații proaste ale vieții de zi cu zi! [2].
O altă posibilitate pe care clonarea vocii o deschide este citirea automată a milioane de cărți în vocea autorului sau în vocea unui actor vocal dorit, cu o eficiență extremă din punctul de vedere al muncii umane. Totuși, trebuie luat în considerare faptul că tehnologiile de tip DeepFake pot reprezenta un instrument periculos în cazul folosirii abuzive, cum ar fi în cazul imitării vocii cuiva fără consimțământul acestuia. Vocile clonate în acest proiect sunt din cadrul corpusului public RSS [11], iar cele din cadrul corpusului SWARA (Stan *et al.*, 2017).
Modelul folosit presupune un modul de sinteză din text în spectrogramă mel (sintetizator) și unul de inferență a vorbirii din spectrogramă mel (vocoder). Tacotron 2 [10] antrenat pe setul de date LJ Speech (*The LJ Speech Dataset*, 2017) a fost ales pentru sintetizator, iar WaveGlow [1], antrenat tot în limba engleză, a fost vocoderul ales. S-au ales parametri audio identici pentru ambele modele, corespunzători cu parametrii LJ Speech. Seturile de date în limba română au fost preprocesate, ambele modele fiind antrenate pe acestea. Sintetizatorul a fost antrenat pe un set de date corespunzător unui singur vorbitor, în timp ce vocoderul a fost antrenat pe un set de date cu mai mulți vorbitori.

Rezultate statistice și interpretare

În vederea analizării performanței au fost colectate opinii despre clipurile cu voce naturală sintetizate. Au fost sintetizate 10 propozitii relativ scurte din știri recente, iar participanților (cei invitați să participe la acest studiu) li s-au dat două clipuri din corpusul RSS împreună cu două clipuri sintetizate pentru a le nota gradul de naturalețe. Pentru consistența cu alte studii, evaluarea a fost făcută pe scara Likert (5-Excellent, 4-Good, 3-Fair, 2-Poor, 1-Bad). Toți au fost vorbitori nativi de română și cunoscători de limba engleză. Participanții au folosit mijloace proprii de ascultare a clipurilor, fiind relevant faptul la doar o minoritate au folosit căști audio pentru evaluare. S-a putut astfel efectua o comparație directă cu alte rezultate, lucru care a oferit o privire de ansamblu mai bună asupra calității vorbirii, indiferent de mediul de ascultare.

Un al doilea studiu a fost făcut în condiții similare, participanții fiind rugați să evalueze gradul de naturalețe a vocii clonate. Rezultatele studiilor, împreună cu alte rezultate din literatură, sunt prezentate în Tabelul 1.

[1] http://www.vc-challenge.org
[2] FreudBot – Aplicații pe Google Play

Tabelul 1: Evaluarea comparativă pentru TTS

Model	**Scor mediu de opinie**	**Limba**
Voci naturale	4.80 ± 0.14	română
TTS Eletron (din această lucrare)	4.47 ± 0.15	română
Clonare vocală - 18 minute	4.24 ± 0.23	română
Clonare vocală - 6 minute	3.89 ± 0.24	română
Clonare vocală - două minute	3.64 ± 0.23	română
Voci naturale(Chen et al., 2019)	4.89 ± 0.045	germană și franceză
Învățare prin transfer din engleză folosind Tacotron - 25 minute	4.01 ± 0.085	germană și franceză
Învățare prin transfer din engleză folosind Tacotron – 15 minute	3.48 ± 0.119	germană și franceză
Voci naturale (Shen et al., 2018)	4.582 ± 0.053	engleză
Tacotron 2	4.526 ± 0.066	engleză
Metoda prin concatenare	4.166 ± 0.091	engleză

Pentru a compara rezultatele cu alte sisteme de TTS în română, au fost strânse date întrebând subiecții ce clip preferă dintre unul sintetizat cu TTS Eletron, altul folosind TTS-ul din cadrul proiectului SWARA (*Romanian TTS - Online text-to-speech system*) și TTS-ul din cadrul Google Translate (*Google Traducere*). Din experimentele efectuate, reiese că metoda principală prezentată este optimă pentru sinteza generală, date fiind seturile de date disponibile. Moduri de a implementa sistemul educațional inteligent propus în Ouatu & Gîfu (2020) și integrarea lucrării de față în acesta reprezintă direcții viitoare de cercetare. Vom explora, de asemenea, modalități de îmbunătățire a ambelor tipuri de sisteme de sinteză. În vederea obținerii de rezultate superioare, implementarea modelelor prezentate de Liu et al., (2019) și Zhu et al. (2019) sunt prioritare. Sinteza emoțională a vocii și copierea stilului de vorbire constituie o altă direcție de cercetare.

3.8 **Concluzii**

Obiectivul principal al proiectului ReTeRom a fost dezvoltarea de resurse multimodale aliniate și a unor tehnologii compatibile. Atingerea acestui obiectiv și calitatea acestor resurse și tehnologii pot fi cel mai bine demonstrate prin utilizarea lor în contexte realiste, răspunzând unor necesități ale comunității de cercetători, studenți și chiar ale publicului larg.

Rapoartele 3.4 și 4.1 prezintă o serie de 8 aplicații care au ca obiectiv valorificarea resurselor și tehnologiilor dezvoltate în ReTeRom. Stadiile lor de dezvoltare și progresele făcute chiar înainte de finalizarea proiectului, în toate cazurile cu studenți sau alte persoane voluntare, arată interesul comunității în valorificarea acestor resurse.

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

Bibliografie

[1] Prenger, R., Valle, R., & Catanzaro, B. (2018). *Waveglow: A Flow-Based Generative Network for Speech Synthesis*.

[2] Procter, M. (n.d.). *Freudbot: An Investigation of Chatbot Technology in Distance Education*.

[3] Radu, I. Raport Activitate A1.5: *Definirea specificațiilor funcționale și arhitecturale ale platformei integrate și configurabile de prelucrare a textelor*, proiectul ReTeRom

[4] Radu, I. Raport Activitate A1.6: *Definirea modulelor software și a serviciilor oferite de proiect; identificarea adaptărilor pentru modulele NLP existente și a modulelor noi necesare,* proiectul ReTeRom.

[5] Boroș, T., Dumitrescu, Ș., Pais, V. *Tools and resources for Romanian text-to-speech and speech-to-text applications*, 2018.

[6] Zamfirescu, A.N., Rebedea, T.E. *Identificarea entităţilor, citatelor şi evenimentelor în ştiri şi texte din Web-ul social în limba română*, în Revista Română de Interacţiune Om-Calculator 6 (2) 2013, 169-192.

[7] Burileanu, C., Cucu, H. Raport Activitate A1.11: *Studiul metodelor din literatură pentru alinierea transcrierilor aproximative cu semnalul de vorbire*, proiectul ReTeRom.

[8] Aldabbagh, O., Mohsen, K. Design and Implementation of Online Location Based Services Using Google Maps for Android Mobile. International Journal of Computer Networks and Communications Security. 2. pp. 113-118, 2014

[9] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (Vols. 2018-April). https://doi.org/10.1109/ICASSP.2018.8461368

[10] Stan, A., Dinescu, F., Tiple, C., Meza, S., Orza, B., Chirila, M., & Giurgiu, M. (2017). The SWARA speech corpus: A large parallel Romanian read speech dataset. *2017 9th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2017*, 1–6. https://doi.org/10.1109/SPED.2017.7990428

[11] Stan, A., Yamagishi, J., King, S., & Aylett, M. (2011). The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication*, *53*(3), 442–450. https://doi.org/10.1016/j.specom.2010.12.002

[12] Pistol, I., Scutelnicu, A., Onofrei M., Gîfu D., Boghiu Ș., Raport Activitate 3.4: *Proiectare de aplicații de exploatare a corpusului bimodal și a tehnologiilor de prelucrări textuale și voce, create în proiectele P2, P3, P4,* proiectul RETEROM

[13] Ouatu, B., Gîfu, D. *Chatbot, the Future of Learning?* In: Proceedings of the 5th International Conference on Smart Learning Ecosystems and Regional Development (SLERD 2020), in Ludic, Co-design and Tools Supporting Smart Learning Ecosystems and Smart Education, Springer, 2020, pp. 263-268.

Raport ştiinţific - tehnic proiect component TEPROLIN

Date de identificare proiect

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

1. Introducere

Proiectul „Tehnologii pentru procesarea limbajului natural - text” (TEPROLIN), proiect component al proiectului complex nr. 73PCCDI/2018 „Resurse și tehnologii pentru dezvoltarea interfețelor om-mașină în limba română” (ReTeRom), a avut ca obiectiv dezvoltarea unui set de tehnologii avansate pentru procesarea textelor din corpusurile bimodale în limba română: analiză morfologică și sintactică a textelor (lemă, parte de vorbire, relație de dependență sintactică, etc.) aplicată textelor colectate într-un alt proiect component al ReTeRom, anume proiectul „Corpus bimodal pentru limba română adnotat pe multiple niveluri” (COBILIRO). Pentru a facilita dezvoltarea sistemelor de recunoaștere a vorbirii și sinteză a vorbirii planificată în ReTeRom, au fost necesare procesări fonetice alte textelor cum ar fi transcrierea fonetică a cuvintelor, silabația cu detectarea silabei accentuate.

