Developments on Text to Speech Synthesis in Romanian at the Speech Processing Group in Technical University of Cluj-Napoca

Prof. Mircea Giurgiu
Speech Processing Research Group (UTCN)
http://speech.utcluj.ro

ConsILR 2020

Speech processing GROUP

OUTLINE

1. Speech processing research laboratory

2. Introduction on Text to Speech (TTS) synthesis

3. Natural Language Processing (NLP) tasks related to TTS

4. Acoustic modelling using deep neural networks (DNN) for high quality TTS, speaker adaptation, and expressive speech synthesis

5 – Conclusions and future work

1. Speech processing research laboratory

1. Speech processsing research laboratory (http://speech.utcluj.ro)

Laboratory accreditation (UTCN) in 2003

FP7, Coordinated by University of Cambridge
UTCN: ASR in noisy environments, RO-GRID

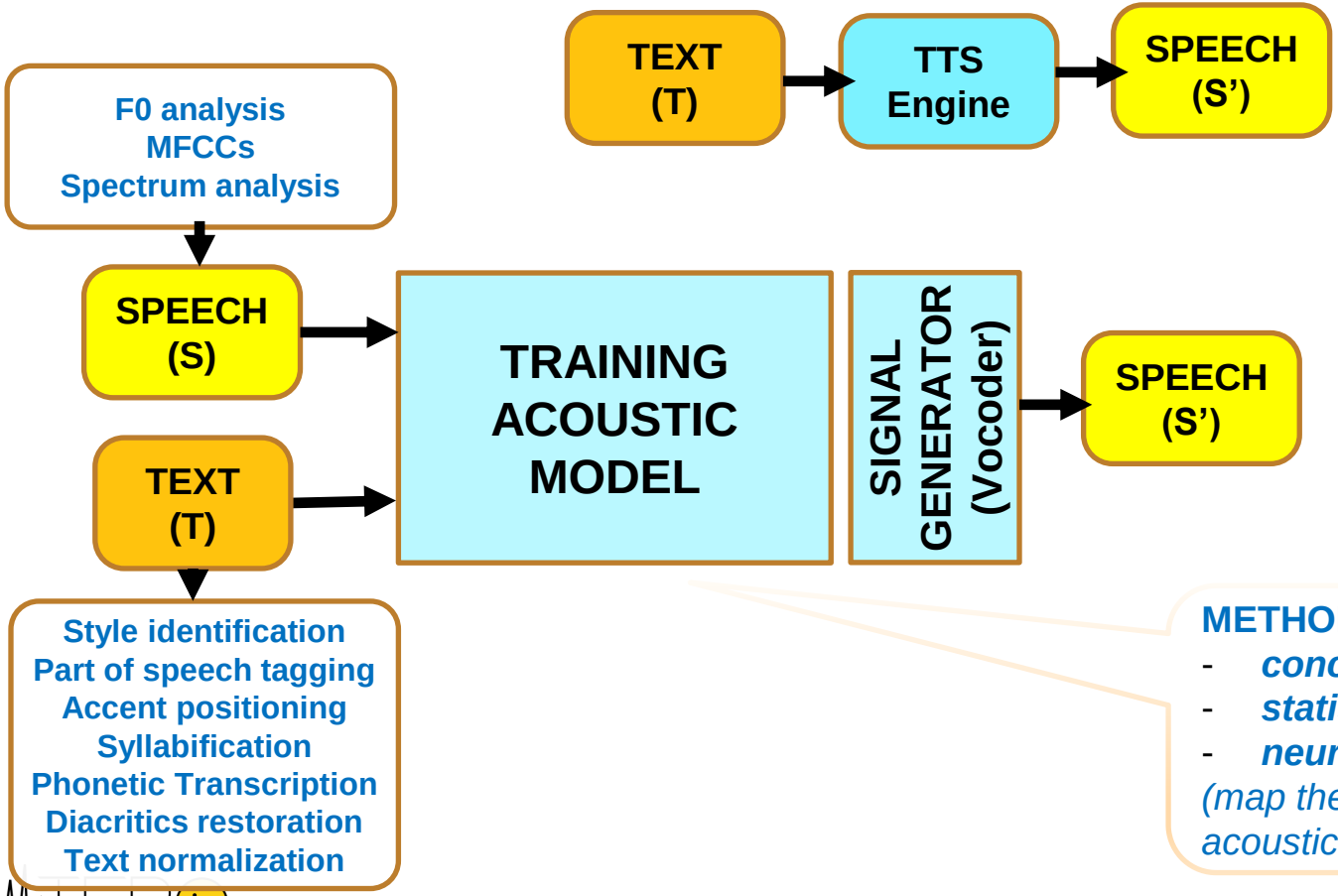
FP7, Coordinated by University of Edinburgh,
UTCN: Supervised and unsupervised tools for TTS

PCCA, Coordinated by UTCN
UTCN: TTS assistive tools for laryngectomized patients

PCCDI, Coordinated by RACAI
UTCN: Neural based TTS, Expressive TTS

2. Introduction on Text to Speech Synthesis

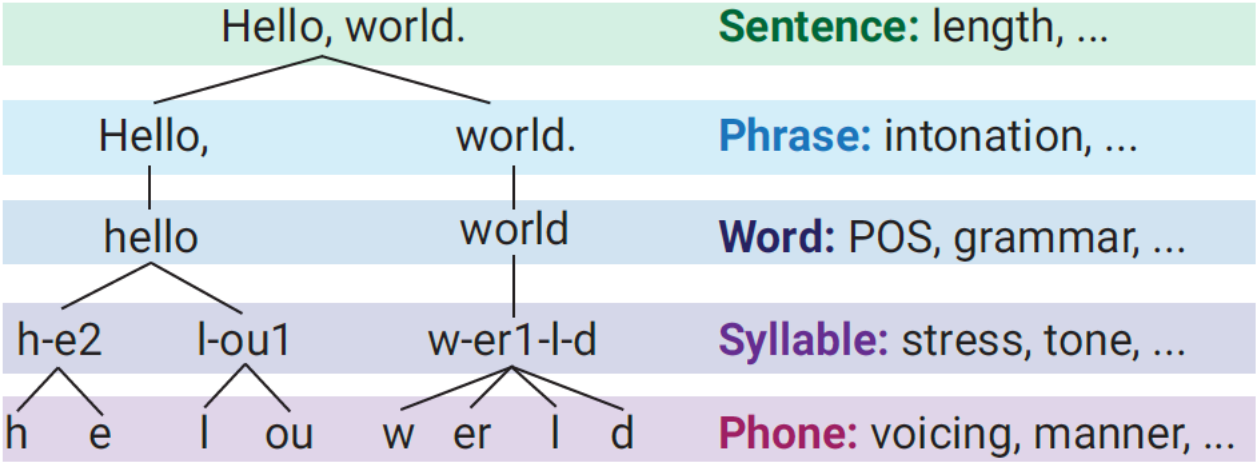
2.1. The main components of a TTS system



METHODS:
- ***concatenative***
- ***statistical (HMM)***
- ***neural (DNN)***

(map the linguistic features with acoustic features)

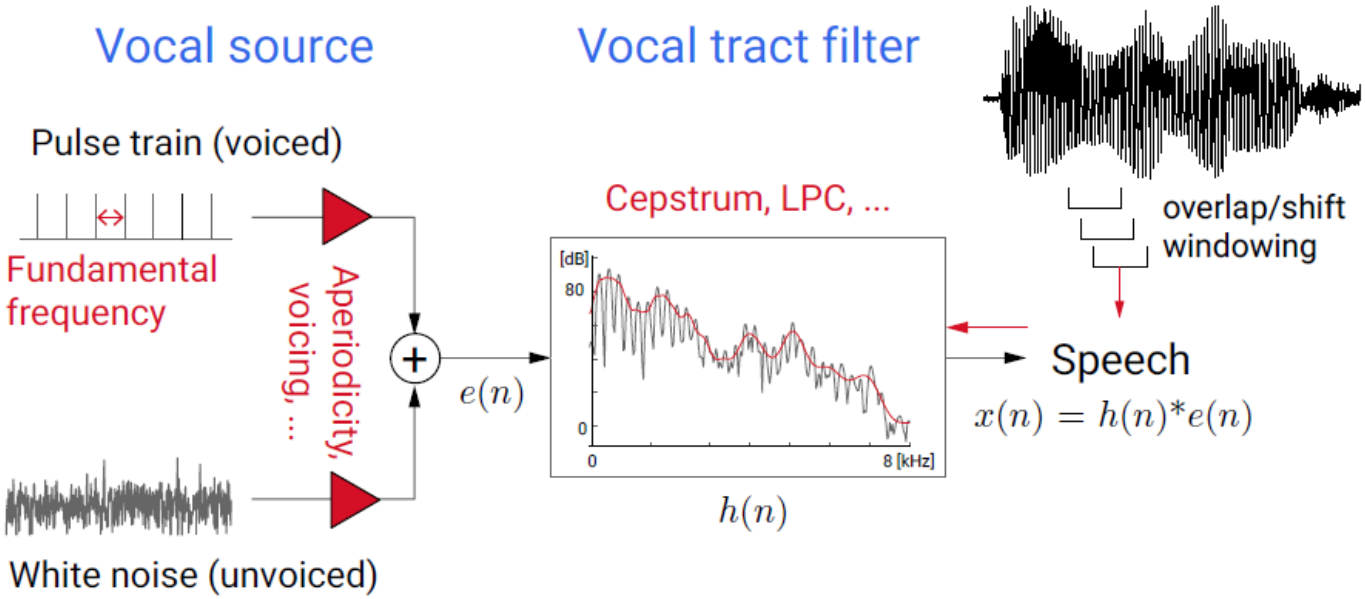
2.2. Representation of the linguistic features



