D3.15. Dezvoltarea unei noi tehnologii pentru adaptarea vocii sintetice la stilul și expresivitatea unui nou vorbitor

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”

© 2018-2020 – SINTERO
Acest document este proprietatea organizațiilor participante în proiect și nu poate fi reprodus, distribuit sau diseminat către terți, fără acordul prealabil al autorilor.

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:	**D3.15. Dezvoltarea unei noi tehnologii pentru adaptarea vocii sintetice la stilul și expresivitatea unui nou vorbitor**
Termen:	**Noiembrie 2020**
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 metodele și rezultatele experimentale obținute în urma implementării tehnologiei de sinteză text-vorbire bazată pe adaptarea vorbirii sintetice la stilul și expresivitatea unui nou vorbitor. Această adaptare a fost realizată prin intermediul a trei noi metode: adaptarea de la un sistem cu vorbitori multipli la un sistem cu un singur vorbitor folosind un set redus de date, adaptarea de la un sistem cu voce ne-expresivă la un sistem cu voce expresivă, adaptarea sistemului de sinteză la un nou vorbitor printr-o procedură de adaptare bazată pe post-filtrare. Pentru a asigura implementarea acestor tehnologii de adaptare, a fost necesară extinderea corpusului deja existent SWARA cu noi voci, corpusul fiind numit SWARA 2.0 și conține aproximativ 65 de ore de vorbire. Crearea corpusului SWARA 2.0 s-a realizat cu aplicația software RecoApy care permite înregistrarea secvențială a textelor și nu necesită segmentare ulterioară. Această aplicație suportă mai multe limbi, inclusiv pentru transcrierea fonetică. Pentru a beneficia de datele disponibile pentru mai mulți vorbitori, dar și pentru a adapta sistemele de sinteză text - vorbire la stilul și expresivitatea unui nou vorbitor, am proiectat o serie de metode de modelare acustică și de sinteză de semnal bazate de diferite structuri de rețele neuronale profunde (CNN, Transformer, DC-TTS, Tacotron, Tacotron2, Flowtron, etc), le-am implementat pe 2 servere de procesare paralelă a datelor, fiecare dotate cu plăci GPU de ultimă generație, și am efectuat o serie de experimente pentru vorbitori multipli. Rezultatele evaluării obiective a sistemelor sunt detaliate, iar rezultate audio ale acestor sisteme sunt indicate în diferite pagini web aferente proiectului SINTERO.

Cuprins

1. Introducere

Acest raport prezintă setul de dezvoltări și experimente realizate în vederea extinderii resurselor de date audio folosite anterior acestei etape de raportare și utilizarea acestora pentru antrenarea sistemelor de sinteză text-vorbire în scopul demonstrării tehnologiei de adaptare atât la pentru un vorbitor unic, precum și pentru vorbitori multipli.

În primul rând, raportul prezintă noul corpus cu semnal vocal SWARA2.0 colectat pentru a extinde adaptarea sistemului de sinteză la noi vorbitori, aplicația software RecoApy utilizată pentru crearea acestui corpus audio, precum și experimentele realizate în vederea minimizării necesarului de date de antrenare a sistemului de sinteză, dar totuși cu îmbunătățirea calității sintezei în sistemele cu vorbitori multipli, folosind o funcție de cost suplimentară bazată pe rata de eroare egală (en. *EER - Equal Error Rate)*.

Pentru validarea noii tehnologii de sinteză text-vorbire, s-au realizat o serie de experimente care pun în evidență adaptarea prozodiei sistemului de sinteză la stilul și expresivitatea noului vorbitor prin augmentarea datelor disponibile în procesul de antrenare.

2. Corpusul vocal SWARA 2.0 pentru adaptarea sistemului de sinteză la noi vorbitori

În perioada aprilie-iunie 2020 au fost efectuate în mod intensiv înregistrări audio în vederea extinderii corpusului de date deja existent (SWARA) în cadrul proiectului SINTERO. Având în vedere situația sanitară, înregistrările au fost efectuate de către studenți voluntari folosind echipamentele proprii și nu în cadrul studioului de înregistrări de voce ce aparține grupului nostru de cercetare, așa cum s-a înregistrat corpusul inițial SWARA. Pentru a facilita înregistrările pe dispozitivele proprii ale voluntarilor și pentru a minimiza timpul necesar prelucrărilor și segmentărilor manuale, a fost dezvoltat un tool automat, denumit RecoApy, descris în secțiunea următoare și prezentat în cadrul conferinței Interspeech 2020.

Noile înregistrări realizate conțin suplimentar față de datele din versiunea 1 a corpusului audio SWARA și înregistrarea povestirii scurte ”Ivan Turbincă” a lui Ion Creangă, pentru care s-a solicitat voluntarilor generarea vocală a unei intonații expresive față de subsetul audio alcătuit din extrase de știri. A rezultat astfel un set de înregistrări ce conține 51.839 de segmente audio de la 29 de vorbitori: 14 masculini și 15 feminini. Durata totală a înregistrărilor este de aproximativ 65 de ore.

Corpusul audio SWARA 2.0 a fost colectat respectând prevederile protecției datelor cu caracter personal, iar acordul semnat de către vorbitori este disponibil în Anexa 1. Un rezumat al conținutului corpusului SWARA 2.0 este prezentat în Tabelul 1. Împreună cu prima versiune a corpusului SWARA, aceasta reprezintă cea mai mare resursă audio paralelă de înaltă calitate pentru limba română.

