D2.16. Implementarea unui modul de adaptare la un nou vorbitor a 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:	**D2.16. Implementarea unui modul de adaptare la un nou vorbitor a sistemului de sinteză**
Termen:	**Decembrie 2019**
Editor:	**Beáta Lőrincz (Universitatea Tehnică din Cluj-Napoca)**
Adresa de eMail editor:	**Beata.Lorincz@com.utcluj.ro**
Autori, în ordine alfabetică:	**Beáta Lőrincz, Maria Nuțu, Adriana Stan**
Ofițer de proiect:	**Cristian STROE**

Rezumat:

Acest raport prezintă un set de experimente menite să evalueze modul în care cantitatea de date necesară antrenării unui sistem de sinteză text-vorbire de calitate poate fi redusă. Sunt prezentate două metode principale: una bazată pe sisteme de sinteză statistic-parametrice și una bazată pe sisteme de sinteză neurale complete. În cadrul primei metode a fost analizată utilizarea unei rețele de tip post-filtru. Aceasta realizează o mapare între vocea sintetizată de un sistem antrenat pe date puține la vocea naturală. Tot în cadrul acestei categorii de sisteme de sinteză s-a analizat realizarea unei adaptări către un vorbitor țintă folosind o rețea pre-antrenată pe date de la mai mulți vorbitori. Pentru sistemele de sinteză complete, s-a utilizat o arhitectură de tip Tacotron GST în cadrul căreia s-au folosit reprezentări vectoriale ale identității vorbitorilor țintă combinat cu utilizarea unor ponderi de inițializare a rețelei învățate din date ale unui alt vorbitor. Totodată s-a avut în vedere și menținerea expresivității sistemului de sinteză prin fixarea modulului GST din arhitectura sistemului.

Cuprins

1. Introducere

Acest livrabil (D4.2.2. Modul software pentru adaptarea sistemului de sinteză la un nou vorbitor) prezintă rezultatele obținute în activitatea A.4.2.2 din cadrul sub-proiectului P4 SINTERO - Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate.

Pe lângă naturalețea și expresivitatea vocii în sistemele de sinteză, eforturi semnificative au fost depuse și pentru adaptarea sistemelor la un nou vorbitor, folosind date audio cât mai puține. Adaptarea la un nou vorbitor presupune capacitatea sistemului de sinteză de a se adapta rapid (pe baza a cât mai puține mostre audio) pornind, însă, de la un model existent. Această problemă, de adaptare la un nou vorbitor (en. *speaker adaptation*) este întâlnită și în sistemele de recunoaștere a vorbirii, în cadrul cărora acuratețea transcrierilor este îmbunătățită prin ajustarea parametrilor modelului către vorbitorul curent. În cadrul acestui livrabil, ne vom adresa cu precădere adaptării sistemelor de sinteză.

În ultimii ani, în literatura de specialitate cele mai bune performanțe ale sistemelor de sinteză sunt obținute cu rețele neuronale. Acestea înlocuiesc sistemele bazate pe metode parametric-statistice, cum ar fi cele bazate pe modele Markov. În cadrul modelelor ce folosesc rețele neuronale, cea mai des întâlnită abordare referitoare la adaptarea la un nou vorbitor se bazează pe utilizarea unei rețele pre-antrenate. Această rețea folosește date de la vorbitori multipli, iar ponderile sale vor fi ulterior ajustate pe baza datelor vorbitorului țintă. De exemplu, (Wu et al., 2016) prezintă trei metode de adaptare: prima este dată de transformarea spațiului de trăsături; a doua metodă este reprezentată de augmentarea caracteristicilor vorbitorului în cadrul datelor de intrare ale rețelei; iar cea de a treia modifică parametri rețelei. (Öztürk et al., 2019) prezintă o comparație a metodelor de post-filtrare folosind rețele complet conectate (en. *feed forward*), recurente și convoluționale cu antrenare adversarială pentru sisteme parametric-statistice ce pot fi adaptate la un nou vorbitor folosind date audio reduse ca dimensiune (aprox. 5-15 secunde). În articolul (Luong et Yamagishi, 2018), autorii propun o arhitectură de rețea antrenată pe date de la mai mulți vorbitori și care poate fi rapid adaptată la un vorbitor țintă folosind date netranscrise ortografic. Metoda aceasta mai este denumită și adaptare nesupervizată. (Arik et al., 2018) prezintă metode de clonare a vocii cu adaptare către vorbitor și codare a identității vorbitorului. Codarea este realizată cu o rețea antrenată separat și care învață să discrimineze între diverșii vorbitori prin crearea unei reprezentări vectoriale a acestora. Alături de adaptarea la un vorbitor țintă strict pe baza identității vocale a acestuia, există o serie de experimente ce se referă la transferul stilului de vorbire sau a expresivității unui vorbitor (Parker et al., 2018), însă acestea nu fac obiectul acestui raport.

