Raport științific și tehnic
Etapa I-a, an 2018

„Metode de modelare și control a expresivității în sistemele de sinteză text-vorbire”

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

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:	**Raport științific și tehnic (Etapa I-a, 2018)**
Termen:	**Noiembrie 2018**
Editor:	**Mircea Giurgiu (Universitatea Tehnică din Cluj-Napoca)**
Adresa de eMail editor:	**Mircea.Giurgiu@com.utcluj.ro**
Autori, în ordine alfabetică:	**Mircea Giurgiu, Adriana Stan**
Ofițer de proiect:	**Cristian STROE**

Rezumat:

Acest document prezintă o sinteză a realizărilor de natură științifică și tehnică obținute in prima etapă de implementare a sub-proiectului SINTERO din cadrul proiectului PCCDI ReTeRom. Realizările se referă la:

- identificarea pattern-urilor prozodice și corelațiile între text și semnal vocal
- identificarea metodelor de clasificare automată a stilului de exprimare
- analiza metodelor de control și adaptare a expresivității în sistemele de sinteză
- implementarea modulului de control automat al prozodiei

Activitățile de cercetare desfășurate în prima etapă de implementare au condus la obținerea rezultatelor așteptate și ele sunt în concordanță cu obiectivele specifice ale etapei. Astfel, rezultatele raportate în acest document și descrise detaliat în cele 5 livrabile aferente perioadei de raportare, pregătesc cadrul etapei viitoare pentru implementarea componentelor de modelare a prozodiei și adaptare la noi vorbitori a vocilor sintetice. De asemenea, acest raport prezintă detalii referitoare la oferta de servicii de cercetare și tehnologice, activitățile de management și comunicare, modul de valorizare a resursei umane și dezvoltarea acesteia prin activități colaborative la nivelul consorțiului.

Cuprins

1. Activitățile etapei de raportare în contextul general al proiectului

În prima etapa a proiectului SINTERO (2018), etapă cu denumirea *„Metode de modelare și control a expresivității în sistemele de sinteză text-vorbire”*, s-a pornit de la resurse si module software deja existente la partenerii UTCN și ICIA și au fost desfășurate o serie de activități pentru: **a)** identificarea pattern-urilor prozodice și propunerea unei soluții pentru de modelare a prozodiei**, b)** identificarea metodelor de clasificare automata a stilului de exprimare și implementarea algoritmilor pentru reprezentarea vectorială a surselor de date text și audio, **c)** analiza a 3 metode de control și adaptare a expresivității vorbirii artificiale (concatenativ, statistic, neuronal), **d)** implementarea modulului de control automat al prozodiei pe baza unor noi corpusuri de date audio cu diferite stiluri de exprimare (de exemplu stil jurnalistic și stil narativ), cu controlul intonației frazei (3 pattern-uri: declarativ, exclamativ și interogativ), și cu demonstrarea online a rezultatelor, *https://speech.utcluj.ro/sintero/prosody_examples/*.

În etapele următoare ale proiectului SINTERO vom realiza *„Integrarea componentelor pentru modelare prozodie și adaptare la noi vorbitori a vocilor sintetice”* (2019) și în final *„Dezvoltarea unei noi tehnologii pentru sinteza text-vorbire cu expresivitate”* (2020).

2. Gradul de realizare a obiectivelor specifice pentru Etapa I-a

Ob. Pr4.1.15: *Identificarea pattern-urilor prozodice*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 1 set de înregistrări audio pentru evaluare pattern-uri prozodice
- măsurări cantitative și calitative a pattern-urilor accent, întonație, pauze și ritm în fraze declarative, exclamative sau interogative
- evaluarea frecvenței fundamentale, a formanților si duratei vocalelor, diftongilor si triftongilor în funcție de contextul prozodic
- identificarea a 7 pattern-uri prozodice și concluzii privind modul de variație a prozodiei în funcție de gradul de expresivitate a textului.
- un livrabil (D1.15) cu titlul *„Identificarea pattern-urilor prozodice și evidențierea corealțiilor între txt și semnal vocal”.*

Ob. Pr4.1.16: *Identificarea metodelor de clasificare automată a expresivității (text/audio)*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 3 metode candidat pentru reprezentarea vectorială a textelor
- 1 implementare și rezultate preliminare pentru clasificarea stilului din text
- lista cu parametrii acustici relevanți pentru clasificare stil vorbire
- 1 implementare si rezultate preliminare privind clasificarea stilului de vorbire (emotivitate) din datele audio
- 1 livrabil (D1.16) cu titlul *„Identificarea metodelor de clasificare automată a stilului de exprimare din surse de date text si audio”.*

