Raport științific și tehnic
Etapa a II-a, an 2019

„*Implementarea componentelor pentru modelarea prozodiei și adaptarea la noi vorbitori a vocilor sintetice*”

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:	**Raport științific și tehnic (Etapa a II-a, 2019)**
Termen:	**Noiembrie 2019**
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 doua etapă de implementare a sub-proiectului SINTERO din cadrul proiectului PCCDI ReTeRom. Realizările din anul 2019 se referă la:

- implementarea modulului de identificare și codare a stilului de vorbire prin vectori de stil
- implementarea și validarea unei metode de adaptare la noi vorbitori cu set redus de date
- implementarea și testarea unei metode de modificare a prozodiei pentru voci expresive
- îmbunătățirea componentei de control a prozodiei

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

Cuprins

1. Activitățile etapei de raportare în contextul general al proiectului

În a doua etapă (2019) a proiectului SINTERO, etapă cu denumirea *„Implementarea componentelor pentru modelarea prozodiei și adaptarea la noi vorbitori a vocilor sintetice”*, s-a pornit de la resurse și module software dezvoltate deja în etapa 2018 de către partenerii UTCN (corpusuri de date audio și text, sistemul preliminar de sinteză fără expresivitate) și ICIA (module de adnotare a textului disponibile pe platforma Relate[1]) și au fost desfășurate o serie de activități pentru: **a)** implementarea modulului de identificare și codare a stilului de vorbire prin vectori de stil**, b)** implementarea și validarea unei metode de adaptare la noi vorbitori cu set redus de date, **c)** implementarea și testarea unei metode de modificare a prozodiei pentru voci expresive, **d)** îmbunătățirea componentei de control a prozodiei, respectiv activități de testare, validare, diseminare și demonstrare online.

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

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

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

Grad realizare: Obiectiv realizat integral

Rezultate:

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

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

Grad realizare: Obiectiv realizat integral

Rezultate:

- 2 metode de adaptare a sistemului de sinteză la un nou vorbitor: (a) o metodă bazată pe posfiltrarea cu rețele neuronale, (b) o metodă bazată pe arhitectura Tacotron și GST (Global Style Tokens), implementate pe arhitecturi paralele de procesare cu GPU
- 8 sisteme de sinteză text vorbire adaptate la un nou vorbitor prin metoda de postfiltrare
- evaluarea obiectivă și prin teste de ascultare a celor 8 sisteme și pagină web cu mostre audio[3]
- 5 sisteme de sinteză text vorbire bazate pe arhitectura Tacotron GST adaptate la stilul și expresivitatea unui nou vorbitor cu implementare pe sisteme cu GPU, evaluarea sistemelor și pagină web cu mostre audio[4]
- 1 articol care descrie metoda bazată pe postfiltrare
- 1 livrabil (D2.16) cu titlul *„Implementarea unui modul de adaptare la un nou vorbitor a sistemului de sinteză”.*

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

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

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

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

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

Grad realizare: Obiectiv realizat integral

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

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

Grad realizare: Obiectiv realizat integral

Rezultate:
- 1 metodă originală de antrenare a sistemelor de sinteză bazate pe rețele neuronale pentru voci expresive, folosind chiar voce sintetizată cu sisteme de tip parametric
- 5 sisteme de sinteză text vorbire pentru voce expresivă
- web site pentru evaluarea subiectivă a celor 5 sisteme de sinteză[5]
- pagină web cu mostre audio generate cu acest nou model[6]
- 1 livrabil (D2.18.) *„Îmbunătățirea componentei de modelare și control a prozodiei. Activități de testare, validare și demonstrare online a modulelor implementate”.*

Ob. Pr4.2.19: *Diseminarea rezultatelor intermediare*

Grad realizare: Obiectiv realizat integral

Rezultate:
- realizarea și actualizarea web site-ului proiectului[7]
- pagini web cu demonstratoare cu vocile sintetizate
- 1 livrabil referitor la activitățile de diseminare (D2.19).

3. Rezultatele etapei și descrierea lor științifică și tehnică

3.1. Implementarea modulului de identificare a stilului de vorbire

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

Metodele au fost aplicate asupra unui set de date text extrase din corpusul CoRoLa al Institutului de Inteligență Artificială al Academiei Române din București. Setul de date conține text în stilurile: *beletristic, științific, publicistic, memorialistic și juridic*. Pentru fiecare subset am avut la dispoziție aproximativ 1 milion de cuvinte, organizate în 40.000 de fraze. Media numărului de cuvinte dintr-o frază este de 20.