Prezentul raport de cercetare descrie activitățile care s-au desfășurat în anul 2021 în cadrul proiectului TEPROLIN

2. Module software și servicii oferite de TEPROLIN

În cele ce urmează, vom inventaria și detalia modulele software existente la partenerii proiectului sau open-source care implementează operațiile enumerate mai sus *împreună cu adaptările și îmbunătățirile operate în proiectul TEPROLIN*, pentru a obține varianta finală a platformei de prelucrări ale textelor.

O decizie de design cu impact major în reducerea timpului de execuție a unei operații a constat în introducerea de dependențe de tip graf între operațiile de prelucrare a textelor. Acest tip de a preciza ce operații trebuie rulate mai întâi pentru a se putea rula operația de prelucrare dorită e mult mai eficient decât tipul de rulare în secvență pe care l-am gândit inițial. De exemplu, pentru a putea rula operația de adnotare cu etichete morfo-sintactice, nu sunt necesare operații precum silabificarea sau detecția accentului. Modulul Python 3 în care sunt precizate aceste operații cât și graful de dependență a acestora este https://github.com/racai-ai/TEPROLIN/blob/master/TeproAlgo.py, iar metoda care construiește graful se numește _assignAlgorithmsToOperations().

2.1 Procesări utile aplicațiilor de recunoaștere automată a vorbirii și de sinteza vorbirii

Operațiile de despărțire în silabe, transcriere fonetică și detectarea silabei accentuate sunt efectuate cu librăria MLPLA (Boroș et al., 2018). Aceasta este scrisă în Java 15 și, pentru a o putea utiliza în TEPROLIN (Python 3), am folosit mecanismul IPC sockets într-un server care expune toate funcționalitățile librăriei pe un socket legat de IP-ul local al mașinii care găzduiește platforma (https://github.com/racai-ai/TEPROLIN/blob/master/ttsops/MLPLAServer/src/main/java/ro/racai/reterom/tts/MLPLAServer.java).

MLPLA primește la intrare o frază în limba română și efectuează următoarele operații: tokenizare, adnotare cu etichete morfosintactice, detectarea grupurilor sintactice nerecursive, despărțire în silabe, transcriere fonetică și detectarea silabei accentuate. Rezultatele ultimelor trei operații sunt preluate de TEPROLIN și transferate tokenilor corespunzători.

2.2 Procesări textuale standard

În această categorie intră operațiile de segmentare la nivel de frază și unitate lexicală, adnotarea cu etichete morfosintactice, lematizarea, analiza cu relații de dependență sintactică și analiza sintactică de suprafață. TEPROLIN conține următoarele aplicații care furnizează aceste operații: NLP-Cube (Boroș et al. 20018b), TTL (Ion, 2007) și UD-Pipe (Straka et al., 2016). Ele sunt descrise în raportul final in extenso al proiectului TEPROLIN.

Pornind de la prima integrare, s-a urmărit realizarea unei integrări la un nivel superior a modulelor disponibile prin intermediul serviciul web TEPROLIN cu diferitele resurse adiționale existente, cum ar fi interfețele de căutare din cadrul corpusului CoRoLa (Barbu et al., 2018), reprezentările distribuționale ale cuvintelor antrenate automat pe corpusul CoRoLa (Păiș, Tufiș. 2018), WordNet-ul românesc (Tufiș și Mititelu, 2014). Totodată, platforma a urmărit expunerea funcționalităților prin intermediul unei interfețe grafice disponibilă online de tip „portal web”, care să permită interacțiunea facilă a utilizatorilor cu tehnologiile expuse, fără a fi necesare cunoștințe de programare.

Având în vedere necesitatea prelucrării automate a unui corpus de mari dimensiuni, s-a urmărit și paralelizarea lanțurilor de procesare utilizate, în scopul reducerii timpului total de prelucrare. Acest lucru a condus la implementarea unui mecanism de control al procesărilor sub formă de joburi care pot fi apoi distribuite la nivelul unui server pe mai multe procese sau la nivelul unei rețele de calcul pe mai multe noduri.

3. Soluții „ready-to-use” ale platformei de prelucrare a textelor TEPROLIN

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

1. Pentru testare cu fraze scurte (pentru evaluarea performanțelor sau analiza unor fraze) cu efectuarea tuturor operațiilor disponibile, se poate accesa link-ul https://relate.racai.ro/index.php?path=teprolin/complete și se pot vizualiza adnotările făcute;
2. Pentru rularea unor operații la alegere pe fraze scurse, folosind algoritmii preferați, se poate accesa link-ul https://relate.racai.ro/index.php?path=teprolin/custom;

3. Pentru adnotarea corpusurilor cu mai mult de 1000 de cuvinte, se poate solicita acces la platforma RELATE care rulează containerul TEPROLIN pe mai multe fire de execuție;
4. Ca modul Python 3, clonând repository-ul https://github.com/racai-ai/TEPROLIN și urmând indicațiile de instalare din fișierul https://github.com/racai-ai/TEPROLIN/blob/master/README.md. Recomandăm ca toate pachetele necesare să fie instalate într-un mediu dedicat Python 3 (eng. „virtual environment”), executând comenzile:

   ```
   python3 -m venv /calea/către/mediul/dedicat/teprolin
   pip3 install -r requirements.txt
   ```
5. Ca un container Docker în care toate operațiile de instalare care sunt descrise în fișierul README.md au fost deja implementate în fișierul https://github.com/racai-ai/TEPROLIN/blob/master/Dockerfile. Pentru o utilizare rapidă a containerului (fără a mai fi construit), se poate executa comanda `docker pull raduion/teprolin:1.1`. Dacă se dorește recompilarea containerului, se va executa comanda `docker build --pull --rm -f "Dockerfile" -t teprolin:1.1 "."`.

4. Crearea și validarea unui lexicon specific corpusului bimodal colectat în ReTeRom

Această activitate a avut rolul de a extrage lexicoane din corpusurile orale existente la membrii proiectului, de a le corecta atunci când există erori de diverse feluri și de a le face accesibile publicului larg, pe site-ul proiectului. Rezultatele activității, un raport asupra modului de lucru și a lexicoanelor extrase din diversele sub-corpusuri precum și lexiconul extras au fost puse la dispoziție public pe site-ul proiectului.

4.1 Crearea lexiconului specific corpusurilor bimodale existente la membrii proiectului ReTeRom

Transcrierile (componenta textuală a) corpusurilor bimodale disponibile la toți partenerii proiectului au fost colectate și analizate. Corpusul provenind de la ICIA și UAIC (**CoRoLa - oral**) conține texte provenind din RoWikipedia, știri, interviuri pe teme de actualitate, povești. Transcrierile înregistrărilor sunt tokenizate, lematizate și adnotate morfologic, dar calitatea transcrierii și implicit a adnotării nu este omogenă. S-au observat erori de ortografie și punctuație, precum și convenții diferite de transcriere (de exemplu, cuvintele în limbi străine sunt uneori transcrise așa cum se aude, alteori conform convențiilor de scriere din limba respectivă).

Corpusul provenit de la UPB (grupând corpusurile **RSC, SSC-train și SSC-eval**) conține texte ce transcriu înregistrări de știri, talkshow-uri, interviuri, literatură, vorbire spontană. Textele sunt corecte ortografic, dar nu conțin majuscule și nici punctuație. În plus, o mare parte din corpus reprezintă transcrieri automate din emisiuni de știri, bazate pe segmente audio produse tot automat (vezi Cucu et al., 2015) astfel încât să filtreze secvențele non-speech și să conțină rostirea unui singur vorbitor. Transcrierile rezultate sunt arareori propoziții complete. În plus, în altă secțiune a corpusului predomină rostiri ale unor liste de cuvinte, fiecare într-un enunț separat. În consecință, cea mai mare parte a unităților de transcriere din corpusul UPB

reprezintă secvențe sub-propoziționale, fără majuscule și punctuație, motiv pentru care nu se pretează unei adnotări automate cu un tagger.

Corpusul provenit de la partenerul UTCN conține sub-corpusurile **SWARA** (înregistrări în studio cu transcriere ortografică de calitate), **MARA** (din înregistrarea nuvelei Mara de către un vorbitor profesionist de gen feminin, cu transcriere ortografică aproximativă) și **Adevărul.ro** (provenit dintr-un corpus de text compus din articole din ziarul Adevărul - versiunea online, calitate ortografică). Pentru că toate transcrierile sunt ortografice, am putut folosi un instrument de tokenizare, lematizare și adnotare morfologică automată.

