D3.16. Dezvoltarea unei noi metode de adaptare rapidă a vocii sintetice folosind date audio atipice

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:	**D3.16. Dezvoltarea unei noi metode de adaptare rapidă a vocii sintetice folosind date audio atipice.**
Termen:	**Noiembrie 2020**
Editor:	**Adriana Stan**
Adresa de eMail editor:	**adriana.stan@com.utcluj.ro**
Autori, în ordine alfabetică:	**Mircea Giurgiu, Beáta Lőrincz, Maria Nuțu, Adriana Stan**
Ofițer de proiect:	**Cristian STROE**

Rezumat:

Cercetările privind dezvoltarea unui sistem de sinteză text – vorbire folosind date atipice au debutat prin explorarea modului în care augmentarea datelor text de la intrare cu informații lexicale (transcriere fonetică, silabilificare, accent) ar putea să ajute la sinteza de voce folosind date atipice. Primele arhitecturi de rețele neuronale folosite au fost alcătuite din straturi recurente sau convoluționale într-o structură de tip secvență-la-secvență. Prezicerea concurentă a celor trei informații lexicale a obținut o rată de eroare la nivel de cuvânt de 13.36% pe setul de date MaRePhor. Pornind de la aceste rezultate am exploatat prezicerea concurentă a transcrierii fonetice, silabificării și accentului lexical cu arhitecturi de tip Transformer, prin care au obținut o rată de eroare de 3.06% pe setul de date RoLEX. Acest dataset a fost creat și verificat în cadrul proiectului P2. Este studiată, de asemenea, și o arhitectură folosită pentru prezicerea lemei. În continuare au fost antrenate sisteme cu vorbitor unic sau cu vorbitori multipli și care folosesc transcrierea fonetică augmentată sau/și cu accentul lexical. Adăugarea acestor informații a fost verificată prin măsuri obiective (MCD, t-SNE și curbe F0). Sistemele de sinteză sunt bazate pe rețele convoluționale sau recurente. În sistemele antrenate au fost incluse și date atipice, care nu sunt neapărat de bună calitate, conțin zgomot sau nu au o transcriere corectă din punct de vedere a textului. Pentru a evalua folosirea acestor date au fost antrenate două sisteme de sinteză, unul folosind date ale unei prezentatoare de știri, și celălalt o voce masculină extrasă dintr-o resursă colectată în P1 și segmentată și transcrisă în P3. Aceste rezultate dovedesc faptul că sistemele de sinteză analizate au o flexibilitate destul de crescută în ceea ce privește calitatea datelor de intrare și pot genera voci sintetizate de o calitate acceptabilă folosind o cantitate redusă de date de antrenare. Mostre audio cu semnal vocal sintetizat pot fi ascultate în diferite pagini web aferente proiectului Sintero.

Cuprins

1. Introducere

În cadrul acestui raport sunt prezentate rezultatele aferente etapei de valorificare a resurselor dezvoltate în cadrul proiectelor P1, P2 și P3, in cadrul proiectului P4 prin dezvoltarea unei metode de adaptarea a vocii sintetizate pe bază de date atipice.

De exemplu, rezultatele proiectului 2 sunt valorificate prin utilizarea resursei lexicale generate în acest proeict în scopul augmentării datelor de intrare din sistemul de sinteză prin includerea silabificării și a accentului lexical în secvența fonetică a textului de intrare și implementarea unor sisteme de sinteză text – vorbire.

De asemenea, rezultatele colectării datelor în P1 segmentate și transcrise în P3 sunt utilizate pentru generarea unor sisteme de sinteză text-vorbire ce utilizează date atipice, ce nu au fost dezvoltate cu scopul specific de a antrena sisteme de sinteză.

Suplimentar, folosind noile resurse audio generate în P4 și descrise în livrabilul D3.15, în speță corpusul SWARA 2.0, acestea sunt utilizate pentru a realiza transferul expresivității unui vorbitor în cadrul unui sistem de sinteză bazat pe fluxuri de normalizare (en. Normalizing Flows). Secțiunile următoare descriu sinteza aceste rezultate.

2. Augmentarea datelor de intrare de tip text prin predicția informației lexicale de nivel înalt.

Una din problemele esențiale care se referă la date atipice este lipsa, sau prezența doar parțială, a adnotărilor necesare pentru sistemul de antrenare. Astfel, pentru a analiza folosirea informațiilor lingvistice de nivel înalt în sistemele de sinteză text – vorbire am dezvoltat o serie de modele și am implementat aplicațiile software aferente predicției automate a informațiilor de transcriere fonetică, silabificare, accent lexical și lemă a unui cuvânt. În continuare, descriem aceste metode și experimentele aferente.

2.1 Predicția simultană a informațiilor lexicale

În cadrul acestui set de experimente s-a urmărit utilizarea informației lexicale de nivel înalt: transcriere fonetică, silabificare și accent lexical în codificarea textului de intrare pentru sistemul de sinteză. Pornind de la resursele existente în Consoțiu (ICIA): RoSyllabiDict, MaRePhor, DEX online, s-au antrenat o serie de rețele neuronale ce au ca obiectiv predicția concurentă a celor 3 informații lexicale.

Într-o primă fază, s-a utilizat setul de date MaRePhor (Toma et al., 2017) combinat cu RoSyllabiDict (Barbu, 2008) și baza de date Dex online (DEX). A rezultat astfel un set de date în care pentru fiecare cuvânt din lista MaRePhor s-a generat o reprezentare a textului ce conține transcrierea fonetică, silabificarea și accentul lexical. Un exemplu de astfel de transcriere este prezentată în Tabelul 1.

Pentru a prezice concurent cele 3 informații lexicale: transcrierea fonetică, silabificarea și accentul lexical am examinat 4 arhitecturi diferite de rețele neuronale de tip S2S (Sequence to Sequence) pentru un set de date în limba română și în limba engleză. Aceste arhitecturi sunt bazate pe straturi recurente de tip Long Short Term Memory - LSTM și Bidirectional LSTM - BLSTM, precum și straturi convolutionale de tip Convolutional Neural Network - CNN. De asemenea, s-a folosit un mecanism de atenție pentru a combina ieșirea codorului și a decodorului. Arhitectura care a obținut cele mai bune rezultate este ilustrată în Figura 1.

