Raport de activitate SINTERO

Etapa a III-a (2020)

Prof. Mircea Giurgiu, Dr. Adriana Stan, Beáta Lőrincz, Maria Nuțu

Speech processing GROUP

Cuprins

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

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

D3.17 - Integrare tehnologie nouă și demonstrarea în realizarea interfețelor om-mașină pentru sinteza text – vorbire

Diseminare

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

Cuprins

D3.15.1. - Corpusul vocal SWARA 2.0 pentru adaptarea sistemului de sinteză la noi vorbitori

D3.15.2. - RecoApy - o nouă aplicație software pentru automatizarea înregistrărilor audio necesare în sistemele moderne de sinteză de voce de tip E2E

D3.15.3. - Rezultate experimentale privind adaptarea sistemului de sinteză la un nou vorbitor și la un nou stil de vorbire

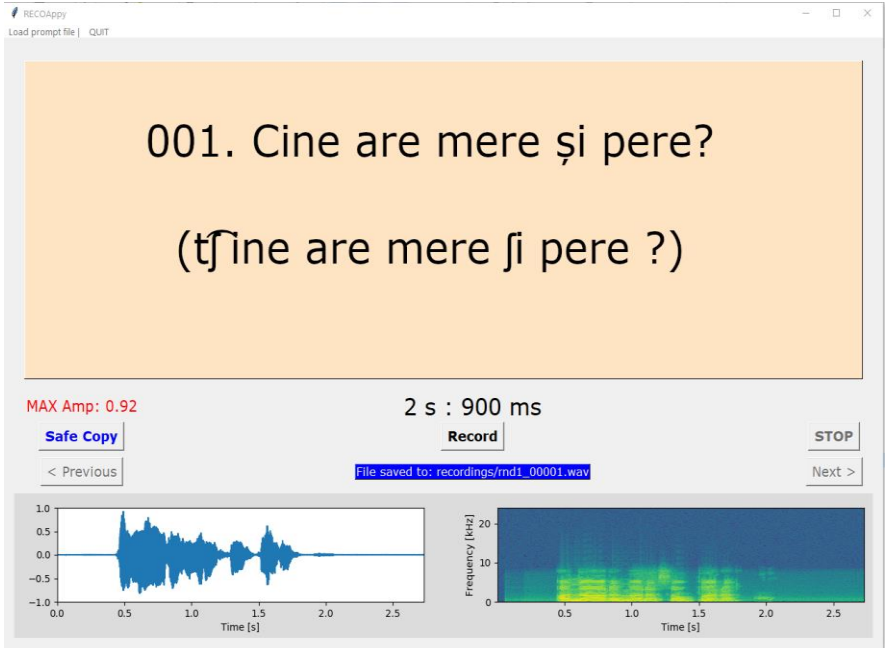
D2.15.4. - O nouă metodă de adaptare la vorbitor folosind post-filtrarea

D3.15.1. - Corpusul vocal SWARA 2.0 pentru adaptarea sistemului de sinteză la noi vorbitori

- 51.839 de segmente audio
- 29 de vorbitori: 14 masculini, 15 feminini.
- 65 de ore de vorbire
- Date paralele (text & voce)
- 48kHz, 16bps
- Condiții de înregistrare semi-profesionale

Documente suport: (1) Acord pentru utilizarea și prelucrarea semnalului vocal din partea vorbitorilor, (2) Formular de licențiere.

D3.15.2. - RecoApy - o nouă aplicație software pentru automatizarea înregistrărilor audio necesare în sistemele moderne de sinteză de voce de tip E2E. Aplicatia.



