D2.17. Implementarea unui modul de transplantare a prozodiei unui vorbitor în sistemului de sinteză

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:	**D2.17. Implementarea unui modul de transplantare a prozodiei unui vorbitor în sistemului de sinteză**
Termen:	**Decembrie 2019**
Editor:	**Maria Nuțu (Universitatea Tehnică din Cluj-Napoca)**
Adresa de eMail editor:	**Maria.Nutu@com.utcluj.ro**
Autori, în ordine alfabetică:	**Beáta Lőrincz, Maria Nuțu, Adriana Stan**
Ofițer de proiect:	**Cristian STROE**

Rezumat:

Acest raport prezintă rezultatele obținute în cadrul proiectului SINTERO în vederea transferului informației prozodice de la un vorbitor sursă la un vorbitor țintă în cadrul sistemelor de sinteză text-vorbire în limba română. Raportul descrie două abordări distincte: prima abordare analizează transferul simplu de la un vorbitor la altul; iar cea de a doua are în vedere transferul prozodiei unui vorbitor către un sistem antrenat cu date de la mai mulți vorbitori, astfel încât toate vocile sintetizate să utilizeze prozodia vorbitorului țintă.

Raportul detaliază arhitectura sistemelor de sinteză utilizate, modul de antrenare a acestora și datele utilizate, precum și rezultatele obținute.

Cuprins

1. Introducere

Acest livrabil *D2.17. Implementarea unui modul de transplantare a prozodiei unui vorbitor în sistemului de sinteză* prezintă rezultatele obținute în cadrul activității A.4.2.3 Implementarea unui modul de transplantare a prozodiei unui vorbitor în sistemul de sinteză, în cadrul subproiectului *P4 (SINTERO) - Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate.*

Naturalețea vocilor generate cu ajutorul sistemelor de sinteză actuale, bazate pe rețelele neuronale, este apropiată de cea a vocii umane (Shen et al., 2018). Totuși, aceste modele nu tratează și problema expresivității sau a adaptării prozodiei în funcție de stilul de vorbire redat. În lipsa unor date/înregistrări expresive, vocea generată de aceste sisteme de sinteză are în general o prozodie neutră.

Expresivitatea vocii depinde de contextul în care sistemul de sinteză este folosit. De exemplu, dacă sistemul este folosit pentru a reda știri sau informații de interes general, înțelegerea mesajului transmis este mult mai importantă decât prozodia folosită. Pe de altă parte, dacă vocea generată este utilizată în redarea poveștilor pentru copii, existența elementelor expresive este foarte importantă.

Totodată, este dificil de specificat exact care sunt elemente care contribuie la apariția expresivității în vocea naturală. Factori precum starea emoțională a vorbitorului, condițiile culturale, etnice, sociale și educaționale ale vorbitorului, care pot influența prozodia vorbirii, sunt dificil de modelat, acești factori fiind relativi și subiectivi. Pe de altă parte, lipsa unor măsuri obiective de măsurare a expresivității unei voci îngreunează modelarea și analiza prozodiei. În absența unor metode obiective automate de măsurare a expresivității, se recurge la teste de ascultare (en. *listening test*), prin care participanții voluntari analizează, pe diferite scale de măsură (1-5, 1-100) caracteristici ale vocii precum naturalețea, expresivitatea, inteligibilitatea, similitudinea cu vorbitorul, etc.. Din nou, fiind implicată resursa umană, rezultatele sunt dificil de cuantificat obiectiv și de generalizat.

În literatura de specialitate există o multitudine de studii care abordează expresivitatea vocii generate de un sistem de vorbire. Indexăm în continuare unele dintre cele mai importante și actuale astfel de studii. (Skerry-Ryan et al., 2018) implementează o extensie a sistemului de sinteză bazat pe rețele neuronale dezvoltat de Google, numit **Tacotron** (Shen et al., 2018). Folosind o referință audio ce conține tipul de prozodie dorit, sistemul propus învață o reprezentare vectorială (en. *embedding)* a prozodiei, pe care o utilizează ulterior în cadrul etapei de sinteză. Astfel, sistemul reține informații prozodice (tipul emoției, intonație, etc.) care nu pot fi extrase din componenta text și nici din identitatea vorbitorului, dar necesare în transferul prozodiei.