Ob. Pr4.1.17: *Analiza metodelor de control și adaptare automată a expresivității*

Grad realizare: Obiectiv realizat integral

Rezultate:

- raport cu metodele de control a expresivității în sisteme concatenative
- raport cu metodele de control a expresivității în sisteme statistice HMM
- raport cu metodele de control a expresivității în sisteme DNN
- 1 livrabil (D1.17) cu titlul *„Analiza metodelor de control și adaptare automată a expresivității vorbitorilor în sistemele de sinteză text-vorbire”.*

Ob. Pr4.1.18: *Realizarea unui modul de control automat al prozodiei*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 1 metodă și interfață funcțională pentru controlul manual al prozodiei prin modificarea liniară a duratei și a intonației în propoziție
- 3 noi voci sintetice pentru prozodie cu expresivitate neutră, stil jurnalistic de prezentator de știri, stil narativ de tip audio book
- implementarea metodei de adaptare CSMALPR pentru adaptarea vocii neutre la 2 stiluri de vorbire (jurnalistic, narativ)
- 1 demonstrator online pentru controlul automat al prozodiei[1]
- 1 livrabil (D1.18.) *„Implementarea modulului de control automat al prozodiei”.*

Ob. Pr4.1.19: *Diseminarea rezultatelor intermediare*

Grad realizare: Obiectiv realizat integral

Rezultate:

- realizarea și actualizarea web site-ului proiectului[2]
- 1 pagină cu demonstrator online pentru control și adaptare prozodie
- 1 articol la conferința CONSILR 2018.
- 1 livrabil referitor la activitățile de diseminare (D1.19).

3. Rezultatele etapei și descrierea lor științifică și tehnică

3.1. Identificarea pattern-urilor prozodice și corelațiile între text și semnal vocal

Rezultatele raportate în aceasta secțiune corespund obiectivului Pr4.1.15, iar ele sunt descrise in extenso in livrabilul D1.15. Ca fundament pentru cercetările raportate în acest livrabil sunt rezultatele anterioare obținute de partenerii CO-ICIA (procesarea limbajului natural) și P1-UTCN (analiza unităților acustice din semnalul vocal), care pun în evidență principalii factori de natură lingvistică prin care se manifestă modificările prozodice în forma de undă: accentul, intonația în vorbire, silabificarea, pauzele, ritmul vorbirii, respectiv elemente de morfologie și sintaxă în interacțiune. Pornind de aici s-au ramificat două direcții de cercetare: identificarea modului de manifestare a prozodiei în parametrii semnalului vocal, respectiv corelația parametrilor prozodici cu caracteristici extrase din text.

În primul rând sunt prezentate rezultatele experimentale privind variația parametrilor prozodici frecvență fundamentală pentru vocale (Tabelul 4.1[3]), frecvența fundamentală în funcție de accent, frecvență fundamentală în funcție de intonația din propoziție, variația frecvenței formanților pentru diferiți vorbitori, respectiv rolul duratei și a pauzelor în modelarea pattern-urilor prozodice. Analiza s-a realizat pe un corpus de semnal vocal înregistrat în acest scop.

De exemplu, pentru unitățile acustice diftongi (Tabelul 4.2), pattern-urile prozodice indica faptul ca frecvențele fundamentale suferă variații atunci când diftongii (respectiv vocalele) sunt încadrați în cuvinte (Fig.4.1); F0 maxim scade atunci când avem grupuri de vocale încadrate împreună în cuvânt, iar energia acestor diftongi încadrați în cuvinte este sensibil mai mică decât cea a diftongilor, triftongilor izolați (Fig.4.2). Similar s-au obținut rezultate pentru diferite categorii de unități acustice. Un alt exemplu este pentru accent.