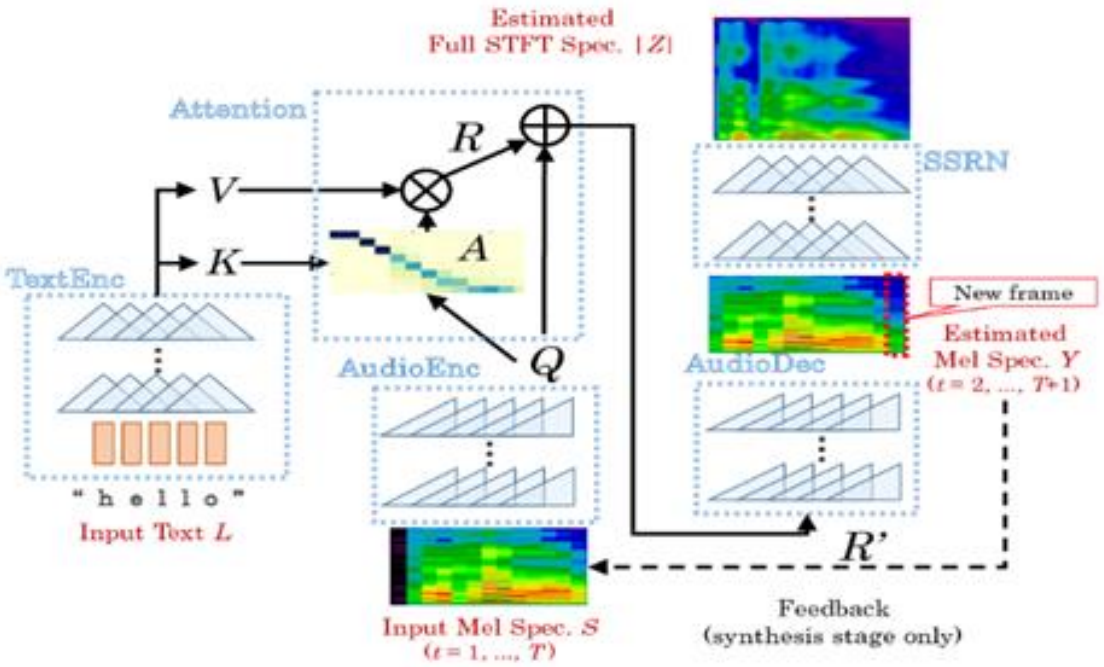
D3.15.2. - RecoApy - o nouă aplicație software pentru automatizarea înregistrărilor audio necesare în sistemele moderne de sinteză de voce de tip E2E . Rezultate transcrieri în diverse limbi.

Lang	Lexicon	Entries	Unique entries	Phonetic symbols	Model	G1	G2	G3	G4	G5	G6	G7	G8	G9	WER	PER
EN	**CMUdict**	132,585	123,874	39	CNN	2	128	2	128	3	128/64/32	ReLU	RMSp	512	29.82	11.41
					Transformer	4	3	64	4	0.01	512	64	-	-	**23.16**	**8.03**
	Wiktionary	71,332	48,773	39	CNN	2	128	2	128	3	128/64/32	ReLU	RMSp	256	28.92	12.39
					Transformer	4	4	32	4	0.01	128	128	-	-	**22.50**	**8.23**
RO	**MaRePhor**	72,375	72,375	40	CNN	3	64	2	32	3	64/32/32	Lin	Adam	128	2.64	0.5
					Transformer	2	4	32	2	0.05	64	64	-	-	**2.30**	**0.42**
	Wiktionary	63,013	62,733	32	CNN	3	128	2	32	3	128/64/32	Lin	Adam	512	**3.00**	**0.50**
					Transformer	3	2	64	2	0.05	64	256	-	-	3.58	0.71
CZ	Wiktionary	42,014	41,419	41	CNN	2	32	4	128	3	64/32/32	Lin	RMSp	128	11.69	3.84
					Transformer	2	2	32	2	0.05	64	32	-	-	**9.45**	**2.37**
DE	Wiktionary	327,296	315,793	51	CNN	3	128	3	32	3	128/64/32	ReLU	Adam	512	**5.50**	**1.43**
					Transformer	4	2	64	2	0.05	32	64	-	-	8.80	2.24
ES	Wiktionary	49,346	42,732	31	CNN	3	128	4	64	2	128/64	ReLU	Adam	128	**9.81**	**2.20**
					Transformer	2	4	32	4	0.05	32	32	-	-	11.90	2.95
FR	Wiktionary	1,121,714	1,115,343	35	CNN	3	128	3	32	3	128/64/32	ReLU	Adam	512	**4.38**	1.02
					Transformer	2	3	64	2	0.05	128	64	-	-	4.78	**0.97**
IT	Wiktionary	29,826	29,242	28	CNN	2	128	4	128	2	64/32	ReLU	RMSp	256	**18.67**	**4.44**
					Transformer	2	2	64	2	0.01	512	64	-	-	19.04	5.00
PL	Wiktionary	35,646	35,544	48	CNN	4	64	2	128	2	128/64	ReLU	Adam	128	3.59	1.84
					Transformer	3	2	64	4	0.05	1024	128	-	-	**2.98**	**1.34**

D3.15.3. - Rezultate experimentale privind adaptarea sistemului de sinteză la un nou vorbitor și la un nou stil de vorbire. Metoda propusă.