În completare, (Wang et al., 2018 a) introduce „Global Style Tokens" (Tacotron GST) pentru a modela prozodia. În timpul antrenării, sistemul de sinteză Tacotron GST învață reprezentări vectoriale ale stilurilor de vorbire prezente în datele de antrenare (en. *style tokens*), fără ca acestea să fie etichetate anterior (învățare nesupervizată). Stilurile de vorbire identificate de către acești tokeni reprezintă, de fapt, dimensiunea de variabilitate maximă a datelor de antrenare. În etapa de sinteză, transferul prozodiei se poate realiza în două moduri:

1. Se pot modifica (manual) valorile/ponderile acestor tokeni pentru a obține tipul de prozodie dorit (din cele existente în datele de antrenare). În acest fel se poate modela prozodia fără a fi necesară o referință audio suplimentară.

2. Alături de textul care se dorește a fi sintetizat, sistemul primește și o referință audio (care poate sau nu să facă parte din setul de antrenare) din care se extrage informația de prozodie cuantificată ca o combinație liniară a tokenilor de stil.

Tot pe baza sistemului Tacotron, (Stanton et al., 2018) introduce „Text Predicting Global Style Token”: reprezentări latente ale stilului extrase direct din textul de intrare și folosite ulterior în procesul de sinteză.

Toate studiile enumerate anterior folosesc sistemele preantrenate cu date ce provin de la mai mulți vorbitori. În (Parker et al., 2018) este propusă o arhitectură neuronală antrenată pentru un singur vorbitor. Pentru a se realiza transferul expresivității către un vorbitor nou, modelul e augmentat cu un strat ascuns complet conectat (en. *fully connected*) responsabil cu învățarea caracteristicilor noi ale vorbitorului țintă. Astfel, sistemul preantrenat va fi folosit pentru a estima caracteristicile acustice, în timp ce neuronii din stratul ascuns vor învăța diferențele dintre vorbitorul sursă și cel nou.

(Yosinski et al., 2014) introduce ideea adaptării straturilor unei rețele neuronale (en. *layer adaptation*), conform căreia primele straturi ale unei rețele neuronale sunt responsabile de reprezentarea și învățarea caracteristicilor datelor de intrare (prozodie, distribuție, etc.) mai degrabă decât de rezolvarea unor sarcini specifice pentru care este folosită acea rețea neuronală. Pornind de la acest studiu, (Kulkami et al., 2019) propune transferul de cunoștințe (en *transfer learning*) și adaptarea straturilor unei rețele neuronale ca soluții pentru transferul prozodiei unui vorbitor către un altul pentru care există doar date neutre. Datele expresive folosite pentru antrenare fac parte din 5 categorii de emoții (bucurie, surpriză, frică, tristețe și dezgust). Acest model se bazează pe utilitarul Merlin (Wu et al., 2016), un sistem de sinteză ce folosește rețele neuronale complexe multistrat (en. *deep neural networks*).

Deși, în general, studiile de specialitate tratează expresivitatea din perspectiva emoțiilor de bază (bucurie, surpriză, frică, tristețe, dezgust), în experimentele proprii am analizat expresivitatea pornind de la stilul de vorbire (neutru, narativ, etc.) independent de emoții. Datele sunt în limba română, iar sistemul nu cunoaște apriori din ce categorie prozodică fac parte. Totul este învățat într-un proces complet (en. *end-to-end*) care folosește date de intrare de tip perechi de fișier audio - text corespunzător. Rețeaua neuronală învață astfel, o mapare a secvențelor de text la secvențe audio parametrizate sau direct forma de undă.

În continuare va fi prezentată arhitectura folosită în cadrul experimentelor SINTERO din această etapă. De asemenea, vor fi analizate rezultatele obținute în urma antrenării sistemelor de sinteză text-vorbire.

2. Descrierea arhitecturii GST

Deoarece Tacotron GST (Wang et al., 2018 a) oferă posibilitatea de a reține și de a modela prozodia unui vorbitor prin intermediul Global Style Tokens, am ales această arhitectură pentru experimentele prezentate în acest raport. Toate cele 8 modele descrise au fost antrenate folosind implementarea Mozilla a arhitecturii Tacotron GST[1].