Pornind de la analiza stării actuale a cercetărilor din domeniu, studiile noastre s-au axat fie pe utilizarea unui proces de post-filtrare, fie pe baza utilizării reprezentărilor vectoriale a identității vorbitorilor în cadrul unei rețele antrenate cu date mixte. Secțiunile următoare descriu aceste două metode, experimentele efectuate și rezultatele lor.

2. Adaptarea la un nou vorbitor folosind sisteme de sinteză statistic-parametrice și metoda de post-filtrare

2.1. Descrierea metodei

În ultimii ani, sistemele de sinteză text-vorbire obțin o calitate a vocii apropiată de cea umană, însă aceste sisteme sunt de cele mai multe ori antrenate pe un număr foarte mare de date. Dacă aceste date nu sunt disponibile, este necesară utilizarea unor metode ce fac posibilă antrenarea sistemelor pornind de la un set redus de date. Cu mențiunea că s-ar putea să apară anumite pierderi de calitate.

Luând în considerare această necesitate de a reduce necesarul de date audio, în cadrul acestui experiment s-a optat pentru utilizarea unui sistem de sinteză statistic-parametric. Acesta implică realizarea unei conversii a textului în caracteristici lingvistice la nivel de fonem și maparea acestor caracteristici într-un vector acustic de durată mică (în general 5-10ms). Caracteristicile lingvistice includ transcrierea fonetică, stabilirea accentului lexical, silabificare, identificarea părților de vorbire, precum și caracteristici contextuale cum ar fi numărul cuvintelor, sau al silabelor din textul de intrare, etc. Pentru reprezentarea acustică, se utilizează de cele

mai multe ori o parametrizare bazată pe coeficienți cepstrali, frecvență fundamentală și coeficienți de aperiodicitate. Această reprezentare este obținută cu ajutorul vocoderului WORLD (Morise, Yokomori, & Ozawa, 2016).

În continuare vom prezenta două metode pentru obținerea vocilor sintetice cu date reduse, utilizate în experimentele raportate în cadrul acestui livrabil: metoda de post-filtrare și metoda de adaptare de voce pornind de la o rețea pre-antrenată.

Metoda bazată pe principiul de post-filtrare presupune crearea unui sistem de sinteză minimal cu cantități reduse de date. Ieșirea acestui sistem va fi ulterior filtrată de o rețea neuronală antrenată pentru conversia vocii sintetizate în voce naturală. O schiță a acestei metode este ilustrată pe Figura 1.