(1) Implementarea LDA (Fig. 1), ca model probabilist, este capabilă să modeleze în mod ierarhic fiecare stil de exprimare ca o combinație finită de probabilități de stiluri de exprimare,

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

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

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

din cele disponibile în mod latent în setul de antrenare. Această modelare este potrivită, mai ales având în vedere spectrul larg de posibilități de exprimare în diferitele stiluri de vorbire.

Rezultatele sunt vizibile sub formă grafică (Fig. 2), sub forma probabilităților de clasificare a unui text necunoscut către unul dintre stilurile din corpusul de antrenare (mai jos grafic rezultat din 4 stiluri), precum și a scorului de coerență ale modelelor (ca măsură a gradului de separabilitate ale modelelor).

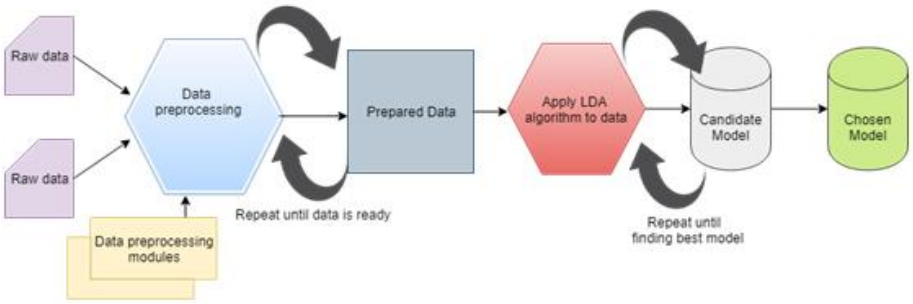

Fig. 1. Fluxul de procesări pentru metoda LDA (vezi detalii în livrabilul D2.15)

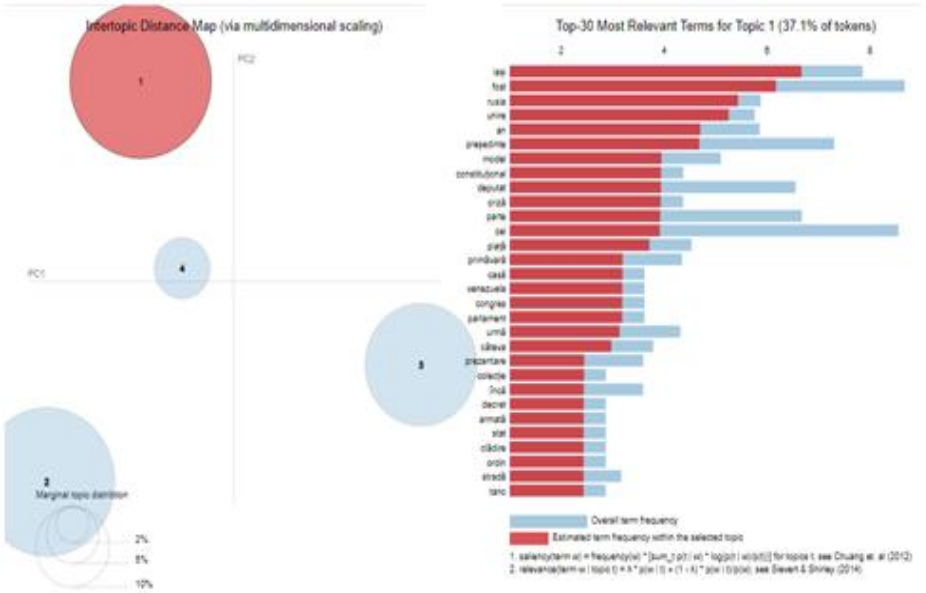

Fig. 2. Grafice pentru gradul de separabilitate a celor 4 modele și cuvintele definitorii

(2) Implementarea cu rețele neuronale convoluționale (CNN – Convolutional Neural Networks) s-a bazat pe un studiu detaliat[8] și pe experimente preliminare de pre-procesare și reprezentare vectorială a textului, care sunt descrise detaliat în livrabilul D2.15. Arhitectura rețelei[9] este inspirată din aceste studii și conține următoarele niveluri de prelucrare (codul este disponibil online[10]): 1) reprezentare vectorială compactă, 2) reducere număr neuroni prin tehnica dropout, 3) strat convoluțional cu filtre de dimensiune 3, 4, sau 5, 4) nivel cu conectare totală de tip dens, 5) reducere neuroni prin dropout și funcție de activare ReLU (Rectified Linear Unit), 6) nivel final de clasificare cu conectare totală a neuronilor și funcție de activare de tip SoftMax.