Tabelul 1. Exemple de intrare-ieșire pentru rețelele de predicție concurentă a transcrierii fonetice, silabificării și accentului lexical. Punctul marchează silabele, apostroful accentul, iar transcrierea fonetică este dată în format SAMPA.

Intrare	Ieșire
abandonarăți	*a . b a n . d o . n a' . r @ ts i_0*
basculantei	*b a s . k u . l a' n . t e j*
ciclopul	*tS i . k l o' . p u l*

Rezultatele au fost evaluate la nivel de acuratețe, adică numărul secvențelor de ieșire prezise corect împărțit la numărul total de intrări din setul de test. S-a calculat suplimentar și acuratețea fără a considera accentul sau silabificarea prezisă. Accentul a fost prezis incorect cel mai des, fapt ce rezultă și din complexitatea acestei informații atât pentru limba română cât și pentru limba engleză. Pentru limba română o acuratețe de 86.64% a fost obținută folosind straturi convoluționale pentru prezicerea celor 3 informații. Fără accentul lexical, acesta fiind ignorat în secvența prezisă, iar acuratețea fiind calculată numai pentru transcrierea fonetică și silabificare, valoarea ajunge la 93.84%. Pentru limba engleză acuratețea cea mai bună este de 58.96%, iar fără accent de 64%. La nivel de caracter, acuratețea este de 94.94% folosind o arhitectură bazată pe straturi recurente. Rezultatele cele mai importante sunt sumarizate în Tabelul 2.

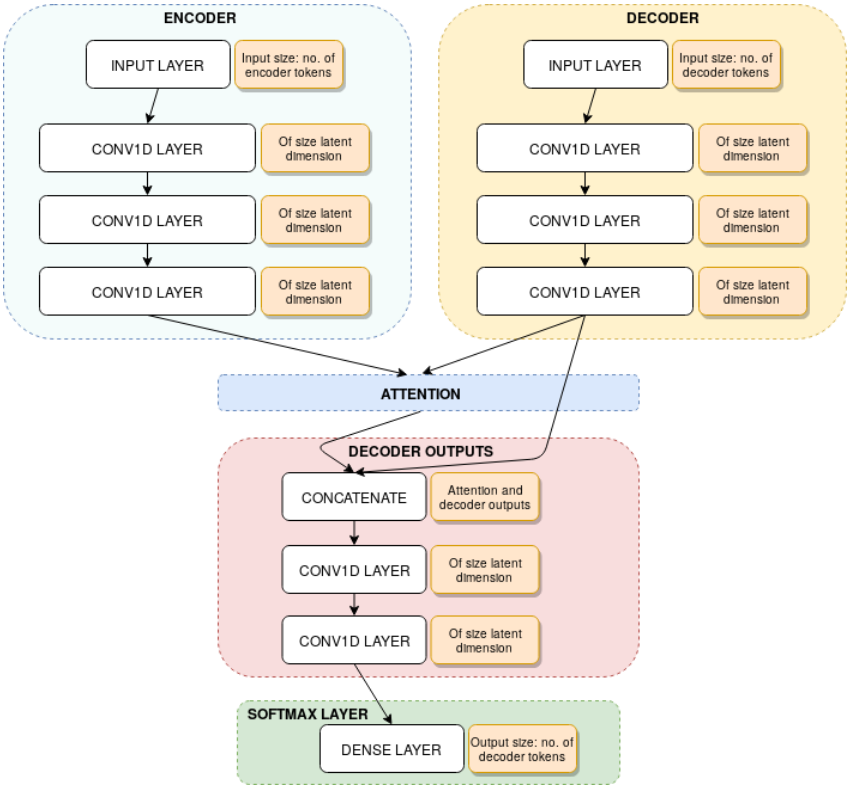

Fig.1 Arhitectura sistemului CNN cu atenție

Tabelul 2. Acuratețea rețelelor neuronale recurente, respectiv convoluționale, ce utilizează mecanismul de atenție în predicția concurentă a informației lexicale derivată din setul de date MaRePhor pentru română și CMUDict pentru engleză.

Arhitectură	**Limba**	**Acuratețe [%]**			
		3 informații lexicale	**fără silabificare**	**fără accent**	**la nivel de caracter**
CNN cu atenție	*română*	***86.64***	***88.83***	***93.84***	-
LSTM cu atenție	*română*	*86.26*	*88.68*	*92.87*	-
LSTM	*română*	*84.60*	*86.89*	*91.19*	-
BLSTM	*română*	*86.10*	*88.13*	*92.73*	-
CNN cu atenție	*engleză*	***58.96***	***59.70***	***64.00***	*85.53*
LSTM cu atenție	*engleză*	*53.79*	*54.43*	*57.24*	*81.15*
LSTM	*engleză*	*52.81*	*53.41*	*56.33*	*90.30*
BLSTM	*engleză*	*56.02*	*56.72*	*59.71*	***94.94***

Aceste experimente au fost prezentate în cadrul articolului: *Beáta Lőrincz, "Concurrent phonetic transcription, lexical stress assignment and syllabification with deep neural networks", Proceedings of the 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems KES 2020, 2020*. Prezentarea video a acestor rezultate este disponibilă la următoarea adresă: https://youtu.be/-iLf2ZvTeKY.

Pornind de la rezultatele anterioare, s-a dorit dezvoltarea unor structuri neuronale mai performante, precum și utilizarea resurselor dezvoltate la parteneri (ICIA). Resursa lingvistică generată de ICIA, descrisă în livrabilul D3.6 și care conține peste 330.000 de cuvinte adnotate cu informații despre lema, descriptori morfo-sintactici, silabificare, accent lexical și transcriere fonetică a fost utilizată pentru predicția concurentă a silabificării, accentului lexical și transcrierii fonetice, pornind fie doar de la transcrierea ortografică a cuvântului, fie de la transcrierea ortografică plus descriptorii morfo-sintactici.

În ceea ce privește rețeaua neuronală, s-a optat pentru o rețea de tip Transformer (Vaswani et al., 2017). Rezultatele predicției concurente sunt prezentate în Figura 2 și utilizează ca măsură a acurateții de predicție, rata de eroare la nivel de cuvânt (en. *Word Error Rate* - WER). Au fost analizate o serie de scenarii și partiționări ale resursei lexicale în vederea stabilirii utilității dezvoltării unor astfel de resurse de dimensiuni extinse și pentru alte limbi. S-a pornit de la utilizarea unui subset selectat aleator de doar 5.000 de cuvinte la care au fost adăugate incremental cuvinte, generând subseturi de 50.000, 100.000, 150.000 și 330.000 de cuvinte.

