# D4.6. Evaluare și distribuție finală a tehnologiei pentru interfețe de sinteză a vorbirii

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 |

MINISTERUL EDUCAȚIEI NAȚIONALE

UEFISCDI

**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: | **D4.6. Evaluare și distribuție finală a tehnologiei pentru interfețe de sinteză a vorbirii** |
| Termen: | **Aprilie 2021** |
| Editor: | **Beáta Lőrincz** |
| Adresa de eMail editor: | **beata.lorincz@com.utcluj.ro** |
| Autori, în ordine alfabetică: | **Mircea Giurgiu, Beáta Lőrincz, Maria Nuțu, Adriana Stan** |
| Ofițer de proiect: | **Cristian STROE** |

**Rezumat:**

Acest raport prezintă rezultatele finale ale sistemelor de sinteză dezvoltate în cadrul proiectului P4. Sintero, precum și interfața web denumită RoNNA - Romanian Neural Network API ce pune la dispoziția utilizatorilor aceste sisteme de sinteză bazate pe rețele neuronale, folosind multiple identități vocale.

**Cuprins**

# 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[1] 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[2]. 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.

[1] https://github.com/NVIDIA/tacotron2
[2] https://github.com/NVIDIA/mellotron

*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 cele din SWARA 2.0 în afara condițiilor de studio, pentru care vorbitorii au folosit instrumentul RecoApy[3] 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

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

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[4] 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 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 |
| **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 |

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

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **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[5] 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[6] 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**

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[7], 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:

[5] https://github.com/tugstugi/pytorch-dc-tts
[6] https://github.com/CSTR-Edinburgh/ophelia
[7] http://sox.sourceforge.net/

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.

Figura 1 prezintă reprezentări vectoriale pentru fișierele naturale și augmentate vizualizate cu algoritmul t-SNE (t-Distributed Stochastic Neighbour Embedding) (Van der Maaten & Hinton, 2008).

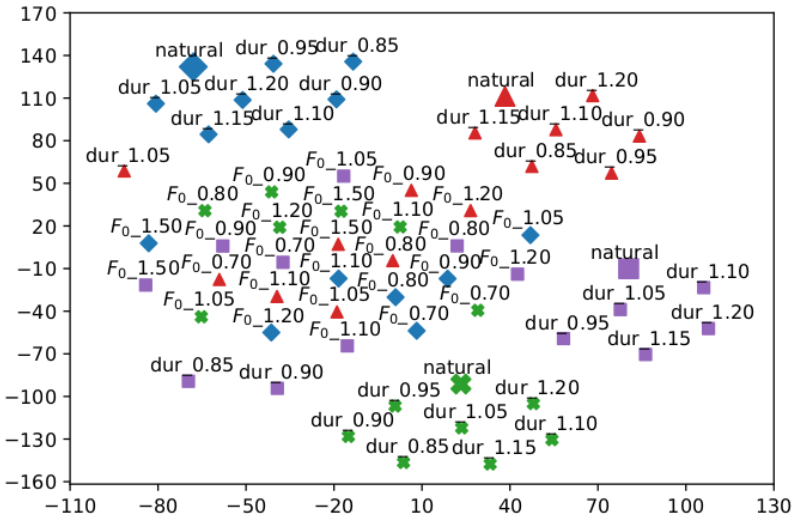


*Fig. 1. Vizualizarea t-SNE a reprezentărilor vectoriale pentru propozițiile augmentate și naturale*

*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[8] (MUlti Stimulus test with Hidden Reference and Anchor) fiind completat de 27 de ascultători.

*Rezultate*

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

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.

*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.

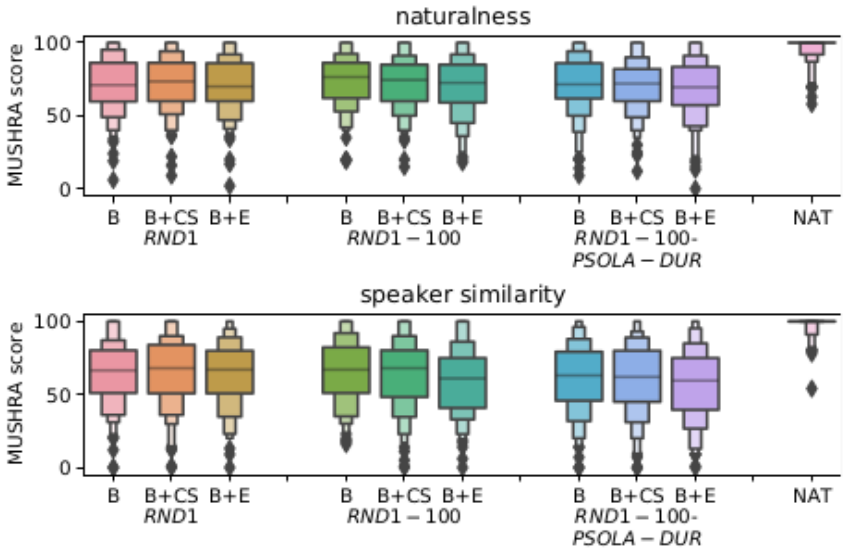


*Fig. 2. Scorurile MuSHRA vizualizate cu diagrame letter-value*

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.

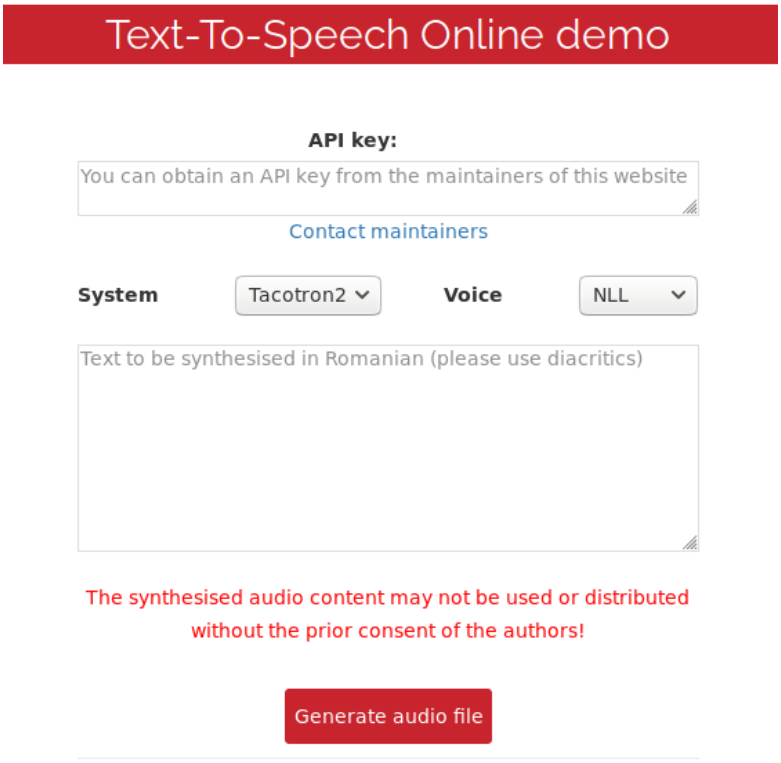


*Fig. 3. Interfața web a API-ului RoNNA [www.speech.utcluj.ro/ronna/](www.speech.utcluj.ro/ronna/)*

## 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. 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).