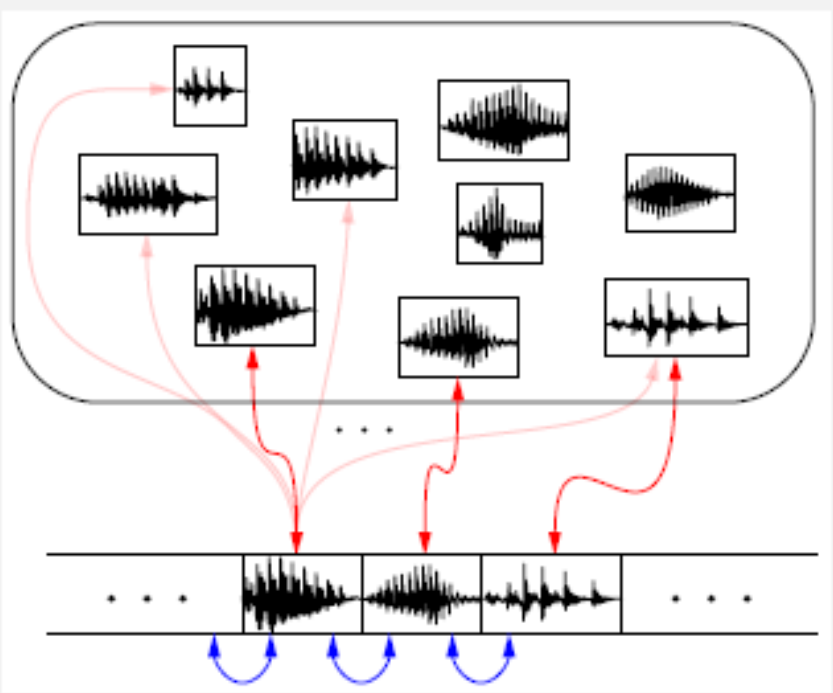
→ Based on knowledge about spoken language

- Lexicon, letter-to-sound rules
- Tokenizer, tagger, parser
- Phonology rules

2.3. Representation of the acoustic features

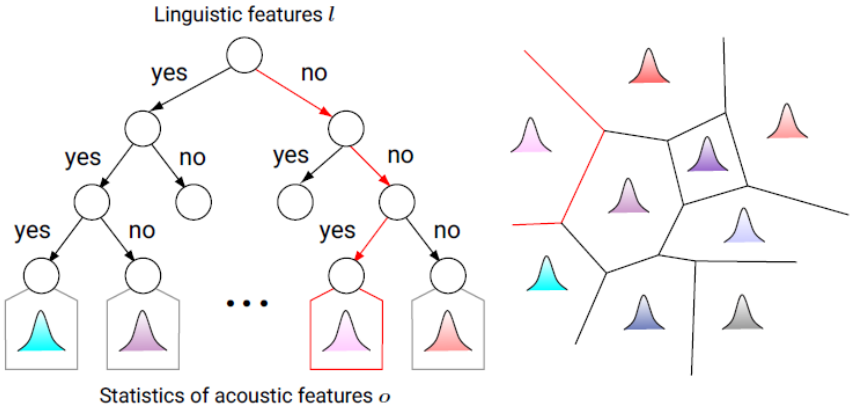


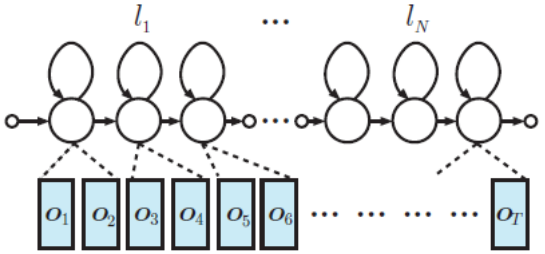
2.4. Concatenative TTS



2.5. Statistical parametric TTS (HMM-based)

HMM: Handle variable length & alignment
Decision tree: Map linguistic → acoustic



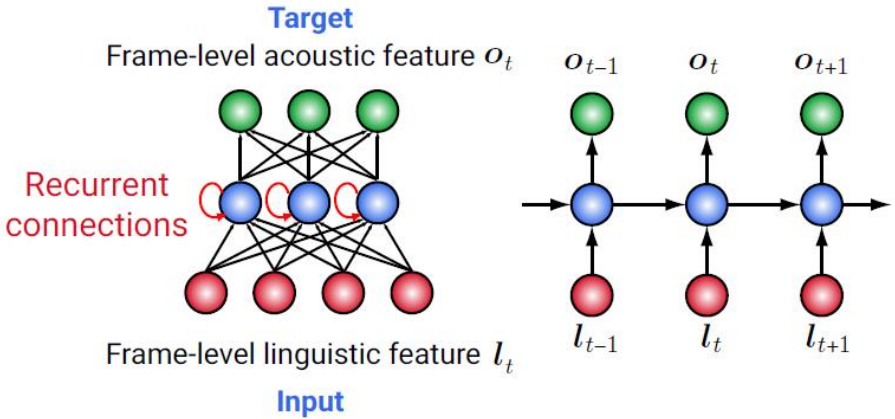


$$p(\boldsymbol{o} \mid \boldsymbol{l}, \lambda) = \sum_{\forall \boldsymbol{q}} \prod_{t=1}^{T} p(o_t \mid q_t, \lambda) P(\boldsymbol{q} \mid \boldsymbol{l}, \lambda) \quad q_t\text{: hidden state at } t$$

$$= \sum_{\forall \boldsymbol{q}} \prod_{t=1}^{T} \mathcal{N}(o_t; \boldsymbol{\mu}_{q_t}, \boldsymbol{\Sigma}_{q_t}) P(\boldsymbol{q} \mid \boldsymbol{l}, \lambda)$$

Regression tree: linguistic features → Stats. of acoustic features

2.6. Neural network – based TTS



$$h_t = g\left(W_{hl} l_t + W_{hh} h_{t-1} + b_h\right) \qquad \hat{\lambda} = \arg\min_{\lambda} \sum_t \|o_t - \hat{o}_t\|_2$$

$$\hat{o}_t = W_{oh} h_t + b_o \qquad \lambda = \{W_{hl}, W_{hh}, W_{oh}, b_h, b_o\}$$

FFNN: $\hat{o}_t \approx \mathbb{E}\left[o_t \mid l_t\right]$ RNN: $\hat{o}_t \approx \mathbb{E}\left[o_t \mid l_1, \ldots, l_t\right]$

$\hat{o}_t \approx \mathbb{E}\left[o_t \mid l_t\right] \rightarrow$ Replace decision trees & Gaussian distributions

3. Developments on Natural Language Processing (NLP) tasks related to TTS

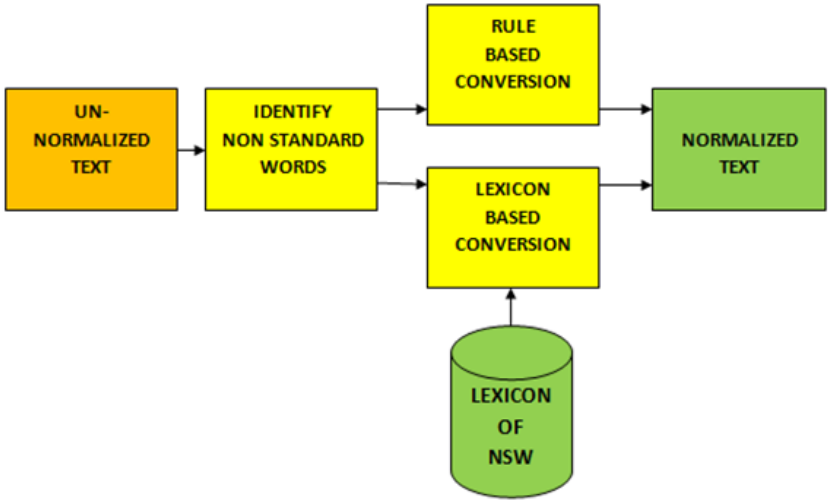
3.1. Text normalization

Text Normalization – overview

- Input text may contain Non Standard Word (NSW) – examples. Taxonomy (Sproat, 1999). TN is not a trivial task.

Categorie	*Codare*	*Specificație*	*Exemplu*
ALPHA	EXPN	abreviere	Adv., dvs., Dvs., Str., CJ.
	LSEQ	secvență de litere	CD., PP., CIA
	ASWD	citit ca un cuvânt separat	NATO, Nume proprii
	MSPL	ortografiere greșită	Siteză, senal vocal
NUMBERS	NUM	număr cardinal	12, 45, ½, 0.4, 15.23, 15,23
	NORD	număr ordinal	Al 25-lea, a 7-a zi
	NTEL	număr de telefon sau o parte din el	0742 551111, 2452
	NDIG	numar ca digiți	Camera 121, etajul 11
	NADDR	numar ca adresă stradă	15 Barițiu str., Barițiu 26
	NZIP	nod postal	400568 Cluj-Napoca
	NTIME	ora	3.20, 03:20
	NDATE	data	2/2/99, 2/2/1999, 2/02/99, 2.02.1999
	NYEAR	anul	2015, '90,
	MONEY	nume de bani (1)	12,45RON, €200,09, €200K
	BMONEY	nume de bani (2)	12,45 Mil RON, €200,09 mld
	PRCT	procent	74%, 0,34%
	SPLT	secvente mixate sau separate	WS99, x2F, 2Fh, 3-a, II-a
	ROM	litere romane	IV, IV-a, MMCX2L
MISC	SLNT	nu se pronunță	(…)
	PUNC	nu se pronunță, sfârșit frază	., (punct)
	FNSP	interjecții	Oaaauuuuuu, ahhh,
	URL	adresa site web	http://com.com
	NONE	ignoră	Caractere nedorite

Text Normalization. Solution (1) – a hybrid approach(*)



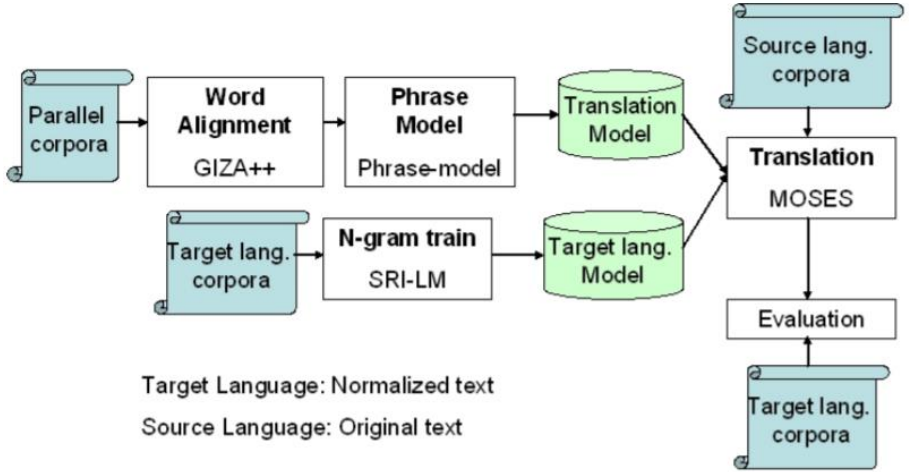
- (a) Rule-based (numerical strings) + (b) Lexicon (@, #, & ..., acronims).
- Cloud-based TTS demonstrator (//speech.utcluj.ro/swara/)