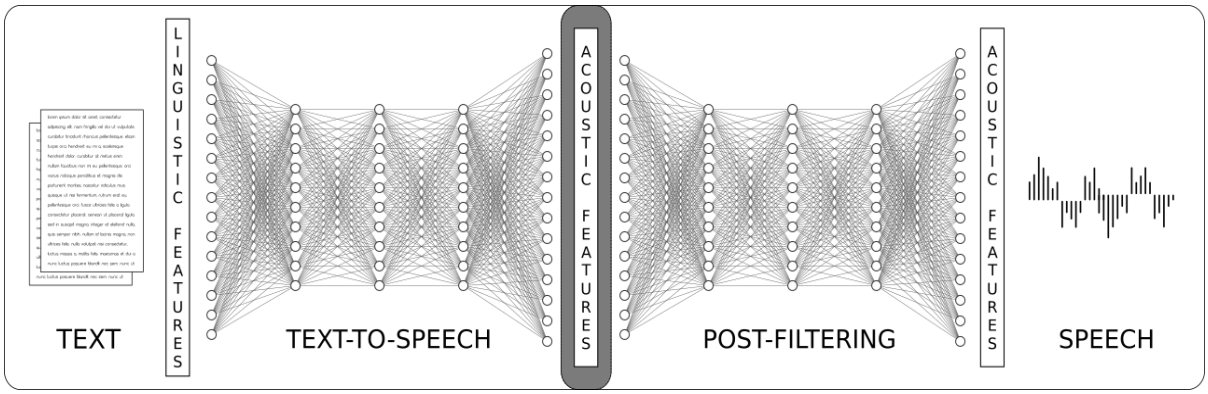

Figura 1. Structura rețelei de sinteză de voce și post-filtrare

Pentru sistemul de post-filtrare, caracteristicile acustice ale semnalului audio sintetizat au fost împerecheate cu caracteristicile acustice ale semnalului natural. Deoarece s-ar putea să apară diferențe temporale între datele sintetizate și cele naturale, s-a utilizat o metodă de aliniere temporală de tipul Dynamic Time Warping (DTW) la nivel de coeficienți cepstrali. Aceasta permite un mai bun control al perechilor de caracteristici acustice prezentate rețelei în pasul de antrenare.

A doua metodă folosită, adaptarea la un nou vorbitor pornind de la o rețea pre-antrenată se referă la antrenarea unei rețele neuronale folosind date audio de dimensiuni mai extinse, însă obținute de la mai mulți vorbitori. Această rețea este ulterior adaptată către un vorbitor țintă. Metoda aceasta permite rețelei să învețe statistici referitoare la caracteristicile acustice ale vorbirii propriu-zise, urmând ca pentru identitatea vorbitorului, ponderile rețelei să fie ajustate pe baza unui set de date mai redus.

2.2. Evaluarea metodei

Date audio

Pentru antrenarea sistemelor de sinteză am folosit corpusul SWARA (Stan et al., 2017) ce conține date de la 17 vorbitori (aproximativ 21 ore de vorbire). Dintre acești vorbitori am selectat 8 voci feminine (*BAS, CAU, EME, DCS, DDM, HTM, PMM* și *SAM*) și am ales subsetul de *RND1* de la fiecare vorbitor. Corpusul nu conține și seturi de test, iar din această cauză am înregistrat încă două voci feminine (*BEA* și *MAR*) în condiții similare de studio. Niciunul dintre vorbitorii selectați nu este orator profesionist.

Datele audio sunt eșantionate la 48kHz cu 16bps și au fost segmentate manual la nivel de propoziție. Alinierea semi-automată a datelor audio la nivel de fonem a fost obținută cu ajutorul unui algoritm iterativ bazat pe modele Markov. Caracteristicile lingvistice ale textului de intrare au fost extrase cu ajutorul sistemului procesare de text din RomanianTTS.[1]

[1] http://romaniantts.com

Experimente

Experimentele au fost rulate cu ajutorul toolkit-ului Merlin (Wu, Watts, & King, 2016) folosind implementarea oficială[2] accesibilă online.

A) Post-filtrarea datelor sintetizate

Pentru sistemele de sinteză ce includ și pasul de post-filtrare, rețeaua neuronală este alcătuită din straturi complet conectate combinate cu straturi recurente de tip Long-Short Term Memory cu număr variabil de noduri pe strat (256, 512 sau 1024) și cu funcție de activare de tip tangentă hiperbolică (*tanh)* sau unitate liniară rectificată (*ReLU)*. Pentru o mai bună analiză a rezultatelor post-filtrării, s-au folosit cantități de date diferite, atât pentru antrenarea sistemului de sinteză, cât și pentru pasul de post-filtrare. Lista sistemelor analizate este prezentată în Tabelul 1.

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

Tabel 1. Lista sistemelor de sinteză antrenate pentru metoda de post-filtrare

Pe lângă sistemele antrenate cu cantități variate de date de antrenare, s-a analizat și un sisteme a cărui rețea de post-filtrare a fost antrenată cu date de la mai mulți vorbitori (M100_PF_MSPK). În cadrul sistemului M100Db_PF100Db, datele existente pentru vorbitor au fost dublate, pentru a simula dublarea artificială a cantității de date disponibile.