- Adaptarea unui sistem de sinteză cu vorbitori multipli folosind o funcție de cost suplimentară
  a. Funcția de cost suplimentară obținută:
     i. prin calculul similarității spectrale
     ii. cu rata de eroare egală (EER - Equal Error Rate) folosind un sistem de verificare de vorbitor
  b. Sistem de sinteză bazat pe rețele convoluționale (DCTTS)
  c. Învățarea identității vorbitorilor este realizată printr-o strategie de învățare a contribuției reprezentării vectoriale la canalele de informație din rețea

D3.15.3. - Rezultate experimentale privind adaptarea sistemului de sinteză la un nou vorbitor și la un nou stil de vorbire. Arhitectură sistem.



D3.15.3. - Rezultate experimentale privind adaptarea sistemului de sinteză la un nou vorbitor și la un nou stil de vorbire. Scenarii testare.

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***	*4.86*	***6***	*2.43*

D3.15.3. - Rezultate experimentale privind adaptarea sistemului de sinteză la un nou vorbitor și la un nou stil de vorbire. Reprezentări vectoriale vorbitori

D3.15.3. - Rezultate experimentale privind adaptarea sistemului de sinteză la un nou vorbitor și la un nou stil de vorbire

Mostre audio:

https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/Sistem-DC-TTS-Adaptare-de-vorbitor

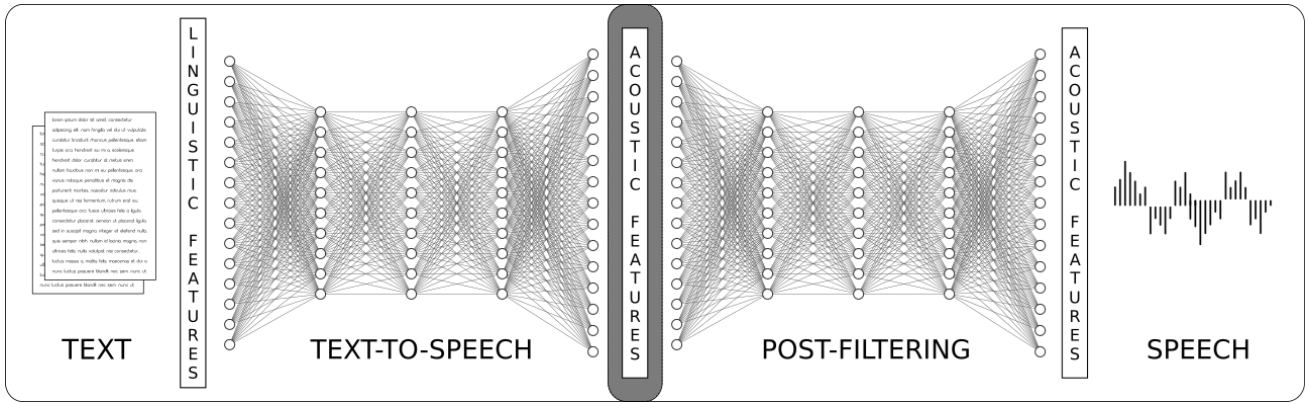
Codul sursă:

https://github.com/lorinczb/pytorch-dc-tts

D2.15.4. - O nouă metodă de adaptare la vorbitor folosind post-filtrarea

Strategii:

a. Sistem minimal de sinteză - aplicând metoda de post-filtru



Arhitectura rețelelor

a. Antrenarea cu vorbitori multipli (cu date extinse) - ajustat cu date reduse la vorbitorul nou

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

Cuprins

D3.16.1. - Augmentarea datelor de intrare de tip text prin predicția informației lexicale de nivel înalt.

Intrare	Ieșire
abandonărăți	*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*

D3.16.1. - Augmentarea datelor de intrare de tip text prin predicția informației lexicale de nivel înalt. Rezultate.

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***

D3.16.1. - Augmentarea datelor de intrare de tip text prin predicția informației lexicale de nivel înalt (cont.)