Pe lângă aceste partiționări, s-a analizat și utilitatea descriptorilor morfo-sintactici sau doar a părții de vorbire ca și caracteristică suplimentară în datele de intrare. Pentru o analiză mai bună asupra influenței erorilor generate de către cele trei informații lingvistice prezise concurent s-au calculat ratele de eroare a predicțiilor atunci când din secvența de ieșire sunt eliminate informațiile de silabificare sau cele de accent sau ambele simultan. O analiză și discuție mai detaliată a acestor rezultate este prezentată în livrabilul D3.6. Rezultatele acestor experimente sunt în proces de transmitere spre evaluare în articolul: *Beáta Lőrincz, Elena Irimia, Adriana Stan, "RoLEX: An extended Romanian lexical dataset and its evaluation for predicting concurrent lexical information",* IEEE Signal Processing Letters.

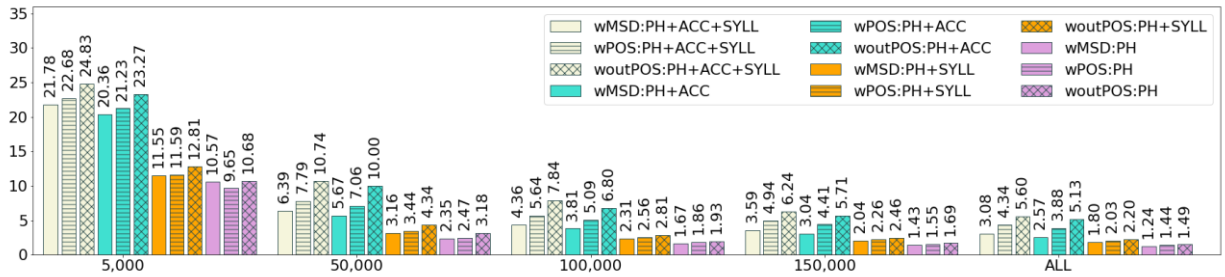

Fig. 2. Rata de eroare la nivel de cuvânt pentru predicția concurentă a informației lexicale folosind rețeaua de tip Transformer și setul de date RoLEX.

Ca urmare a rezultatelor foarte bune obținute de către rețeaua neuronală de tip Transformer pentru predicția concurentă a celor 3 informații lingvistice, am realizat antrenarea unui set de sisteme de sinteză text-vorbire care să preia la intrare aceste predicții concurente, prin comparație cu sisteme de sinteză ce utilizează doar transcrierea ortografică a textului. S-au analizat două arhitecturi de sinteză de voce diferite: una bazată pe rețele convoluționale (DC-TTS) (Tachibana et al., 2018) și una bazată pe rețele recurente (Tacotron2) (Shen et al., 2018). În ambele arhitecturi, textul setului de date de antrenare a fost transcris cu rețelele de predicție concurentă antrenate anterior. Aceste sisteme sunt descrise în secțiunile următoare.

2.2 Predicția automată a lemei

Pentru a îmbogățirea descriptorilor pentru textul de intrare în procesul de sinteză text-vorbire s-a fost dezvoltat un modul de predicție automată a lemei unui cuvânt. Au fost examinate arhitecturi bazate pe rețele neuronale recurente (LSTM) și convoluționale (CNN). Sistemele au fost antrenate folosind setul de date DEX online precum și un subset din setul de date CoRoLa (Mititelu et al., 2014).

Inițial, lema a fost prezisă folosind ca date de intrare perechi formate din cuvânt și lema atribuită. Ulterior, au fost adăugate informații contextuale prin adnotarea datelor de intrare cu partea de vorbire a cuvintelor (POS tagging) sau prin folosirea trigramelor (aceasta doar pentru subsetului CoRoLa). În cazul trigramelor, au fost testate două scenarii: sistemele prezic lema pentru întregul trigram sau doar pentru al doilea cuvânt al secvenței. De asemenea, trigramele au fost analizate cu și fără adnotarea părților de vorbire ale cuvintelor componente.

Rezultatele au fost evaluate din punct de vedere al acurateței prezicerii. Astfel, pentru setul de date colectat din DEX, sistemele bazate pe rețele LSTM ating o acuratețe de 99.43% la nivel de caracter și folosind adnotarea POS, iar pentru subsetul CoRoLa cea mai mare acuratețe este de 99.69% pentru sistemul bazat pe rețele convoluționale antrenate folosind trigrame adnotate cu partea de vorbire, care prezic lema tuturor cuvintelor din trigram. Rezultatele acestor experimente vor fi diseminate în perioada imediat următoare într-un articol într-un jurnal indexat ISI.

2.3 Utilizarea informației lexicale pentru augmentarea reprezentării textului în sisteme de sinteză text – vorbire.

Pentru sistemul DC-TTS s-a ales arhitectura prezentată în (Tachibana et al., 2018) care conține două componente bazate pe rețele neuronale: Text2Mel - prezice din text componentele spectrogramei mel de rezoluție redusă, respectiv SSRN (Spectrogram Super Resolution Network) – generează o spectrogramă de înaltă rezoluție. În experimente s-au folosit date în limbile română și engleză.

(A) Rezultate pentru antrenarea cu date audio / text în limba română.

Date audio. Au fost analizate atât sistemele antrenate cu date aparținând vorbitorului din corpusul MARA - cât și sistemele cu date audio provenind de la mai mulți vorbitori din corpusul SWARA.