[1] *http://speech.utcluj.ro/sintero/prosody_examples*

[2] http://speech.utcluj.ro/sintero/

[3] *Tabelele și figurile sunt indexate sub forma 4.x, corespunzător numărului subproiectului SINTERO (# 4).*

Una din concluziile importante ale studiului se referă la o creștere a frecventei fundamentale pentru silabele (sau vocalele) accentuate, față de cele neaccentuate în medie cu 5%..20% (în 90% din cazuri creșterea s-a plasat în intervalul 9%..12%). Conform studiilor realizate pană acum s-a arătat ca în general silabele accentuate au tendința de a avea frecventa fundamentala, durata și amplitudine mai ridicata decât silabele neaccentuate (Tabel 4.3). Însă, exista și cazuri în care numai unul sau doi din acești parametri este mai ridicat, precum și situații în care tendința silabelor accentuate este de a-si reduce fundamentala sau ceilalți parametrii.Merită făcută și observația ca au existat și câteva cazuri în care accentuarea unei silabe nu a adus nici un fel de diferențiere din punctul de vedere al valorii F0. Similar sunt prezentate rezultate pentru formanți (Tabel 4.4, 4.5, 4.6), respectiv evaluarea duratei unităților acustice în funcție de accent (Tabel 4.6).

Tabelul 4.1. Variația F0 pentru vocalele unui vorbitor feminin

Parametru	Vocala						
	/ a /	/ e /	/ i /	/ o /	/ u /	/ ă /	/ î /
F_0 minim [Hz]	*161*	*175*	*207*	*103*	*100*	*203*	*220*
F_0 mediu [Hz]	*186*	*191*	*214*	*197*	*135*	*224*	*276*
F_0 maxim [Hz]	*261*	*217*	*229*	*232*	*235*	*273*	*289*
Intensitate [dB]	*73*	*60*	*68*	*72*	*80*	*80*	*67*

Tabelul 4.2. Variația F_0 pentru diftongi / triftongi în pronunție izolată, respectiv în context

Parametru	Diftong / Triftong						
	/au/	/ae/	/ai/	/oi/	/ie/	/aie/	/iau/
F_0 minim [Hz] - izolat	*129*	*157*	*168*	*93*	*128*	*168*	*203*
F_0 minim [Hz] - context	*176*	*174*	*167*	*170*	*137*	*163*	*175*
F_0 maxim [Hz] - izolat	*237*	*272*	*228*	*202*	*226*	*236*	*217*
F_0 maxim [Hz] – context	*217*	*227*	*229*	*222*	*202*	*258*	*243*
Intensitate prima vocală	*73*	*60*	*73*	*72*	*68*	*65*	*59*
Int a doua vocală [dB]	*80*	*73*	*68*	*68*	*60*	*67*	*67*
Intensitate context [dB]	*55*	*59*	*59*	*63*	*53*	*52*	*53*