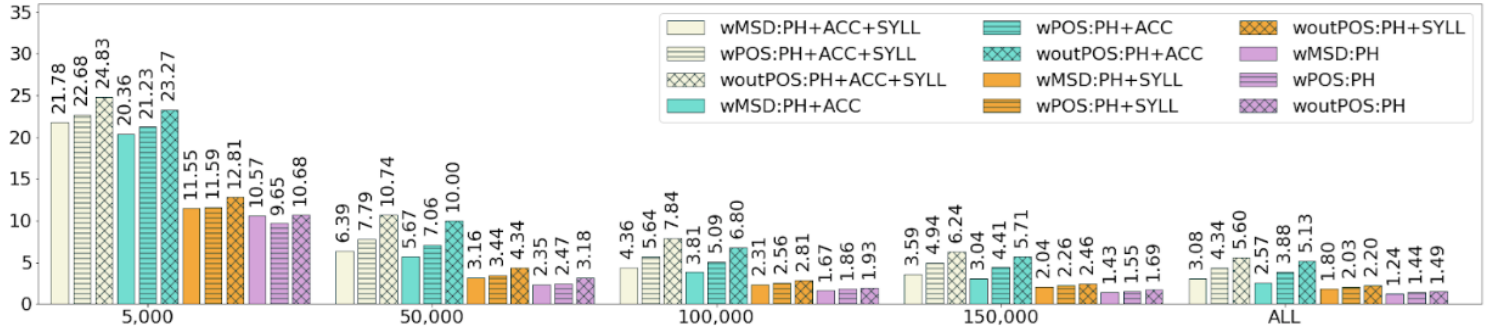


Fig. 2. Rata de eroare la nivel de cuvânt pentru predicția concurentă a informației lexicale folosind rețeaua de tip Transformer și setul de date RoLEX.

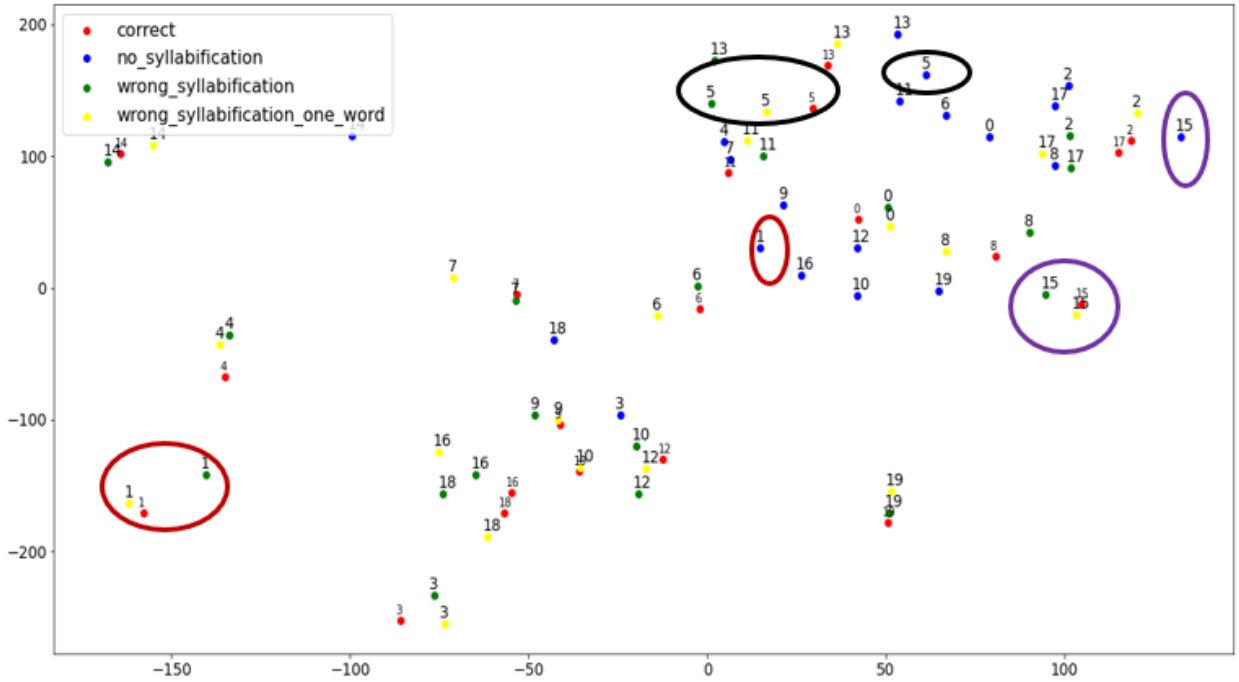
D3.16.1. - Augmentarea datelor de intrare de tip text prin predicția informației lexicale de nivel înalt (cont.)