Tabel 1. Descriere conținut corpus SWARA 2.0

Nr.	ID vorbitor	Sex	Durata înregistrării
1	***BAL***	*M*	*2h:4’*
2	***BGL***	*F*	*2h:17’*
3	***BIM***	*M*	*1h:45’*
4	***BMM***	*F*	*2h:3’*

5	***BVL***	*M*	*3h:42'*
6	***CCL***	*F*	*2h:24'*
7	***CMM***	*F*	*1h:57'*
8	***DLL***	*F*	*2h:20'*
9	***DOL***	*F*	*2h:7'*
10	***GAM***	*F*	*2h:19'*
11	***GIM***	*F*	*2h:0'*
12	***GNM***	*F*	*1h:52'*
13	***HRM***	*M*	*1h:53'*
14	***LMM***	*F*	*2h:32'*
15	***MAL***	*F*	*2h:14'*
16	***MAM***	*M*	*2h:39''*
17	***MGL***	*M*	*2h:21''*
18	***MRL***	*F*	*2h:13'*
19	***MSM***	*M*	*1h:56''*
20	***MXM***	*M*	*2h:25''*
21	***NAM***	*M*	*1h:29'*
22	***NLL***	*M*	*1h:57'*
23	***OGL***	*F*	*2h:13'*
24	***PBL***	*F*	*2h:2'*
25	***PDL***	*M*	*2h:4'*
26	***PTL***	*M*	*2h:39'*
27	***SMM***	*F*	*2h:10'*
28	***SRL***	*M*	*2h:11'*
29	***ZPL***	*M*	*2h:18'*

În această etapă de raportare (2020) corpusul SWARA 2.0 a fost utilizat pentru a antrena un vocoder de tip WaveGlow, precum și o serie de noi voci pentru a valida tehnologia de adaptare a sistemului de sinteză la noi vorbitori. Mostre audio aferente noilor voci dezvoltate pe baza acestui corpus sunt disponibile la adresa: https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home.

Vocile au fost antrenate folosind o arhitectură de rețea neuronală profundă de tip Tacotron2 (Shen et al., 2018) și vocoder-ul WaveGlow (Prenger et al., 2019). Pentru a permite antrenarea cu un set de date cât mai redus, s-a aplicat metoda de transfer a învățării (en. *Transfer Learning*) în cadrul căreia s-a antrenat inițial o voce pe baza corpusului audio Mara (15 ore de vorbire de înaltă calitate de la un singur vorbitor), iar apoi ponderile rețelei au fost antrenate prin Transfer Learning către un nou vorbitor.

Suplimentar, s-a cercetat modul de comportare a vocoderului WaveGlow antrenat cu date mixte (feminin, masculin) și s-a demonstrat că acest vocoder prezintă anumite erori de generare a vocii sintetizate, în special pentru vocile masculine. Prin urmare, s-a renunțat la antrenarea cu date mixte (feminin, masculin) și s-a realizat antrenarea acestui vocoder cu date provenite doar de la vorbitori masculini din ambele versiuni ale corpusului SWARA. În total s-au utilizat 20 de vorbitori, 30.584 de segmente audio, cu o durată totală de 35 de ore și 40 de minute. În consecință, vocoderul WaveGlow este în prezent utilizat pentru a genera vocea sintetică a vorbitorilor masculini din API-ul RoNNA: www.speech.utcluj.ro/ronna.

3. RecoApy - o nouă aplicație software pentru automatizarea înregistrărilor audio necesare în sistemele moderne de sinteză de voce de tip E2E (End-to-End)

Aplicația software RecoApy prezintă o interfață grafică ușor de utilizat pentru realizarea înregistrărilor audio necesare în antrenarea sistemelor de sinteză text-vorbire. Tool-ul este implementat în Python, iar o captură de ecran este prezentată în Figura 1.

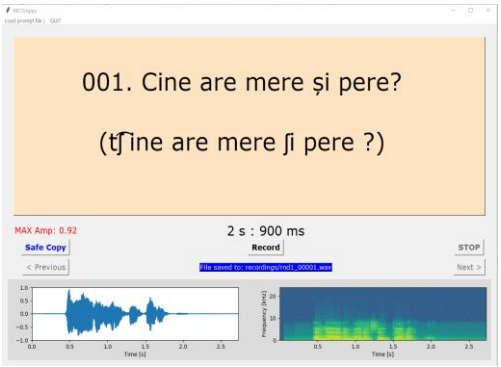

Fig.1. Interfața aplicației software RecoApy

RecoApy permite înregistrarea secvențială de către vorbitori a propozițiilor prezentate acestora sub formă de prompt-uri. Ca atare, vorbitorul poate să înregistreze câte un segment audio distinct pentru fiecare propoziție sau frază de intrare, iar dacă a greșit citirea, poate relua înregistrarea. Adițional, se permite salvarea unei copii în cazul în care vorbitorul sau operatorul înregistrărilor nu este sigur de corectitudinea citirii prompt-ului. La finalul înregistrărilor se poate realiza automat normalizarea la 0dB și eliminarea pauzelor de început și final.

Spre deosebire de alte aplicații software de înregistrare audio, RecoApy nu înregistrează un flux continuu ce ar necesita ulterior verificarea și segmentarea manuală, ci înregistrări scurte aferente fiecărei propoziții din prompt. Ca o caracteristică suplimentară și utilă pentru înregistrările realizate, mai ales în limbi ce au o complexitate fonetică mai mare, precum limba engleză, RecoApy permite generarea și afișarea transcrierii fonetice a textului. Generarea transcrierii fonetice a textului este realizată folosind rețele neuronale convoluționale (CNN - Convolutional Neural Networks) și de tip Transformer (Vaswani et al., 2017). Aceste rețele neuronale au fost antrenate pe baza datelor extrase automat din resursa Wiktionary[1], iar hiper parametrii au fost optimizați cu ajutorul unei strategii evolutive.