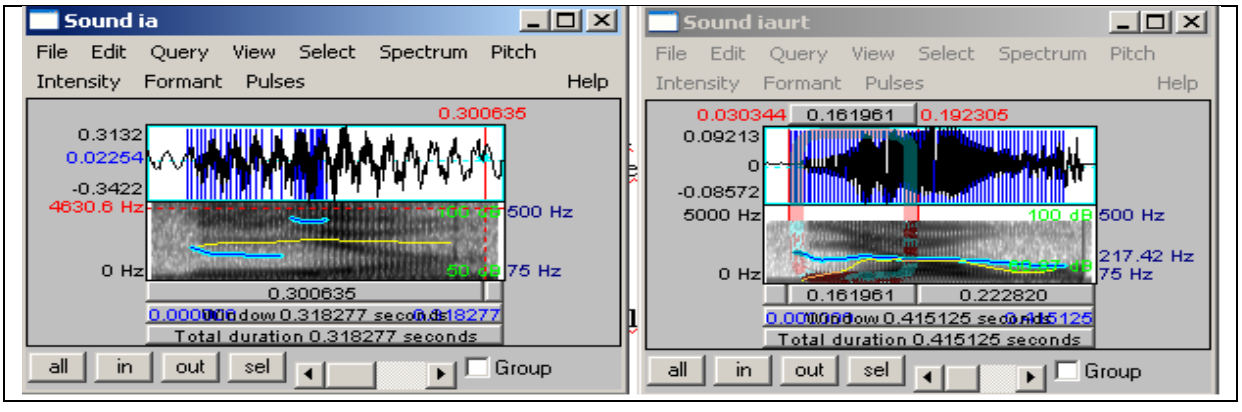

Fig.4.1. Diftongul /au/ izolat, respectiv în context

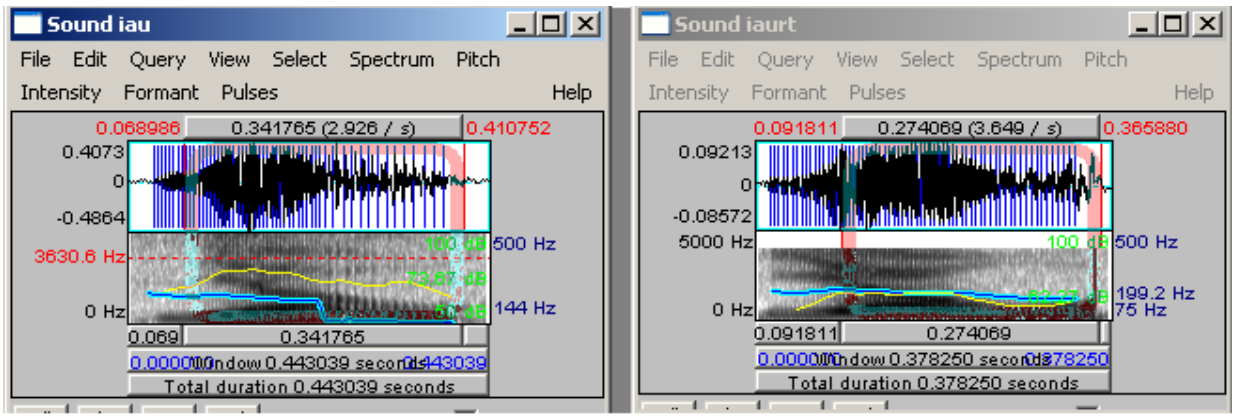

Fig. 4.2. Variația parametrilor prozodici pentru triftongul /iau/

Tabelul 4.3. Tabel sintetic privind variația parametrilor prozodici în funcție de accent

Cuvânt	Silabe purtătoare de accent	Valoarea de referință a F0 (neaccentuat) [Hz]	Valoarea lui F0 pe silaba accentuata [Hz]	Valoarea de referință intensitate[dB] (neaccentuat)	Valoarea intensității [dB] pe silaba accentuată
Fac**tu'**ra	tu	*172*	241	*55*	57
Factu**ra**'	ra	*195*	231	*54*	60
Hai**'**na	i	*193*	205	*54*	53
Lu'mina	lu	*199*	245	*57*	59
Lu**mi'**na	mi	*177*	257	*55*	59
Ve**se**'la	se	*183*	201	*57*	59

Tabelul 4.4. Pattern-uri pentru formanții vocalelor pentru vorbitorii feminini din corpus

Vocala	F1 [Hz]			F2 [Hz]			F3 [Hz]		
	minim	mediu	maxim	minim	mediu	maxim	minim	mediu	maxim
/a/	*156*	*827*	*2182*	*1257*	*1792*	*3135*	*1825*	*2487*	*4357*
/e/	*81*	*867*	*2247*	*838*	*1706*	*3032*	*1928*	*2853*	*4208*
/i/	*251*	*634*	*2158*	*653*	*2392*	*3227*	*2388*	*3259*	*4030*
/o/	*178*	*966*	*1937*	*940*	*1518*	*3328*	*1407*	*2981*	*3971*
/u/	*169*	*647*	*1942*	*635*	*1245*	*3293*	*1360*	*2986*	*4107*

Tabelul 4.5. Pattern-uri pentru formanții vocalelor pentru vorbitorii masculini din corpus

Vocala	F1 [Hz]			F2 [Hz]			F3 [Hz]		
	minim	mediu	maxim	minim	mediu	maxim	minim	mediu	maxim
/a/	*92*	*837*	*1979*	*676*	*1735*	*3043*	*1621*	*2523*	*4106*
/e/	*87*	*791*	*2366*	*441*	*1951*	*3217*	*1914*	*1951*	*3217*
/i/	*92*	*634*	*2158*	*653*	*2392*	*3227*	*2388*	*3259*	*4030*
/o/	*84*	*761*	*1718*	*526*	*1321*	*2924*	*1028*	*2408*	*2845*
/u/	*83*	*574*	*1738*	*572*	*1227*	*2916*	*1021*	*2420*	*3980*

Tabelul 4.6 Legătura între durata cuvintelor / difonemelor și intonație (frecvența fundamentală)