Principalele componente ale sistemului sunt enumerate mai jos, iar o schemă a acestuia este prezentată în Figura 1:

1. Modulul Tacotron (Wang et al., 2018 b) - sistemul de sinteză de bază;

[1] "Mozilla-TTS - GitHub." https://github.com/mozilla/TTS.

2. Modulul Style Tokens - învață reprezentarea latentă a prozodiei prin două componente:
   2.1. Un codor pentru referința audio (en. *reference encoder*) - format din rețele neuronale recurente și convoluționale;
   2.2. Un modul denumit Style Token - format dintr-un set de neuroni denumiți tokeni și un strat de atenție ce folosește acești tokeni pentru a genera o reprezentare vectorială a stilului pornind de la codarea referinței audio.

În timpul procesului de **inferență**, sistemul poate primi date de intrare sub forma unei referințe audio a cărei prozodie trebuie reprodusă de sistem. Pe de altă parte, se poate renunța la referința audio, modelarea prozodiei realizându-se prin setarea manuală a valorilor reprezentărilor de stil (style tokens) învățate în timpul antrenării.

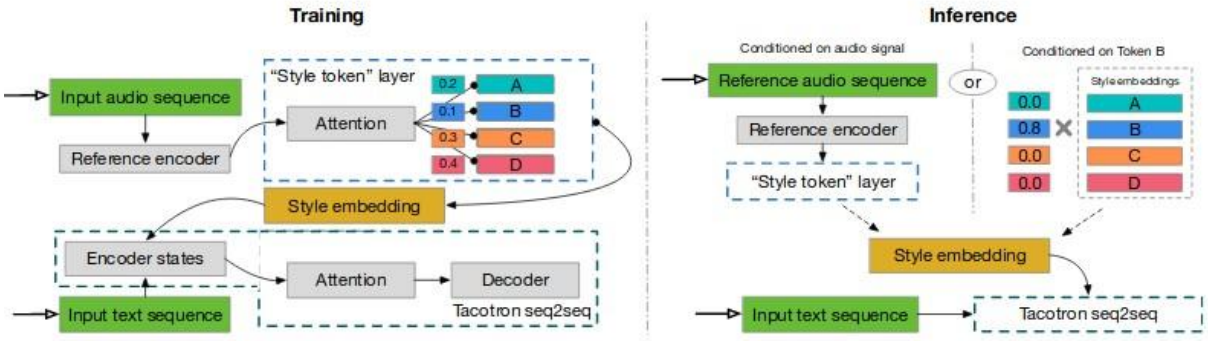

Figura 1. Arhitectura TacotronGST (Wang et al., 2018 a)

3. Experimente

Sistemele recente de sinteză text-vorbire bazate pe rețele neuronale generează voci a căror naturalețe este comparabilă cu vocea umană. Din păcate această naturalețe este bazată în principal pe o cantitate mare de date de antrenare. Reducerea numărului de înregistrări audio de calitate utilizate în antrenare atrage după sine și o reducere a calității sistemului de sinteză. De asemenea, dacă scopul sistemului este de a genera vorbire expresivă, aceste date de antrenare trebuie să conțină astfel de înregistrări, preferabil achiziționate de la același vorbitor.

Pentru limba română, corpusurile de date audio expresive disponibile au o dimensiune redusă, astfel încât antrenarea unui sistem de sinteză bazat pe rețele neuronale complexe ridică anumite probleme. S-a analizat astfel posibilitatea de a transfera prozodia unui vorbitor țintă folosind un număr redus de date. Au fost urmărite două abordări, ambele având la bază un sistem preantrenat cu date acustice expresive. Într-un prim pas, am urmărit transferul prozodic către un singur vorbitor pentru care există doar înregistrări cu prozodie neutră. În al doilea pas, a fost analizat și cazul în care sistemul țintă către care se realizează transferul prozodiei este antrenat cu date ce provin de la mai mulți vorbitori.