Date text. În ceea ce privește textul de intrare s-au urmărit cinci scenarii diferite, obținute prin îmbogățirea treptată a informației lexicale și prezentate în tabelul următor:

Forma ortografică	E însă altceva la mijloc.
Forma transcrisă fonetic	e <> a@ n s @ <> a l t ch e v a <> l a <> m i zh l o k <.>
Forma transcrisă fonetic cu despărțire în silabe	e <> a@ n * s @ <> a l t * ch e * v a <> l a <> m i zh *bl o k <.>
Forma transcrisă fonetic cu accent	e0 <> a@1 n s @0 <> a1 l t ch e0 v a0 <> l a0 <> m i0 zh l o0 k <.>
Forma transcrisă fonetic cu silabificare și accent	e0 <> a@1 n * s @0 <> a1 l t * ch e0 * v a0 <> l a0 <> m i0 zh * l o0 k <.>

Transcrierile datelor text de intrare s-au realizat folosind rezultatele din (Lorincz, 2020b). Semnul grafic * delimitează silabele, semnele <> marchează începutul și sfârșitul unui cuvânt sau ale unui semn de punctuație, în timp ce prezența sau lipsa accentului este marcată cu 1, respectiv 0.

Experimente. Deoarece doar informația textuală este cea care suferă modificări, a fost antrenată o singură rețea SSRN pentru toate cele 5 scenarii analizate. Atât rețelele Text2Mel cât și rețeaua SSRN au fost antrenate în 3.000 de epoci.

Evaluare. În cadrul etapei de evaluare s-a analizat în ce măsură corectitudinea informațiilor lexicale suplimentare influențează sinteza sistemelor în următoarele scenarii de testare:

- silabificare și/sau accent corecte pentru întreaga propoziție
- silabificare/accent incorecte pentru un singur cuvânt
- silabificare/accent absente pentru un singur cuvânt
- silabificare/accent incorecte pentru toate cuvintele (alocate aleator)
- silabificare/accent absente pentru toate cuvintele (alocate aleator)

Sistemele descrise mai sus au fost validate folosind măsura de distorsiune cepstrală (en. *Mel Cepstral Distortion* - MCD). Pentru calcularea valorilor MCD au fost sintetizate câte 20 de propoziții lungi cu maximum 20 de cuvinte și 20 de propoziții scurte de 5 cuvine, pentru fiecare sistem, în toate cele 5 scenarii de testare. Textele au fost alese din setul de date inițial pentru a avea disponibile și mostrele audio cu vocea naturală. Astfel, valorile MCD au fost calculate între fișierele cu voce naturală și fișierele sintetizate cu fiecare sistem în parte, folosind același text. Rezultatele sunt prezentate în Figura 3.

Mostre audio pentru aceste sisteme pot fi accesate și ascultate aici: https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home .

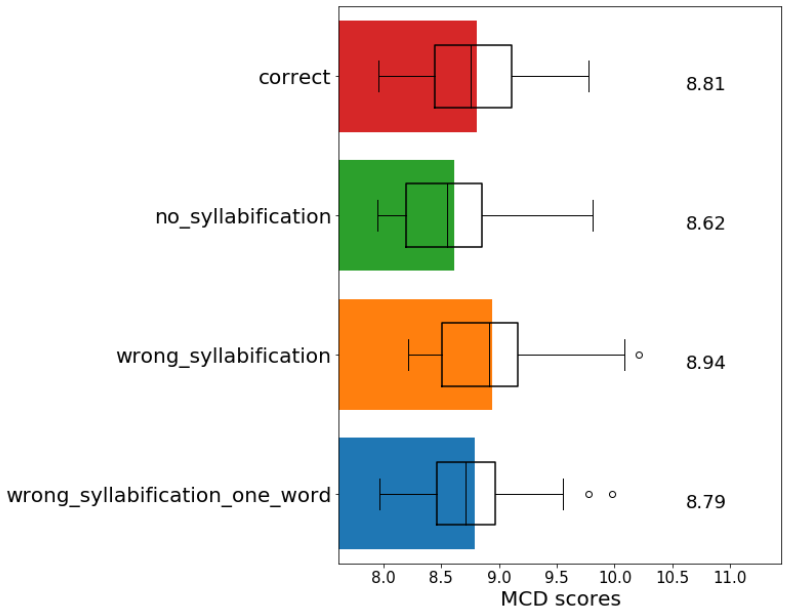

*Fig. 3 Valorile MCD calculate pentru sistemele cu un **singur vorbitor** în **limba română** antrenate folosind text transcris fonetic îmbogățit cu silabificare.*

La nivel de text, pentru a ilustra comportamentul sistemelor în cele 5 scenarii de testare au fost calculate reprezentări vectoriale ale textului (en. *embedding)* la nivel de propoziție. Acestea au fost calculate cu ajutorul codorului de text din cadrul rețelei Text2Mel. Rezultatele au fost reprezentate grafic folosind metoda t-Distributed Stochastic Neighbour Embedding (t-SNE) (Maaten & Hinton, 2008) și sunt prezentate în Figura 4.

La nivel acoustic, au fost analizate contururile curbelor F0 pentru fiecare mostră audio din cele trei scenarii de testare. Frecvența F0 a fost extrasă din fișiere ce conțin același text, dar au accent și silabificare diferită (greșită/lipsă).