B) Sistem de adaptare la un nou vorbitor

Acest sistem a fost antrenat, de asemenea, cu cantități diferite de date: 100 (aproximativ 10 minute), respectiv 500 de propoziții (aproximativ 50 de minute) de la fiecare dintre cei 8 vorbitori selectați din corpusul SWARA și cu adăugarea celor două voci înregistrate adițional. Pentru adaptarea sistemului s-au folosit vocile *BEA* și *MAR*, din care am selectat 100, respectiv 500 de propoziții pentru experimente. Experimente desfășurate pentru adaptare de voce sunt rezumate în Tabelul 2. Identificatorii sistemului specifică numărul de propoziții utilizate de la fiecare vorbitor în pasul de antrenare, precum și numărul de propoziții utilizate pentru adaptare. Astfel că, de exemplu, pentru sistemul SPKA100_E100 s-au folosit câte 100 de propoziții de la fiecare vorbitor, adaptarea realizându-se cu 100 de propoziții de la vorbitorul țintă.

Exemple audio pentru aceste sisteme pot fi accesate și ascultate aici: https://speech.utcluj.ro/pf_lrec2020/.

Acronim de sistem	Numărul de propoziții folosite de la fiecare vorbitor	Numărul de propoziții folosit de la vorbitorul țintă
SPKA100_E100	10x100	100
SPKA100_E500	10x500	100

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

SPKA500_E500	10x500	500

Tabel 2. Lista sistemelor antrenate pentru adaptare de voce

Rezultate

Sistemele descrise anterior au fost validate cu metode obiective și subiective. Pentru metoda obiectivă am folosit măsura de distorsiune cepstrală (en. *Mel Cepstral Distortion (MCD)*) (Kubichek, 1993). 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. Aceste propoziții nu au fost incluse în datele de antrenare. Figura 2 rezumă rezultatele pentru vorbitorul MAR. 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.

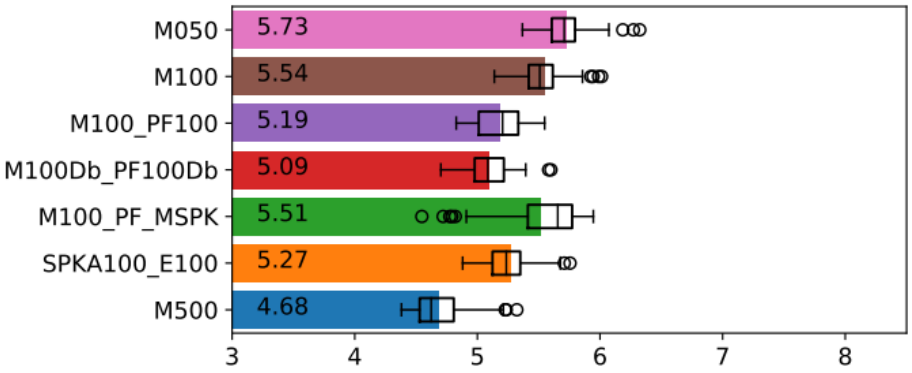

Figura 2. Valorile MCD pentru vorbitorul MAR

Deși sunt multe studii care analizează metode de măsurare obiectivă a calității vocii sintetizate, încă nu există măsuri care se apropie de evaluarea perceptuală a vorbirii sintetizate. Din această cauză, sunt necesare teste de ascultare subiective. Pentru evaluarea sistemelor am ales mostre pentru ambele voci. Testul de ascultare a fost compus din 4 secțiuni:

- naturalețe - evaluată pe scală de Mean Opinion Score (MOS) cu 5 puncte (1-nenatural, 5-natural)
- asemănarea cu vorbitorul țintă - evaluată pe scale de MOS cu 5 puncte (1-deloc similar, 5-foarte similar)
- inteligibilitate - evaluată pe baza ratei de eroare de transcriere a cuvintelor (en. Word Error Rate)
- naturalețe ABX - evaluată prin selectarea unei mostre din două prezentate ascultătorului.