(*) *Implemented in SWARA project: http://speech.utcluj.ro/swara/*

Text Normalization. Solution (1) - results

Tip NSW	Intrare	Ieșire
Numere mai mici de 10^12	123	o sută douăzeci și trei
Numere mai mari de 10^12	750580558282384	șapte cinci zero cinci opt zero cinci cinci opt doi opt doi trei opt patru
Numere de telefon	0758123456	zero șapte cinci opt zero unu doi trei patru cinci șase
Numere cu semn	-123	minus o sută douăzeci și trei
Numere cu virgulă	123,5	o sută douăzeci și trei virgulă cinci
Ora ăn format HH:mm	Ședința este la ora 12:30	Ședința este la ora douăsprezece și treizeci de minute
Ora în format hh:mm:ss	11:35:40	Ora unsprezece treizeci și cinci de minute și patruzeci de secunde
Data in formatul d.mon.y	Noul cod fiscal va intra în vigoare la 1.ian.2016	Noul cod fiscal va intra în vigoare de la întâi ianuarie două mii șaisprezece.
Data în formatul d.m.y	Noul cod fiscal va intra în vigoare de la 01.01.2016	Noul cod fiscal va intra în vigoare de la întâi ianuarie două mii șaisprezece
Data în formatul d.mon	Noul an școlar începe în data de 15.sept.	Noul an școlar începe în data de cinsprezece septembrie
Data în format m.y	09.1944	septembrie o mie nouă sute patruzeci și patru

Text Normalization. Solution (2) – Language independent number transcription using statistical machine translation (SMT)(*)



- Text normalization is based on data only, instead of on expert rules.
- **(1)** a tokenizer, **(2)** a phrase-based translation, **(3)** a post-processing

(*) *R. San-Segundo, J.M. Montero, M. Giurgiu, I. Muresan, S. King, "Multilingual Number Transcription for Text-to-Speech Conversion", In Proc. of The 8th Speech Synthesis Workshop, Barcelona, September, 2013.*

Text Normalization. Solution (2) – Language independent number transcription using statistical machine translation (SMT)(*)

Tokenization			
EN	BLEU	WER	SER
1st alternative	97.9	1.6	7.4
2nd alternative	**98.5**	**0.8**	**6.8**
ES	BLEU	WER	SER
1st alternative	97.8	1.9	6.8
2nd alternative	**98.2**	**0.9**	**6.1**
RO	BLEU	WER	SER
1st alternative	98.5	0.9	5.4
2nd alternative	**99.2**	**0.5**	**4.6**

Aligment for training the translation model			
EN	BLEU	WER	SER
grow-diag-final	98.5	0.8	6.8
srctotgt	**99.4**	**0.4**	**4.4**
tgttosrc	98.1	0.7	6.4
ES	BLEU	WER	SER
grow-diag-final	98.2	0.9	6.1
srctotgt	98.2	0.9	6.1
tgttosrc	**98.5**	**0.7**	**4.5**
RO	BLEU	WER	SER
grow-diag-final	**99.2**	**0.5**	**4.6**
srctotgt	98.9	0.5	5.1
tgttosrc	98.5	0.8	7.8

Training Set Size			
EN	BLEU	WER	SER
200 numbers	98.3	0.8	6.4
400 numbers	99.3	0.4	4.6
800 numbers	99.4	0.4	4.4
4000 numbers	**99.9**	**0.2**	**1.8**
ES	BLEU	WER	SER
200 numbers	97.3	1.5	8.8
400 numbers	98.2	0.9	6.0
800 numbers	98.5	0.7	4.5
4000 numbers	**99.6**	**0.2**	**1.0**
RO	BLEU	WER	SER
200 numbers	95.2	2.3	20.2
400 numbers	97.7	1.3	11.2
800 numbers	99.2	0.5	4.6
4000 numbers	**99.7**	**0.2**	**2.4**

- ***(+)***
- *Use the SMT, without any rule or language specific interventions*
- *Small post-corrections (0.1%)*
- *Small training datasets (200 samples)*
- *Solution expanded for Twitter messages*
- *Tested on 3 languages (ES, EN, RO)*

(*) *"NORMA Toolkit" - http://simple4all.org/product/norma/index.html*

3. Developments on Natural Language Processing (NLP) tasks related to TTS

3.2. Diacritics restoration

Diacritics Restoration. Overview for Romanian

- *[Tufis, 1999] - 75% may be deterministically corrected. DIAC system (93% bigram, 98,85% for 8 gram).*

- [Burileanu, 2010] - n-gram language model, multilevel prediction, 1 M words, (96,93% - 99,63%).

- *[Boros, 2013] - extends the DIAC with more tokens, errors - proper names, editing, (96,89% - 99,31%)*

- [Mihalcea, 2002] – the method is Instance Base Learning , (96,14% - 99,69%) - global rate of 98,17%. We started from this approach, incl. CART

Diacritics Restoration. (1) TIMBL – SWARA project(*)

- **Corpora:** RomParl (2011-2014), RomLit (2006 – 2015), RomWiki (2014). **RomParl** has been used, about 420.000 instances.
- Window size (context of 5, 7, 9, 11 letters), Classification: TIMBL

Vector (context 3 litere), diacritic							Cuvânt proveniență
i,	*n,*	*v,*	*t,*	*a,*	*m,*	*ă*	*-învățăm- (învățământului)*
n,	*v,*	*a,*	*a,*	*m,*	*a,*	*ț*	*-nvățămâ- (învățământului)*
v,	*a,*	*t,*	*m,*	*a,*	*n,*	*ă*	*-vățămân- (învățământului)*
t,	*a,*	*m,*	*n,*	*t,*	*u,*	*â*	*-țământu- (învățământului)*
t,	*<SP>,*	*c,*	*r,*	*u,*	*i,*	*ă*	*-t cărui – (cărui)*
u,	*l,*	*t,*	*t,*	*i,*	*l,*	*ă*	*-ultățil- (facultățile)*
l,	*t,*	*a,*	*i,*	*l,*	*e,*	*ț*	*-ltățile- (facultățile)*

Information Gain / atribut (linie 1) Pondere atribut / atribut (linie 2)
0,09 / **0,44** / **0,53** / 0,10 0,02 / **0,12** / **0,15** / *0,02*
0,04 / 0,13 / **0,56** / **0,66** / 0,14 / 0,05 0,01 / 0,03 / **0,14** / **0,17** / *0,03* / 0,01
0,02 / 0,06 / 0,15 / **0,58** / **0,71** / 0,16 / 0,06 / 0,03 0,01 / 0,01 / 0,03 / **0,14** / **0,18** / 0,04 / 0,01 / 0,01
0,01 / 0,03 / 0,08 / 0,15 / **0,58** / **0,71** / 0,17 / 0,07 / 0,03 /0,01 0,00 / 0,00 / 0,01 / 0,03 / **0,14** / **0,18** / 0,04 / 0,01 / 0,00 / 0,00

(*) *M. Giurgiu, A. Stan, Livrabil proiect: D1.2 Sistem preliminar de sinteză text vorbire (2016), Swara project..*

Diacritics Restoration. (1) TIMBL - SWARA project

Diacritic	Dimens vector (context)	Vectori U / W (Unic / Within features)		
		Nr. vectori antrenare	Entropie	Performanța [%]
a – ă	5 + 1 + 5	66.000	0,79	94,93
a - â	5 + 1 + 5	53.000	0,31	99,60
a – ă - â	5 + 1 + 5	69.000	1,02	94,95
i - î	5 + 1 + 5	65.000	0,39	99,48
s - ș	5 + 1 + 5	25.000	0,72	98,20
t - ț	5 + 1 + 5	42.000	0,69	98.44
toate	5 + 1 + 5	99.900	2,62	97,63

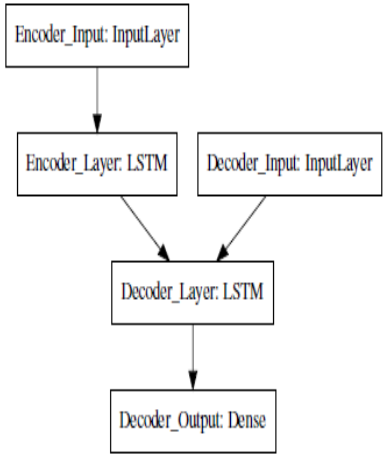
Tip eroare	Predicții incorecte
a prezis ca ă	*gafă, că, pictă, arăta,attilă, listă, contră, urmă*
ă prezis ca a	*dusa, ramân, varga, doina, amarui*
a prezis ca â	*cândida, mânea, până, român, atânasiu, bârbu, ruxândra, jucât, câsnică, rugât*
â prezis ca a	*varf, dansii, franeze, carpi, stalpilor, coborat*
i prezis ca î	*înegal, întemperii, îndependent, învers*
î prezis ca i	*reintregi, indestulat, impreunare, inclin, intreg, indrazneț*
s prezis ca ș	*clașifica, aușter, școasă, prăpaștia, folоși*
ș prezis ca s	*sinele, defineste, usurință, desarte, sefii, sapte*
t prezis ca ț	*pacință, simțe, președințe, oponență, plățindu-l*
ț prezis ca t	*piata, știti, amenințați, ședintelor, fortelor, finante, prezentă*