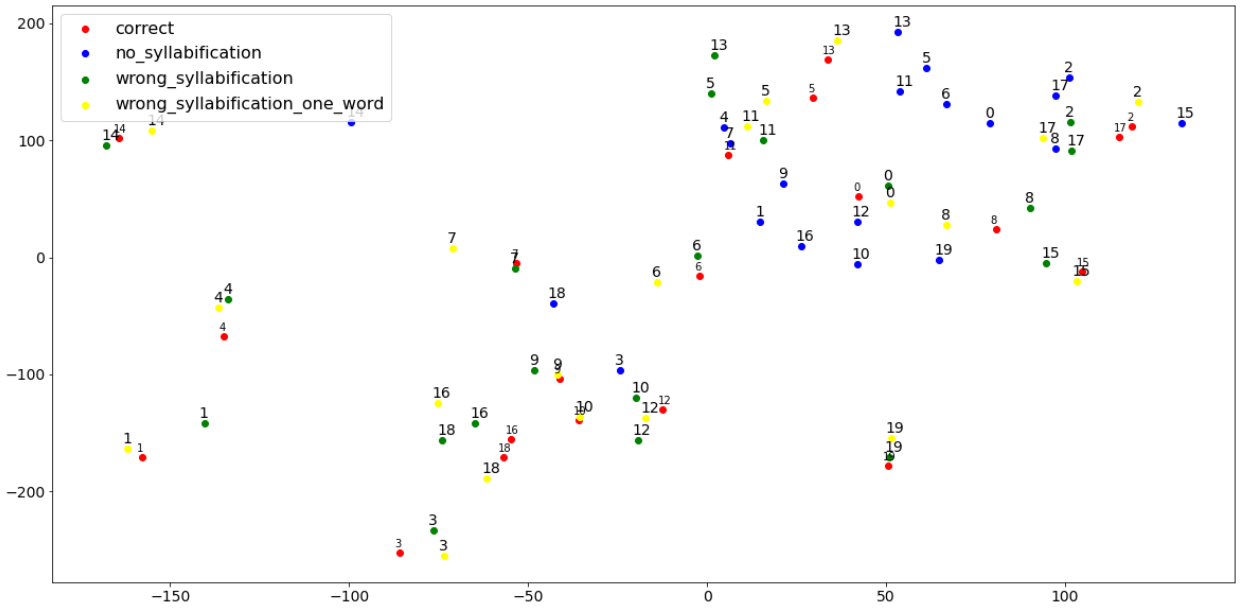

*Fig. 4 Valorile embeddingurilor de text calculate pentru sistemele cu un **singur vorbitor** în **limba română** antrenate folosind text transcris fonetic îmbogățit cu silabificare. Numerele reprezintă indicele propozițiilor folosite pentru sinteză în scenariul de test.*

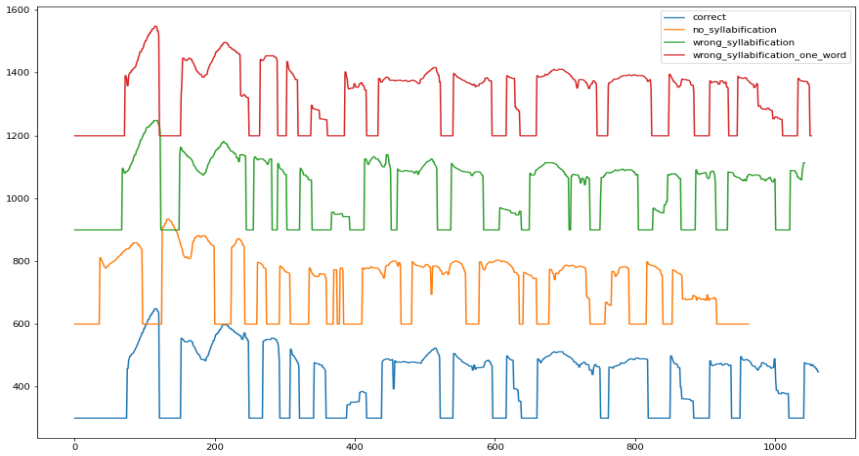

*Fig. 5 Valorile curbelor F0 calculate pentru sistemele cu un **singur vorbitor** în **limba română** antrenate folosind text transcris fonetic îmbogățit cu silabificare*

(B) Rezultate pentru antrenarea cu date audio / text în limba engleză.

Pentru antrenarea sistemelor cu un singur vorbitor s-a folosit setul de date LJSpeech[1] ce conține 24h de înregistrări aparținând unui vorbitor profesionist. Pentru scenariul sistemelor cu vorbitori multipli a fost ales setul de date LibriTTS[2] descris în (Zen et al., 2019) din care au fost selectate aproximativ 27h de înregistrări ce conțin voci feminine și aproximativ 26h de înregistrări ce conțin voci masculine.

Similar experimentelor în limba română, textele datelor de intrare au fost transcrise fonetic, fiind îmbogățite cu accent și silabificare. Această preprocesare a textului a fost realizată folosind Festival[3]. Testarea și evaluarea sistemelor antrenate în limba engleză s-au realizat în aceleași condiții descrise pentru experimentele din limba română.

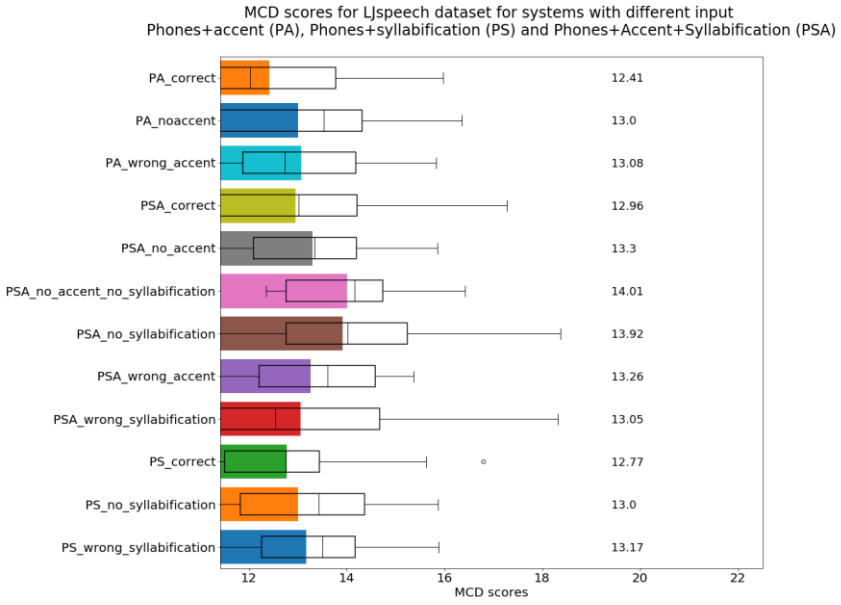

*Fig. 6 Valorile MCD calculate pentru sistemele cu un **singur vorbitor** în **limba engleză***

[1] https://keithito.com/LJ-Speech-Dataset/
[2] https://research.google/tools/datasets/libri-tts/
[3] http://www.cstr.ed.ac.uk/projects/festival/