4.2 Validarea lexiconului

Această etapă de lucru presupune **evaluarea manuală** a fiecărui cuvânt și, dacă acesta este cuvânt corect în limba română (i.e. există sau este o creație ad-hoc posibilă conform regulilor morfologice ale limbii, are toate diacriticele necesare), stabilirea a trei elemente:

- lema sa (forma de dicționar);
- descrierea morfosintactică (PoS tag);
- restul paradigmei flexionare: pentru fiecare formă din paradigmă se notează lema și descrierea morfosintactică.

Erorile frecvent întâlnite și care au dus la eliminarea cuvintelor găsite în corpus din lexiconul creat au fost:

- scrieri greșite: litere inversate, litere lipsă, litere în plus;
- probleme cu diacriticele: lipsa totală a acestora sau existența parțială;
- probleme de segmentare: scrierea împreună a două cuvinte, introducerea unui blanc în interiorul cuvântului.

4.3 Îmbogățirea lexiconului și corectarea erorilor detectate

Lexiconul a fost completat cu informația de segmentare în silabe, de accent și cu transcrierea ortografică. A rezultat o resursă cu 346.074 intrări cuprinzând informație standardizată asociată unei forme ocurență: lema (forma de dicționar a cuvântului), eticheta morfo-sintactică în format MSD (Erjavec, 2004), împărțirea în silabe a formei ocurență, marcarea accentului (printr-un apostrof) în fața vocalei accentuate și transcrierea fonetică a formei ocurență, între paranteze drepte. Pentru corectarea erorilor de transcriere fonetică s-au implementat numeroase reguli (vezi raportul detaliat).

Rețeaua neuronală folosită pentru evaluarea lexiconului are ca scop predicția (în format text) concurentă a silabificării, accentului și transcrierii fonetice pornind doar de la forma (ortografică a) cuvântului, în lipsa unui context de utilizare.

În familia arhitecturilor neuronale utilizate pentru a învăța funcții „secvență la secvență” (eng. „sequence-to-sequence”), rețelele neuronale convoluționale (CNN, Gehring et al. 2017) și rețelele bazate pe atenție (Vaswani et al., 2017) sunt cele care au demonstrat cea mai bună acuratețe în probleme de procesare a limbajului natural (vezi (Stan, 2020) și raportul detaliat pentru detalii despre implementarea acestora).

În procesul de evaluare, setul de date a fost divizat în loturi de de 70%-10%-20%, dedicate proceselor de antrenare, validare și, respectiv, testare. Seturile de antrenare și

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

Ne-a interesat îmbunătățirea performanței sistemului de predicție atunci când setul de date crește gradat și identificarea momentului de platou al acestei îmbunătățiri. În acest scop, au fost selectate partiții aleatoare ale datelor pornind de la un minim de 5.000 de intrări.

Am urmărit de asemenea îmbunătățirea performanței în funcție de cantitatea de date atunci când selecția unui set redus de date se face asigurându-ne că dispunem de întreaga acoperire a lexiconului la nivel de leme.

Drept metrici de evaluare, am folosit *word error rate* (WER, „rata de eroare la nivel de cuvânt") și *character error rate* (CER, „rata de eroare la nivel de caracter"), evaluate pe șiruri de caractere care includ marcajele de silabificare și accent. CER a fost măsurată folosind distanța Levenshtein (Levenshtein, 1966) între secvența de caractere prezisă și cea țintă. În plus, ne-am dorit să evaluăm erorile introduse de fiecare sarcină de predicție, motiv pentru care am calculat WER și CER eliminând din predicția concurentă informația furnizată de una sau mai multe dintre sarcini.

Evaluarea prealabilă, în care comparam resursa noastră cu cele folosite anterior, ne arată că, cu lexiconul ReTeRom, putem obține o rată a erorii **WER de 3,08** și o rată **CER de 1,08** pentru predicția concurențială, atunci când folosim toată informația disponibilă (inclusiv eticheta morfosintactică), ceea ce reprezintă o reducere importantă a ratei erorii de la 10,47 WER și 3,93 CER, obținute cu resursele utilizate anterior la UTCN. Asta demonstrează încă o dată valoarea unei resurse lexicale de dimensiuni mari, cu adnotări complexe și validată manual.

4.4 Analiza erorilor sistemului RAV

Evaluarea la nivel de cuvânt arată, așa cum era de așteptat, o îmbunătățire a tuturor criteriilor de analiză: cuvinte recunoscute greșit, cuvinte inserate eronat, cuvinte netranscrise.

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

- Nume de entități: am extras din diverse surse liste separate de nume de persoane, nume de locuri, nume de organizații/firme/etc., cu scopul de a îmbogăți lexiconul sistemului de ASR. Listele create sunt disponibile pe site-ul proiectului, la categoria Resurse.
- Cuvinte cu cratimă: folosind un lexicon intern ICIA, am validat o parte dintre cuvintele cu cratimă, adică acelea corecte, dar absente din lexiconul ASR.
- Cuvinte obișnuite: s-a folosit aceeași procedură ca la cuvintele cu cratimă. Pentru aceasta, am folosit o măsură de similaritate a tuturor tipurilor de cuvinte (Levenshtein distance) pentru a indica forma cea mai apropiată din listele de nume și lexiconul folosite. Aceste aproximări au fost validate manual și s-au creat liste de corecturi propuse pentru îmbunătățirea transcrierilor existente în corpusul din platforma CoBiLiRo.

5. Concluzii

Rezultatele **imediat utilizabile** ale proiectului TEPROLIN sunt următoarele:

1. Platforma integratoare de tehnologii de prelucrare a limbii române RELATE, accesibilă la https://relate.racai.ro. Gândită inițial pentru a facilita testarea platformei de prelucrare a textelor TEPROLIN, a ajuns rapid la un portal de prezentare și testare a ultimelor tehnologii de prelucrare a limbii române dezvoltate la Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu” al Academiei Române.
2. Containerul Docker (https://hub.docker.com/r/raduion/teprolin) și repository-ul TEPROLIN (https://github.com/racai-ai/TEPROLIN) cu care se poate utiliza platforma de prelucrare a textelor românești TEPROLIN dezvoltată în proiect. Această platformă a fost foarte bine testată și este o platformă de procesare a textelor stabilă și configurabilă.
3. Un lexicon (https://www.racai.ro/p/reterom/rapoarte/rezultate.total.final.zip), **validat manual**, cu transcrieri fonetice și silabație pentru mai mult de 340 de mii de cuvinte în limba română. Un astfel de lexicon fie poate fi folosit direct pentru aflarea transcrierilor fonetice fie poate fi utilizat în programe de învățare automată care să genereze automat transcrieri fonetice pentru cuvinte necunoscute.

Pe lângă cele trei rezultate enumerate mai sus, proiectul TEPROLIN a fost prezent în cadrul a 18 modalități de diseminare din care două cărți, zece articole publicate la conferințe (din care șapte indexate ISI), un articol publicat într-o revistă indexată ISI, patru conferințe invitate și un articol de popularizare a științei.

Având în vedere toate cele prezentate mai sus, considerăm că proiectul TEPROLIN și-a atins toate obiectivele propuse, a produs rezultatele scontate și, nu în ultimul rând, a format o echipă nouă, extinsă, cu toți membrii proiectelor componente ale ReTeRom care își propune să continue inovația în ce privește tehnologiile de prelucrare a limbii române scrise și vorbite.

6. Structura ofertei de servicii de cercetare şi tehnologice

ICIA oferă pe platforma ERRIS serviciile de cercetare şi tehnologice enumerate în Tabelul 1.

Tabelul 1. *Servicii de cercetare şi tehnologice oferite de ICIA*

Serviciu	**Detalii**
Containerul Docker cu fluxul de prelucrari TEPROLIN	https://hub.docker.com/r/raduion/teprolin
Repository-ul TEPROLIN	https://github.com/racai-ai/TEPROLIN
Un lexicon unic pentru tehnologia limbii române	https://www.racai.ro/p/reterom/rapoarte/rezultate.total.final.zip
Acces la portalul tehnologiilor pentru prelucrarea limbii române	https://relate.racai.ro/index.php

7. Locuri de muncă susţinute prin program

Echipa de cercetare a Universităţii Politehnica din Bucureşti pentru proiectul component TADARAV este prezentată în Tabelul 2.

Tabelul 2. *Echipa de cercetare ICIA*

Nr.	Nume	Calitatea	Poziţia	Normă
1	Dan Tufis	CS1	Director Proiect complex	Parţială
2	Verginica Barbu Mititelu	CS2	Membru cercetător	Parţială
3	Radu Ion	CS2	Membru cercetător	Parţială
4	Elena Irimia	CS3	Membru cercetător	Parţială
5	Maria Carp (Mitrofan)	CS3	Membru cercetător	Parţială
6				

Situaţia noului cercetător, cu normă întreagă

Începând cu data de 31.11.2020 au încetat raporturile de muncă în cadrul proiectului ReTeRom pentru noul cercetător Vasile Păiș. Pe 01.12.2020 Vasile Păiș a fost angajat în ICIA cu normă întreagă, continuând să activeze și în proiectul ReTeRom.