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

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

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

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

Tabel 1. Rezultatele identificării stilului din text folosind rețele CNN

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

Pentru preprocesarea textului în aceste două metode de clasificare au fost incluse și modulele de restaurare automată a diacriticelor (utilă în special pentru texte fără diacritice colectate din mediul online), respectiv de detecție automată a părții de vorbire (pentru realizarea dezambiguării înțelesului unui cuvânt). Descrierea completă a sistemelor este realizată în cele două articole publicate la conferința ICCP 2019 (vezi livrabil D2.19 – Diseminare). De asemenea, tot în domeniul prelucrării textului au fost experimentate diferite metode de reprezentare a caracteristicilor textuale folosind informații auxiliare de natură lingvistică (de exemplu accent, silabificare), cu publicarea unui articol la conferința SPED 2019.

Cercetările demonstrează faptul că prin analiza textului de intrare se poate determina stilul și expresivitatea în vorbire, cu scopul de a controla modul de generare a semnalului vocal.

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

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

(1)Metoda de adaptare folosind postfiltrarea. Ideea acestei metode este de a crea un sistem de sinteză text vorbire generic, antrenat cu mai multe voci pentru a mări volumul de date necesar antrenării, și bazat pe arhitecturi DNN (Deep Neural Network). Ieșirea acestui sistem este adaptată către noul vorbitor tot prin intermediul unei rețele neuronale, cu rol de postfiltrare

(vezi Fig 2). Acest postfiltru are rolul de a condiționa caracteristicile acustice, sintetice, generate la ieșire, către caracteristicile acustice naturale ale noului vorbitor. Antrenarea sistemului de sinteză s-a realizat cu 8 voci feminine din corpusul SWARA pe o arhitectură de tip Merlin[11]. Pentru partea de postfiltrare s-a implementat o rețea neuronală cu nivele total conectate, combinate cu nivele recurente de tip LSTM (Long Short Term Memory), cu număr variabil de neuroni în fiecare strat (256, 512 sau 1024) și cu diferite funcții de activare (tanh, ReLu).

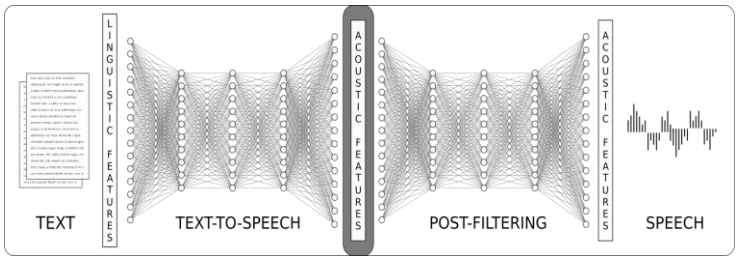

Fig. 3. Structura rețelei neuronale de sinteză de voce și post-filtrare

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

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

Sistemele prezentate mai sus au fost validate cu metode obiective și subiective. Pentru metoda obiectivă s-a folosit măsura de distorsiune cepstrală (en. *Mel Cepstral Distortion (MCD)*). Acuratețea alinierilor pe nivel de stare nu este cunoscută, motiv pentru care valoarea MCD a fost obținută folosind un pas de aliniere a datelor cu ajutorul algoritmului DTW. Pentru calcularea valorilor MCD au fost sintetizate 50 de propoziții cu fiecare sistem (vezi Fig. 4). Aceste propoziții nu au fost incluse în datele de antrenare. Sistemele M050, M100 și M500 sunt sisteme de sinteză minimale, ce utilizează 50, 100 și 500 de propoziții de la un singur vorbitor fără post-filtrare. Cea mai bună valoare pentru sistemele antrenate pe 100 de propoziții este obținută cu ajutorul post-filtrării, urmată de adaptarea de vorbitor.