Testul de ascultare a fost completat de 20 de ascultători și arată că metoda de post-filtrare și adaptare la vorbitor rezultă într-o voce mai naturală și inteligibilă. Figura 3 ilustrează aceste rezultate ale testului de ascultare din secțiunea naturalețe pentru sistemele de post-filtrare și adaptare de voce. 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.

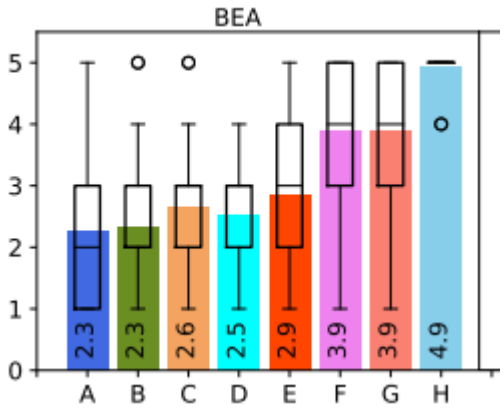

Figura 3. Rezultatele testului de ascultare pentru vorbitorul BEA în secțiunea naturalețe

Rezultatele măsurii obiective și testele de ascultare atestă că metoda de 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.

3. Adaptarea la un nou vorbitor folosind sisteme de sinteză neuronale complete

3.1. Descrierea metodei

Pentru antrenarea unui model cu vorbitori multiplii, ținând cont de scopul de a antrena un sistem de sinteză care conține mai mulți vorbitori și care poate fi adaptat la vorbitori noi, dar și de abilitatea de a învăța expresivitatea, am ales arhitectura propusă de (Wang et al., 2018) numită Tacotron GST. Tacotron și Tacotron2 sunt sisteme de sinteză des folosite pentru că obțin o calitate bună de vorbire sintetizată care este aproape de vorbirea naturală. Tacotron GST (GST fiind abrevierea pentru Global Style Token), implementat folosind sistemul Tacotron, este o extensie a acestuia cu un modul de control al prozodie care poate fi folosit și la învățarea de identități ale vorbitorilor.

Principalele componente ale sistemului sunt enumerate mai jos, iar o schemă a acestuia este prezentată în Figura 4:

1. Sistemul de sinteză Tacotron (Wang et al., 2017)
2. Modulul Global Style Tokens (învață reprezentarea latentă a prozodiei sau identitatea vorbitorilor) este compus din:
   2.1. Codor pentru referința audio (en. Reference encoder) - format din rețele neuronale recurente și convoluționale,
   2.2. Submodul Global 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 inferenței, sistemul poate primi ca date de intrare o referință audio pentru a sintetiza textul pornind de la codarea acestei referințe. Pe de altă parte, se poate renunța la referința audio, modelarea prozodiei sau reprezentarea vectorială a vorbitorilor realizându-se prin setarea manuală a tokenilor învățați în timpul antrenării. Acest model, Tacotron GST, a fost extins suplimentar cu un strat de reprezentări vectoriale ale vorbitori din setul de date de antrenare.