Valorificarea şi îmbunătăţirea competenţelor şi resurselor existente la nivelul consorțiului

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

Toate obiectivele incluse în plan pentru această etapă si global pe acest subproiect au fost realizate integral.

6. Referințe bibliografice

Adda-Decker, M., and L. Lamel, 2000. The use of lexica in automatic speech recognition. In F. van Eynde & D. Gibbon (Eds.), *Lexicon Development for Speech and Language Processing*, Kluwer Academic Publishers.

Barbu, Ana-Maria (2008) "Romanian Lexical Databases: Inflected and Syllabic Forms Dictionaries." In Proceedings of LREC 2008.

Barbu Mititelu, Verginica, Ion, Radu, Simionescu, Radu, Irimia, Elena and Perez, Cenel-Augusto (2016). The Romanian Treebank Annotated According to Universal Dependencies. In Proceedings of HrTAL2016, Dubrovnik, Croatia, 29 September - 1 October 2016.

V. Barbu Mititelu, D. Tufiș, E. Irimia (2018) The Reference Corpus of Contemporary Romanian Language (CoRoLa). In *Proceedings of the 11th Language Resources and Evaluation Conference – LREC'18*, Miyazaki, Japan, European Language Resources Association (ELRA).

Boroș, Tiberiu, Ștefan Daniel Dumitrescu and Vasile Păiș. (2018). *Tools and resources for Romanian text-to-speech and speech-to-text applications*. arXiv:1802.05583v1 [cs.CL]

Boroș, Tiberiu, Ștefan Daniel Dumitrescu and Ruxandra Burtica. (2018b). *NLP-Cube: End-to-End Raw Text Processing With Neural Networks*. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics. pp. 171--179. October 2018.

Buchholz, Sabine and Marsi, Erwin (2006). CoNLL-X shared task on Multilingual Dependency Parsing. In Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X), pages 149–164, New York City, June 2006. © 2006 Association for Computational Linguistics.

Dan Cristea, Cristian Pădurariu, Șerban Boghiu, Daniela Gîfu, Mihaela Onofrei, Diana Trandabăț, Ionuț Cristian Pistol, Anca Bibiri and Andrei Scutelnicu, The CoBiLiRo

project: Building and Distributing a Bimodal Corpora for the Romanian Language In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language – CONSILR 2019, pages 13-24.

Cucu, Horia, Andi Buzo, Laurent Besacier, Corneliu Burileanu (2015). Enhancing ASR Systems for Under-Resourced Languages through a Novel Unsupervised Acoustic Model Training Technique, in *Advances in Electrical and Computer Engineering*, vol. 15, no. 1, pp. Dan

Cristea, Cristian Pădurariu, Șerban Boghiu, Daniela Gîfu, Mihaela Onofrei, Diana Trandabăț, Ionuț Cristian Pistol, Anca Bibiri and Andrei Scutelnicu, The CoBiLiRo project: Building and Distributing a Bimodal Corpora for the Romanian Language In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language – CONSILR 2019, pages 13-24.63-68, ISSN: 1582-7445, doi:10.4316/AECE.2015.01009.

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

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

Fielding, Roy Thomas (2000). "Chapter 5: Representational State Transfer (REST)". *Architectural Styles and the Design of Network-based Software Architectures* (Ph.D.). University of California, Irvine.

Finkel, Jenny Rose, Trond Grenager și Christopher Manning. 2005. *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling*. In Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363--370.

Ion, Radu (2007). Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză și română. Teză de doctorat, Academia Română, mai 2007, 148 de pagini.

Ion, Radu, Valentin Gabriel BADEA, George CIOROIU, Verginica BARBU MITITELU, Elena IRIMIA, Maria MITROFAN, Dan TUFIȘ, 2020. A Dialog Manager for Micro-Worlds, *Studies in Informatics and Control*, ISSN 1220-1766, vol. 29(4), pp. 411-420, https://doi.org/10.24846/v29i4y202003

Ion, Radu, 2018. “TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian.” In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018), November 22-23, 2018, Iași, Romania. (2018)

Klavans, J.L., and E. Tzoukermann, 1994. Machine-readable Dictionaries in Text-to-speech Systems. In *Proceedings of the 15th Conference on Computational Linguistics* - Volume 2 (COLING), p. 971-975.

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

Liberman, M. Y. and K. W. Church, 1992. Text Analysis and Word Pronunciation in Text-to-speech Synthesis. In S. Furui, M.M. Sondhi (eds.), *Advances in Speech Signal Processing*, New York: Marcel Dekker, p. 791-831.

Mitrofan, Maria, Verginica Barbu Mititelu și Grigorina Mitrofan. 2019. *MoNERo: a Biomedical Gold Standard Corpus for the Romanian Language*. În Proceedings of the BioNLP 2019 workshop, pages 71–79Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics.

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

Păiș, Vasile (2019). Contributions to semantic processing of texts; Identification of entities and relations between textual units; Case study on Romanian language. PhD Thesis, Romanian Academy, December 9th, 2019, 123 pages.

Păiș, V., D. Tufiș, R. Ion (2019). ”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.

Păiș, V., Tufiș, D. (2018) Computing distributed representations of words using the COROLA corpus. In *Proceedings of the Romanian Academy*, Series A, Volume 19, Number 2/2018, pp. 403–409.

Stan, Adriana, Junichi Yamagishi, Simon King, and Matthew Aylett. 2011. „The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate” In Speech Communication vol.53 442-450.

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

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

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, “Brat: a Web-based Tool for NLP-Assisted Text Annotation”, in Proceedings of the Demonstrations Session at EACL 2012.

Straka, Milan, Jan Hajič and Jana Straková. (2016). *UD-Pipe: trainable pipeline for processing CoNLL-Ufiles performing tokenization, morphological analysis, POS tagging and parsing*. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association, Portorož, Slovenia.

Toma, Ştefan-Adrian, Adriana Stan, Mihai-Lică Pura, and Traian Bârsan. 2017. "MaRePhoR—An open access machine-readable phonetic dictionary for Romanian." International Conference on Speech Technology and Human-Computer Dialogue (SpeD). IEEE. Tufiș, Dan (1999). *Tiered Tagging and Combined Language Models Classifiers*. În TSD '99 Proceedings of the Second International Workshop on Text, Speech and Dialogue, pp 28--33, Springer-Verlag London, ISBN:3-540-66494-7.

Tufiș, Dan (2000). Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. International Conference on Language Resources and Evaluation LREC’2000, Athens, 2000, pp. 1105-1112

Tufiș, D., and Mititelu, V.B. (2014) The Lexical Ontology for Romanian, pages 491–504. Springer.

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

Zhang, Aston, Zachary C. Lipton, Mu Li, and Alexander J. Smola. 2020. Dive into Deep Learning. https://d2l.ai.

Raport ştiinţific - tehnic proiect component TADARAV

Date de identificare proiect

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

1. Rezumatul etapei

A patra etapă a proiectului TADARAV a avut o singură activitate (Activitatea 4.5 Management şi diseminare) cu două obiective ce au fost realizate în proporție de 100%:

1. redactarea și depunerea a cel putin un articol științific la conferințe sau reviste indexate ISI
2. redactarea și depunerea unei cereri de brevet

În cadrul ultimei activități a proiectului TADARAV metodele de estimare a încrederii proiectate în cadrul etapei 3/2020 au fost aplicate și evaluate sistematic pe seturile de date în limba română. Rezultatele acestei activități au fost diseminate într-un articol științific transmis spre evaluare și publicare în jurnalul ISI Buletinul științific UPB, Seria C. Recenziile primite pentru acest articol sunt pozitive și solicită o revizie minoră (modificări de exprimare și redactare), urmând ca articolul să fie publicat până la jumătatea anului.

Metodele îmbunătățite de normalizare a textelor și creare a modelelor de limbă pentru recunoașterea automată a vorbirii proiectate în etapa anterioară a proiectului au fost prezentate într-un articol științific prezentat la Conferința Linguistics Resources and Tools for Natural Language Processing ce a avut loc la București în luna decembrie 2020.

Metodele de estimare a încrederii și metodologia de integrarea a lor într-un sistem mai complex care să permită adnotarea automată a seturilor de date de vorbire au făcut obiectul unei cereri de brevet depusă la Oficiul de Stat pentru Invenții și Mărci (OSIM).

Astfel, în urma activității A4.5 din etapa 4/2021 a proiectului TADARAV, au rezultat toate livrabilele asumate de consorţiu la începutul acestei etape:

- Un articol științific la conferința ConsILR 2020;
- Un articol științific în revista Buletinul științific UPB, Seria C (indexată ISI);
- O cerere de brevet.