Aplicația software a fost dezvoltată pentru 8 limbi: română, engleză, spaniolă, franceză, germană, italiană, cehă și poloneză. Performanța transcrierii fonetice cu această aplicație este le nivelul metodelor de dată recentă din domeniu. Aplicația software este este disponibilă în mod gratuit la adresa: https://gitlab.utcluj.ro/sadriana/recoapy. O prezentare mai detaliată a rezultatelor RecoApy a fost realizată în cadrul conferinței Interspeech 2020, iar această prezentare este indicată în pagina proiectului: www.speech.utcluj.ro/sintero/.

[1] *https://www.wiktionary.org*

4. Rezultate experimentale privind adaptarea sistemului de sinteză la un nou vorbitor și la un nou stil de vorbire

Pentru adaptarea unui sistem de sinteză la un nou vorbitor există două direcții, una axată pe adaptarea vorbitorului, iar a doua pe adaptarea de stil și expresivitate. Adaptarea sistemelor de sinteză la un nou vorbitor presupune ajustarea unui model existent, folosind cât mai puține date audio de la vorbitorul nou. Adaptarea sistemelor de sinteză pentru stil și expresivitate are în vedere modificare stilului de exprimare a unei voci sintetice la un nou stil, modificând caracteristicile vorbirii, cum ar fi intonația, ritmul, durata, pauzele din vorbire, dar fără a modifica identitatea vorbitorului.

În secțiunile următoare sunt descrise experimentele realizate în cadrul proiectului SINTERO în vederea obținerii acestor tipuri de adaptări ale sistemelor de sinteză text-vorbire.

4.1. Rezultate experimentale privind adaptarea unui sistem de sinteză cu vorbitori multipli pe baza unei funcții de cost suplimentare

Sistemele de sinteză antrenate cu semnal vocal de la mai mulți vorbitori au capacitatea de a sintetiza vocea pentru fiecare vorbitor în parte. Cele mai des folosite sisteme de sinteză la ora actuală sunt cele de tip end-to-end (doar semnal vocal, respectiv transcrierea ortografică a acestuia, fără alte adnotări suplimentare) bazate pe rețele neuronale. Totuși, de cele mai mai multe ori în sistemele cu vorbitori multipli, identitatea vorbitorului în procesul de sinteză nu este păstrată foarte bine. Acest fapt se datorează, printre altele, uniformizării caracteristicilor acustice în procesul de modelare acustică prin antrenarea rețelei neuronale.

Este important de notat faptul că o altă sarcină legată de procesarea de voce, o reprezintă recunoașterea sau verificarea identității unui vorbitor. Pe baza acestei metode (eg. de recunoaștere a vorbitorilor) se poate dezvolta un sistem de sinteză care să încerce să reproducă identitatea vorbitorilor din setul de date de antrenare, menținând în același timp și inteligibilitatea și naturalețea vorbirii sintetizate.

În consecință, pentru îmbunătățirea învățării identității vorbitorului și adaptarea unui sistem de vorbitori multipli la un nou vorbitor, s-a propus folosirea unui sistem de identificare a vorbitorului combinată cu un sistem de sinteză text-vorbire, prin adăugarea unei funcții de cost suplimentare bazată pe reprezentările interne ale sistemului de identificare a vorbitorului.

4.1.1 Descrierea metodei de adaptare folosind o funcție de cost suplimentară

Sistemul de sinteză cu vorbitori multipli este bazat pe arhitectura prezentată în (Tachibana et al., 2018) și utilizează rețele neuronale convoluționale (denumit DC-TTS în continuare). Figura 2 ilustrează arhitectura sistemului.

Sistemul de sinteză DC-TTS conține două componente care se antrenează separat. Prima componentă (Text2Mel) primește ca și intrare text, iar la ieșire prezice parametrii spectrogramei Mel reduse (un set redus de parametri Mel). A doua componentă generează din spectrograma Mel redusă, o spectrogramă completă (en: Spectrogram Super-Resolution Network - SSRN). Din spectrograma completă SSRN se generează forma de undă pe baza algoritmului iterativ Griffin-Lim (Griffin & Lim, 1984).

Componenta Text2Mel este compusă din 3 module principale: a) codorul de text, b) codorul audio și c) decodorul de audio.

Componenta SSRN conține un set de straturi convoluționale cu dilatare prin care s-a urmărit creșterea rezoluției în frecvență.

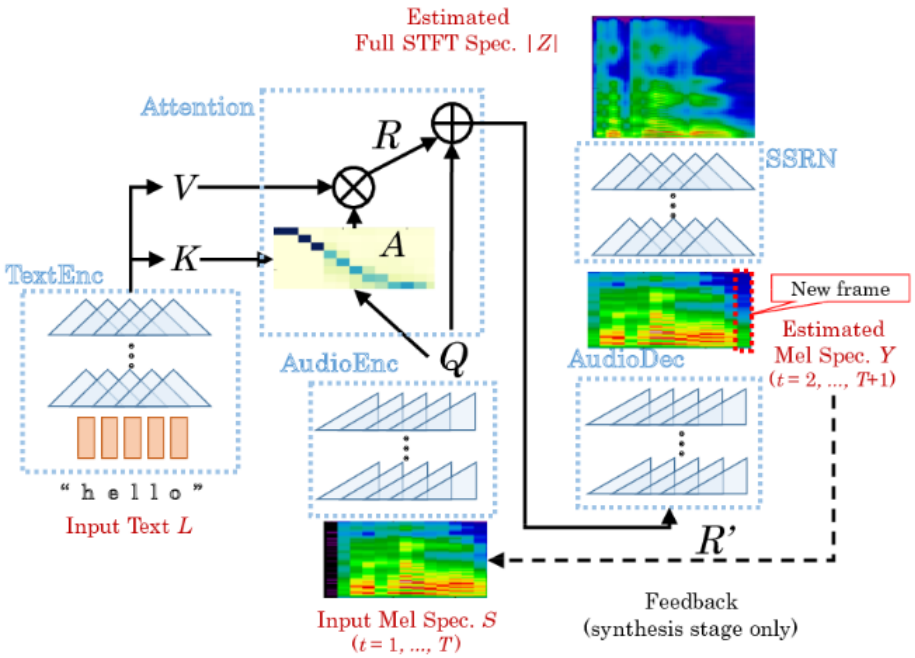

Fig. 2. Arhitectura sistemului DC-TTS (Tachibana et al., 2018)

În prima etapă s-a adoptat o implementare în Tensorflow a sistemului DC-TTS, pornind de la implementarea realizată în grupul de cercetare de la Universitatea din Edinburgh (CSTR)[2]. Prin mai multe experimente, am constatat că există probleme în integrarea rețelei de identificare a vorbitorilor pe care noi am propus-o.