Testul de ascultare a fost completat de 20 de ascultători și arată că metoda de post-filtrare și adaptare la vorbitor conduce la o voce mai naturală și inteligibilă (vezi Fig. 5). Identificatorii sistemelor sunt următorii: A - Sistem de *sinteză* antrenat cu 50 de propoziții; B - sistem de sinteză antrenat cu 100 de propoziții; C - sistem de sinteză antrenat cu 100 de propoziții și post-filtrare cu 100 de propoziții; D - sistem de sinteză și post-filtrare antrenate cu 100 de propoziții dublate artificial; E - sistem de sinteză antrenat cu 100 de propoziții și post-filtrare multi-vorbitor; F - sistem de adaptare la vorbitor pornind de la o rețea preantrenată cu date multi-vorbitor; G - sistem de sinteză antrenat cu 500 de propoziții; H - vocea naturală. Un demonstrator online pentru aceste voci se găsește la adresa *https://speech.utcluj.ro/pf_lrec2020/*

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

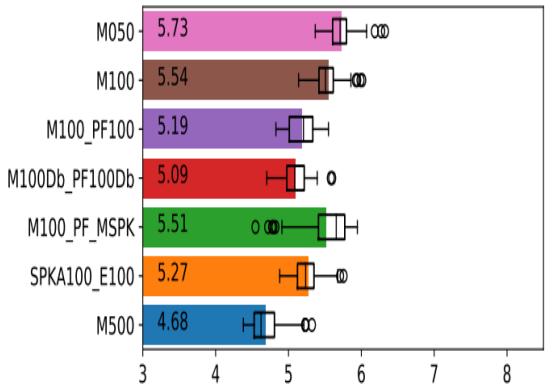

Fig. 4. Valorile MCD pentru vorbitorul MAR

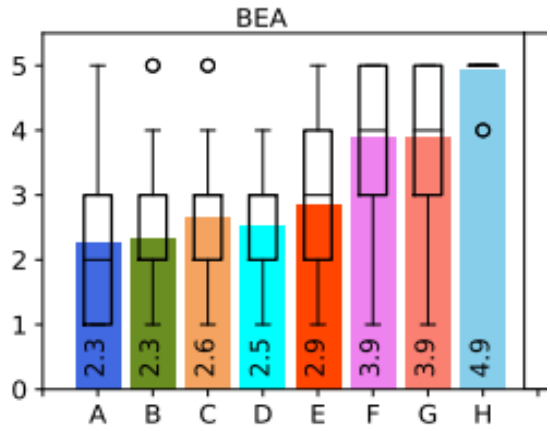

Fig. 5. Test ascultare vorbitor BEA

(2)Metoda de adaptare folosind arhitectura Tacotron și GST (Global Style Tokens). Această arhitectură[12] permite antrenarea cu mai mulți vorbitori, dar mai ales permite învățarea automată a expresivității din vorbire prin intermediul unor vectori (GST – Gloal Style Tokens) incluși în această arhitectură. În plus, modulul de control al prozodiei poate fi folosit și la învățarea identității vorbitorilor. Cerința principală a unui astfel de sistem este volumul foarte mare de date audio pentru antrenare.

În consecință, pentru evaluarea metodei s-au folosit corpusurile MARA[13] (1 vorbitor, 11 ore de vorbire preluată din audiobook) și SWARA[14] (17 vorbitori, 21 de ore de vorbire, din care au fost preluate aproximativ 50 de minute în configurație de 10 vorbitori feminine, respectiv 5 feminini și 5 masculini). În livrabilul D2.16 sunt prezentate în detaliu rezultatele experimentale. Concluziile arată că adaptarea la un nou vorbitor se face relativ rapid (aproximativ 10 epoci, 5-10 minute de rulare pe sistem cu o singură placă GPU). Identitatea vorbitorilor este controlată prin intermediul unui strat de reprezentări vectoriale, învățate tot în cadrul antrenării sistemului. Exemple audio pentru aceste sisteme sunt disponibile și pot fi ascultate aici: *https://speech.utcluj.ro/sintero/dnn-samples/*. Pe această direcție, se propune ca în viitor să fie explorate și alte metode de învățare, de exemplu doar cu o singură mostră audio sau cu un set redus de mostre audio.