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

- C. Manolache, A.-L. Georgescu, H. Cucu, V. Barbu Mititelu, C. Burileanu, *"Improved Text Normalization and Language Models for SpeeD's Automatic Speech Recognition System,"* in the Proceedings of the 14th International Conference on Linguistics Resources and Tools for Natural Language Processing, Bucharest, Romania, 2020.
- A. Caranica, D. Oneaţă, H. Cucu, C. Burileanu, *"Confidence Estimation For Lattice-based And Lattice-free Automatic Speech Recognition,"* in UPB Scientific Bulletin, Series C, 2021.

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 1.

Tabelul 1. *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 2.

Tabelul 2. *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

Situaţia celor trei posturi de noi cercetători, cu normă întreagă

Începând cu data de 31.11.2020 au încetat raporturile de muncă în cadrul proiectului ReTeRom pentru noii cercetători: Dan Theodor ONEAŢĂ, Gheorghe POP și Cristian MANOLACHE. Pe 01.12.2020 Dan Theodor ONEAŢĂ a fost angajat în UPB cu normă întreagă în cadrul proiectului VORBIS (contract PD97/ 2020), iar Cristian MANOLACHE a fost angajat în UPB cu normă întreagă în cadrul proiectului CLARA (contract 295PED/ 2020). Domnul Gheorghe POP nu a dorit continuarea raporturile de muncă cu UPB, postul respectiv lui fiind scos la concurs. Postul a fost ocupat începând cu data de 01.04.2021 de Lucian-Alexandru GEORGESCU şi va fi finanţat până la finalizarea perioadei de sustenabilitate a proiectului (până pe data de 30.11.2022) din Programul UPB Proof-of-Concept.

4. **Valorificarea şi îmbunătăţirea competenţelor şi resurselor existente la nivelul** consorțiului

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

Raport ştiinţific - tehnic proiect component SINTERO

Date de identificare proiect

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

1. Introducere

Proiectul SINTERO: Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate a avut ca obiectiv principal crearea de sisteme de sinteză text-vorbire în limba română folosind tehnologii bazate pe rețele neuronale profunde. Pe lângă crearea acestor sisteme, s-a avut în vedere și extinderea resurselor audio disponibile în vederea antrenării lor, precum și studiul modalităților prin care necesarul de resurse audio de la un vorbitor poate fi redus prin adaptarea rapidă a modelelor acustice.

În cadrul acestui raport vor fi detaliate cele mai recente metode de adaptare dezvoltate și diseminate în cadrul SINTERO. Rezultatele metodelor de adaptare au fost trimise spre diseminare la conferințele EUSIPCO 2021 și KES 2021, iar articolele aferente sunt anexate la acest raport. În partea a doua a raportului este descrisă metoda de facilitare a accesului utilizatorilor la sistemele create prin intermediul unei interfețe web denumită RoNNA - Romanian Neural Network API și care vine în continuarea interfeței Romanian TTS (www.romaniantts.com) ce prezenta sistemele bazate pe modele Markov. O primă versiune a RoNNA a fost prezentată în cadrul livrabilului D3.17, ea fiind ulterior extinsă cu o serie de alte identități vocale și tehnologii de sinteză, precum și cu posibilitatea de a realiza prelucrarea doar la nivel de text prin expunerea transcrierii fonetice, a silabificării și a accentului lexical.

2. Adaptarea sistemelor de sinteză la o nouă identitate vocală

Sistemele de sinteză pot fi antrenate în scenariul de singur sau vorbitori multipli. Acesta din urmă are avantajul de a incorpora într-un singur model mai multe voci, iar în timpul inferenței vocea dorită pentru sinteză poate fi selectata. Pentru a crea astfel de sisteme cea mai frecvent folosită metodă este cea de reprezentări vectoriale pentru vorbitori. Aceste reprezentări sunt învățate în timpul antrenării.

Sistemele de sinteză bazată pe rețele neuronale pentru a obține voci de bună calitate au nevoie de cantități mari de date. Acest fapt este valabil și pentru sistemele de vorbitori multipli. Când aceste date nu sunt disponibile de la fiecare vorbitor antrenarea unui sistem de vorbitori multipli este greu de realizat sau rezultă într-un grad de similaritate mai scăzută pentru vorbitori și naturalețe redusă.

Pentru a aborda această problemă mai multe sisteme de sinteză sunt antrenate cu diferite arhitecturi și cantități de date. Arhitecturile analizate sunt bazate pe rețele neuronale profunde de tip recurent (sistemul Tacotron2) și de tip convoluțional (DC-TTS). Acestea sunt evaluate în scenariul de sistem de sinteză pentru vorbitori multipli pentru a beneficia de vocile disponibile în corpusurile SWARA și SWARA 2.0. Suplimentar, Tacotron2 este analizat și în scenariul de utilizare a diferitelor tipuri de reprezentare a informației textuale.

2.1. Adaptarea identității vocale folosind sistemul Tacotron2

Sistemul de sinteză Tacotron2 (Shen et al., 2018) este bazat pe rețele recurente și a raportat un scor MOS de 4.53 care este foarte aproape de vorbirea umană. În cadrul experimentelor bazate pe arhitectura de Tacotron2 implementarea făcută disponibilă de către cei de la NVIDIA pentru acest instrument[3] a fost punctul de plecare. Această implementare a fost extinsă cu funcționalități de antrenare cu vorbitori multipli pe baza reprezentărilor vectoriale ale vorbitorilor (en. embedding). Aceste reprezentări vectoriale sunt învățate în timpul antrenării, și sunt anexate la ieșirea codorul de text, care este folosită mai apoi la intrarea decodorului audio. Adăugarea de embedding de vorbitori este inspirată din implementarea instrumentului Mellotron realizat la fel de echipa NVIDIA[4]. Vocoderul folosit pentru experimente este WaveGlow (Prenger et al., 2019), un vocoder bazat pe fluxuri de normalizare și care poate realiza sinteza de forma de undă mai rapid decât timp real.

2.1.1. Scenarii de antrenare

Sistemele de sinteză bazată pe Tacotron2 folosind date de la mai mulți vorbitori sunt antrenate și evaluate în două scenarii fiecare folosind trei tipuri diferite de reprezentare a textului de intrare: transcriere ortografică, transcriere fonetică și transcriere fonetică augmentată cu informații de silabificare și accent lexical.

1. Scenariul 1 (ID: **MSPK**): antrenare de sistem de vorbitori multipli pe baza de embedding de vorbitor. Identitatea vorbitorului este anexată textului de intrare pentru fiecare propoziție.
2. Scenariul 2 (ID: **ADAPT**): antrenarea unui sistem ce folosește datele audio de la toți vorbitorii, dar nu specifică identitățile fiecăruia, creând astfel o voce medie a identităților văzute în setul de antrenare. Sistemul astfel antrenat pentru aproximativ 200 de epoci, este mai apoi adaptat către fiecare vorbitor. Adaptarea constă în antrenarea în continuare a modelului pentru un număr predefinit de epoci. Pentru adaptare am folosit diferite cantități de date (5, 50, respectiv 200 de propoziții) de la fiecare vorbitor, și antrenarea a fost continuată pentru 50 sau 100 de epoci.

Date de antrenare

Corpusurile audio SWARA și SWARA 2.0 au fost folosite pentru antrenarea sistemelor de sinteză. Un număr de 41 de vorbitori au fost selectați dintre care 18 aparținând corpusului SWARA și restul de 23 aparținând SWARA 2.0. Dintre aceste voci 25 sunt feminine (11 din SWARA și 14 din SWARA 2.0) și 16 masculine (7 din SWARA și 9 din SWARA 2.0). 37 dintre vorbitori au fost folosiți pentru antrenare, o voce feminină și una masculină din fiecare corpus a fost rezervată pentru validare. Datele din SWARA au fost înregistrate în condiții de studio, iar

[3] https://github.com/NVIDIA/tacotron2

[4] https://github.com/NVIDIA/mellotron

cele din SWARA 2.0 în afara condițiilor de studio, pentru care vorbitorii au folosit instrumentul RecoApy[5] și echipamentele de înregistrare personale.

În antrenarea sistemelor s-a utilizat același set de propoziții de la fiecare vorbitor, astfel încât să putem evalua mai exact influența timbrului vocal și a condițiilor de înregistrare asupra calității sistemului de sinteză.

Date audio folosite pentru cele două scenarii:

1. **MSPK**: 500 pronunții paralele pentru fiecare vorbitor.
2. **ADAPT**: 500 de pronunții paralele selectate de la fiecare vorbitor pentru pre-antrenarea modelului acustic. Și 5, 50 sau 200 de pronunții paralele de la fiecare vorbitor pentru adaptarea modelului.