În a doua etapă, am recurs la o implementare PyTorch a sistemului DC-TTS[3]. Din nou, printr-o analiză mai profundă a modelelor acustice generate și a modului în care acestea sunt implementate, am constatat că nu putem generaliza modulele software pentru a rezolva problema sintezei cu vorbitori multipli.

În consecință, pornind de la dezvoltările software realizate CSTR s-a extins rețeaua DC-TTS din Pytorch cu posibilitatea de concatenare a unei reprezentări vectoriale a identității vorbitorului (numită *embedding* în continuare) la cele 3 module ale componentei Text2Mel, respectiv prin utilizarea reprezentării vectoriale într-o strategie de învățare a contribuției reprezentării vectoriale la canalele de informație din rețea (en. *channel contributions*). Astfel, este posibilă inserarea identității vorbitorului în mai multe puncte ale rețelei. Având la dispoziție aceste modificări ale sistemului de sinteză, s-a trecut la implementarea unei funcții de cost suplimentare utilizate în antrenarea sistemului și care să ducă antrenarea mai aproape de identitatea vorbitorilor din setul de antrenare.

Astfel că, sistemul cu vorbitor multipli a fost antrenat în 3 scenarii, listate și detaliate după cum urmează:

1) **Scenariul 1:** sistem cu vorbitor unic sau vorbitori multipli considerat sistemul de bază (ID: **B**)
2) **Scenariul 2**: sistem cu vorbitor unic sau vorbitori multipli și adăugarea unei funcții de cost suplimentare obținută prin calculul similarității spectrale (en: Cosine Similarity) (ID: **B+CS**)
3) **Scenariul 3**: sistem cu vorbitor unic sau vorbitori multipli cu adăugarea unei funcții de cost suplimentare obținută cu rata de eroare egală (en: Equal Error Rate) folosind sistemul de verificare de vorbitor (ID: **B+E**)

[2] http://www.cstr.ed.ac.uk/
[3] https://github.com/tugstugi/pytorch-dc-tts

Cele 3 scenarii au fost antrenate cu diferite seturi de date:

- ALL - folosind toate datele de la 18 vorbitori din corpusul SWARA: între 1.000 și 1.500 de pronunții / propoziții de la fiecare vorbitor, în total 21.302 pronunții (ID: **ALL**)
- RND1 - folosind un subset de aproximativ 500 de pronunții / vorbitor de la 18 vorbitori din corpusul SWARA: în total 8.932 de pronunții (ID: **RND1**)
- RND1-100 - folosind un subset de aproximativ 100 de pronunții / vorbitor de la 18 vorbitori din corpusul SWARA: în total 1.787 de pronunții (ID: **RND1-100**)
- RND1-SAM - folosind un subset de date de la un singur vorbitor: în total 500 de pronunții (ID: **RND1-SAM**)

Scenariul 1 - B. S-au antrenat mai multe sisteme de sinteză folosind două metode, a) metoda de codare prin embedding a vorbitorilor, b) metoda bazată pe contribuția canalelor de învățare. Analizând rezultatele, am selectat a doua metodă pentru sistemul de sinteză cu vorbitori multipli, deoarece rezultatele în sinteză la sistemele bazate pe embedding de vorbitori nu puneau clar în evidență identitatea vorbitorilor. Ca atare, sistemul de referință (B - Baseline) este antrenat cu funcția de cost L1 și strat de atenție între spectrograma prezisă și cea naturală.

Scenariul 2 - B+CS. Similar cu Scenariul 1, sistemul cu vorbitori multipli a fost antrenat prin metoda contribuției la canalele de învățare, doar că funcția de cost a fost extinsă cu o componentă dependentă de vorbitor. Am folosit metrica de similitudine a cosinusului pentru spectrograma prezisă și cea naturală, calculată pentru fiecare vorbitor. Cu această metrică am pornit de la premiza că sistemul cu vorbitori multipli va genera voce sintetizată cu asemănare spectrală mai mare cu cea a vorbitorului dorit.

Scenariul 3 - B+E. Și aici, s-a procedat similar cu scenariul 2, dar folosind funcția de cost rata de eroare egală (EER - Equal Error Rate). Pentru a calcula aceasta valoare, este implicată noua componentă introdusă de către noi, și anume sistemul de verificare a vorbitorului. Acest sistem a fost antrenat cu corpusul SWARA folosind o implementare[4] de la Clova AI Research. Valoarea EER este un bun indicator pentru punctul în care valoarea de rata de acceptare falsă este egală cu rata de respingere falsă. Acesta este o metrică folosită des pentru a calcula eficiența unui sistem de recunoaștere și verificare de vorbitor. Pentru a calcula acest cost suplimentar am folosit un model de verificare de vorbitor anterior salvat și prin care am comparat fiecare spectrogramă prezisă cu câteva mostre naturale de la același sau alți vorbitori. Aceste mostre au fost selectate aleator, adică un număr predefinit (aproximativ numărul de vorbitori) de mostre de la același vorbitor și câte o mostră de la ceilalți vorbitori.

