

Proiect component TADARAV

- Raport științific și tehnic în extenso 2021 -

Lucian Georgescu, Horia Cucu, Dragoș Burileanu, Corneliu Burileanu

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

Proiect complex: Resurse și tehnologii pentru dezvoltarea interfețelor om-mașină în limba română (ReTeRom)

Proiect component: Tehnologii pentru adnotarea automată a datelor audio și pentru realizarea interfețelor de recunoaștere automată a vorbirii (TADARAV)

Data: 30.04.2021

Etapa: 4 / 2021

Activitatea / activitățile:

- Activitatea 4.5 Management și diseminare

Număr contract: 73PCCDI / 2018

Acord de colaborare: 30/20.02.2018 ICIA, 4726/01.03.2018 UTCN, 3950/07.03.2018 UPB, 3805/06.03.2018 UAIC

Autoritatea contractantă: Unitatea Executivă pentru Finanțarea Învățământului Superior, a Cercetării, Dezvoltării și Inovării

Conducător proiect component: Universitatea POLITEHNICA din București

Conducător proiect complex: ICIA

Responsabil proiect component: Conf. Horia Cucu

Responsabil proiect complex: Prof. Corneliu Burileanu

1 Rezumatul etapei

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

- redactarea și depunerea a cel puțin un articol științific la conferințe sau reviste indexate ISI
- 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 crearea 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 consorțiu 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. Oneață, 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 și tehnologice

Laboratorul de cercetare *Speech and Dialogue* (Speed) din cadrul Universității Politehnica din București (UPB), reprezentantul UPB în proiectul TADARAV, oferă pe platforma ERRIS serviciile de cercetare și tehnologice enumerate în Tabelul 1.

Tabelul 1. Servicii de cercetare și tehnologice oferite de Laboratorul de cercetare *Speech and Dialogue*

Serviciu	Detalii
Serviciu și aplicație web de transcriere de documente ce conțin vorbire în limba română	https://transcriptions.speed.pub.ro
Serviciu și aplicație web de identificare de cuvinte cheie în documente ce conțin vorbire în limba română	https://keywords.speed.pub.ro
Serviciu și aplicație web de restaurare de diacritice în limba română	https://diacritics.speed.pub.ro

Proiectarea și implementarea de aplicații personalizate de transcriere a vorbirii continue	La cerere
Proiectarea și implementarea de aplicații personalizate de identificare de cuvinte și termeni de interes	La cerere
Proiectarea și implementarea de aplicații personalizate de sinteză de vorbire pornind de la text	La cerere
Proiectarea și implementarea de sisteme de recunoaștere de pattern-uri folosind inteligență artificială	La cerere

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

3 Locuri de muncă susținute prin program

Echipe de cercetare a Universității Politehnica din București pentru proiectul component TADARAV este prezentată în Tabelul 2.

Tabelul 2. Echipa de cercetare UPB

Nr.	Nume	Calitatea	Poziția	Normă
1	Horia CUCU	Conf. Univ.	Responsabil proiect component	Parțială
2	Corneliu BURILEANU	Prof. Univ.	Membru cercetător	Parțială
3	Dragoș BURILEANU	Prof. Univ.	Membru cercetător	Parțială
4	Alexandru-Lucian GEORGESCU	ACS	Membru cercetător	Parțială

Situația 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 ONEAȚĂ, Gheorghe POP și Cristian MANOLACHE. Pe 01.12.2020 Dan Theodor ONEAȚĂ 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 raporturilor 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 și va fi finanțat până la finalizarea perioadei de sustenabilitate a proiectului (până pe data de 30.11.2022) din Programul UPB Proof-of-Concept.

4 Valorificarea și îmbunătățirea competențelor și resurselor existente la nivelul consorțiului

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