3.1 Descrierea seturilor de date utilizate în antrenarea sistemelor

Pentru sistemele de sinteză testate am utilizat date de antrenare din două corpusuri în limba română: **MARA** și **SWARA** (Stan et al., 2017).

Corpusul **MARA** conține redarea audio a întregului roman „Mara” de Ioan Slavici, înregistrat în condiții de studio de către o actriță profesionistă. Datele sunt eșantionate la 44kHz

cu 16bps și conțin aproximativ 11 ore de înregistrări. Pentru a reduce timpul de procesare al acestui corpus, datele audio au fost ulterior sub-eșantionate la 16kHz. Segmentarea corpusului la nivel de propoziție sau frază de dimensiuni reduse a fost realizată manual. Alinierea la nivel de fonem s-a realizat semi-automat folosind un sistem bazat pe modele Markov și antrenat pe întreg corpusul audio.

Corpusul **SWARA** conține date neutre de la 17 vorbitori (aproximativ 21 ore de vorbire), segmentate manual la nivel de propoziție și aliniate semi-automat la nivel de fonem. Datele audio din SWARA sunt eșantionate la 16kHz cu 16bps. Pentru experimentele prezentate în acest raport au fost selectați 10 vorbitori în două configurații:

- 10 vorbitori feminini - *BAS, CAU, DCS, DDM, EME, HTM, PMM, SAM, TIM* și *MAR. MAR* este un vorbitor înregistrat ulterior, folosind condiții similare de studio;
- 5 voci feminine - *CAU, DCS, DDM, EME, SAM* și 5 voci masculine - *FDS, IPS, PCS, PSS, TSS.*

Pentru fiecare vorbitor au fost selectate cele 500 de propoziții ale subsetului *RND1* (aproximativ 50 minute de vorbire). Niciunul dintre acești vorbitori nu este orator profesionist.

3.2 Sistemele de sinteză antrenate pentru transferul prozodiei

În această secțiune va fi descris modul în care arhitectura Tacotron GST și corpusurile audio au fost combinate pentru a evalua transferul prozodiei în cadrul sistemelor de sinteză text-vorbire în limba română bazate pe rețele neuronale. Sunt prezentate în total un număr de 8 modele neuronale fiecare având ca obiectiv un scenariu specific de transfer prozodic.

M1. Modelul MARA - modelul de bază

Ca prim pas, arhitectura TacotronGST a fost antrenată cu întreg corpusul expresiv MARA, pentru a obține un sistem de referință (denumit simplu MARA) și un punct de pornire pentru sistemele următoare. Modelul a fost antrenat 800 epoci cu un lot de dimensiune 16. Deoarece datele de antrenare conțin un grad ridicat de expresivitate, modulul GST a reușit să surprindă această variabilitate prozodică a datelor audio. S-au folosit două configurații: cu 10 tokeni și cu 5 tokeni de stil. În fiecare din cele două configurații, tokenii au reținut stiluri diferite de prozodie.

În etapa de sinteză, au fost analizate două abordări.

- **setarea manuală a tokenilor de stil**: pentru a urmări și evidenția conținutul învățat de fiecare token de stil, au fost generate date audio pentru fiecare token în parte. Astfel, în reprezentarea vectorială dată de modulul GST, ponderea tokenului analizat a fost setată manual cu o valoare din intervalul (0,1), în timp ce ponderile celorlalți tokeni au primit valoarea 0. Aceasta abordare este ilustrată în Figura 1., în partea de inferență, sub numele de *conditioned on token B*. Se observă că fiecare token a învățat un stil prozodic diferit. Astfel, prin această reglare manuală a valorilor tokenilor de stil poate fi modelat și controlat stilul prozodic al vocii generate.
- **utilizarea unei referințe audio**: în articolul (Wang et al., 2018 a) alături de textul care se dorește a fi sintetizat, sistemul primește și o referință audio (care poate sau nu să facă parte din setul de antrenare) din care să fie extrasă prozodia. Pentru a analiză această modalitate de transfer al prozodiei, am folosit mai multe tipuri de voce în fișierele audio de referință: voce masculină, voce cântată acapela sau voce feminină neexpresivă, cu intonație neutră. Rezultatele audio au

arătat că referința audio furnizată la intrare în timpul procesului de sinteză nu influențează în mod semnificativ prozodia semnalului de ieșire.

