Raport științific și tehnic
Etapa a IV-a, an 2021

„*Evaluare și distribuție finală tehnologii realizate in Proiectul SINTERO*”

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:
“SINTERO: Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate”

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

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 interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate”**
Titlu livrabil:	**Raport științific și tehnic (Etapa a IV-a, 2021)**
Termen:	**Aprilie 2021**
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 Nuțu, Adriana Stan**
Ofițer de proiect:	**Cristian STROE**

Rezumat:

Acest document prezintă o sinteză a realizărilor de natură științifică și tehnică obținute în a IV-a etapă de implementare a sub-proiectului SINTERO din cadrul proiectului PCCDI ReTeRom. Realizările din anul 2021 și se referă la:

- teste finale și prezentare tehnologie de sinteză bazată pe Tacotron 2
- teste finale și prezentare tehnologie de sinteză bazată pe DC TTS
- disponibilizare online a tehnologiei de sinteză text vorbire în limba română prin interfața RONNA
- diseminare și exploatare rezultate

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 în cele 2 livrabile 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.

Cuprins

1. Activitățile etapei de raportare în contextul general al proiectului

În a IV-a etapă (2021) a proiectului SINTERO, etapă cu denumirea *„Evaluare și distribuție finală tehnologii realizate în Proiectul 4„*, s-au finalizat două arhitecturi de sinteză text vorbire în limba română, au fost evaluate și sunt facute disponibile online ca serviciu demonstratuv de sintezăp cu expresivitate. Prin acestea, s-au incheiat activitățile planificate în proiect.

2. Gradul de realizare a obiectivelor specifice pentru Etapa a IV-a, 2021

Ob. Pr4.4.6: *Evaluare și ditribuție finală tehnologii dezvoltate în Proiectul 4*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 2 tehnologii dezvoltate și finalizate pentru sinteza text vorbire cu expresivitate în limba romînă
- 21 de teste de evaluare obiectivă pentru 21 de voci sintetizate folosind metricilie EER (Equal Error Rate), respectiv WER (Word Error Rate)
- 42 de teste (21 de voci x 2 tehnologii de sinteză) de evaluare subiectivă folosind metodologia standard MUSHRA.
- un livrabil (D4.6) cu titlul *„Distribuție finală tehnologie pentru interfețe de sinteză a vorbirii”.*

Ob. Pr4.4.7: *Diseminarea rezultatelor finale*

Grad realizare: Obiectiv realizat integral

Rezultate:

- realizarea și actualizarea web site-ului proiectului[1]
- pagini web cu demonstratoare cu vocile sintetizate
- 1 livrabil referitor la activitățile de diseminare (D4.7).

3. Rezultatele etapei și descrierea lor științifică și tehnică

3.1. Evaluare și distribuție finală a tehnologiei de sinteză text vorbire bazată pe Tacotron2

Rezultatele raportate în această secțiune corespund obiectivului Pr4.4.6 și ele sunt descrise în extenso in livrabilul D4.6. Implementarea inițială a arhitecturii Tacotron2 a fost extinsă în tehnologia finală cu funcționalități de antrenare cu vorbitori multipli pe baza reprezentărilor vectoriale ale vorbitorilor. Adăugarea de embedding de vorbitori este inspirată din implementarea instrumentului Mellotron. Sinteza de semnal este realizată cu vocoderul WaveGlow.

Scenarii de evaluare. Sistemele de sinteză bazată pe Tacotron2 folosind date de la mai mulți vorbitori au fost 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. În 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. În 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

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

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 folosite în evaluare. 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. Datele din SWARA au fost înregistrate în condiții de studio, iar cele din SWARA 2.0 în afara condițiilor de studio. Pentru primul scenariu – set MSPK: 500 pronunții paralele pentru fiecare vorbitor. Pentru al doilea scenariu – set ADAPT: 500 de pronunții paralele selectate de la fiecare vorbitor pentru pre-antrenarea modelului acustic. Adaptarea - 5, 50 sau 200 de pronunții paralele de la fiecare vorbitor.

Rezultate obținute pe baza metricilor 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. 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. 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 vorbitor2 antrenat pe un număr de 5594 de vorbitori. Tabelul 1 sumarizează 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 uterante	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
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.

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

*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.

3.2. Evaluare și distribuție finală a texhnologiei de sinteză text vorbire bazată pe DCTTS

Acestă tehnologie de sinteză text vorbire 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 pentru obținerea formelor de undă. Ca punct de plecare am folosit o implementare 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 implementarea3 și care e o versiune TensorFlow a aceleiași arhitecturi.

Scenarii de evaluare. Scenariul 1 (ID: B): sistem de sinteză antrenat cu vorbitori multipli. 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 între spectrograma generată în timpul antrenării și spectrograma fișierului natural corespunzător.