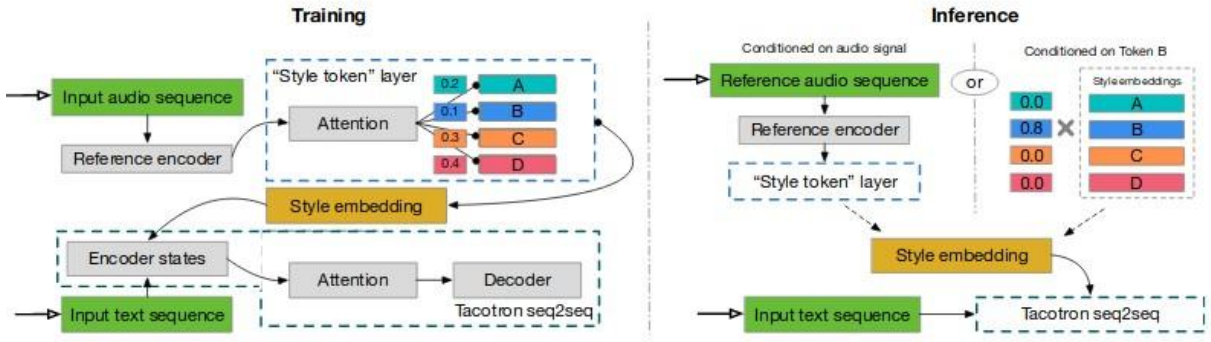

*Figura 4. Arhitectura Tacotron GST (*Wang et al., 2018*)*

3.2. Evaluarea metodei

Date audio

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

Corpusul MARA conține înregistrarea întregului roman „Mara" de Ioan Slavici, (aproximativ 11 ore de vorbire) realizată în condiții de studio de o actriță profesionistă. Datele au fost segmentate manual la nivel de propoziție sau fraze de dimensiuni reduse și au fost verificate inconsistențele dintre fișierele audio și textul romanului. Datele sunt aliniate semi-automat la nivel de fonem, folosind un aliniator bazat pe Modele Markov antrenat pe întreg setul de date.

Corpusul SWARA este compus din date de la 17 vorbitori (aproximativ 21 ore de vorbire). Pentru experimentele desfășurate am folosit subsetul *RND1* al corpusului (500 de propoziții de la fiecare vorbitor - aproximativ 50 de minute de vorbire). Vorbitorii au fost selectați în două configurații:

- 10 vorbitori feminini (*BAS, CAU, DCS, DDM, EME, HTM, PMM, SAM, TIM* și *MAR* adăugat ulterior)
- 5 voci feminine (*CAU, DCS, DDM, EME, SAM*) - 5 voci masculine (*FDS, IPS, PCS, PSS, TSS*).

Corpusul MARA conține vorbire cu expresivitate, iar corpusul SWARA conține o intonație mai degrabă plată.

Experimente

Experimentele au fost rulate pornind de la implementarea Mozilla[3] a arhitecturilor Tacotron și Tacotron GST. Alte implementări disponibile pe GitHub[4] au fost încercate, dar acestea nu au avut timp de convergență similari și s-a renunțat la folosirea lor.

Antrenarea sistemelor a fost realizată pe un GPU GeForce GTX 970 (driver version: 430.26, CUDA: 10.2) cu memorie de 4 GB. Acest hardware ne-a permis folosirea unui batch size de dimensiune 16. Lista experimentelor realizate folosind arhitectura și datele descrise anterior este prezentată în continuare, cu evidențierea rezultatelor atât din punct de vedere al prozodiei sistemului, cât și a identității vocii rezultate.

[3] https://github.com/mozilla/TTS
[4] https://github.com/

M1. SWARA model

Scopul experimentului:

- învățarea identității vorbitorului prin reprezentări vectoriale,
- învățarea prozodiei prin modulul GST.

Rezultatul experimentului:

- vorbire neinteligibilă, eșuat.

Acest experiment cu Tacotron GST a fost rulat cu 10 voci selectate din corpusul SWARA, conținând 5 voci feminine și 5 masculine. Modelul Tacotron GST a fost extins cu un strat de reprezentări vectoriale (en. *embedding)* ale identității vorbitorilor. Antrenarea modelului a fost oprită pentru că sinteza nu genera vorbire inteligibilă după mai mult de 1000 pe epoci, deși identitatea vorbitorilor putea fi recunoscută în mostrele de audio generate la inferență.

M2. SWARA-peste-MARA model

Scopul experimentului:

- învățarea identității vorbitorului prin reprezentări vectoriale,
- învățarea prozodiei prin stratul GST,
- utilizarea unei rețele cu ponderi pre-antrenate.

Rezultatul experimentului:

- identitatea vorbitorului este controlabilă prin reprezentările vectoriale învățate pentru fiecare vorbitor
- vorbire inteligibilă, neutră (cu puține caracteristici prozodice).

Acest experiment, folosind date de la mai mulți vorbitori din corpusul SWARA, a fost rulat peste o rețea a cărei ponderi au fost învățate din datele corpusului MARA. Antrenarea a fost rulată pentru aproximativ 1000 de epoci cu 10 ore de vorbire de la 5 vorbitori feminini și 5 masculini ce aparțin corpusului SWARA.