Exemple audio pentru acest model sunt disponibile și pot fi ascultate aici: https://speech.utcluj.ro/sintero/prosody_examples_2019/.

M2. Modelele IPS-peste-MARA și EME-peste-MARA

Pornind de la modelul de bază MARA, următorul pas a fost să analizăm dacă prozodia vorbitorului inițial este menținută atunci când ponderile rețelei sunt actualizate folosind date audio de la un vorbitor cu prozodie neutră. Scopul fiind de a combina identitatea noului vorbitor cu prozodia celui anterior. Pentru acest experiment au fost alese 2 voci din SWARA: una masculină (IPS) și una feminină (EME).

Cele două noi sisteme sunt denumite sugestiv ***IPS-peste-MARA*** și ***EME-peste-MARA*****.** Chiar după prima epocă de antrenare, ***modelul MARA*** s-a adaptat către identitatea noului vorbitor impus (IPS, respectiv EME), însă vocile generate au păstrat doar parțial din caracteristicile prozodiei din MARA. Exemple audio pentru aceste două modele pot fi accesate și ascultate aici: https://speech.utcluj.ro/sintero/prosody_examples_2019/.

M3. Modelul SWARA-peste-MARA

Pornind de la ideea modelului anterior, de a transfera prozodia vocii din MARA către un vorbitor cu o voce neutră, s-a analizat și transferul prozodic către un model mixt, antrenat cu date de la mai mulți vorbitori. Scopul fiind același: de a transfera informațiile de prozodie extrase de modulul GST din datele audio MARA către alți vorbitori ce nu includ înregistrări audio expresive. Astfel, pentru a obține un sistem de sinteză expresiv cu mai mulți vorbitori (en. *multi-speaker*), rețeaua antrenată cu datele din corpusul **MARA** a fost antrenată în continuare cu date mixte de la 10 vorbitori din corpusul SWARA. S-a observat că, deși în modelul inițial **MARA** tokenii de stil realizau o discriminare a stilurilor prozodice din datele de intrare, în modelul **SWARA-peste-MARA** aceeași tokeni surprind identitatea vorbitorilor din datele de antrenare.

La sinteză, forțând valoarea unui anumit token (valoare setată manual între 0-1, ceilalți nouă tokeni fiind 0), vocea generată conține caracteristicile unui anumit vorbitor mai degrabă decât un stil prozodic. Exemple audio pentru acest model sunt disponibile aici: https://speech.utcluj.ro/sintero/prosody_examples_2019/.

A fost testat și scenariul în care cei 10 vorbitori SWARA fac parte din genuri diferite. Astfel, au fost alese 5 voci feminine și 5 voci masculine. Și în acest caz, tokenii au reținut caracteristici ale identității vorbitorilor iar nu stilul de prozodie.

M4. Modelele SWARA-peste-MARA cu ponderile stratului GST fixate (cu 10 și 15 tokeni de stil)

Pe baza rezultatelor modelului anterior și în principal pornind de la observația că modulul GST reține dimensiunea de variabilitate maximă din datele de antrenare, următorul pas al experimentelor noastre a avut în vedere păstrarea tokenilor învățați din datele audio MARA. Astfel că, a fost antrenat un model nou Tacotron GST în cadrul căruia ponderile pentru cei 10 tokeni de stil au fost fixate la valorile din modelul MARA (en. *freezed*). Restul ponderilor modulului GST putând fi modificate de procesul de antrenare cu noile date audio. Și în acest context, deși ponderile tokenilor de stil erau fixate, restul ponderilor modulului GST s-au adaptat pentru a prioritiza identitatea vorbitorilor, ignorând prozodia învățată anterior.

În același context, s-a avut în vedere extinderea numărului de tokeni de stil. Un nou model Tacotron GST a fost antrenat, de data aceasta folosind 15 tokeni de stil. Primii 10 tokeni au fost inițializați și fixați la valorile tokenilor din modelul **MARA** (pentru a păstra prozodia). Ultimii 5 tokeni au fost inițializați aleatoriu și au putut fi actualizați de procesul de învățare al rețelei neuronale. Au fost folosite datele audio de la aceiași 10 vorbitori din corpusul SWARA.