4.1.2 Evaluarea metodei

Sistemele au fost evaluate obiectiv calculând valoarea EER cu un model pre-antrenat pe un număr mare de vorbitori cu sistemul de verificare de vorbitor. Sistemele de sinteză antrenate folosind același volum de date audio s-au comparat cu un test set care conține 144 de pronunții sintetizate cu aceste sisteme de sinteză. Fiecare semnal sintetizat este comparat cu o mostră naturală de la un același vorbitor și de la un alt vorbitor ales aleator. Avem în vedere diseminarea acestor rezultate după ce vom realiza și evaluarea subiectivă. Tabelul 2 prezintă rezultatele obținute până în prezent.

Valorile numerice pot fi comparate pe coloană, sistemele fiind antrenate pentru același număr de epoci (1.000) și același volum de date. Folosind toate datele sau doar 100 de pronunții de la fiecare vorbitor sistemul *B+E* obține cele mai bune rezultate, iar pentru coloana *RND1* rezultatul cel mai bun fiind atins de sistemul *B+CS*.

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

Tabel 2. Valoarea ratei de eroare egală (EER) pentru sistemele Baseline, CosSim și EER, antrenate cu diferite cantități de date

Sistem	**ALL** (EER)	**RND1** (EER)	**RND1-100** (EER)	**RND1-SAM** (EER)
B	*6.94*	*4.86*	*8.33*	*2.43*
B+CS	*6.25*	***4.66***	*6.25*	*2.43*
B+E	***4.66***	*8*	***6***	*2.43*

În continuare, rezultatele obținute prin aplicarea metodei de codare a vorbitorilor prin embedding pot fi vizualizate cu ajutorul metodei de t-Distributed Stochastic Neighbour Embedding (t-SNE) (Maaten & Hinton, 2008) care este o modalitate de a ilustra valorile de embedding multidimensionale. Acestea sunt calculate pentru fiecare vorbitor cu ajutorul rețelei de verificare de vorbitor.

Figura 3 arată aceste valori calculate din spectrograma prezisă de sistemul *B,* care este cel de bază fără funcții de cost suplimentare. Figura 4 prezintă aceste valori pentru sistemul *B+E,* care are ca și funcție de cost suplimentar valoarea de EER. Am ales același set de vorbitori pentru cele 2 figuri, valorile de embedding generate din mostrele sintetizate cu sistemele antrenate sunt marcate cu puncte, iar cele extrase din mostrele naturale cu cruciulițe.

Se poate observa că valorile de embedding pentru același vorbitor apar în grupuri, atât dacă sunt generate după spectograme sintetice, cât și pentru voce naturală, dar acestea nu sunt neapărat apropiate. În sistemul *B+E* aceste grupuri se apropie pentru un număr de vorbitori, dar nu pentru toți. Aceste rezultate necesită investigații suplimentare.

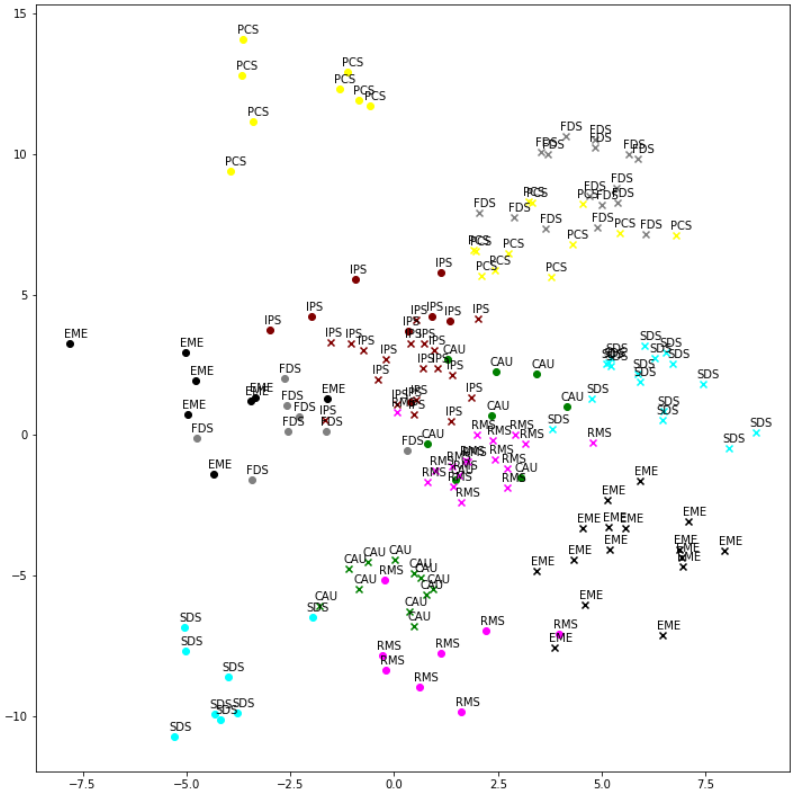

Fig. 3. Vizualizare de embedding de vorbitori cu t-SNE pentru sistemul B (Baseline)