Modelul pre-antrenat pe corpusul MARA a produs vorbire inteligibilă, de bună calitate și conține caracteristici expresive. Pornind de la acest model, rețeaua a fost antrenată în continuare cu date din corpusul SWARA. Pe lângă arhitectura originală, s-a adăugat și un strat de reprezentări vectoriale pentru identitatea vorbitorilor. Aceste reprezentări vectoriale au fost învățate într-un timp relativ scurt (aprox. 10 epoci: 5-10 minute de antrenare pe hardware-ul descris mai sus). În acest fel, adaptarea către un nou vorbitor poate fi rezolvată folosind acest model, care în timp scurt poate învăța noi voci. În procesul de inferență, ID-ul vorbitorului poate fi specificat, iar vocea sintetizată va folosi identitatea vorbitorului respectiv. Antrenarea acestui model a folosit, ca în experimentul anterior arhitectura Tacotron GST. Modulul GST a învățat identitatea vorbitorilor, la inferență vorbitorul putând fi controlat și prin setarea manuală a parametrilor acestui modul.

M3. SWARA-peste-MARA model fără reprezentări vectoriale ale vorbitorilor

Scopul experimentului:

- testarea sistemului de sinteză fără reprezentări vectoriale pentru vorbitori

Rezultatul experimentului:

- stratul de GST a învățat identitatea vorbitorilor, sinteza fiind controlabilă prin modificarea ponderilor din acest modul;
- identitatea vorbitorilor a fost învățată într-un mod nesupervizat, identitatea vorbitorilor nefiind dată în procesul de antrenare
- vorbire inteligibilă, neutră (cu puține caracteristici prozodice).

Pornind de la modelul pre-antrenat cu date din corpusul MARA am antrenat Tacotron GST cu 10 voci feminine (500 de propoziții de la fiecare vorbitor), dar fără stratul de reprezentări vectoriale ale identității vorbitorilor. În cazul acesta, modulul GST, care a învățat prozodia în experimentele rulate cu corpusul MARA, a început să învețe identitatea vorbitorilor. În experimentul acesta identitatea vorbitorului este controlabilă prin intermediul modulului GST.

M4. SWARA-peste-MARA model cu modulul GST fixat și fără reprezentări vectoriale ale vorbitorilor

Scopul experimentului:

- testarea sistemului de sinteză fără reprezentări vectoriale pentru vorbitori,
- păstrarea prozodiei din modelul pre-antrenat pe date expresive prin fixarea modulului GST

Rezultatul experimentului:

- identitatea vorbitorilor este învățată parțial, schimbarea valorilor din tokeni afectează în mod minimal vocea sintetică generată
- vorbire inteligibilă, dar neutră (cu puține caracteristici prozodice).

Pentru a obține un model expresiv și cu vorbitori multipli am încercat să fixăm ponderile din modulul GST, care în cazul modelului antrenat pe vorbire expresivă a reușit învățarea a diferite stiluri de vorbire. După antrenarea modelului, rețeaua pierde expresivitatea, stratul fixat fiind ignorat la inferență. Acest experiment va fi continuat cu exploatarea metodelor de învățare continuă în care vom încerca alte metode de a păstra prozodia și transplantarea acesteia la vorbitori noi chiar și fără înregistrări expresive.

M5. SWARA-peste-MARA model cu stratul GST fixat și cu reprezentări vectoriale ale vorbitorilor

Scopul experimentului:

- învățarea identității vorbitorului prin reprezentări vectoriale,
- păstrarea prozodiei din modelul pre-antrenat pe date expresive prin fixarea modulului GST

Rezultatul experimentului:

- vorbire neinteligibilă, eșuat.

Experimentul diferă de cel de dinainte prin adăugarea unui strat de reprezentări vectoriale pentru vorbitori. Această antrenare a fost oprită după 1000 de epoci pentru că nu a produs vorbire inteligibilă în procesul de inferență. Stratul GST fixat și cu reprezentări vectoriale pentru vorbitori din implementarea curentă nu este utilizabil.

Tabelul 3. rezumă sistemele descrise anterior. Exemple audio pentru aceste sisteme sunt disponibile și pot fi ascultate aici: https://speech.utcluj.ro/sintero/dnn-samples/.