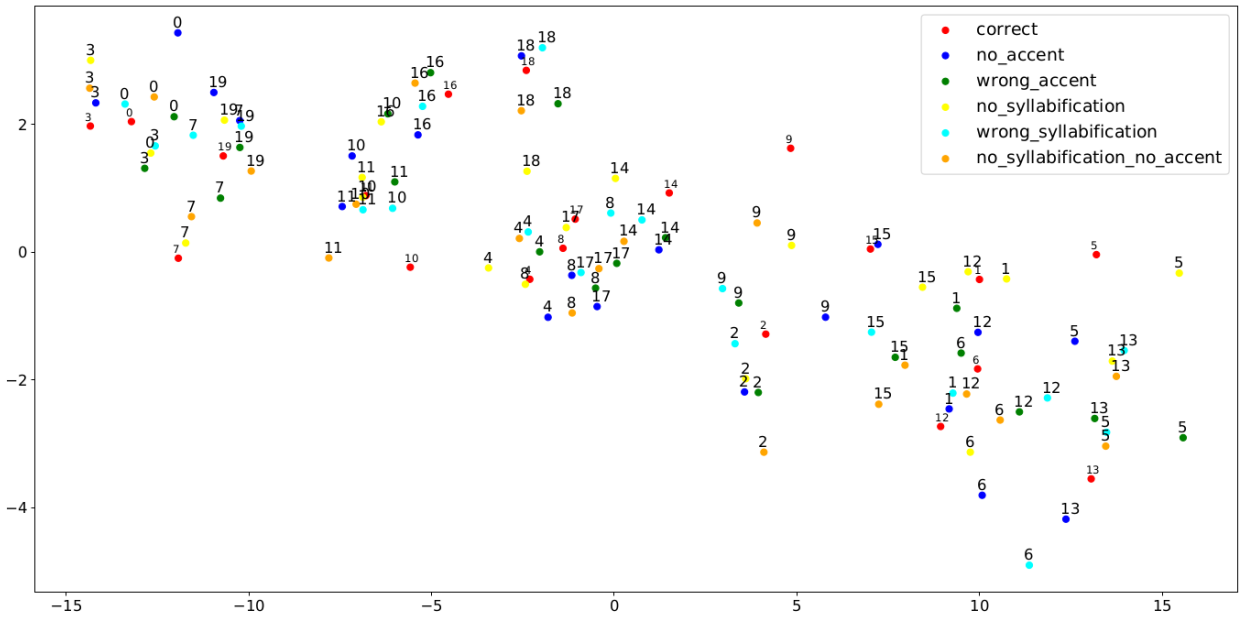

*Fig. 7. Valorile embeddingurilor de text calculate pentru sistemele cu un **singur vorbitor** în **limba engleză** antrenate folosind text transcris fonetic îmbogățit cu silabificare și marcarea silabelor accentuate. Numerele reprezintă indicele propozițiilor folosite pentru sinteză în scenariul de test.*

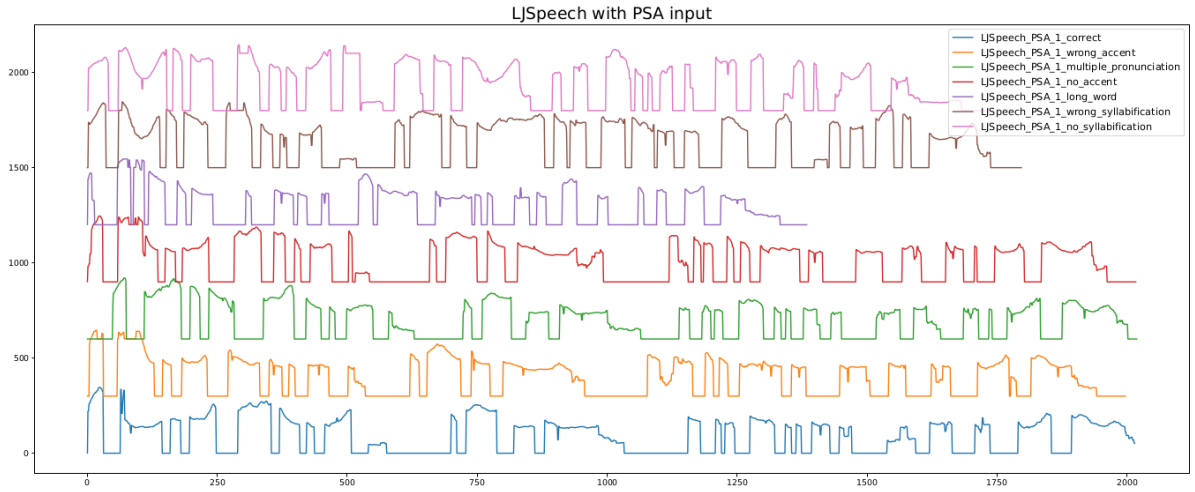

*Fig. 8 Valorile curbelor F0 calculate pentru sistemele cu un **singur vorbitor** în **limba engleză** antrenate folosind text transcris fonetic (P) îmbogățit cu silabificare (S) și marcarea silabelor accentuate (A)*

Rezultate ale acestor sisteme sunt disponibile la adresa: https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home .

Pentru sistemul Tacotron2, s-au utilizat transcrierile obținute cu ajutorul rețelei de tip Transformer. Pentru acest sistem, întreg corpusul audio Mara (aproximativ 15 ore) a fost transcris. Transcrierile nu au fost verificate manual, dar pe baza evaluării rețelei se consideră că acestea sunt în proporție de 95% corecte, cu cele mai multe erori derivând din poziționarea incorectă a accentului lexical. Rezultate ale acestor experimente sunt disponibile la următoarea adresă: https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home și vor fi diseminate într-un articol de jurnal indexat ISI în perioada imediat următoare.

3. Adaptarea vocii sintetice pe baza transferului stilului de vorbire cu Flowtron

Având în vedere dezvoltarea rapidă a domeniului de sinteză text-vorbire, transferul stilului de vorbire poate fi realizat relativ ușor și folosind o cantitate redusă de date audio, prin intermediul metodei de transfer a învățării (en. *Transfer Learning*). Cea mai recentă provocare în cadrul sistemelor de sinteză devenind acum așa numita metodă de *One-Shot Learning*, ceea ce înseamnă utilizarea unui singur exemplu audio de la un anumit vorbitor pe baza căruia rețeaua să realizeze o adaptare rapidă a ieșirii fără a fi nevoie de un proces de învățare suplimentar.