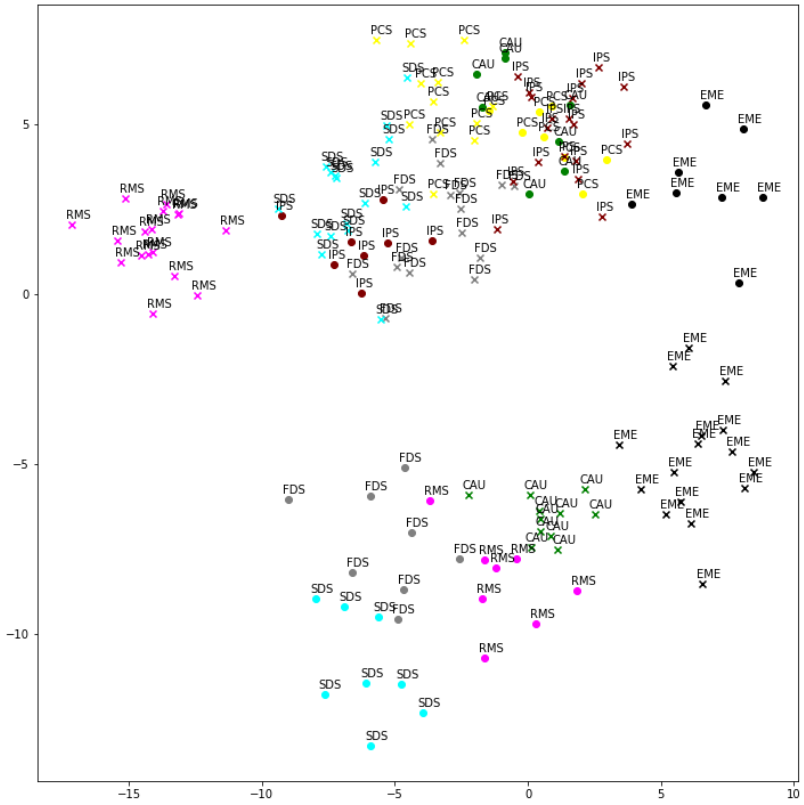

Fig. 4. Vizualizare de embedding de vorbitori cu t-SNE pentru sistemul B+E (Baseline + EER)

4.2 Rezultate experimentale privind adaptarea vorbirii sintetizate de la date audio inexpresive la date audio cu expresivitate

4.2.1 Descrierea metodei

O metodă experimentată pentru adaptarea de stil a fost antrenarea sistemelor de sinteză cu date neexpresive și continuarea învățării cu date expresive. Sistemul evaluat este bazat pe sistemul DC-TTS, prezentat în secțiunea 4.1.1. Antrenarea sistemelor a folosit implementarea[5] publicată de grupul de cercetare de la Universitatea din Edinburgh (CSTR) care este bazată pe Tensorflow 1.13.

4.2.2 Evaluarea metodei

Date audio. Am efectuat două tipuri de experimente. Pentru primul tip am folosit date de la un singur vorbitor, iar pentru al doilea tip, datele de la un singur vorbitor au fost extinse cu date de la alți 3 vorbitori.

Date pentru Experiment tip A. Datele folosite pentru antrenarea primului tip de experimente aparțin vorbitorului SAM din corpusul SWARA și au fost împărțite în două categorii, în funcție de evaluarea subiectivă a gradului lor de expresivitate. Cele două categorii selectate sunt bazate pe selectare manuală. Astfel, avem următoarele seturi:

- date neexpresive: rnd1, rnd2, rnd3, diph1, diph2 (2.476 de pronunții / propoziții)
- date expresive: Ivan Turbinca, stan (704 de propoziții)

Date pentru Experiment tip B. Pentru al doilea tip de experimente am adăugat date neexpresive de la vorbitorii BEA, EME și IPS (vorbitori selectați din corpusul SWARA). De la acești vorbitori am utilizat câte 40 sau 100 de pronunții selectate din rnd1, fiind considerate date neexpresive.

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

Date text. Textele corespunzător datele audio au fost folosite în trei forme diferite:

Forma ortografică	*Pe de altă parte, conform rezultatelor obținute*
Forma transcrisă fonetic	*p e <> d e <> a l t @ <> p a r t e <,> k o n f o r m <> r e z u l t a t e l o r <> o p ts i n u t e*
Forma transcrisă fonetic cu accent	*p e0 <> d e0 <> a1 l t @0 <> p a1 r t e0 <,> k o0 n f o1 r m <> r e0 z u0 l t a1 t e0 l o0 r <> o0 p ts i0 n u1 t e0*

Semnele <> marchează limitele cuvintelor, iar 1 și 0 indică dacă silabele sunt accentuate sau neaccentuate.

Experimente. Experimentele din prima categorie au fost antrenate pe date neexpresive de la vorbitorul SAM, iar pentru fiecare formă de intrare de text câte un model separat. După 3.000 de epoci antrenarea a fost continuată cu date expresive de la vorbitorul SAM, textul de intrare fiind același cu cel folosit în modelul antrenat pe date neexpresive.

Mostre audio de voce sintetizată pentru aceste sisteme pot fi accesate și ascultate aici: https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home . Sistemele antrenate sunt sumarizate în Tabelul 3.

Tabel 3. Descrierea experimentelor din prima categorie

Date text	Date audio	Epoci	Număr de propoziții	Date audio suplimentare	Epoci suplimentare	Număr de propoziții
Forma ortografică	*neexpresiv*	*3.000*	*2.476*	*Expresiv*	*3.000*	*704*
Forma transcrisă fonetic	*neexpresiv*	*3.000*	*2.476*	*Expresiv*	*7.000*	*704*
Forma transcrisă fonetic cu accent	*neexpresiv*	*3.000*	*2.476*	*Expresiv*	*6.000*	*704*

Experimentele cu sistemele de sinteză din a doua categorie au fost antrenate cu date expresive de la vorbitorul SAM și antrenate în continuare cu date puține (40 sau 100 de propoziții) de la alți trei vorbitori (BEA, EME, IPS) selectați manual din corpusul SWARA.

Textul de intrare a fost furnizat sistemului în forma transcrisă fonetic. Pentru a învăța mai mulți vorbitori, ID-ul de vorbitor a fost concatenat textului de intrare și sistemul a fost antrenat cu vorbitori multipli. Identitatea a mai multor vorbitori în sistem a fost învățată cu ajutorul unei metode numite învățare a contribuției la canalele de informație (en. *learning channel contributions)*. Această metodă adaugă informația de vorbitor în mai multe straturi din fiecare componentă a rețelei.