Diacritics Restoration. (2) CART vs TIMBL - SWARA project

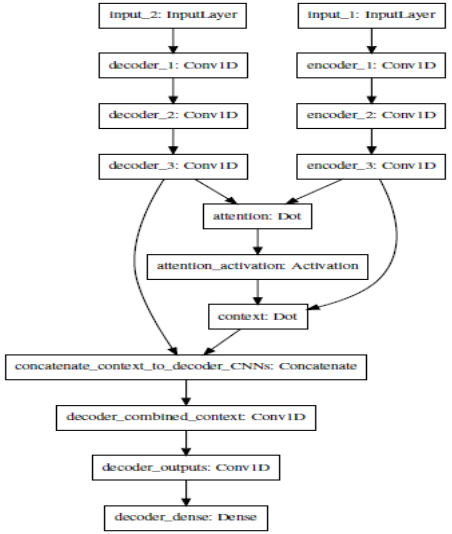
Model pentru predicția perechii …	J 48 cel mai bun	TIMBL	
		(nU), cu diacritic inclus	(U), cu diacritic inclus
	[%]	[%]	[%]
a - ă	*96,04*	*95,27*	92,73
a - â	*99,64*	*99,71*	99,43
i - î	*99,84*	*99,50*	99,29
s - ș	*98,95*	*98,10*	97,20
t - ț	*98,75*	*98,34*	97,09
a – ă - â	*95,10*	*95,06*	92,77
model global (toate)	*98,15*	*97,71*	96,25

Diacritics Restoration. (3) LSTM & CNN – ReTeRom project(*)

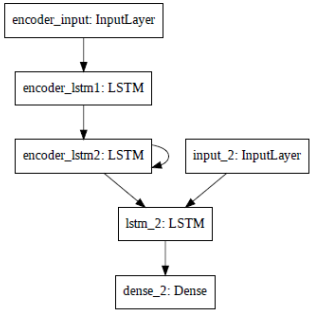
- **CoRoLa**, 1 M Tokens, 63.194 words, Input features: OHE, **4 NN structures**.



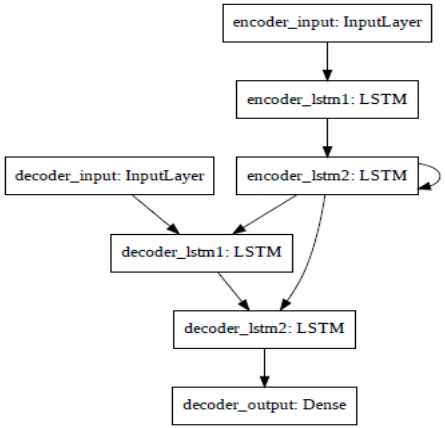
(a) seq2seq_LSTM



(b) seq2seq_CNN



(c) seq2seq_stacked_1_LSTM



(d) seq2seq_stacked_2_LSTM

(*) *M. Nutu et al, "Deep Learning for Automatic Diacritics Restoration in Romanian", ICCP 2019*

Diacritics Restoration. (3) LSTM & CNN – ReTeRom project

Architecture ID	Latent dimension	Batch size	Accuracy		
			3-gram	Word	Character
seq2seq_LSTM	128	512	75.50%	89.98%	71.61%
seq2seq_stacked_1_LSTM	256	128	79%	93 %	78%
seq2seq_stacked_2_LSTM	128	512	84%	94 %	82%
seq2seq_CNN	128	1024	**91%**	**97 %**	**89%**

Architecture ID	Accuracy			
	a-ă-â	i-î	s-ș	t-ț
seq2seq_CNN	93.51 %	99.44 %	98.39 %	97.94 %

3. Developments on Natural Language Processing (NLP) tasks related to TTS

3.3. Phonetic transcription

Phonetic transcription. Overview for Romanian

- Important task in TTS & ASR. Apart of the 9 phonetic rules in Romanian, there are a number of exceptions (diftongs, hiat).
- Syllabification and lexical stress may help the phonetic transcription.
- Previous work - diversity: **different text datasets** / different classification methods

Paper	Level	Accuracy
Burileanu, 2002	word	**98.30 %**
Ordean et al, 2009	word	94.80 %
Toma et al, 2009	word	95.00 %
Toma et al, 2013	word	96.68 %
Boroş et al, 2012	word	93.00 %
Boroş et al, 2013	word	96.29 %
Cucu et al, 2014	word	97.24 %
Boroş et al, 2017	word	95.05 %
Domokos et al, 2011	phone	92.83 %
Toma et al, 2017	phone	99.61 %
Stan et al, 2018	phone	**99.63 %**

Phonetic transcription. (1) Decision trees – NaviRO(*)

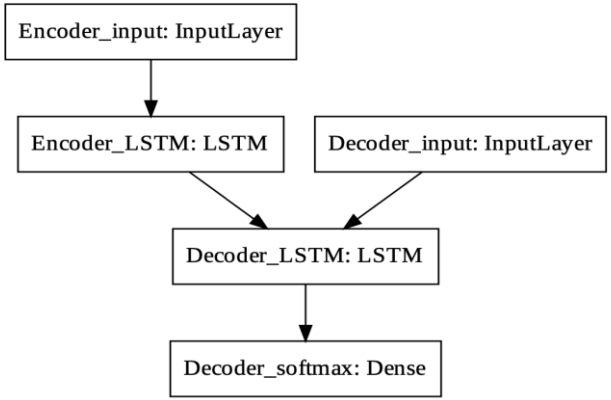
NaviRO – 138.500 words, phonetically transcribed. From the total set of 31 phonemes, prediction have been done for those depending on the context (14ph)

Lungime fereastră (nr. de litere)	Procent date de antrenare [%]	Acuratețe [%]
1	25	92.48
	50	92.38
	100	92.35
3	25	98.35
	50	98.35
	100	98.37
5	25	98.91
	50	99.15
	100	**99.30**
7	25	98.72
	50	99.05
	100	99.28

(*) *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*

Phonetic transcription. (2) DNN approach – s2s (LSTM)(*)

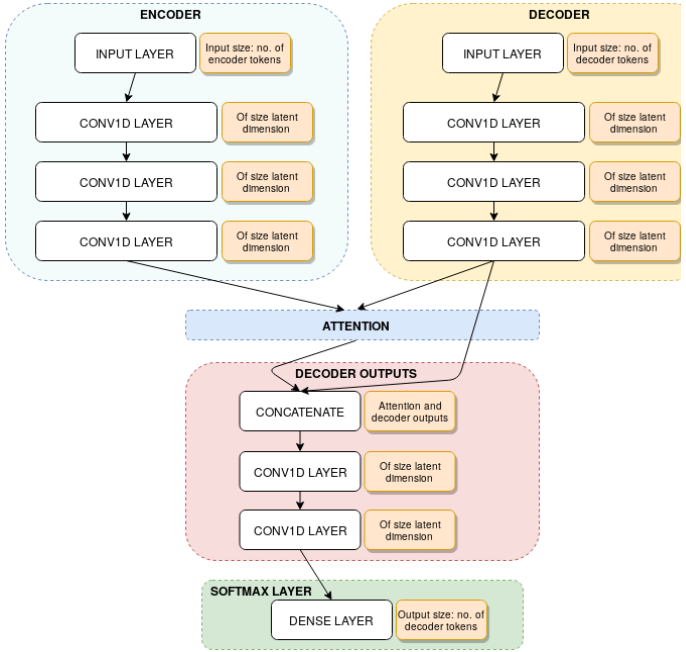
- MaRePhoR datasets: 72.375 words, 591.570 characters (Phonetic)
- DEX: 1.6 M words with their accent (Lexical stress)
- RoSyllabiDict: 507.000 syllabified words (Syllabification)
- ➔ common words: 62.874 words



	SEQ2SEQ						DECISION TREE	
	One-hot encoding		Letter embeddings		Embedding layer		Grapheme window	
Input features	Phone	Word	Phone	Word	Phone	Word	Phone	Word
Ortographic	99.42	97.05	99.51	97.40	98.75	95.40	99.54	96.71
Orthographic+Syll	99.12	96.40	99.30	96.40	98.42	95.26	99.52	96.68
Orthographic+Stress	99.01	95.85	99.56	97.70	99.01	96.00	99.56	96.95
Orthographic+Syll+Stress	99.40	97.45	**99.62**	**97.90**	99.15	96.25	**99.60**	**97.17**

(*) *Adriana Stan, "Input Encoding for Sequence-to-Sequence Learning of Romanian Grapheme-to-Phoneme Conversion", In Proceedings of the 10th IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, 2019*

Simultaneous predictions (Ph / Syll / Stress), (3) CNN/At, BLSTM(*)



Architecture	Data input format	Epochs / Batch size / Latent dimension	Accuracy (%)		
			all tasks	without syllabification	without lexical stress
CNN w Attention	OHE	200 / 256 / 128	**86.64**	**88.83**	**93.84**
LSTM	OHE	200 / 256 / 128	83.17	85.37	90.09
BLSTM	OHE	200 / 256 / 64	85.19	87.64	91.51
LSTM w Attention	OHE	200 / 256 / 128	84.25	86.62	90.90
CNN w Attention	OHE	1000 / 256 / 128	80.91	83.35	88.44
LSTM	OHE	1000 / 256 / 128	84.60	86.89	91.19
BLSTM	OHE	1000 / 256 / 64	86.10	88.13	92.73
CNN w Attention	LE	200 / 256 / 128	85.67	87.82	92.52
LSTM	LE	200 / 256 / 128	83.75	85.89	90.14
BLSTM	LE	200 / 256 / 64	85.26	87.62	91.65
LSTM w Attention	LE	200 / 256 / 128	86.26	88.68	92.87

(*) *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*

3. Developments on Natural Language Processing (NLP) tasks related to TTS

3.4. Syllabification and accent

Syllabification. Results from SWARA project(*)

- RoSyllabiDict – reference sylabification dataset
- **Simple scheme for feature encoding**

Litera	Trasaturi	Clasa
î	0 1 * * * * * î n ţ e l e	no
n	0 0 * * * * î n ţ e l e g	yes
ţ	0 0 * * * î n ţ e l e g i	no
e	0 1 * * î n ţ e l e g i *	yes
l	0 0 * î n ţ e l e g i * *	no
e	1 1 î n ţ e l e g i * * *	no
g	0 0 n ţ e l e g i * * * *	no
i	0 1 ţ e l e g i * * * * *	yes