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

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).

Date folosite în evaluare. 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 (vezi livrabil D4.6 pentru detalii).

Rezultatele evaluărilor. Figura 1 prezintă reprezentări vectoriale pentru fișierele naturale și augmentate vizualizate cu algoritmul t-SNE (t-Distributed Stochastic Neighbour Embedding).

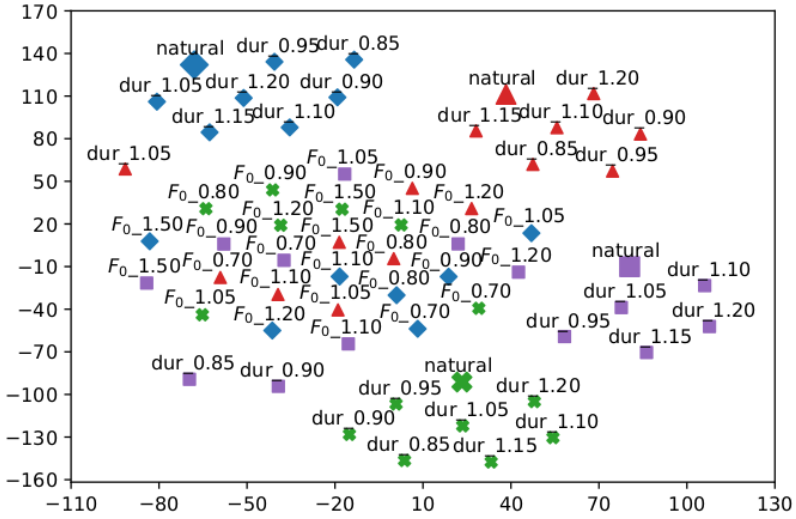

Fig. 1. Vizualizarea t-SNE a reprezentărilor vectoriale pentru propozițiile augmentate și naturale

Tabel 4. 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

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 MuSHRA4 (MUlti Stimulus test with Hidden Reference and Anchor) fiind completat de 27 de ascultători. 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 4. Rezultatele **testului de ascultare** sunt prezentate în Figura 2. Mostre audio pentru sistemele antrenate și pentru augmentare de date sunt disponibile la adresa: https://speech.utcluj.ro/multispeaker_tts/.

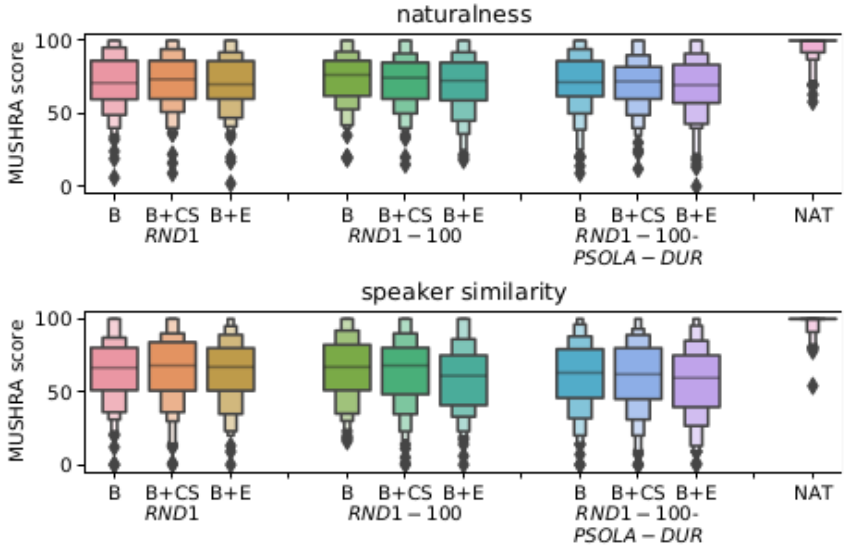

Fig. 2. Scorurile MuSHRA vizualizate cu diagrame letter-value

3.3. Disponibilizare online a tehnologiei de sinteză text vorbire prin interfața RONNA

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 coordonatorul grupului de cercetare al Proiectului 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. Sistemele de sinteză disponibile acum în platforma API:

1. Sistem bazat pe rețele convoluționale (DC-TTS) - voce BEA
2. Sistem bazat pe rețele convoluționale (DC-TTS) - voce CAU
3. Sistem bazat pe rețele convoluționale (DC-TTS) - voce DCS
4. Sistem bazat pe rețele convoluționale (DC-TTS) - voce DDM
5. Sistem bazat pe rețele convoluționale (DC-TTS) - voce EME
6. Sistem bazat pe rețele convoluționale (DC-TTS) - voce FDS
7. Sistem bazat pe rețele convoluționale (DC-TTS) - voce HTM
8. Sistem bazat pe rețele convoluționale (DC-TTS) - voce IPS
9. Sistem bazat pe rețele convoluționale (DC-TTS) - voce MAR
10. Sistem bazat pe rețele convoluționale (DC-TTS) - voce PCS
11. Sistem bazat pe rețele convoluționale (DC-TTS) - voce PMM