Pentru one-shot learning, una dintre cele mai recente arhitecturi flexibile pentru sistemele de sinteză este Flowtron (Valle et al, 2020). Flowtron utilizează așa-numitele fluxuri de normalizare (en. normalizing flow) (Kobyzev et al, 2020) care sunt alcătuite din funcții inversabile și care permit estimarea cu o mai bună precizie a distribuției de probabilitate a datelor de antrenare. Fluxurile de normalizare realizează o proiecție a distribuției de probabilitate a datelor de antrenare într-un spațiu latent a cărui distribuție este cunoscută și de complexitate mică (de exemplu o distribuție gausiană de medie 0 și deviație standard 1). Prin intermediul acestei proiecții inversabile, teoretic orice eșantionare a spațiului latent se va traduce într-un eșantion plauzibil, corect în spațiul inițial al datelor de intrare. Flowtron utilizează aceste structuri de normalizare pentru a realiza o proiecție a datelor audio reprezentate sub formă de Mel-spectrogramă într-o reprezentare de aceeași dimensiune (număr frecvențe Mel * număr maxim de cadre), condiționat de o reprezentare vectorială învățată a textului de intrare.

Datorită acestei proiecții în spațiul latent a datelor audio de intrare, și pornind de la experimentele realizate în domeniul prelucrării secvențelor video (Kingma and Dhariwal, 2018), odată ce spațiul latent poate reprezenta orice punct din distribuția inițială, prin proiectarea unor date audio ce conțin informații diferite de expresivitate sau identitate vocală, se poate obține o reprezentare latentă inversabilă ce va duce segmentul audio generat mai aproape de exemplul oferit.

Pentru a demonstra această capacitate a sistemului Flowtron, am recurs la antrenarea a două sisteme: unul ce folosește doar corpusul Mara (vorbitor unic) și unul ce utilizează versiunea 1 a corpusului SWARA (vorbitori multipli). Pentru transferul stilului au fost folosite date provenite din buletine de știri, precum și date preluate de la alți vorbitori din corpusul SWARA. Deși Flowtron este capabil să realizeze transferul stilului pe baza modificărilor spațiului latent[4], este nevoie ca sistemele antrenate să folosească un set de date extins și de foarte bună calitate. În experimentele noastre, transferul stilului nu se realizează în mod perfect și este nevoie de o îmbunătățire a metodei de calcul a proiecției latente, precum și o mai bună antrenare a sistemului inițial. Se poate observa în Figurile 9 și 10 că alinierea text-audio nu este realizată în mod ideal, ceea ce introduce erori de pronunție și artefacte audio. Această lipsă de aliniere a putut fi observată în majoritatea mostrelor audio provenite din sistemul Mara, cât și pentru o parte din vorbitorii din sistemul SWARA.

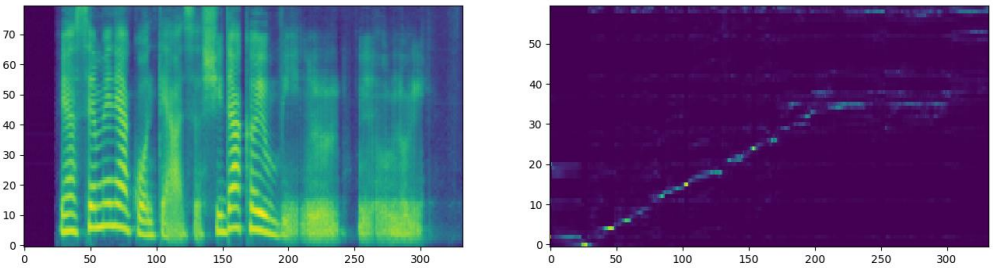

Fig. 9. Exemplu Mel-spectrogramă și aliniere text-audio în sistemul Flowtron-Mara

[4] https://nv-adlr.github.io/Flowtron

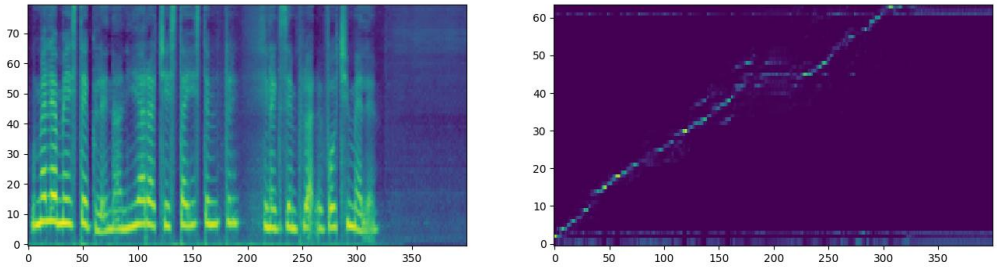

Fig. 10. Exemplu Mel-spectrogramă și aliniere text-audio în sistemul Flowtron-Swara - vorbitorul 14.

Rezultate audio ale acestor experimente sunt disponibile la adresa: https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home . Suplimentar s-a observat faptul că pentru vorbitorii de gen masculin, vocoderul Waveglow introduce o serie de erori suplimentare față de cele date de sistemul de text-Mel. Astfel că s-a recurs la o serie de antrenări ale acestui vocoder folosind date doar de la vorbitori masculini. Deocamdată, însă, rezultatele nu sunt apropiate de cele obținute pentru vorbitori feminini și investigăm modalități de îmbunătățire a lor.

4. Rezultate ale metodei de adaptare folosind date atipice

Utilizarea datelor disponibile în mod liber pe internet pentru antrenarea sistemelor de sinteză text-vorbire reprezintă o problemă de mare interes pentru comunitatea științifică. Față de un corpus dezvoltat specific pentru sistemele de sinteză, un set de date audio disponibil pe Internet poate conține o serie de probleme, precum: transcrieri inexacte, lipsa segmentărilor la nivel de propoziție, condiții de zgomot diferite, etc. Toate aceste probleme se transpun în probleme și calitate scăzută a ieșirii sistemului de sinteză. Cu toate acestea, în condițiile lipsei unor resurse de calitate, este importantă dezvoltarea unor metode ce reduc influența acestor date atipice în vocea sintetizată. Aceste metode pornesc de la o analiză mai bună a erorilor din sinteză.

Pentru acest scop, am antrenat 2 sisteme ce utilizează fie o voce a unei prezentatoare de știri și pentru care erau disponibile doar 20 de minute de audio, fie o voce masculină extrasă dintr-o resursă colectată în P1 și transcrisă în P3. Pentru vocea din urmă, transcrierile nu sunt 100% corecte, astfel că se poate analiza influența erorilor de transcriere.