Setul de antrenare	Clasificator	Acuratete de clasificare	Precizie
Set 1	Random Forest	**99,46%**	99,5%
	SMO	96,27 %	96,3 %
	Naive Bayes	87,03 %	87,5 %
	Ada Boost	80,92 %	81,3 %
Set 2	Random Forest	**99,00 %**	99.0%
	SMO	96,04%	96,1 %
	Naive Bayes	87,09 %	87,6%
	Ada Boost	80,92 %	81,3%
Set 3	Random Forest	**98,93 %**	98,9%
	SMO	96,02 %	96,0%
	Naive Bayes	87,13 %	87,7%
	Ada Boost	80,92 %	81,3%

(*) *Livrabil D1.15. Sistem de sinteza text vorbire preliminar (2016), Swara Project*
(Joint research with Prof. Potolea Rodica from the Computer Science Departament)

Syllabification & Lexical stress

Performanța de clasificare obtinuta de RF pentru poziționarea accentului

Test	Nr. instante	Nr. cuvinte	Acuratete (la nivel litera)
Set 1	58.434	13.210	96,37 %
Set 2	58.626	13.209	96,37 %
Set 3	58.207	13.210	96,26 %
Set 4	58.293	13.211	96,58 %
Set 5	58.288	13.210	93,53 %

Tabelul 2.4.7. Performanta de clasificare obtinuta de RF pe sarcina de pozitionare a accentului, dimensiuni variabile ale multimii de evaluare

Nr. Test	*Nr. cuvinte*	*Nr. instante*	*Acuratete*	*Cuvinte cu lipsa predictie accent*
1	1.100	4.860	96,04 %	143
2	2.200	9.720	95,72 %	318
3	3.300	14.410	96,12 %	389

(*) *Livrabil D1.15. Sistem de sinteza text vorbire preliminar (2016), Swara Project*
(Joint research with Prof. Potolea Rodica from the Computer Science Departament)

3. Developments on Natural Language Processing (NLP) tasks related to TTS

3.5. Part of speech tagging

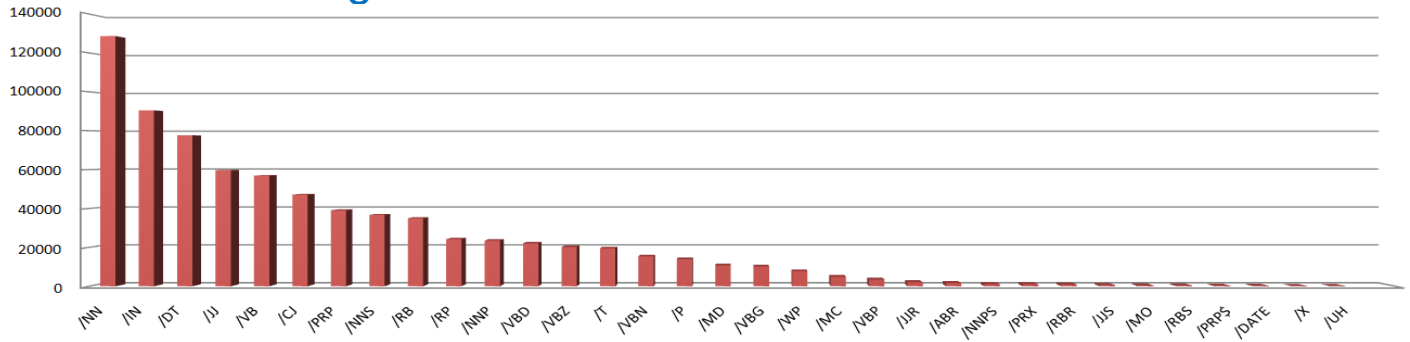
Part of Speech tagging (POS). Overview of results, RO

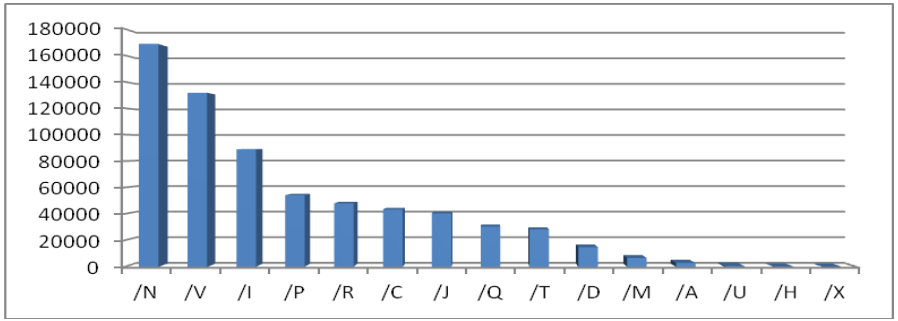
- POS is important for expressive TTS
- POS tags: MSD, C-tagset, root POS (first letter of MSD tag)

Authors	Method	Accuracy	Tagset
Tufis & Mason [2]	Probabilistic	98.39%	MSD
Boros & Dumitrescu [3]	Deep Neural Networks	98.19%	MSD
Simionescu [6]	Probabilistic & Rule-based	97.03%	MSD
Teodorescu et al. [7]	Probabilistic	96.12%	Root POS
Frunza et al. [8]	Machine Learning	95.30%	Root POS

Part of Speech tagging. Results from SWARA project(*)

- NAACL 2003 English – Romanian dataset





(*) *Toth Grigore, "Part of speech tagging for Romanian", Diploma project, 2016*
Swara Project, D1.15. Sistem preliminar de sintza text vorbire, 2016

Part of Speech tagging. Results from SWARA project(*)

Tool\Training Corpus (in lines)	1.000 linii	5.000 linii	10.000 linii	20.000 linii	30.000 linii	40.000 linii
Kytea - SVM	76.41%	86.48%	89.30%	91.57%	93.20%	**97.14%**
Kytea - LR	**76.43%**	**86.55%**	89.27%	**91.67%**	**93.25%**	97.08%
HMM Tagger	68.99%	84.69%	87.84%	90.00%	92.09%	95.29%
NLTK - TnT	76.23%	86.47%	**89.43%**	91.54%	93.07%	96.66%

(*) *Toth Grigore, "Part of speech tagging for Romanian", Diploma project, 2016*
Swara Project, D1.15. Sistem preliminar de sintza text vorbire, 2016

Part of Speech tagging. Results from ReTeRom project (DNN)(*)

- Datasets: **a)** CoRoLa (180Kw), **b)** DEX (1.9Mw), **c)** WPT (1.9Mw).
- Individual words, no context
- Encoding: OHE + Letter Embedding (Gensim)
- **(1)** LSTM with dense layers, **(2)** LSTM sequence to sequence (80%-20%)

System ID	Dataset	Tag	Network type	Character encoding	Latent dimension	Batch size	Epochs	Accuracy
1	WPT	RPOS	LSTM + Dense (*)	OHE	256	512	50	**99.18%**
2	WPT	RPOS	LSTM + Dense	OHE	256	512	50	94.85%
3	WPT	RPOS	LSTM + Dense	LE	256	256	25	54.80%
4	WPT	RPOS	seq2seq LSTM	LE	256	256	25	94.99%
5	WPT	RPOS	seq2seq + Embedding layer	OHE	256	256	20	93.88%
6	WPT	MSD	seq2seq LSTM (*)	OHE	512	1024	50	98.25%
7	WPT	MSD	seq2seq LSTM	OHE	512	1024	50	75.28%
8	WPT	MSD	seq2seq + Embedding layer	OHE	256	512	50	76.62%
9	DEX	RPOS	LSTM + Dense	OHE	256	512	50	94%
10	CoRoLa	CTAG	seq2seq LSTM	OHE	256	512	100	97.15%

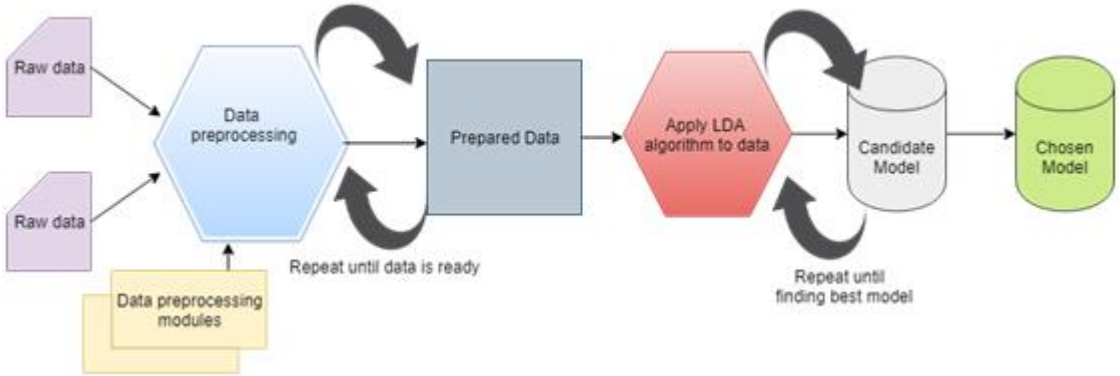
(*) *Beáta Lőrincz, Maria Nuțu, Adriana Stan, "Romanian Part of Speech Tagging using LSTM Networks", In Proceedings of the IEEE 15th International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, 2019,*

3. Developments on Natural Language Processing (NLP) tasks related to TTS

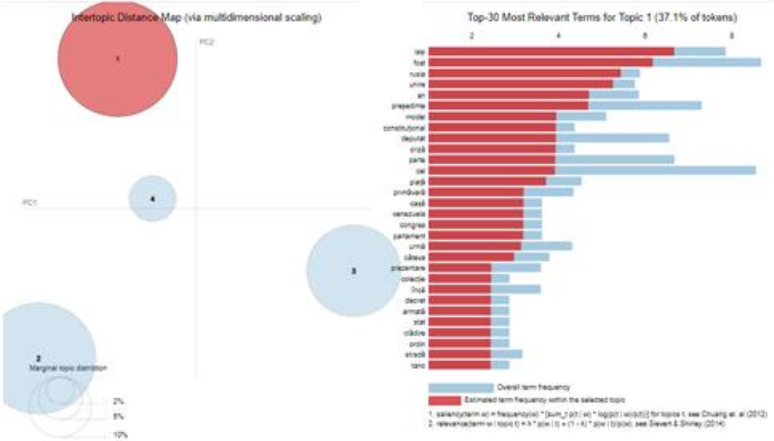
3.6. Text style identification

Text Style Identification – ReTeRom (2018)

- Important task - to control the expressivity in TTS
- Few research for Romanian (Univ. Bucharest, Romanian Academy).
- We have proposed 2 methods: LDA, CNN classifiers
- Dataset: CoRola
- Text styles: **(1)** Beletrictic: 251.927 words, **(2)** Judicial: 337.143, **(3)** Memorialistic: 315.213, **(4)** Political: 374.680, **(5)** Jurnalistic: 325.191



Text Style Identification. LDA (Latent Dirichlet Allocation) (*)

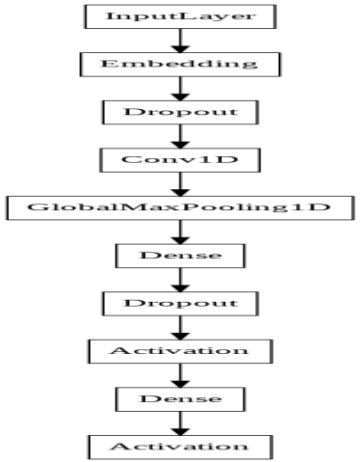


Drawback – the method gives poor result for short texts (sentences, or even paragraphs), so longer context is required to identify the style.