3.3. Implementarea modulului de transfer a prozodiei

Rezultatele raportate în această secțiune corespund obiectivului Pr.4.2.17 și se referă la posibilitatea de a modifica prozodia generată de un sistem de sinteză text vorbire la prozodia unui anumit vorbitor. Pentru implementarea modulului de transfer a prozodiei s-a ales arhitectura Tacotron GST (vezi Fig. 6), deoarece aceasta permite codificarea prozodiei prin intermediul vectorilor GST (Global Style Tokens). Modulul este compus din următoarele componente

- componenta Tacotron, adică sistemul de sinteză text vorbire de bază
- componenta Style Tokens prin intermediul căreia se codează și reprezintă latent prozodia din vorbire. Această componentă e compusă din (a) codorul parametrilor acustici prin intermediul unei rețele neuronale cu nivele recurente, respectiv convoluționale și (b) codorul

[12] *https://github.com/mozilla/TTS*
[13] *https://speech.utcluj.ro/marasc/*
[14] *https://speech.utcluj.ro/swarasc/*

pentru stil și prozodie, bazat pe un nivel cu neuroni de atenție, capabili să exploreze corelații pe termen lung (prozodie) în parametrii acustici

- componenta de decodare, capabilă să producă - prin intermediul unei rețele neuronale și pe baza codurilor generate în procesul de antrenare – semnalul sintetizat. La intrarea decodorului se prezintă pe de o parte textul de sintetizat, pe de altă parte, fie (a) un semnal referință a cărui prozodie se dorește a fi transferată pe semnalul sintetizat, fie (b) o combinație a vectorilor prin care s-a codat prozodia în procesul de antrenare.

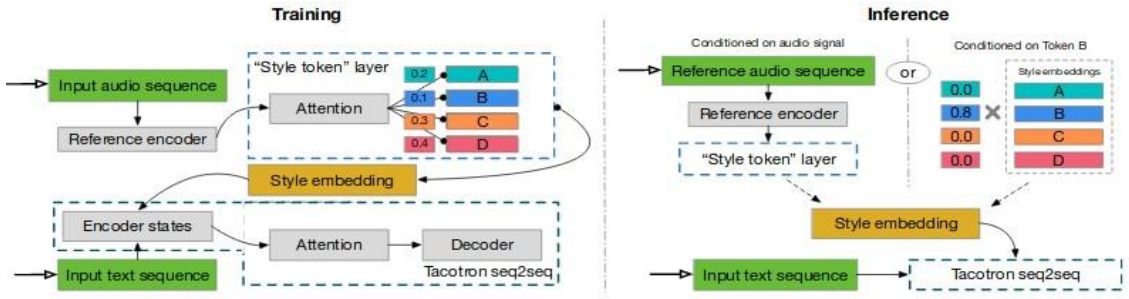

Fig 6. Arhitectura generală a sistemului Tacotron GST (conform cu https://github.com/mozilla/TTS)

Deoarece sistemele de sinteză bazate pe arhitectura Tacotron necesită un volum foarte mare de date audio, s-a abordat o strategie prin care s-a folosit un sistem deja pre-antrenat cu date audio cu expresivitate colectate din audiobook-ul Mara. Modelul a fost antrenat 800 de epoci, suficient pentru ca modulul GST (în configurație cu 10, respectiv 5 tokeni de stil) să surprindă variabilitatea prozodică a datelor audio. În etapa de sinteză s-au aplicat 2 strategii:

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

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

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

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

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

3.4. Îmbunătăţirea componentei de modelare şi control a prozodiei

Rezultatele raportate în această secţiune corespund obiectivului Pr.4.2.18 şi sunt descrise în detaliu in articolul transmis spre publicare la conferinţa internaţională Language Resources an Evaluation 2020, titlu ”*The MARA Corpus: Expressivity in end-to-end TTS systems using synthesised speech data”*. Articolul este prezentat în Anexa la livrabilul D2.18 „Îmbunătăţirea componentei de modelare şi control a prozodiei”. Componenta de modelare şi control a prozodiei a fost evaluată conform cu cele descrise în secţiunea anterioară şi s-a evidenţiat faptul că pentru transferul prozodiei sunt necesare date de antrenare bogate în conţinut prozodic.

Astfel, în lipsa acestor date cu caracteristici prozodice variate, s-a propus o idee originală, şi anume posibilitatea utilizării datelor audio sintetizate ca date de antrenare ale acestor modele. Datele sintetizate au fost obţinute cu ajutorul sistemelor de sinteză parametrice bazate pe modele Markov sau reţele neuronale cu straturi complet conectate unidirecţionale. Datele sintetizate au utilizat conturul frecvenţei fundamentale şi durata la nivel de fonem extrase din înregistrările audio originale şi în care există o mare variabilitate prozodică.