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

12. Sistem bazat pe rețele convoluționale (DC-TTS) - voce PSS
13. Sistem bazat pe rețele convoluționale (DC-TTS) - voce RMS
14. Sistem bazat pe rețele convoluționale (DC-TTS) - voce SAM
15. Sistem bazat pe rețele convoluționale (DC-TTS) - voce SDS
16. Sistem bazat pe rețele convoluționale (DC-TTS) - voce SGS
17. Sistem bazat pe rețele convoluționale (DC-TTS) - voce TSS
18. Sistem bazat pe rețele recurente (Tacotron2) și vocoder Waveglow - voce DOL
19. Sistem bazat pe rețele recurente (Tacotron2) și vocoder Waveglow - voce EME
20. Sistem bazat pe rețele recurente (Tacotron2) și vocoder Waveglow - voce MARA
21. Sistem bazat pe rețele recurente (Tacotron2) și vocoder Waveglow - voce NLL
22. 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.

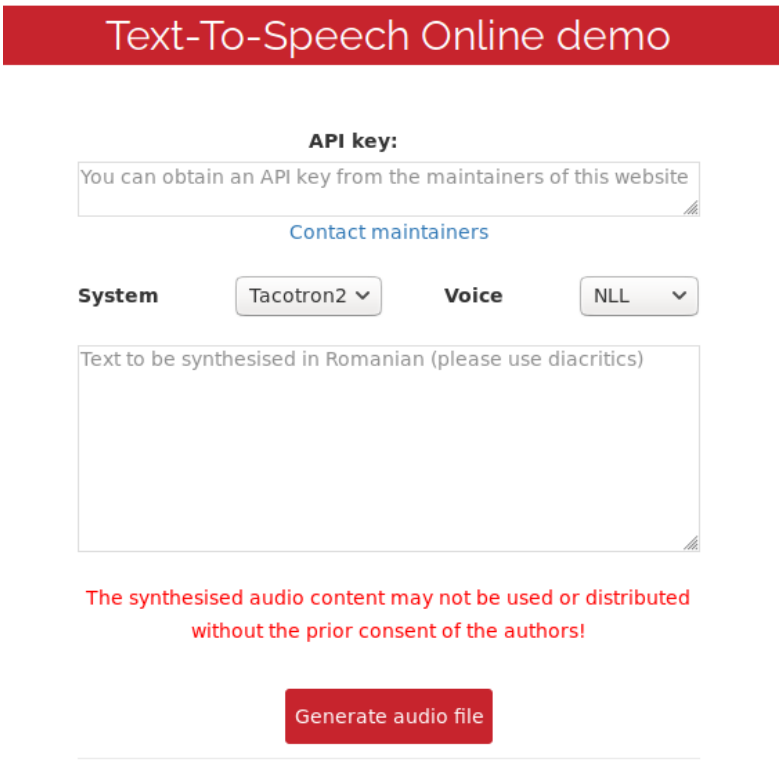

Fig. 3. Interfața web a API-ului RoNNA www.speech.utcluj.ro/ronna/

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

Tabel 5. 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 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 umană nou angajată în UTCN	Conform acordului de grant au fost angajate 2 ACS, doctoranzi, începând cu 1 ianuarie 2019.
Valorificare resurse în parteneriat	● UTCN a preluat de la ICIA resurse de date text (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”
Cecuri	● nu au putut fi folosite în această perioadă.

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 între cercetătorii din proiect. Din punct de vedere administrativ s-a primit 1 tranșă de avans. Resursele financiare alocate UTCN pentru anul 2021 nu au fost utilizate integral deoarece o nou angajată a trecut în concediu de maternitate, iar alte resurse dedicate pentru costuri de capital nu au fost folosite din cauza întârzierilor locale în aprobarea investițiilor pe anul 2021.

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 paginii web cu tehnologia de sinteză text vorbire disponibilă online, speech.utcluj.ro/ronna/.

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

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.

7. Concluzii

Activitățile de cercetare desfășurate în etapa a IV-a de implementare a proiectului (2021) 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 2 livrabile aferente perioadei de raportare (vezi Secțiunea 8 a acestui raport), asigură finalizarea cu succes a proiectului SINTERO.

8. Referințe la livrabilele aferente etapei 2021 (Anexe la raport)

[1] Livrabil D4.6:	*„Distribuție finală tehnologie pentru interfețe de sinteză a vorbirii",* Aprilie 2021.
[2] Livrabil D4.7:	*„Diseminare și exploatare rezultate pe anul 2021",* Aprilie 2021.