Din nou, deși ponderile tokenilor de stil au fost fixate, restul ponderilor modulului GST s-au adaptat pentru a reține caracteristici ale identității vorbitorilor. Ultimii 5 tokeni au fost ignorați. O posibilă interpretare a acestui rezultat poate fi aceea că, modulul GST învățând variația cea mai mare din datele de antrenare, rețeaua renunță la expresivitate în favoarea identității vorbitorilor. Exemple audio pentru acest model sunt disponibile aici: https://speech.utcluj.ro/sintero/prosody_examples_2019/.

M5. Modelul SWARA-peste-MARA cu întreg modulul GST fixat

Având în vedere faptul că simpla fixare a ponderilor tokenilor de stil din modulul GST nu a fost suficientă pentru a menține prozodia din corpusul MARA și de a o transfera către vorbitorii din SWARA, următorul pas în experimentele noastre a fost de a fixa ponderile întregului modul GST. Restul arhitecturii putând fi ajustat în timpul învățării. Din păcate și în acest scenariu, rețeaua neuronală a ignorat reprezentările vectoriale date de modulul GST, iar prozodia vorbitorului MARA a fost din nou suprascrisă de prozodia vorbitorilor din SWARA.

M6. Modelul SWARA-peste-MARA cu ponderile stratului GST fixate și date de antrenare îmbogățite cu mostre audio din corpusul MARA

Pornind de la rezultatele experimentului anterior și având ca obiectiv evitarea fenomenului de uitare a prozodiei învățate de către rețea din corpusul MARA, s-a optat pentru extinderea datelor extrase din SWARA cu date expresive ale vorbitorului MARA. Astfel, din corpusul MARA au fost selectate acele fișiere audio care prezintă o expresivitate mai mare, expresivitate măsurată în funcție de variația frecvenței fundamentale la nivel de propoziție.

Arhitectura TacotronGST a fost antrenată în continuare, pornind de la ponderile modelului MARA. Ca date de antrenare au fost folosite înregistrările de la cei 10 vorbitori din corpusul SWARA, la care au fost adăugate propozițiile expresive din MARA. Similar cu experimentul anterior, s-au folosit 15 tokeni de stil, dintre care primii 10 au fost inițializați și fixați la valorile tokenilor din modelul MARA preantrenat, iar ultimii 5 au fost inițializați cu valori arbitrare, aceștia putând fi ajustați în procesul de învățare.

La fel ca în cazul modelelor cu vorbitori multipli descrise anterior, în ciuda faptului că sistemul a avut acces la date din corpusul inițial (MARA) pentru a-și „reaminti” prozodia, tokenii au învățat identitatea vorbitorului în defavoarea prozodiei. Într-o oarecare măsură, însă, aceștia au păstrat și caracteristici ale stilurilor prozodice. Exemple audio pentru acest model sunt disponibile aici: https://speech.utcluj.ro/sintero/prosody_examples_2019/.

Toate sistemele descrise anterior sunt rezumate în Tabelele 1 și 2 pentru o sumarizare mai bună a rezultatelor descrise în cadrul acestui raport.

Sistem	Preantrenare	Antrenare	Număr propoziții antrenare	Inițializări	Număr tokeni	Informația din tokeni
MARA	--	Mara	7932	--	10	stil prozodic
IPS-peste-MARA	Mara	IPS	500	ponderi tokeni de stil model MARA	5	parțial prozodie
EME-peste-MARA	Mara	EME	500	ponderi tokeni de stil model MARA	10	parțial prozodie

*Tabel 1. Descrierea sistemelor antrenate pentru transferul prozodiei către **un singur vorbitor***