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 <.>

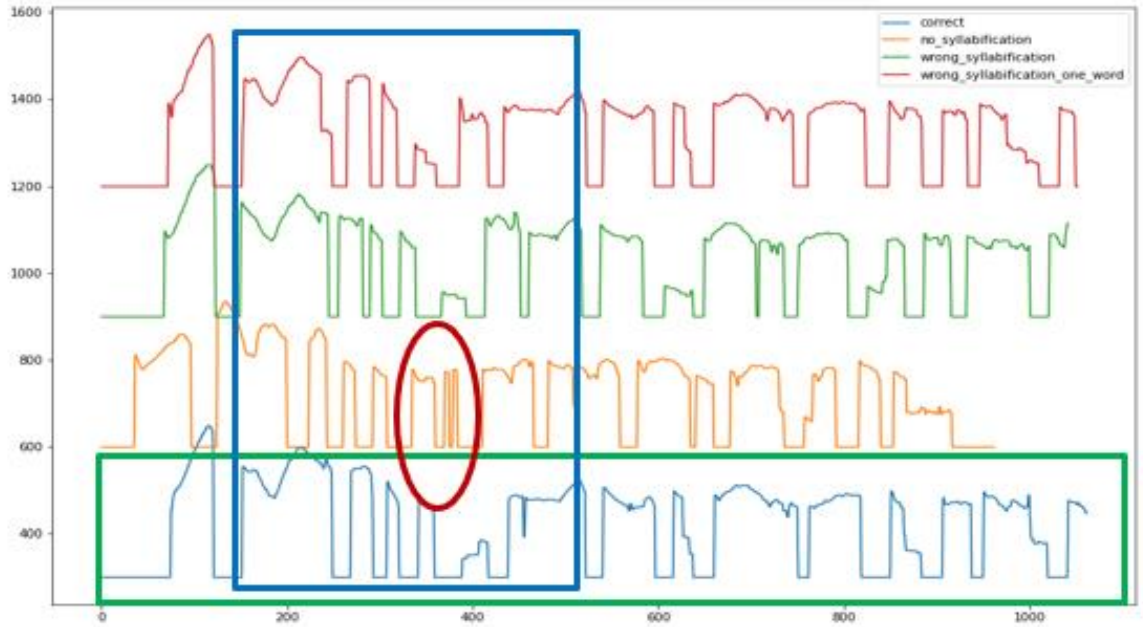
D3.16.1. - Augmentarea datelor de intrare de tip text prin predicția informației lexicale de nivel înalt (cont.)

- 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)

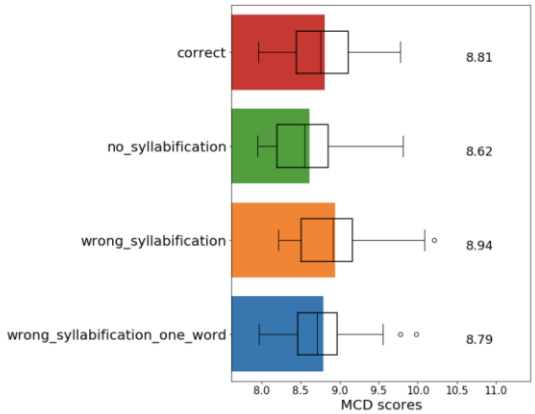
D3.16.1. - Augmentarea datelor de intrare de tip text prin predicția informației lexicale de nivel înalt (cont.). Reprezentări vectoriale ale propozițiilor

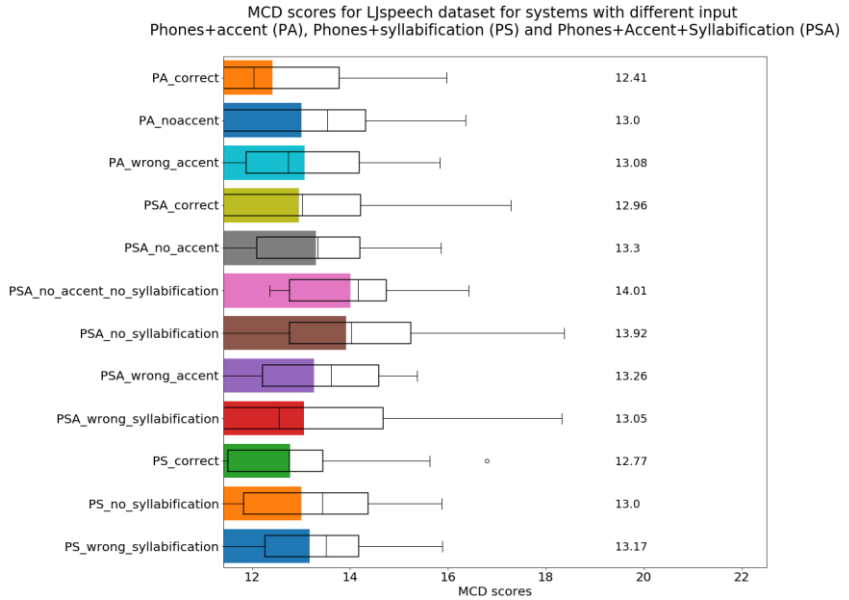


D3.16.1. - Augmentarea datelor de intrare de tip text prin predicția informației lexicale de nivel înalt (cont.). Frecvența fundamentală.