Sistem	Pre-antrenare	Număr mostre preantrenare	Antrenare	Număr mostre antrenare	Informația din tokeni/Rezultate
SWARA	-	-	SWARA	10x500	identitatea vorbitorilor / experiment eșuat (vorbire neinteligibilă)

SWARA-peste-MARA cu reprezentări vectoriale pentru vorbitori	MARA	8134	SWARA	10x500	identitatea vorbitorilor
SWARA-peste-MARA fără reprezentări vectoriale pentru vorbitori	MARA	8134	SWARA	10x500	identitatea vorbitorilor
SWARA-peste-MARA cu stratul GST fixat și fără reprezentări vectoriale	MARA	8134	SWARA	10x500	identitatea vorbitorilor
SWARA-peste-MARA cu stratul GST fixat și cu reprezentări vectoriale	MARA	8134	SWARA	10x500	identitatea vorbitorilor / experiment eșuat (vorbire neinteligibilă)

Tabel 3. Lista sistemelor antrenate pentru adaptarea vorbitorilor noi

4. Concluzii

Acest raport a prezentat experimentele rulate în studiul modului de adaptare la un nou vorbitor a sistemelor de sinteză. Pentru a implementa sisteme de sinteză care pot fi adaptate la un nou vorbitor am folosit două metode.

Prima metodă este reprezentată de utilizarea unor modele de sinteză statistic-parametrice în cadrul cărora s-a dorit utilizarea a cât mai puține date de antrenare. Sistemele au fost fie antrenate cu date de la mai mulți vorbitori, adaptarea realizându-se ulterior prin ajustarea ponderilor rețelei neuronale, fie pornind de la date puține de la un singur vorbitor și utilizarea unei rețele de post-filtrare antrenată pentru a face conversia din voce sintetizată în voce naturală.

Cea de-a doua metodă se referă la utilizarea unui sistem complet neuronal ce nu utilizează reprezentări parametrice ale formei de undă, TacotronGST. În cadrul acestui sistem s-a evaluat utilizarea unei rețele pre-antrenate pe date de la un singur vorbitor și adaptarea ulterioară a ponderilor rețelei către unul sau mai mulți vorbitori cu date audio mai puține. În cadrul acestei metode, adaptarea se face relativ rapid (aprox. 10 epoci - 5-10 minute de rulare cu hardware-ul descris mai sus), iar identitatea vorbitorilor poate fi controlată prin intermediul unui strat de reprezentări vectoriale ale acestei identități, învățare tot în cadrul antrenării sistemului.

În experimentele următoare dorim să exploatăm metode de învățare a unei noi voci pornind de la un singur exemplu acustic (en. *one-shot inference)*, pentru ca identitatea vocii să fie învățată folosind numai câteva mostre de la vorbitorul nou. Totodată vom avea în vedere utilizarea principiilor de învățare continuă (en. *continual learning)* pentru a putea face atât transferul identității vorbitorilor, cât și a expresivității sau prozodiei acestora.

5. Bibliografie

Arik et al., 2018 Arik, S., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). Neural voice cloning with a few samples. In *Advances in Neural Information Processing Systems* (pp. 10019-10029).

Kubichek, 1993 Kubichek, R. (1993, May). Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on*

Communications Computers and Signal Processing (Vol. 1, pp. 125-128). IEEE.

Luong et Yamagishi, 2018 — Luong, H. T., & Yamagishi, J. (2018). Multimodal speech synthesis architecture for unsupervised speaker adaptation. *arXiv preprint arXiv:1808.06288*.

Morise, Yokomori, & Ozawa, 2016 — Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, *99*(7), 1877-1884.

Öztürk et al., 2019 — M. G. Öztürk, O. Ulusoy and C. Demiroglu, "DNN-based Speaker-adaptive Postfiltering with Limited Adaptation Data for Statistical Speech Synthesis Systems," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 7030-7034.

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).

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.

Wang et al., 2017 — 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*.

Wang et al., 2018 — Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., & Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*.

Wu et al., 2015 — Wu, Z., Swietojanski, P., Veaux, C., Renals, S., & King, S. (2015). A study of speaker adaptation for DNN-based speech synthesis. In *Sixteenth Annual Conference of the International Speech Communication Association*.

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