Mostre audio de voce sintetizată de către aceste sisteme pot fi accesate și ascultate aici: https://rb.gy/6pbnny. Sistemele antrenate sunt listate în Tabelul 4.

Tabel 4. Descrierea experimentelor din a doua categorie

Date text	Date audio	Epoci	Număr de propoziții	Vorbitor	Date audio suplimentare	Epoci suplimentare	Număr de propoziții	Vorbitor
Forma transcrisă fonetic	*Expresiv*	*2.000*	*704*	*SAM*	*neexpresiv*	*1.500*	*40*	*BEA*
Forma transcrisă fonetic	*Expresiv*	*2.000*	*704*	*SAM*	*neexpresiv*	*1.500*	*100*	*BEA*
Forma transcrisă fonetic	*Expresiv*	*2.000*	*704*	*SAM*	*neexpresiv*	*1.500*	*40*	*EME*
Forma transcrisă fonetic	*Expresiv*	*2.000*	*704*	*SAM*	*neexpresiv*	*1.500*	*100*	*EME*
Forma transcrisă fonetic	*Expresiv*	*2000*	*704*	*SAM*	*neexpresiv*	*1.500*	*40*	*IPS*
Forma transcrisă fonetic	*Expresiv*	*2000*	*704*	*SAM*	*neexpresiv*	*1.500*	*100*	*IPS*

Aceste categorii de experimente au avut rezultate modeste, identitatea vorbitorului fiind păstrată sau învățată, dar caracterul de expresivitate al vorbirii a fost pierdut în acest process de învățare.

5. O nouă metodă de adaptare la vorbitor folosind post-filtrarea

Extinzând diferitele experimente realizate în laborator pentru sistemele de sinteză text - vorbire bazate pe rețele neuronale, am ajuns la concluzia că există o serie de limitări în adaptarea la un nou vorbitor pe acest tip de structuri. În special, atunci când se dorește adaptarea rețelei dintr-un set redus de date problemele sunt și mai complexe, iar calitatea semnalului generat nu este satisfăcătoare.

Astfel, noi am propus o nouă metodă de adaptare a sistemului de sinteză la un nou vorbitor folosind post-filtrarea semnalului vocal sintetizat cu o rețea pre-antrenată cu date puține, tot cu o structură de rețea neuronală care este antrenată să mapeze semnalul vocal sintetizat în semnal vocal natural. Ca atare, metoda bazată pe post-filtrare presupune crearea unui sistem de sinteză minimal antrenat cu cantități reduse de date. Ieșirea acestui sistem va fi ulterior filtrată de o rețea neuronală antrenată pentru conversia vocii sintetizate în voce naturală.

Această soluției este pretabilă pentru adaptarea la un nou vorbitor, deoarece putem porni de la o rețea pre-antrenată cu date audio de la mai mulți vorbitori, iar această rețea este adaptată ulterior către un vorbitor țintă.

Sistemele de sinteză text vorbire care includ post-filtrarea folosesc rețele neuronale, compuse din straturi recurente. Pentru o mai bună analiză a rezultatelor post-filtrării și adaptării de vorbitor, s-au folosit cantități de date diferite, atât pentru antrenarea sistemului de sinteză, cât și pentru pasul de post-filtrare sau adaptare.

Sistemele de sinteză implementate prin această metodă în laborator au fost validate cu atât prin metode obiective (calculul metricii de similaritate MSD - Mel Spectral Distance) și

subiective (evaluatori umani, cu aprecierea calității semnalului sintetizat pe scala MOS - Mean Opinion Score). Rezultatele măsurărilor obiective și testele de ascultare atestă că metoda de post-filtrare și adaptare de voce pornind de la o rețea pre-antrenată cu mai mulți vorbitori și folosind puține mostre (chiar și 10 minute de vorbire) poate fi folosită la crearea vocilor sintetice cu o calitate relativ bună.

Experimentele și rezultatele au fost prezentate la conferința Intelligent Systems 2020. O înregistrare a prezentării este disponibilă pe adresa: https://youtu.be/OLAJGaQmjqA.

6. Concluzii

Acest raport a prezentat metodele și rezultatele experimentale obținute în urma implementării tehnologiei de sinteză text-vorbire bazată pe adaptarea vorbirii sintetice la stilul și expresivitatea unui nou vorbitor. Această adaptare a fost realizată prin intermediul a trei noi metode: adaptarea de la un sistem cu vorbitori multipli la un sistem cu un singur vorbitor folosind un set redus de date, adaptarea de la un sistem cu voce ne-expresivă la un sistem cu voce expresivă, adaptarea sistemului de sinteză la un nou vorbitor printr-o procedură de adaptare bazată pe post-filtrare. Pentru a asigura implementarea acestor tehnologii de adaptare, a fost necesară extinderea corpusului deja existent SWARA cu noi voci, corpusul fiind numit SWARA 2.0. Pe lângă datele cuprinse în SWARA s-a solicitat și înregistrarea unei povești în stil de vorbire cu expresivitate (nuvela Ivan Turbincă). A rezultat astfel un corpus de 29 de vorbitori (14 masculini și 15 feminini) conținând aproximativ 65 de ore de date audio. Crearea corpusului SWARA 2.0 s-a realizat cu aplicația software RecoApy care permite înregistrarea secvențială a textelor și nu necesită segmentare ulterioară. Această aplicație suportă mai multe limbi, inclusiv pentru transunui sistem la crierea fonetică.