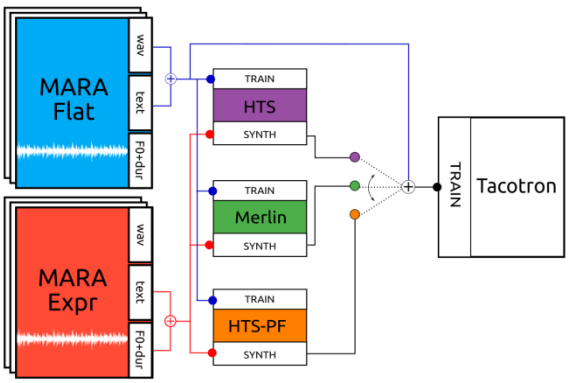

Fig. 7. Arhitectura generală a sistemului de modelare prozodică şi transfer prozodie

Au fost evaluate 5 sisteme de sinteză complete ce folosesc date expresive provenite fie de la vorbitorul original (MARA Expr în figura de mai sus), fie din sistemele statistic-parametrice anterior dezvoltate (HTS). Rezultatele acestor 5 sisteme au fost evaluate atât din punct de vedere obiectiv, folosind o măsură a distorsiunii spectrogramei pe scală Mel, precum şi din punct de vedere subiectiv folosind teste de ascultare. Testele de ascultare au inclus două secţiuni: naturaleţe şi expresivitate în format MuSHRA (TU-R Recommendation BS.1534-1). În urma evaluării nu au fost determinate diferenţe statistice semnificative între sisteme. Pentru

testele de ascultare a fost creată o pagină web distincta la adresa http://romaniantts.com/lrec/ unde se pot asculta și mostrele de test[15].

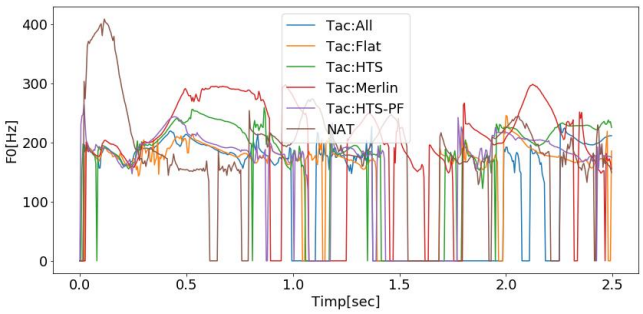

Fig.8. Variația conturului F0 pentru expresivitate

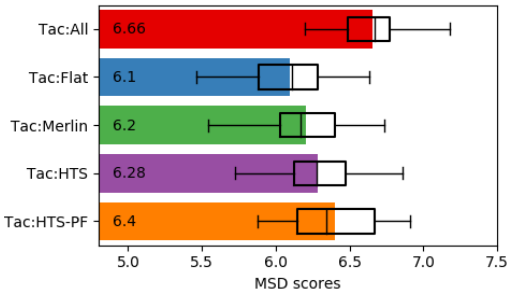

Fig. 9. Metrica obiectivă MSD

Variația conturului F0 pentru metoda propusă demonstrează faptul că acest contur este foarte apropiat de conturul F0 al vorbirii naturale expresive, iar măsurarea obiectivă a distanței spectrale între semnalul original și semnalul sintetizat indică valori care sunt tipice pentru sisteme de sinteză de înaltă calitate.

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

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

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

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

5. Management și comunicare

Activitățile de management au fost orientate în special către managementul proiectului complex în vederea integrării diferitelor grupuri de cercetare și a resurselor tehnice ale acestora. S-au organizat mai multe conferințe Skype și o reuniune a parteneriatului în 18 Noiembrie 2019 la Cluj-Napoca. Este de notat faptul că s-a asigurat de către ICIA (prin responsabilul de achiziții) o bună comunicare și coordonare pentru realizarea planului de achiziții global, respectiv pentru documentația de raportare etapă. Din punct de vedere administrativ s-au primit 4 tranșe de avans cu o regularitate adecvată. Resursele financiare alocate UTCN pentru anul 2019 au fost utilizate în majoritate, cu excepția unei sume în categoria cecuri, care a trecut la economii.

6. Diseminarea rezultatelor

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

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

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

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

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

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

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

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

7. Concluzii

Activitățile de cercetare desfășurate în etapa a II-a de implementare a proiectului (2019) au condus la obținerea rezultatelor așteptate și ele sunt în concordanță cu obiectivele specifice ale etapei. Astfel, rezultatele raportate în acest document și descrise detaliat în cele 5 livrabile aferente perioadei de raportare (vezi Secțiunea 8 a acestui raport), asigură modulele software pentru etapa finală a proiectului.

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

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