Sistem	Preantrenare	Antrenare	Număr propoziții antrenare	Inițializări	Număr tokeni	Informația din tokeni
SWARA-peste-MARA	Mara	Swara	10 x 500	ponderi model MARA	10	identitate vorbitori
SWARA-peste-MARA cu ponderile tokenilor de stil din stratul GST fixate	Mara	Swara	10 x 500	ponderi tokeni de stil fixate din modelul MARA	10	identitate vorbitori
SWARA-peste-MARA cu ponderile tokenilor din stratul GST fixate	Mara	Swara	10 x 500	ponderi tokeni de stil fixate din modelul MARA	15	identitate vorbitori
SWARA-peste-MARA cu întreg modulul GST fixat	Mara	Swara	10 x 500	modul GST fixat din modelul MARA	15	identitate vorbitori
SWARA-peste-MARA cu întreg modulul GST fixat și date de antrenare îmbogățite cu date din Mara	Mara	Swara + 500 Mara	10 x 500 + 500 MARA expresive	modul GST fixat din modelul MARA	15	identitate vorbitori + parțial prozodie

*Tabel 2. Descrierea sistemelor antrenate pentru transferul prozodiei către **mai mulți vorbitori***

4. Concluzii

În acest raport a fost prezentat un set de experimente realizata în vederea transferului prozodiei de la un vorbitor expresiv către unul sau mai mulți vorbitori neutri. S-a folosit o arhitectură de rețele neuronale recurente și convoluționale, Tacotron GST, care permite modelarea prozodiei unui vorbitor prin manipularea unor reprezentări latente ale stilurilor de vorbire (tokeni de stil). În experimentele derulate de noi, s-a urmărit în principal fixarea unor componente de rețea în vederea transferului de cunoștințe de la o etapă de antrenare către următoarea. Din păcate, datorită complexității rețelei, aceasta poate să își adapteze rapid ponderile în noile etape de antrenare, astfel încât aceasta poate să ignore complet ceea ce învățase anterior. Totodată, s-a putut observa faptul că în arhitectura modulului GST, tokenii rețin dimensiunea de variabilitate maximă a datelor de antrenare (ex. prozodia pentru un singur vorbitor, respectiv identitatea vorbitorilor pentru sisteme antrenate cu date de la mai mulți vorbitori). Ca urmare, păstrarea informației anterioare în cadrul modulului GST nu este fezabilă.

În dezvoltările următoare, pentru a îmbunătăți transferul prozodiei, vor fi abordate alte tehnici care folosesc rețelele neuronale: învățarea continuă (en. *continual learning*), învățarea folosind puține eșantioane (en. *few-shots/ one-shot learning*). O altă metodă ar fi augmentarea setului de date neutre de antrenare cu date sintetice generate de o voce expresivă.

5. Bibliografie

Kulkarni et al., 2019 — Kulkarni, A., Colotte, V., & Jouvet, D. (2019, May). Layer adaptation for transfer of expressivity in speech synthesis. (online)

Parker et al., 2018 — Parker, J., Stylianou, Y., & Cipolla, R. (2018). Adaptation of an expressive single speaker deep neural network speech synthesis system. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5309-5313) (online)

Shen et al., 2018 — Shen Jonathan, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu (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. (online)

Skerry-Ryan et al., 2018 — Skerry-Ryan, R. J., Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, Rif A. Saurous (2018). Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. *arXiv preprint arXiv:1803.09047*.

Stan et al., 2017 — Stan, A., Dinescu, F., Ţiple, C., Meza, Ş., Orza, B., Chirilă, M., & Giurgiu, M. (2017, July). The SWARA speech corpus: A large parallel Romanian read speech dataset. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (pp. 1-6). IEEE. (online)

Stanton et al., 2018 — Stanton, D., Wang, Y., & Skerry-Ryan, R. J. (2018, December). Predicting expressive speaking style from text in end-to-end speech synthesis. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 595-602). IEEE.

(online)

Wang et al., 2018 a

Wang, Yuxuan, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, Rif A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis." https://arxiv.org/abs/1803.09017 (2018).

Wang et al., 2018 b

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Le, Q. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Wu et al., 2016

Wu, Z., Watts, O., & King, S. (2016, September). Merlin: An Open Source Neural Network Speech Synthesis System. In *SSW* (pp. 202-207). (online)

Yosinski et al., 2014

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. In *Advances in neural information processing systems* (pp. 3320-3328). (onlne)