Date text folosite pentru antrenare:

1. Grafeme (ID: **GR**): forma ortografică a textului (ex. *Acesta se referă însă doar la proprietățile din capitală*.)
2. Foneme (ID: **PH**): forma transcrisă fonetic a textului (ex. *aCesta se refer@ 1ns@ dPar la proprietățile din kapital@*.)
3. Foneme cu silabificare și accent (ID: **EXT**): forma transcrisă fonetic cu limita silabelor și accentul marcat (ex. *a-CEs-ta se re-fE-r@ Ân-s@ dPar la pro-pri-e-tĂ-ți-le din ka-pi-tA-l@*.)

2.1.2 Rezultate obiective

Mostrele sintetizate pentru fiecare vorbitor au fost evaluate obiectiv cu funcția de cost rata de eroare egală (EER -- en: Equal Error Rate) și cu rata de eroare la nivel de cuvânt (WER, en: Word Error Rate). EER ar trebui în principiu să estimeze similaritatea vocilor sintetice cu vocea naturală, iar WER gradul de inteligibilitate al vorbirii sintetizate. Aceste două măsuri sunt utilizate în mod frecvent în ultima perioadă pentru o primă evaluare a sistemelor de sinteză.

Pentru fiecare vorbitor 12 de propoziții sunt sintetizate cu sistemele cu identitate vocală multiplă, precum și cu sistemele de adaptare folosind diferite cantități de date. Aceste mostre audio sunt transcrise cu ajutorul instrumentului de recunoaștere a vorbirii descris în (Georgescu et al., 2019). Transcrierea fișierelor este comparată cu textul sintetizat pentru calculul WER. Pentru EER, cele 12 propoziții sintetizate pentru fiecare vorbitor sunt comparate cu un fișier audio de la același vorbitor, și un altul de la un alt vorbitor selectat aleator. Valoarea EER este obținută pe baza unui sistem neural de identificare de vorbitor[6] antrenat pe un număr de 5594 de vorbitori.

Tabelul 1 rezumează rezultatele de EER și WER pentru sistemele MSPK și ADAPT pentru cele trei tipuri de intrare de text.

Tabel 1. Rezultatele de WER și EER pentru sistemele MSPK și ADAPT și cele 3 tipuri de reprezentări ale textului: GR - ortografică, PH - fonetică și EXT - fonetică plus silabificare și accent lexical.

Sistem	Număr de uterațe	Număr de propoziții adaptare	Epoci	WER (%)			EER (%)		
				GR	PH	EXT	GR	PH	EXT
MSPK	37x500	N/A	216	28.13	26.87	28.68	17.34	14.41	15.76

[5] https://gitlab.utcluj.ro/sadriana/recoapy

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

ADAPT	37x500	37x200	216+50	29.75	33.35	29.93	15.31	14.63	15.54
ADAPT	37x500	37x200	216+100	27.95	32.85	28.80	14.18	15.31	14.86
ADAPT	37x500	37x50	216+100	27.35	65.88	74.81	15.09	14.41	15.99
ADAPT	37x500	37x5	216+100	29.38	67.40	73.58	15.54	17.11	17.11

Rezultatele pentru modelele sunt analizate și din perspectiva condițiilor de înregistrare, și categorizate pe gen: feminin și masculin în Tabelele 2 și 3.

*Tabel 2. Valori EER pe vorbitor pentru sistemul de vorbitori multipli (**MSPK**)*

Înregistrări în studio (SWARA)								**Înregistrări în afara studioului (SWARA 2.0)**							
Feminin				**Masculin**				**Feminin**				**Masculin**			
ID	**GR**	**PH**	**EXT**	**ID**	**GR**	**PH**	**EXT**	**ID**	**GR**	**PH**	**EXT**	**ID**	**GR**	**PH**	**EXT**
BAS	16.66	25	16.66	**FDS**	16.66	16.66	8.33	**BGL**	16.66	16.66	16.66	**BIM**	8.33	8.33	16.66
BEA	8.33	16.66	0	**PSS**	8.33	0	0	**BMM**	0	8.33	16.66	**BVL**	50	41.66	41.66
DCS	16.66	25	16.66	**RMS**	0	0	0	**CCL**	8.33	8.33	0	**MGL**	16.66	16.66	16.66
DDM	8.33	8.33	16.66	**SDS**	8.33	0	16.66	**CMM**	41.66	41.66	41.66	**NLL**	16.66	16.66	0
EME	8.33	8.33	8.33	**SGS**	8.33	0	16.66	**GAM**	50	58.33	58.33	**PDL**	8.33	16.66	25
HTM	8.33	8.33	8.33	**TSS**	16.66	16.66	8.33	**GIM**	16.66	8.33	16.66	**PTL**	25	25	16.66
PCS	0	16.66	0					**GNM**	16.66	16.66	16.66	**SRL**	25	25	16.66
PMM	16.66	16.66	16.66					**MAL**	16.66	25	25	**ZPL**	16.66	8.33	16.66
SAM	0	0	0					**MRL**	33.33	41.66	33.33				
								OGL	0	0	0				
								PBL	16.66	16.66	16.66				
								SMM	16.66	0	0				

Tabel 3. Valori de WER pe vorbitor pentru sistemul de vorbitori multipli

Înregistrari în studio (SWARA)								**Înregistrari în afara studioului (SWARA 2.0)**							
Feminin				**Masculin**				**Feminin**				**Masculin**			
ID	**GR**	**PH**	**EXT**	**ID**	**GR**	**PH**	**EXT**	**ID**	**GR**	**PH**	**EXT**	**ID**	**GR**	**PH**	**EXT**
DCS	13.75	21.47	14.76	**RMS**	10.61	12.21	10.13	**CCL**	30.32	26.22	33.86	**MGL**	16.52	20.73	16.1
DDM	15.36	16.42	15.59	**SDS**	22.05	13.83	21.93	**CMM**	20.54	23.78	20.54	**NLL**	14.76	18.81	17.1
EME	14.32	13.14	17.22	**SGS**	29.95	20.19	21.47	**GAM**	30.19	34.72	38.69	**PDL**	54.41	37.85	54
HTM	12.56	19.66	24.64	**TSS**	20.91	14.29	14.76	**GIM**	34.74	30.31	22.02	**PTL**	25.78	18	34.96
PCS	14.54	12.57	14.69					**GNM**	31.23	30.21	33.58	**SRL**	46.64	26.64	42.4
PMM	11.4	10.96	17.91					**MAL**	20.89	19.16	18	**ZPL**	26.15	27.3	26.24

SAM	9.69	11.05	17.07		**MRL**	37.21	33.55	14.87	
					OGL	26.54	17.27	30.38	
					PBL	67.93	71.94	42.84	
					SMM	23.47	21.4	30.75	

Atât valorile de EER cât și cei de WER sunt în medie mai bune pentru vocile înregistrate în condiții de studio. Din perspectiva valorii EER, tipul de input de text nu influențează similitudinea de vorbitor învățată. În ceea ce privește valoarea WER și aceasta atestă că tipul înregistrării afectează calitatea vorbirii rezultate, obținând valori mai mici în cazul vorbitorilor din corpusul SWARA. Tipul de reprezentare a textului influențează în mod diferit WER. În cazul celor mai mulți vorbitori inputul de PH și EXT obțin valori mai bune decât sistemele de GR, dar cu excepția de un număr mic de vorbitori în cazul cărora sistemul EXT are ceea mai mare valoare pentru vorbitor. Pentru a analiza efectul tipului de intrare de text teste de ascultare sunt necesare pentru a evalua naturalețea și similitudinea de vorbitor în mod subiectiv. O analiză mai elaborată a acestor rezultate poate fi regăsită în articolul atașat acestui raport și trimis pentru evaluare la conferința KES 2021.

2.2. Adaptarea în cadrul sistemului DC-TTS

Sistemul de sinteză DC-TTS este bazat pe arhitectura prezentată în (Tachibana et al., 2018). Acest model folosește rețele convoluționale și conține două componente, prima generează o mel spectrogramă de granularitate mai redusă, urmată de o componentă care produce mel spectrograma finală și care este mai apoi trecută prin algoritmul Griffin-Lim (Griffin & Lim, 1984) pentru obținerea formelor de undă. Ca punct de plecare am folosit o implementare[7] PyTorch a DC-TTS ce poate antrena sisteme folosind o singură identitate vocală. Acest instrument a fost extins pentru a permite învățarea simultană a mai multor identități vocale, pe baza metodei de învățare a contribuției la canalul de informație din implementarea[8] și care e o versiune TensorFlow a aceleiași arhitecturi.