Mostre audio ale acestor voci sunt disponibile în pagina https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home , sistem FEM și sistem COB. Se pot extrage următoarele observații: pentru sistemul FEM, atât calitatea înregistrărilor, cât și cea a transcrierilor este în conformitate cu cerințele sistemului de sinteză. Vocea sintetică rezultată din aceste date fiind de aceeași calitate ca și cele antrenate pe datele din corpusul SWARA. În schimb, pentru vocea COB, au existat o serie de probleme:

1) transcrierile text ale datelor audio nu sunt 100% corecte, existând o serie de inexactități;

2) înregistrările audio, deși realizate într-un studio, prezintă reverberație și zgomot de fond;

3) vorbirea este spontană și apar o serie de întreruperi, repetări, pauze în discurs;

4) segmentarea automată și transcrierea datelor a generat secvențe audio de dimensiuni reduse. Datorită inexactităților din transcriere, nu s-a putut realiza o concatenare a acestor date, astfel încât să rezulte segmente de aproximativ 10 secunde - cerințe de bază ale sistemului de sinteză.

Toate aceste probleme se reflectă în rezultatele audio ale sistemului de sinteză. Cu toate acestea, se poate observa faptul că identitatea vorbitorului este preluată de către sistem și în proporție de 80% mostrele audio au o inteligibilitate și naturalețe ridicată. Acest fapt indică o foarte bună flexibilitate a arhitecturii sistemului în ceea privește adaptarea la date atipice.

Suplimentar, într-o colaborare cu Universitatea din Madeira s-a dorit analizarea utilizării sistemelor de sinteză ca furnizor de informații în posturi de radio locale ale comunităților izolate.

În cadrul articolului: *Kristen M Scott, Simone Ashby, Adriana Stan "Designing a Synthesized Content Feed System for Community Radio", Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, Estonia, 2020* s-a efectuat un prim studiu al percepției ascultătorilor posturilor de radio privind inteligibilitatea, atractivitatea și gradul de încredere pe care o voce sintetizată îl conferă informației transmise automat și care are directă legătură cu comunitatea căreia i se adresează.

5. Concluzii

Cercetările privind dezvoltarea unui sistem de sinteză text – vorbire folosind date atipice au debutat prin explorarea modului în care augmentarea datelor text de la intrare cu informații lexicale (transcriere fonetică, silabilificare, accent) ar putea să ajute la sinteza de voce folosind date atipice. Primele arhitecturi de rețele neuronale folosite au fost alcătuite din straturi recurente sau convoluționale într-o structură de tip secventă-la-secvență. Prezicerea concurentă a celor trei informații lexicale a obținut o rată de eroare la nivel de cuvânt de 13.36% pe setul de date MaRePhor. Pornind de la aceste rezultate am exploatat prezicerea concurentă a transcrierii fonetice, silabificării și accentului lexical cu arhitecturi de tip Transformer, prin care au obținut o rată de eroare de 3.06% pe setul de date RoLEX. Acest dataset a fost creat și verificat în cadrul proiectului P2. Este studiată, de asemenea, și o arhitectură folosită pentru prezicerea lemei.

Pentru a valorifica predicțiile la nivel lexical, am efectuat o serie de experimente care folosesc aceste reprezentări de date în sisteme de sinteză. Au fost antrenate sisteme cu vorbitor unic sau cu vorbitori multipli și care folosesc transcrierea fonetică augmentată sau/și cu accentul lexical. Adăugarea acestor informații a fost verificată prin măsuri obiective (MCD, t-SNE și curbe F0). Sistemele de sinteză sunt bazate pe rețele convoluționale sau recurente.

Un sistem de sinteză apărut în anul 2020, numit Flowtron, care este bazat pe fluxuri de normalizare, este capabil de a realiza transferul de stil și a fost experimentat cu corpusuri cu unic sau cu vorbitori multipli. Aceste experimente vor fi continuate pentru a îmbunătăți transferul de stil.

În sistemele antrenate au fost incluse și date atipice, care nu sunt neapărat de bună calitate, conțin zgomot sau nu au o transcriere corectă din punct de vedere a textului. Pentru a evalua folosirea acestor date au fost antrenate două sisteme de sinteză, unul folosind date ale unei prezentatoare de știri, și celălalt o voce masculină extrasă dintr-o resursă colectată în P1 și segmentată și transcrisă în P3. Aceste rezultate dovedesc faptul că sistemele de sinteză analizate au o flexibilitate destul de crescută în ceea ce privește calitatea datelor de intrare și pot genera voci sintetizate de o calitate acceptabilă folosind o cantitate redusă de date de antrenare.

6. Bibliografie

(Barbu, 2008) Barbu, A. M. (2008, May). Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries. In *LREC*.

(DEX) "The Romanian Explicative Dictionary (DEX) online." [Online].Available: http://www.dexonline.ro, Retrieved 2020, November

(Kingma and Dhariwal, 2018) Diederik P. Kingma, Prafulla Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions", arXiv:1807.03039, 2018

(Kobyzev et al, 2020) Ivan Kobyzev, Simon J.D. Prince, Marcus A. Brubaker, "Normalizing Flows: An Introduction and Review of Current Methods", IEEE Transactions on Pattern Analysis and Machine Intelligence On page(s): 1-16 Print ISSN: 0162-8828 Online ISSN: 0162-8828

(Lorincz, 2020b) Lőrincz, B. (2020). Concurrent phonetic transcription, lexical stress assignment and syllabification with deep neural networks. *Procedia Computer Science*, *176*, 108-117.

(Maaten & Hinton, 2008) Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(Nov), 2579-2605.

(Mititelu et al., 2014) Mititelu, V. B., Irimia, E., & Tufis, D. (2014, May). CoRoLa—The Reference Corpus of Contemporary Romanian Language. In *LREC* (pp. 1235-1239).

(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.

(Toma et al., 2017) Toma, Ş. A., Stan, A., Pura, M. L., & Bârsan, T. (2017, July). MaRePhoR—An open access machine-readable phonetic dictionary for Romanian. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (pp. 1-6). IEEE.

(Valle et al, 2020) Rafael Valle, Kevin Shih, Ryan Prenger, Bryan Catanzaro, "Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis", arXiv:2005.05957, 2020

(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).

(Zen et al., 2019) Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., ... & Wu, Y. (2019). LibriTTS: A corpus derived from LibriSpeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.