(*) *Nicoleta Raiu, "Text style identification for Romanian", Diploma project (2019)*

Text Style Identification. CNN approach(*)

- Inspired by several approaches that use CNN (Kim, 2014)
- (+) Even with a small training data set of 3.000 sentences, > **92%** accuracywork
- Future work: status for larger contexts, to explore the integration into a TTS

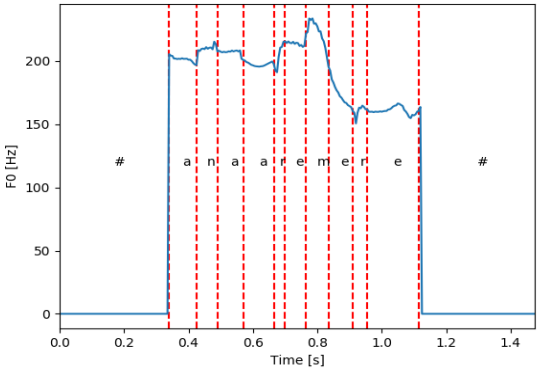


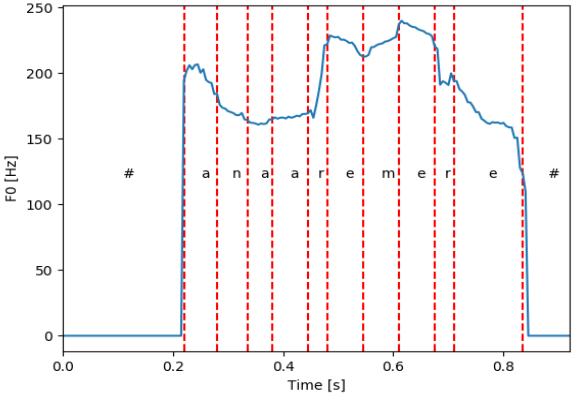
Nr.	Număr propoziții antrenare	Dimensiune conv.	Număr epoci	Acuratețe
1	5*3000	512	25	93.45%
2.			50	93.37%
3.		1024	25	92.77%
4			50	93.46%
5.	5*1000	512	25	91.83%
6.			50	91.81%
7.		1024	25	91.37%
8.			50	91.63%
9.	5*8000	512	25	92.69%
10.			50	92.41%
11.		1024	25	92.72%
12.			50	92.28%
13.	5*38000	512	25	90.10%
14.		1024	25	90.44%

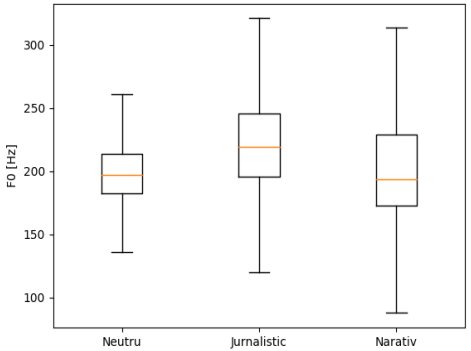
(*) *Adriana Stan, ReTeRom project (2019)*

Speech Styles analyses (*)

- Preliminary studies on: Neutral (SWARA), Journalistic (TV news), Narativ (audiobook) – phoneme duration, pitch

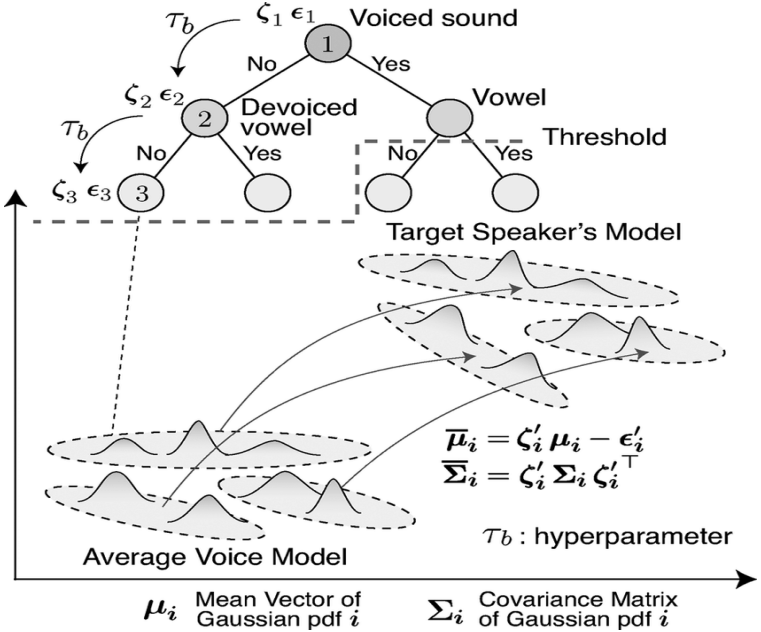






(*) *Adriana Stan, ReTeRom project (2018)*

Adaptation of a statistical TTS for new speaking styles using CSMAPLR (Constrained Maximum Aposteriory Linear Regression) (*)



(*) *J. Yamagishi, et al., "Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm", IEEE trans on Audio Speech and Language Processing, Feb. 2009.*

New speaking styles with CSMAPLR (2018, ReTeRom)

	Neutru	Jurnalistic	Narativ
Voce naturala			
Voce sintetizata			
Adaptare durata			
Adaptare F0			
Adaptare durata si F0			

https://speech.utcluj.ro/sintero/prosody_examples/

4. Acoustic modelling using DNN

4.1. End to end TTS and new voices using Tacotron GST

End to end TTS with Tacotron (*)

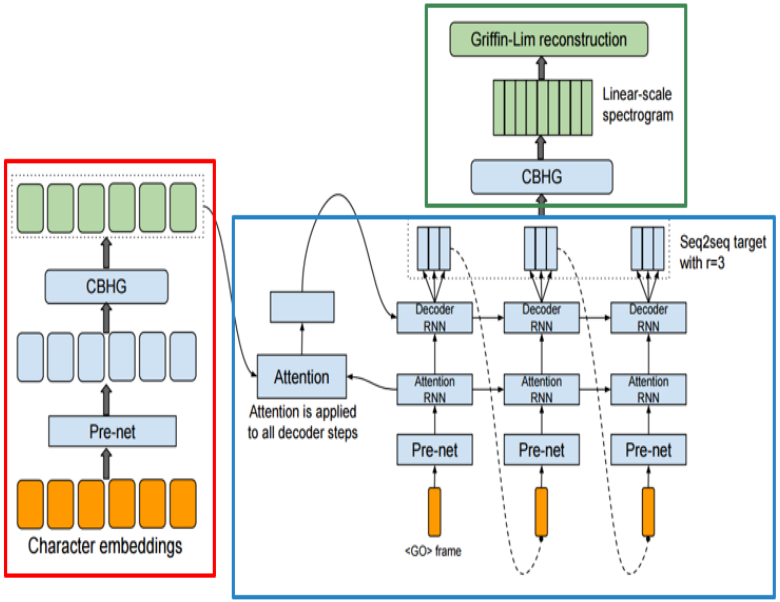
- Original Tacotron: only the pair <text, audio> is necessary to train the model

Truly end-to-end TTS pipeline by Google

- Reduces feature engineering
- Allows conditioning on various attributes

Architecture

- One network based on the sequence-to-sequence with attention paradigm
- Red Encoder
- Blue Decoder
- Green Post-processing net

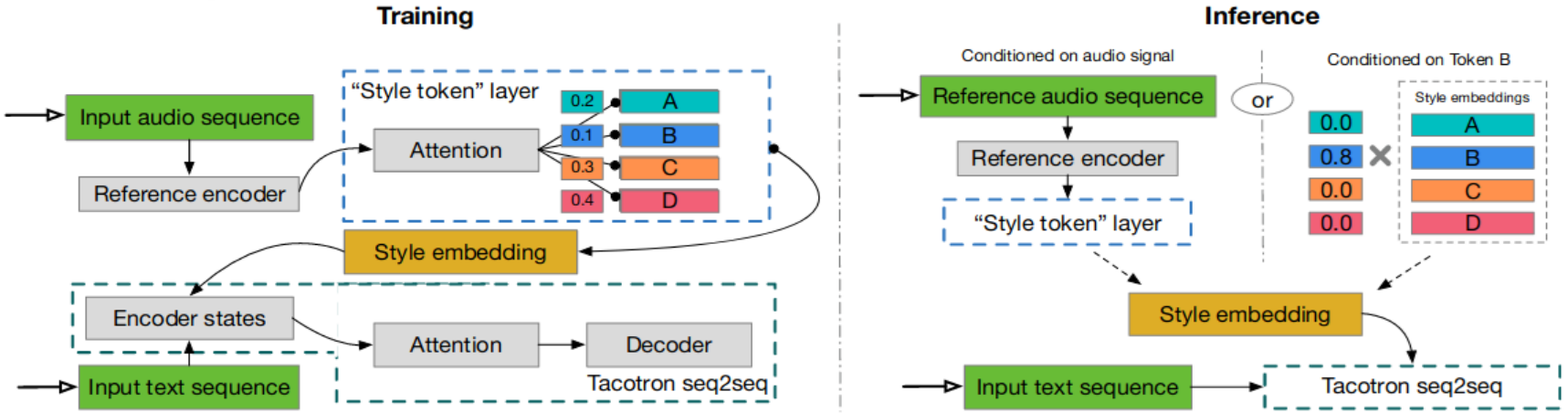


(*) *Yuxuan Wang, et al. "Tacotron: Towards End-to-End Speech Synthesis", Interspeech, 2017*

Tacotron GST (*). Result on using Mara Dataset (about 11hr speech)

- GST – a bank of embeddings jointly trained with Tacotron. In synthesis stage they can control the speed and speaking styles, independent of text. No explicit text labels, yet increased expressiveness.