2.2.1 Scenarii de antrenare

Pentru a facilita învățarea identității de voce sistemele sunt antrenate în trei scenarii:

1. Scenariul 1 (ID: **B**): sistem de sinteză antrenat cu vorbitori multipli.
2. Scenariul 2 (ID: **B+CS**): sistemul de sinteză antrenat cu vorbitori multipli și cu adăugarea unei funcții de cost suplimentare obținută din calculul funcției de similaritate cosinus (en. Cosine Similarity) între spectrograma generată în timpul antrenării și spectrograma fișierului natural corespunzător.
3. Scenariul 3 (ID: **B+E**): sistemul de sinteză cu vorbitori multipli extins cu o funcție de cost suplimentară calculată prin includerea unui sistem de verificare de vorbitor și evaluarea ratei de eroare egală (en. Equal Error Rate).

Sistemele de sinteză cu vorbitori multipli sunt antrenate în aceste trei scenarii cu diferite cantități de date, aceste fiind detaliate în secțiunea următoare.

Date de antrenare

[7] https://github.com/tugstugi/pytorch-dc-tts

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

Sistemele sunt antrenate pe date naturale, folosind toate datele disponibile din SWARA pentru fiecare vorbitor (între 1000 și 1500 de propoziții de la fiecare vorbitor), folosind doar subsetul RND1 (aprox. 500 de propoziții de la fiecare vorbitor) sau folosind 100 de propoziții din RND1 pentru fiecare vorbitor. Urmărind scopul de a îmbunătăți identitatea de vorbitor învățată metode de augmentare de date sunt folosite prin manipularea formelor de undă.

Datele augmentate sunt obține folosind două instrumente:

1. Re-eșantionarea formei de undă efectuată cu SoX[9], așa cum este descris în (Cooper et al., 2020).
2. Manipularea formei de undă folosind algoritmul PSOLA (Pitch Synchronous Overlap and Add) (Moulines & Charpentier, 1990) prin care durata și tonul fiecărei propoziții a vorbitorilor sunt modificate. Comparat cu eșantionarea simplă a formei de undă PSOLA ia în considerare perioadele fundamentale și rezultă astfel segmente audio, deși modificate, mai naturale.

Seturile de date folosite în antrenate care includ și date augmentate sunt următoarele:

1. **RND1-100-UP-DOWN**: date augmentate cu re-eșantionarea formei de undă. De la fiecare vorbitor sunt folosite 100 de propoziții, fiecare fiind re-eșantionată cu 0.95, 0.975, 1.025 și 1.05 din valoarea inițială, rezultând astfel 500 de propoziții pentru fiecare vorbitor.
2. **RND1-100-PSOLA-F0**: date augmentate cu algoritmul PSOLA în domeniul frecvenței. Pentru fiecare vorbitor 100 de propoziții sunt selectate, care au fost augmentate cu raporturi de 0.70, 0.80, 0.90, 1.05, 1.10, 1.20, 1.50. Dintre aceste 7 fișiere cele mai bune 4 sunt selectate rezultând astfel în 500 de propoziții pentru fiecare vorbitor. Selectarea a celor mai bune fișiere s-a realizat prin calcularea distanței Euclidiane față de propoziția naturală a reprezentărilor vectoriale ale fișierelor augmentate. Aceste reprezentări au fost extrase cu ajutorul rețelei de verificare de vorbitor.
3. **RND1-100-PSOLA-DUR**: date augmentate cu algoritmul PSOLA în domeniul timp. Dintre cele 7 fișiere augmentate cu raporturi de 0.85, 0.90, 0.95, 1.05, 1.10, 1.15, 1.20 cele mai bune 4 sunt selectate (după criteriul descris mai sus), rezultând astfel în 500 de propoziții pentru fiecare vorbitor.
4. **RND1-100-PSOLA-MIX**: date augmentate cu algoritmul PSOLA atât în domeniul de timp cât și în domeniul frecvenței. În domeniul de timp raportul folosit este de 0.8 și 1.3, iar în cel de a frecvenței 0.8 și 1.2, rezultând la fel în 500 de propoziții pentru fiecare vorbitor.

2.2.2. Metode de evaluare obiective și subiective

Cele 21 de sisteme de sinteză cu vorbitori multipli au fost evaluate obiectiv cu funcția de cost rata de eroare egală (EER) și cu rata de eroare a cuvintelor (WER) folosind un sistem de recunoaștere de vorbire. Dintre aceste sisteme, 9 au fost selectate pentru evaluare subiectivă. Testul de ascultare efectuat folosește metoda MuSHRA[10] (MUlti Stimulus test with Hidden Reference and Anchor) fiind completat de 27 de ascultători.

Rezultate

Numărul de vorbitori folosit pentru antrenare este de 18 vorbitori, 10 feminini și 8 masculini aparținând setului de date SWARA. Pentru fiecare vorbitor același 8 propoziții sunt sintetizate și evaluate obiectiv. Rezultatele WER și EER sunt prezentate în Tabelul 1.

[9] http://sox.sourceforge.net/

[10] ITU-R Recommendation BS.1534-1

Tabel 1. Rezultatele WER și EER pentru sistemele de sinteză DC-TTS

Date audio	WER (%)			EER (%)		
	B	B+CS	B+E	B	B+CS	B+E
ALL	9.54	7.66	8.26	6.94	4.66	4.66
RND1	9.99	8.67	9.86	4.86	4.00	4.66
RND1-100	11.13	10.21	13.26	5.55	5.33	5.33
RND1-100-UP-DOWN	12.42			8.66		
RND1-100-PSOLA-F0	14.04	15.75	14.18	8.66	10.66	11.33
RND1-100-PSOLA-DUR	11.84	13.62	10.32	8.33	6.25	10.00
RND1-100-PSOLA-MIX	10.05		16.00	9.72		6.94

Rezultatele testului de ascultare sunt prezentate în Figura 2.

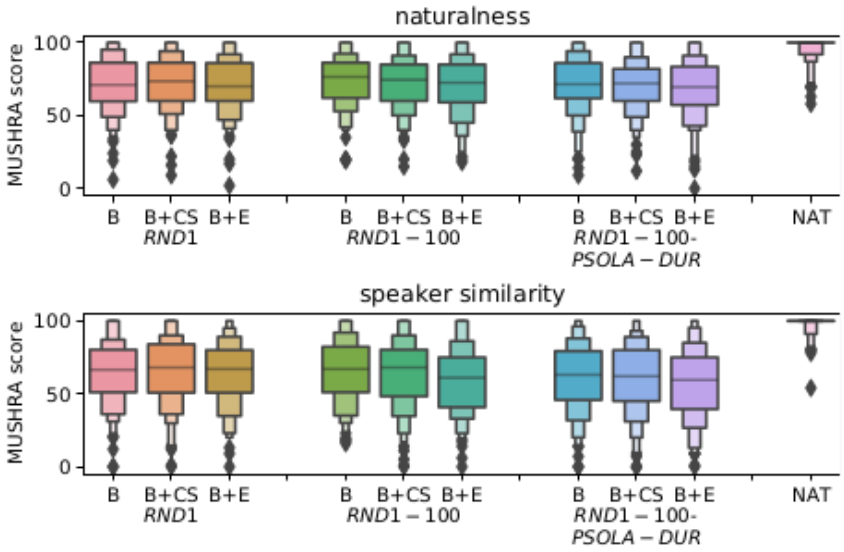

Fig. 2. Scorurile MuSHRA vizualizate cu diagrame letter-value

Prezentarea sistemelor și rezultatelor a fost trimisă la conferința EUSIPCO 2021. Mostre audio pentru sistemele antrenate și pentru augmentare de date sunt disponibile la adresa: https://speech.utcluj.ro/multispeaker_tts/.

2.3 Interfața RoNNA pentru sinteză text-vorbire în limba română

Sistemele de sinteză bazate pe arhitecturile Tacotron2 și DC-TTS sunt accesibile pe pagina RoNNA (Romanian Neural Network API): https://speech.utcluj.ro/ronna/.
Aceasta pagină functionează ca un API, prin care cu ajutorul unei chei obținute de la coordonatorii P4, utilizatorii pot sintetiza text în limba română. Sistemul DC-TTS sau Tacotron2 poate fi selectat, pentru fiecare sistem fiind disponibile un număr de voci (18 pentru DC-TTS și 6 pentru Tacotron2).