Pentru a beneficia de datele disponibile pentru mai mulți vorbitori, dar și pentru a adapta sistemele de sinteză text - vorbire la stilul și expresivitatea unui nou vorbitor, am proiectat o serie de metode de modelare acustică și de sinteză de semnal bazate de diferite structuri de rețele neuronale profunde (CNN, Transformer, DC-TTS, Tacotron, Tacotron2, Flowtron, etc), le-am implementat pe 2 servere de procesare paralelă a datelor, fiecare dotate cu plăci GPU de ultimă generație, și am efectuat o serie de experimente pentru vorbitori multipli.

Prima metodă analizată a fost antrenarea unui sistem cu date de la mai mulți vorbitori și adăugarea unei funcții de cost suplimentare pentru îmbunătățirea și accelerarea învățării identității de vorbitor. Pentru a realiza acest sistem de vorbitori multipli am implicat și un sistem de verificare de vorbitor pentru a calcula costul suplimentar.

Având la dispoziție date expresive și mai puțin expresive am exploatat același sistem de sinteză cu vorbitori multipli folosind datele expresive, antrenat în continuare cu date mai puțin expresive de la un alt vorbitor, astfel încercând adaptarea vorbitorilor la un nou stil. Experimentele au rezultat într-o calitate bună a vocii sintetizate și similaritate de vorbitor acceptabile, dar transferul stilului de la un vorbitor la altul este mai puțin perceptibil.

Adaptarea la un nou vorbitor a fost implementată și cu un sistem statistic-parametric în care s-a folosit metoda de post-filtrare și ajustarea ponderilor rețelei de sinteză am evaluat vocile create cu diferite cantități de date.

7. Bibliografie

(Gehrig et al., 2019) Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin, ”Convolutional Sequence to Sequence Learning”, arXiv:1705.03122

(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.

(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.

(Vaswani et al., 2017) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

ANEXA 1. Consimțământ informat privind prelucrarea datelor audio

Grupul de prelucrare a semnalului vocal
Universitatea Tehnică din Cluj-Napoca
str. G. Barițiu nr. 26-28, sala S2.3,
Cluj-Napoca

Consimțământ informat privind prelucrarea datelor audio

Subsemnatul _____, în perioada aprilie-iunie 2020 am participat în mod voluntar, neremunerat, la o serie de sesiuni de înregistrări ale vocii mele în vederea dezvoltării sistemelor de procesare a vorbirii în limba română. Înregistrările au fost coordonate de către șl.dr.ing. Adriana Stan din cadrul Grupului de Prelucrare a Semnalului Vocal a Universității Tehnice din Cluj-Napoca.

Datele audio vor fi în mod implicit stocate, accesate și prelucrate de către coordonatorul înregistrărilor, iar prin bifarea opțiunilor de mai jos îmi exprim în mod explicit consimțământul ca datele audio înregistrate să fie stocate, accesate și prelucrate suplimentar pentru scopul specificat:

*Bifați toate căsuțele cu care sunteți de acord

Sunt de acord ca datele audio înregistrate să fie stocate și prelucrate de către:	
	membrii grupului de cercetare a semnalului vocal din cadrul UTC-N
	colaboratori direcți ai grupului de cercetare interni UTC-N
	colaboratori direcți ai grupului de cercetare externi UTC-N (ex. parteneri în proiecte de cercetare/dezvoltare)
	colaboratori indirecți externi UTC-N (de ex. persoane externe ce solicită accesul la datele audio pe baza unui acord semnat)
Sunt de acord ca datele audio înregistrate să fie stocate și prelucrate cu scop	
	academic și de cercetare
	comercial
Sunt de acord ca datele audio înregistrate să fie prelucrate în vederea dezvoltării sistemelor de sinteză text-vorbire:	
	în cadrul cărora identitatea mea vocală poate fi direct recunoscută (sisteme antrenate cu o singură voce)
	în cadrul cărora identitatea mea vocală nu poate fi direct recunoscută (sisteme antrenate cu minim două voci distincte)
Sunt de acord ca datele audio înregistrate să fie prelucrate în vederea dezvoltării de:	
	sisteme automate de recunoaștere a vorbirii
	sisteme automate de recunoaștere a vorbitorului
	sisteme automate de anonimizare a vorbitorului
	sisteme automate de clonare a vorbirii
	alte sisteme de procesare de voce ce nu sunt în momentul de față cunoscute, dar care pot să fie generate odată cu avansul cercetării, atât timp cât acest lucru respectă legislația și etica de cercetare
Sunt de acord ca sistemele derivate din înregistrările audio să fie utilizate în scop:	
	academic și de cercetare

	comercial

Am fost informat(ă) asupra faptul că datele înregistrate vor fi păstrate pe o perioadă nedeterminată, iar informații suplimentare privind identitatea mea (precum nume, vârsta, adresa de e-mail sau orice alte date colectate suplimentar în procesul de realizare a înregistrărilor) nu vor fi distribuite către nicio altă persoană internă sau externă grupului de cercetare.

Înțeleg faptul că secțiuni de dimensiuni reduse (maxim 5 propoziții) pot fi utilizate în prezentări/pagini online sau în cadrul unor prezentări orale ale rezultatelor cercetării, precum conferințe, seminarii sau grupuri de lucru.

Am fost informat cu privire la faptul că Universitatea Tehnică din Cluj-Napoca prin intermediul unei persoane desemnate ia toate măsurile necesare pentru a proteja și controla distribuirea în mod involuntar sau malițios către terțe părți a datelor audio înregistrate.

Am fost informat cu privire la dreptul de a revoca accesul sau de a solicita ștergerea datelor personale deținute de grupul de cercetare din Universitatea Tehnică din Cluj-Napoca și de către terțe părți în conformitate cu clauzele Regulamentului european privind prelucrarea datelor cu caracter personal (https://www.dataprotection.ro/).

**Vă rugăm să păstrați o copie a prezentului consimțământ informat pentru a-l putea consulta pe viitor.*

Data

Nume și prenume,

E-mail:

Semnătura,