D3.16.1. - Augmentarea datelor de intrare de tip text prin predicția informației lexicale de nivel înalt (cont.). Măsuri obiective.

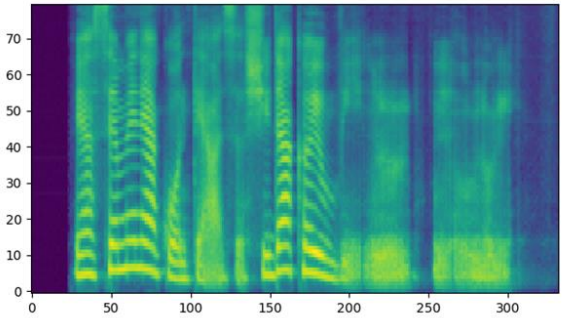


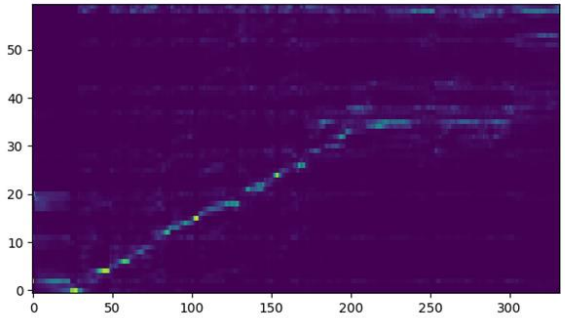


D3.16.2. - Adaptarea vocii sintetice pe baza transferului stilului de vorbire cu Flowtron

https://nv-adlr.github.io/Flowtron

NVIDIA - Flowtron





D3.16.3. - Rezultate ale metodei de adaptare folosind date atipice

Date atipice (transcrieri imperfecte, inregistrari cu zgomot, vorbire spontana).

https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/Date%20atipice

- Vorbitor feminin (FEM) - știri
- Vorbitor masculin (COB) - date extrase din CoBiLiRO - prezentator emisiune

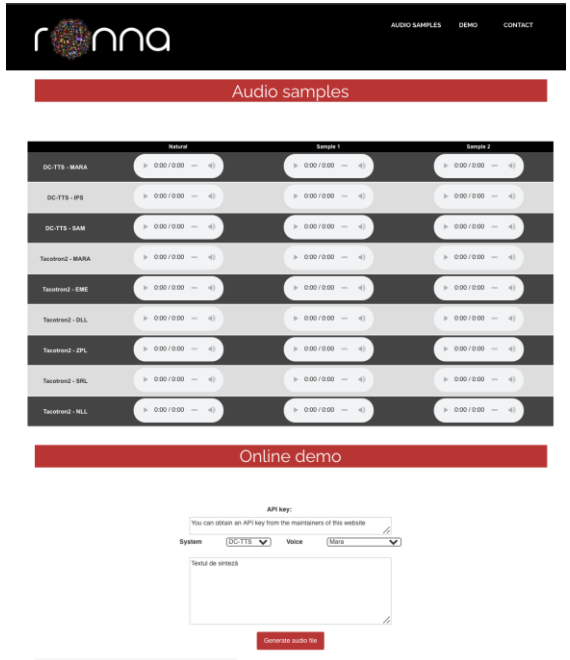
D3.17 - Integrare tehnologie nouă și demonstrarea în realizarea interfețelor om-mașină pentru sinteza text – vorbire

Cuprins

D3.17.1. - Demonstrarea unei interfețe online pentru sinteza text – vorbire în limba română folosind rețele neuronale profunde și procesarea paralelă a datelor pe servere cu GPU (Graphical Processing Unit).