- Natural No_GST GST1 GST2



(*) *Yuxuan Wang, "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis" (2018)*

Adaptation of TTS system to new speakers

(1) **Train Tacotron 2** with a very large dataset: **(a)** single speaker: Mara, ~11hr; **(b)** multiple speakers: Swara1, ~21hr. Result ➔ good modelling of speaker(s), hardly to obtain with small datasets

(2) **Adapt** the pretrained model(s) to new speaker(s) with a small dataset (10-50min)

Adaptation of TTS system to new speakers – sample results

https://speech.utcluj.ro/pf_lrec2020/

- Trained with Mara, adapted to Swara (10 speakers / 500 sentences) – includes vectorial representation of speakers

  SPK0  SPK5  SPK7 

- Trained with Mara, adapted to Swara (10 speakers / 500 sentences) – does not include vectorial representation of speakers

  Tok2  Tok3 

- Trained with Mara, adapted to Swara (10 speakers / 500 sentences) – freezing the GST layer

  FaraTok:  Tok3  Tok7 

Tacotron – based new voices

Text: *"Tezele ar putea fi eliminate în acest an școlar, iar Ministerul Educației analizează mai multe variante.“*

SPK_DOL: 

SPK_EME: 

SPK_NLL: 

Note: Only ortographic transcription and audio are used for training. No other NLP annotation is added. (https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/Sisteme-Tacotron2-voci-noi)

Speech prosody transfer using Global Style Tokens

- Tacotron GST pretrained expressive model (8.000 utt, Mara)
- Task1: Single speaker prosody transfer (500 utt / speaker)
- Task2: Multiple speakers (10) prosody transfer (500 utt / speaker x 10)
- Prosody is controlled: **(a)** reference utterance**, (b)** manual setting of GST weights
- Multispeaker → Mara, adaptation, Swara.

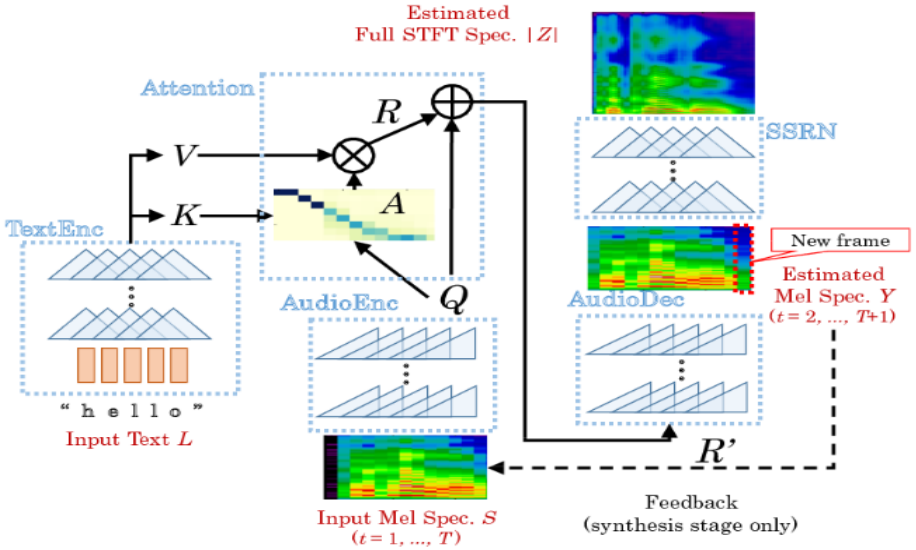
System	Natural		Syntetic				
	Mara	neutral	Eigen voice	Imtermediary	Different GSTs		
MARA		--		---			
MARA adapted to IPS							
MARA adapted to EME							

4. Acoustic modelling using DNN

4.2. Multispeaker TTS using DCTTS implementation

End to end TTS with DC-TTS, Deep Convolutional TTS. (*)

- An alternative to Tacotron (recurrent layers – time and computing power, non-parallelizable) → proposal of a CNN TTS (DC-TTS).



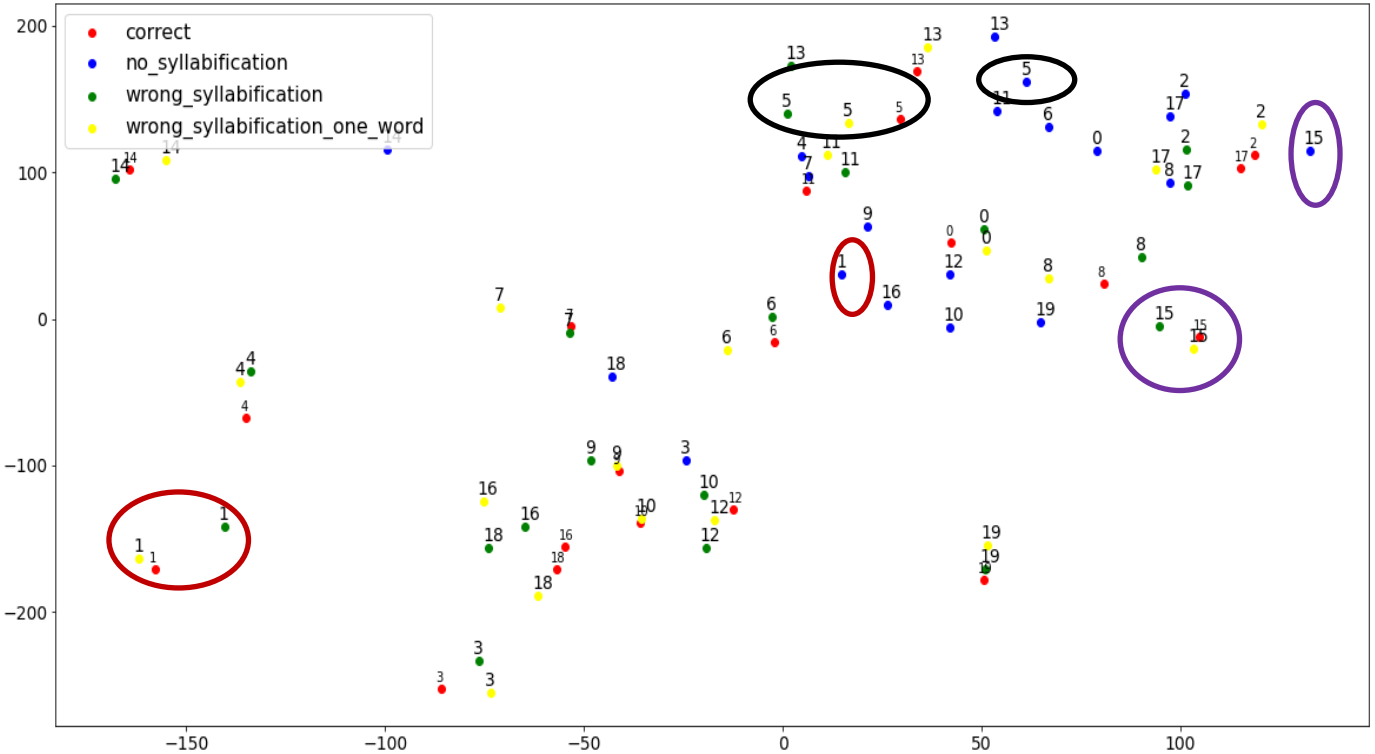
(*) *Hideyuki Tachibana,, "EFFICIENTLY TRAINABLE TEXT-TO-SPEECH SYSTEM BASED ON DEEP CONVOLUTIONAL NETWORKS WITH GUIDED ATTENTION", ICASSP 2018.*

End to end TTS with DCTTS. Experimental setting

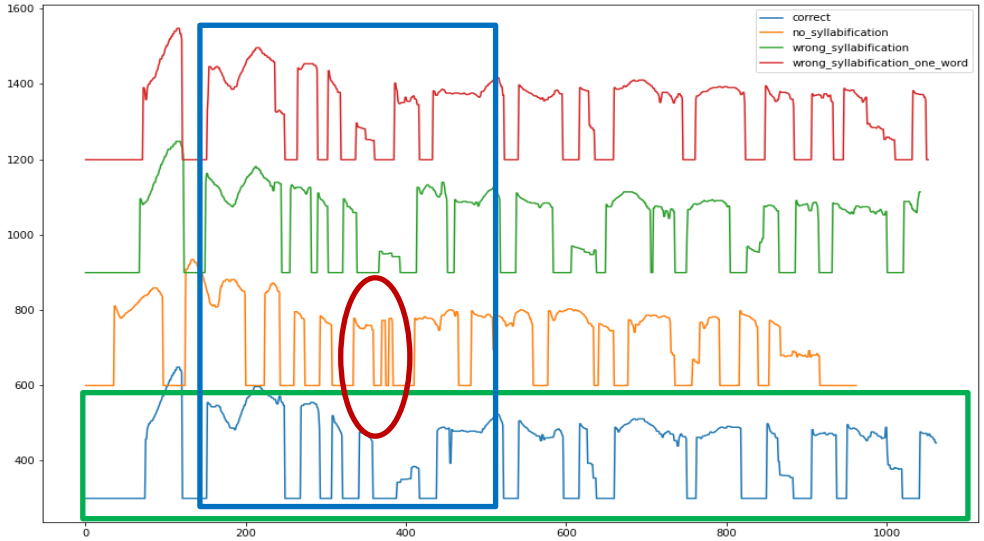
- **Audio** for training: Mara / Some speakers from Swara
- **Text** automatically annotated with existing tools (some errors may appear), in the following scenarios:

Only text	E însă altceva la mijloc.
Phonetic transcription	e <> a@ n s @ <> a l t ch e v a <> l a <> m i zh l o k <.>
Phonetic+syllabification	e <> a@ n * s @ <> a l t * ch e * v a <> l a <> m i zh *bl o k <.>
Phonetic+accent	e0 <> a@1 n s @0 <> a1 l t ch e0 v a0 <> l a0 <> m i0 zh l o0 k <.>
Phonetic+accent+sylab	e0 <> a@1 n * s @0 <> a1 l t * ch e0 * v a0 <> l a0 <> m i0 zh * l o0 k <.>

End to end TTS with DCTTS. Analysis of the text embeddings



End to end TTS with DCTTS. Analysis of F0 trajectories



- https://gitlab.utcluj.ro/speech/tts-samples/-/wikis/Sistem-DC-TTS-Date-lexicale

End to end TTS with DCTTS. Speech samples

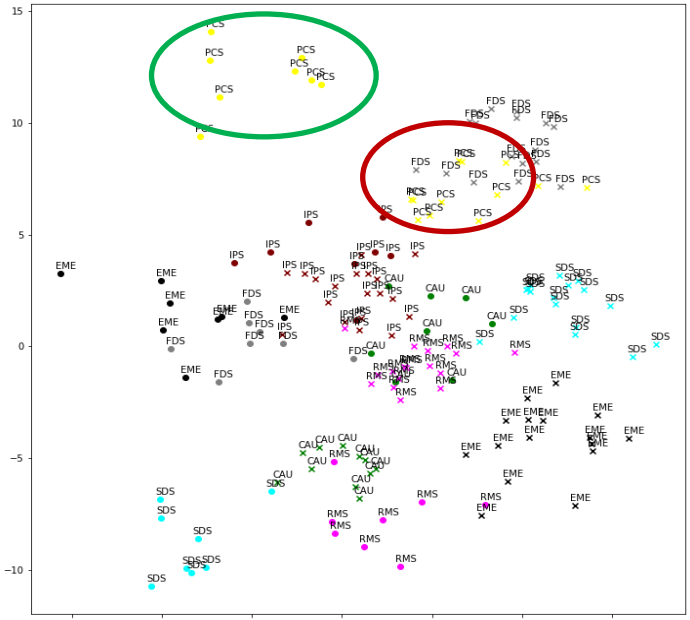
- **Reference speech:**
- **Text to be synthetized:**
- *“Ce se va întâmpla cu aceasta după ce vom ieși din starea de urgență”.*
- *Only ortographic transcription:*
- *Only phonetic transcription:*
- *Phonetic transcription + accent:*
- *Phonetic transcription + syllabification:*
- *Phonetic transcription + accent + syllabification*

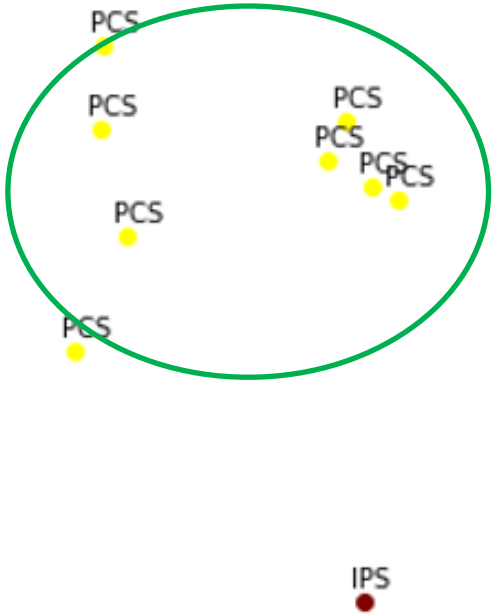
Speaker adaptation using DCTTS and speaker embeddings

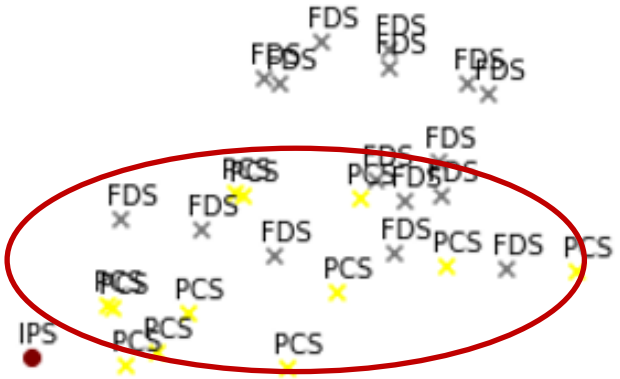
- **Speaker adaptation** is realized using a speaker dependent cost function (Cosine Similarity (CS), respective EER)
- Systems: Baseline (B), B + CS, B + EER

Sistem	ALL (18sp/21.302)	RND1 (18sp/8.900)	RND1-100 (18sp/1.787)	RND1-SAM (1sp/500)
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*

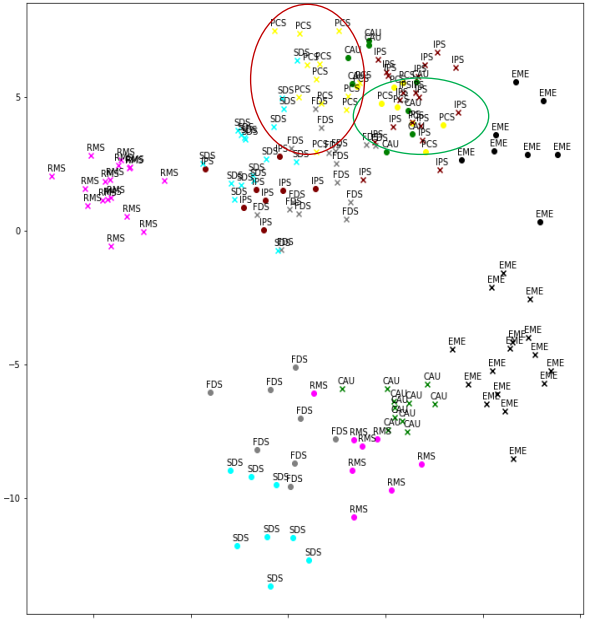
Speaker embeddings (t-SNE) natural vs TTS for the Baseline (B)

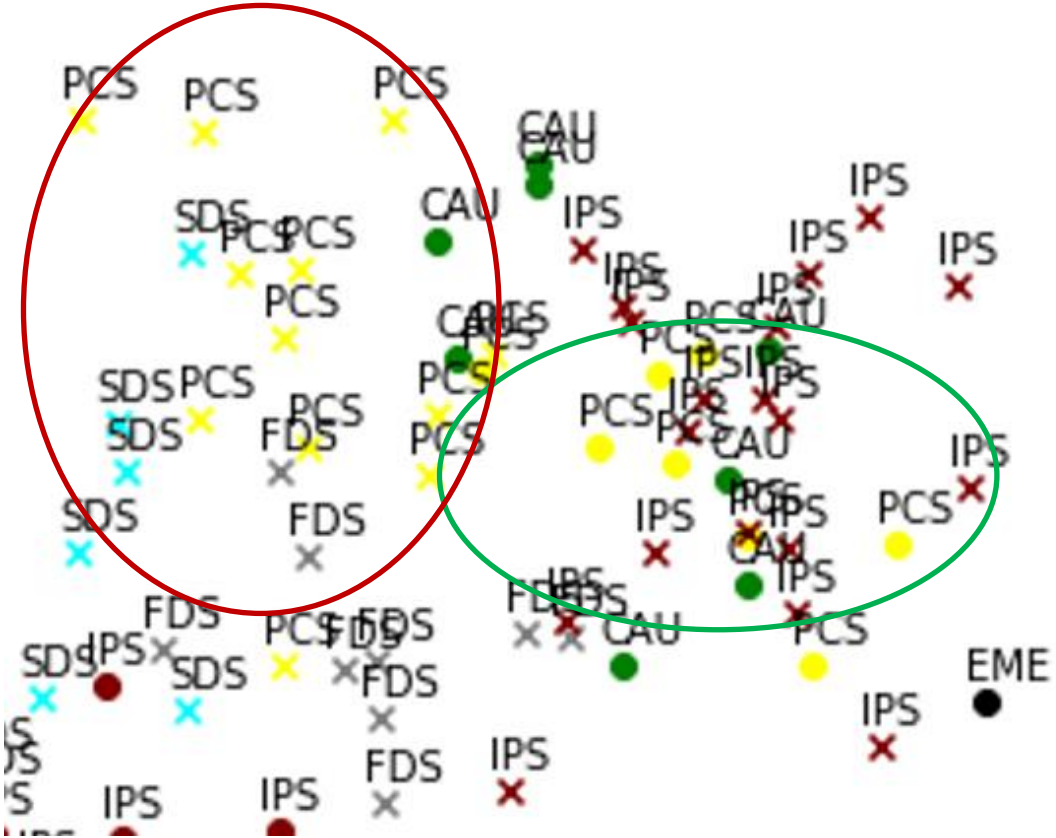






Speaker embeddings natural vs TTS for the B + EER





Speaker adaptation using DCTTS. Samples

Speaker DDM (Swara)

	ALL DATA (1500 utt/spk)	ONLY 100 utt / spk
Baseline		
Baseline + CS		
Baseline + EER	 *"La automatele de vândut cartele se poate achita cu numerar."*	 *"De aceea, spune medicul, autoritățile ar trebui să schimbe tonul discuțiilor."*

New synthetic voices generated from imperfect data

Imperfect data: (a) recorded speech does not match perfectly with the text, (b) pronunciation errors, (c) possible additional spoken artefacts, (d) noisy speech
System: Tacotron 2

	Natural speech	Synthetic
TV speaker		
Interview		

"Tezele ar putea fi eliminate în acest an școlar, iar Ministerul Educației analizează mai multe variante."

5. Conclusions and future work

Conclusions and future work

General aspects…

… discussions with the audience

In particular:

… discussions with the audience

… ?

Many thanks to the research TEAM

Sl.dr.ing. Adriana Stan
drd. Beata Lorincz
drd. Maria Nutu

Previous collaborators:
Andrei Barbos, Ioana Muresan, Andrei Homodi, Dalia Popescu, Cristian Contan, Jozef Domokos, Mihai Ordean.

THANK YOU

FOR YOUR ATTENTION !