Sistemele de sinteză disponibile acum în platforma API:

1. Sistem bazat pe rețele convoluționale (DC-TTS) - voce BAS
2. Sistem bazat pe rețele convoluționale (DC-TTS) - voce BEA
3. Sistem bazat pe rețele convoluționale (DC-TTS) - voce CAU
4. Sistem bazat pe rețele convoluționale (DC-TTS) - voce DCS
5. Sistem bazat pe rețele convoluționale (DC-TTS) - voce DDM
6. Sistem bazat pe rețele convoluționale (DC-TTS) - voce EME
7. Sistem bazat pe rețele convoluționale (DC-TTS) - voce FDS
8. Sistem bazat pe rețele convoluționale (DC-TTS) - voce HTM
9. Sistem bazat pe rețele convoluționale (DC-TTS) - voce IPS
10. Sistem bazat pe rețele convoluționale (DC-TTS) - voce MAR
11. Sistem bazat pe rețele convoluționale (DC-TTS) - voce PCS
12. Sistem bazat pe rețele convoluționale (DC-TTS) - voce PMM
13. Sistem bazat pe rețele convoluționale (DC-TTS) - voce PSS
14. Sistem bazat pe rețele convoluționale (DC-TTS) - voce RMS
15. Sistem bazat pe rețele convoluționale (DC-TTS) - voce SAM
16. Sistem bazat pe rețele convoluționale (DC-TTS) - voce SDS
17. Sistem bazat pe rețele convoluționale (DC-TTS) - voce SGS
18. Sistem bazat pe rețele convoluționale (DC-TTS) - voce TSS
19. Sistem bazat pe rețele recurente (Tacotron2) și vocoder Waveglow - voce DOL
20. Sistem bazat pe rețele recurente (Tacotron2) și vocoder Waveglow - voce EME
21. Sistem bazat pe rețele recurente (Tacotron2) și vocoder Waveglow - voce MARA
22. Sistem bazat pe rețele recurente (Tacotron2) și vocoder Waveglow - voce NLL
23. Sistem bazat pe rețele recurente (Tacotron2) și vocoder Waveglow - voce SWARA

Ultimul sistem bazat pe Tacotron2 cu voce denumită SWARA este antrenat pe corpusul SWARA fără utilizarea identității vocale a fiecărui vorbitor, o voce medie.

Sistemul DC-TTS folosește ca text de intrare forma ortografică a textului, iar textul de intrare pentru Tacotron2 este forma transcrisă fonetic cu silabificare și accent. Pentru cazul din urmă primul pas este pre-procesarea de text, care prezice cu ajutorul unui model de tip Transformer transcrierea fonetică, silabificarea și accentul lexical pentru textul de intrare. Acest text pre-procesat este folosit ca intrare pentru sistemul Tacotron2. Acest pas de pre-procesare este disponibil și pentru utilizatorii RoNNA, care au posibilitatea de a procesa texte de intrare cu scopul de a obține transcrierea fonetică, silabificarea (marcată cu semnul de punct) și accentul lexical (marcat de semnul apostrof), dar fără generarea de audio corespunzător.

O captură de ecran a interfeței RoNNA este prezentată în Figura 3.

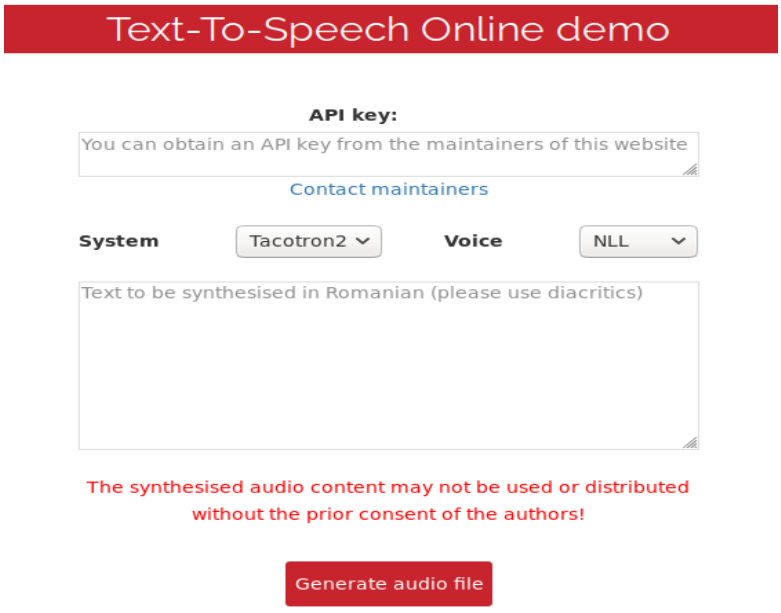

Fig. 3. Interfața web a API-ului RoNNA www.speech.utcluj.ro/ronna/

3. Concluzii

Acest raport a prezentat cele mai recente experimente de adaptare la o nouă identitate vocală folosind 2 arhitecturi de sisteme de sinteză cu rețele neuronale profunde: DC-TTS și Tacotron2. În cadrul celor 2 arhitecturi, au fost aplicate 2 metode de antrenare diferite: pentru DC-TTS s-a adăugat o funcție de eroare suplimentară, bazată pe măsura EER derivată din cadrul unei rețele de identificare de vorbitori; iar pentru sistemul Tacotron2 s-a antrenat o voce ce utilizează toate datele de antrenare de la toți vorbitorii și care nu face distincția între aceștia, iar mai apoi s-a realizat adaptarea cu diferite cantități de date audio și făcând o analiză extinsă a importanței timbrului vocal și a mediului de înregistrare în rezultatul final al sistemului.

Ambele sisteme de sinteză au fost integrate în interfața web RoNNA (www.speech.utcluj.ro/ronna) și sunt disponibile utilizatorilor în urma obținerii unei chei de autentificare de la autorii interfeței. Această metodă a fost aleasă ca urmare a necesității respectării condițiilor de protecție a datelor cu caracter personal disponibile în identitățile vocale utilizate pentru antrenarea sistemelor de sinteză.

4. Diseminare

Publicații elaborate și transmise spre publicare în anul 2021

Dan Oneață, Alexandru Caranica, Adriana Stan, Horia Cucu, "*An Evaluation of Word-level Confidence Estimation for end-to-end Automatic Speech Recognition*", In Proceedings of the 8th IEEE Spoken Language Technology Workshop (SLT 2021), Shenzhen, China, January 2021

Maria Nuțu, ”*Automatic Romanian lemmatization through a deep learning approach*.”, acceptat spre publicare la 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES), 8-10 septembrie 2021, Polonia.

Beata Lorincz, Adriana Stan, Mircea Giurgiu, “*An objective evaluation of the effects of recording conditions and speaker characteristics in multi-speaker deep neural speech synthesis*”,

trimis spre recenzare la The 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES), 8-10 septembrie 2021, Polonia.

Beata Lorincz, Adriana Stan, Mircea Giurgiu, "*Speaker verification-derived loss and data augmentation for DNN-based multispeaker speech synthesis*", trimis spre recenzare la The 29th European Signal Processing Conference, EUSIPCO 2021, Dublin, Ireland.

Beata Lorincz, Elena Irimia, Adriana Stan, Verginica Barbu-Mititelu, "*RoLEX: The development of an extended Romanian lexical dataset and its evaluation at predicting concurrent lexical information*", trimis spre recenzare la Computer Speech and Language.

Adriana Stan, Beata Lorincz, Maria Nuțu, Mircea Giurgiu, "*The MARA corpus: Expressivity in end-to-end TTS systems using synthesised speech data*", trimis spre recenzare la The 11th Conference on Speech Technology and Human-Computer Dialogue, in Bucharest, Romania, 13-15 octombrie 2021.

Lucrări de licență în legătura cu tematica proiectului

- Vlad Crehul - "Improving TTS for Romanian using Tacotron2 based on deep neural networks", lucrare de diplomă, iulie 2021.
- Vlad Hornai - "Adaptation of neural network-based automatic speech recognition system DeepSpeech for a specific application domain", lucrare de diplomă, iulie 2021.
- Bogdan Oros - "Cross-domain NLP adaptation using BERT and Transformers for Twitter text analysis", lucrare de diplomă, iulie 2021.
- Adrian Dobrescu - "Speaker recognition in noisy environments", lucrare de diplomă, iulie 2021.

Stagii de practică pentru studenți

- Oana Ranta (Master) - „Speech denoising with neural networks".

Pagini web ale proiectului SINTERO

Bibliografie

(Cooper et al., 2020) Cooper, E., Lai, C. I., Yasuda, Y., & Yamagishi, J. (2020). Can Speaker Augmentation Improve Multi-Speaker End-to-End TTS?. *arXiv preprint arXiv:2005.01245*.

(Georgescu et al., 2019)	Georgescu, A. L., Cucu, H., & Burileanu, C. (2019, October). Kaldi-based DNN architectures for speech recognition in Romanian. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (pp. 1-6). IEEE.
(Griffin & Lim, 1984)	Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(2), 236-243.
(Maaten & Hinton, 2008)	Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(Nov), 2579-2605.
(Moulines & Charpentier, 1990)	Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, *9*(5-6), 453-467.
(Prenger et al., 2019)	Prenger, R., Valle, R., & Catanzaro, B. (2019, May). Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3617-3621). IEEE.
(Shen et al., 2018)	Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Saurous, R. A. (2018, April). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4779-4783). IEEE.
(Tachibana et al., 2018)	Tachibana, H., Uenoyama, K., & Aihara, S. (2018, April). Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4784-4788). IEEE.
(Van der Maaten & Hinton, 2008)	Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

Data 30.04.2021

Director Proiect Complex,
Dan Tufiș