D3.17.1. - Demonstrarea unei interfețe online pentru sinteza text – vorbire în limba română folosind rețele neuronale profunde și procesarea paralelă a datelor pe servere cu GPU (Graphical Processing Unit).

https://speech.utcluj.ro/ronna/



Diseminare

Articole publicate

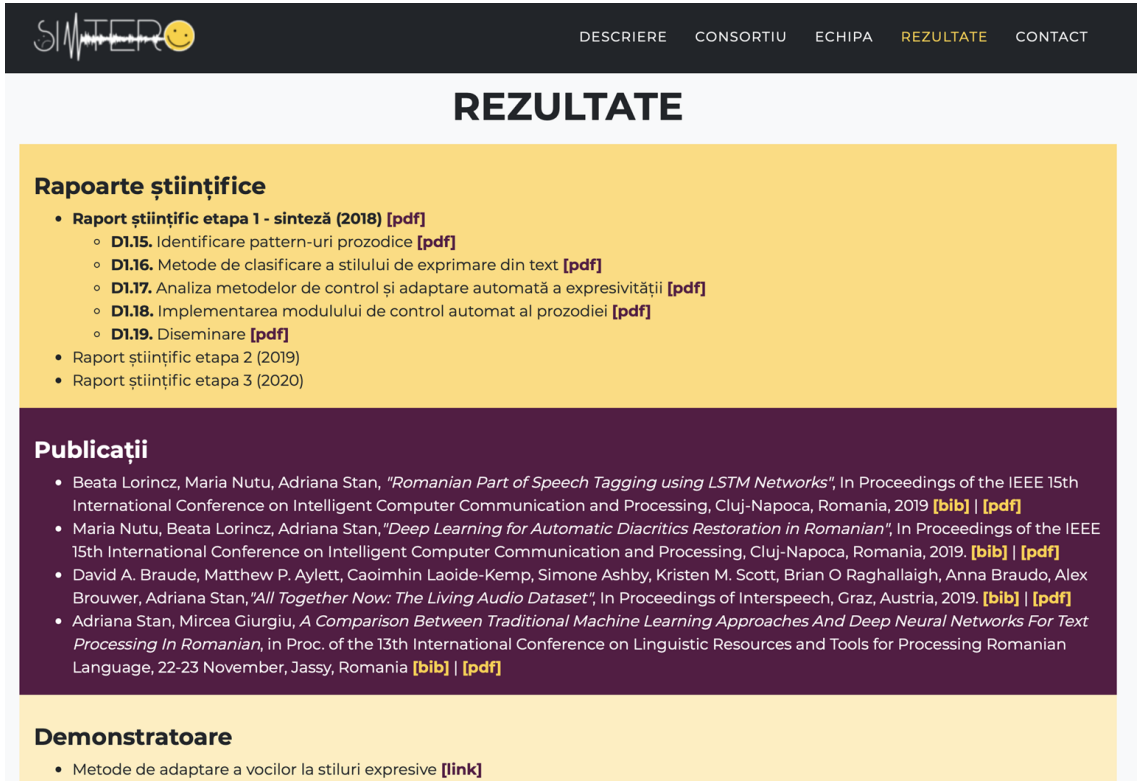
- Beáta Lőrincz, Maria Nutu, Adriana Stan, Mircea Giurgiu "An Evaluation of Postfiltering for Deep Learning Based Speech Synthesis with Limited Data", IEEE 10th International Conference on Intelligent Systems (IS), Bulgaria, 2020
- Beáta Lőrincz, "Concurrent phonetic transcription, lexical stress assignment and syllabification with deep neural networks", Proceedings of the 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems KES2020, 2020
- Adriana Stan, "RECOApy: Data Recording, Pre-Processing and Phonetic Transcription for End-to-End Speech-Based Applications", In Proceedings of the Interspeech, Shanghai, China, 2020
- 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

.

Website

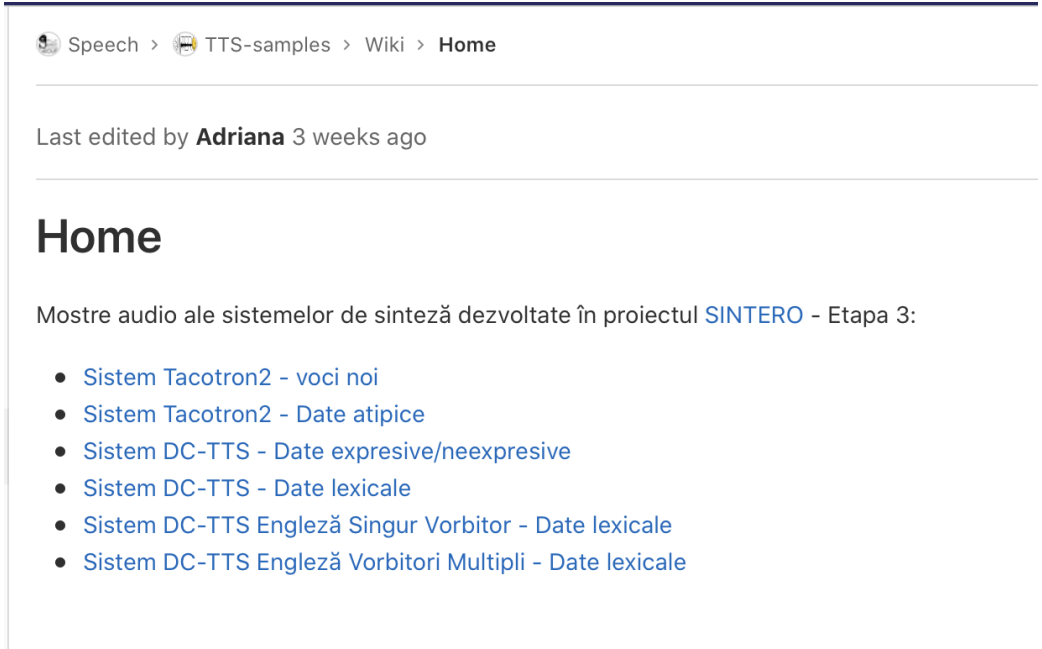

DESCRIERE
CONSORTIU
ECHIPA
REZULTATE
CONTACT
SINTERO: Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate
Proiect finanțat de Ministerul Cercetării și Inovării, Program PN-III-P1-1.2.-PCCDI, nr. 73/2018, durata: 2018-2020
Proiect component al proiectului complex
RETEROM
Proiecte paralele
COBILIRO
TEPROLIN
TADARAV

Website (2)


DESCRIERE
CONSORTIU
ECHIPA
REZULTATE
CONTACT
REZULTATE
Rapoarte științifice
Raport științific etapa 1 - sinteză (2018) [pdf]
D1.15. Identificare pattern-uri prozodice [pdf]
D1.16. Metode de clasificare a stilului de exprimare din text [pdf]
D1.17. Analiza metodelor de control și adaptare automată a expresivității [pdf]
D1.18. Implementarea modulului de control automat al prozodiei [pdf]
D1.19. Diseminare [pdf]
Raport științific etapa 2 (2019)
Raport științific etapa 3 (2020)
Publicații
Beata Lorincz, Maria Nutu, Adriana Stan, "Romanian Part of Speech Tagging using LSTM Networks", In Proceedings of the IEEE 15th International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, 2019 [bib] | [pdf]
Maria Nutu, Beata Lorincz, Adriana Stan, "Deep Learning for Automatic Diacritics Restoration in Romanian", In Proceedings of the IEEE 15th International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, 2019. [bib] | [pdf]
David A. Braude, Matthew P. Aylett, Caoimhin Laoide-Kemp, Simone Ashby, Kristen M. Scott, Brian O Raghallaigh, Anna Braudo, Alex Brouwer, Adriana Stan, "All Together Now: The Living Audio Dataset", In Proceedings of Interspeech, Graz, Austria, 2019. [bib] | [pdf]
Adriana Stan, Mircea Giurgiu, A Comparison Between Traditional Machine Learning Approaches And Deep Neural Networks For Text Processing In Romanian, in Proc. of the 13th International Conference on Linguistic Resources and Tools for Processing Romanian Language, 22-23 November, Jassy, Romania [bib] | [pdf]
Demonstratoare
Metode de adaptare a vocilor la stiluri expresive [link]

Mostre audio

https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/home



Speech › TTS-samples › Wiki › **Home**

Last edited by **Adriana** 3 weeks ago

Home

Mostre audio ale sistemelor de sinteză dezvoltate în proiectul SINTERO - Etapa 3:

- Sistem Tacotron2 - voci noi
- Sistem Tacotron2 - Date atipice
- Sistem DC-TTS - Date expresive/neexpresive
- Sistem DC-TTS - Date lexicale
- Sistem DC-TTS Engleză Singur Vorbitor - Date lexicale
- Sistem DC-TTS Engleză Vorbitori Multipli - Date lexicale

Activități viitoare (2021)

Etapa a IV-a (2021): Evaluare si distributie finala tehnologii realizate in Proiectul 4

- - teste finale în diferite scenarii, de succes;
- - licențiere și drepturi de proprietate intelectuală;
- - manual de utilizare;

Diseminare rezultate pe anul 2021:

- - valorizare cecuri între parteneri (A1) și cu industria (A2) - ?;
- - 2 mobilități / stagii de documentare;
- - publicatii

Vă mulțumim pentru atenție!

Întrebări?