Unitate acustică	*Mama.*		*Mama!*		*Mama?*	
	Fo (Hz)	*durata(ms)*	*Fo (Hz)*	*durata(ms)*	*Fo (Hz)*	*durata(ms)*
_m	*221*	*53*	*245*	*100*	*208*	*81*
ma	*222*	*119*	*260*	*250*	*180*	*200*
am	*225*	*113*	*255*	*235*	*174*	*215*
ma	*229*	*109*	*189*	*177*	*274*	*210*
*a_*	*229*	*85*	*185*	*124*	*205*	*180*

În al doilea rând sunt prezentate rezultate privind analiza caracteristicilor de natura lingvistică ce afectează prozodia, în special la nivel de intonație de propoziție. Sunt identificate un set de 7 pattern-uri intonaționale la nivel de propoziție, dar și efectul prozodic al semnelor de punctuație (vezi Livrabil D1.15).

Cercetările demonstrează faptul că pattern-urile prozodice manifestate la nivelul semnalului vocal au legătură directă și prezintă strânse corelații pe termen scurt sau pe termen lung cu atribute de morfologie și sintaxă aferente textului. Principalele atribute se referă la poziționare accent în cuvinte, silabificare, părți de vorbire, sintaxă, respectiv punctuație. Aceste rezultate prezintă fundamentul pentru dezvoltarea unor noi metode de sinteză expresivă a vorbirii prin intermediul unor module de analiză a expresivității textului (în componenta software de procesare de text), respectiv de modificare automată a prozodiei (în componenta software de sinteză de semnal).

3.2. Identificarea metodelor de clasificare automată a stilului de exprimare

Rezultatele raportate în această secțiune corespund obiectivului Pr.4.1.16, iar ele sunt descrise în extenso în livrabilul D1.16. Acest livrabil prezintă atât rezultate de natură teoretică ce au în vedere identificarea unor soluții de clasificare automată a stilului de exprimare din surse de date text și audio, precum și implementarea modulelor software aferente. Identificarea si clasificarea stilului de exprimare din text este necesară în modulul de procesare a textului din cadrul unui sistem de sinteză text – vorbire cu scopul de a informa generatorul de semnal vocal despre expresivitatea pe care trebuie să o incorpoze la sinteză. Aceasta expresivitate este determinată de conținutul semantic al textului și de polaritatea (ca sentiment +/-) acestuia.

În primă etapă au fost identificate cîteva metode de reprezentare vectorială a textelor. Acestea se referă la reprezentări de tipul Bag of Words, VSM (Vector Space Models) și LSA (Latent Semantic Analysis). Din punct de vedere practic s-au implementat și apoi testat experimental fluxurile de procesări care realizează reprezentările amintite și prin care s-a verificat posibilitatea de clasificare a mai multor stiluri de vorbire similar identificarii automate a topicurilor din discursul de tip text (Fig.4.5). Rezultatele preliminare s-au obținut pe un corpus redus (Fig. 4.3, 4.4), dar avem in vedere utilizarea corpusurilor (belestristic, științific, jurnalistic, narativ) obținute de la Partenerul ICIA.

Document ID	Text
0	*Cu două degete muiate în apă poți să stingi o lumânare. Închizând pleoapele, poți stinge o rază de soare. Dar noapte nu se face.*
1	*Căci noapte nici nu poate fi. Nici noaptea pamântului, noaptea cea mare, nu e noapte, ci doar o umbră într-un univers de lumină.*
2	*Ușor nu e nici cântecul. Zi si noapte nimic nu e ușor pe pamânt; căci roua este sudoarea privighetorilor ce s-au ostenit toată noaptea cântând.*
3	*Și dacă se întâmplă pe tine să te văz, Desigur că la noapte un tei am să visez.*
4	*Și dacă se întâmplă să întâlnesc un tei, Desigur toată noaptea visez la ochii tăi.*
5	*Vremea va fi în general închisă și se va răci iar noaptea va fi geroasă. Cerul va fi mai mult noros.*
6	*Cerul va fi noros și va ninge pe arii extinse. Vântul va sufla slab la moderat.*
7	*Vremea se menține închisa. Cerul va fi mai mult noros și va ninge pe arii relativ extinse în cursul zilei; noaptea va mai ninge la munte.*

Fig 4.3. Stil beletristic (doc 0-4), stil meteo (doc 5-7)

Word ID	Word
0	tei
1	arii
2	visez
3	întâmplă
4	noaptea
5	cerul
6	noapte
7	desigur
8	vremea
9	ninge
10	extinse
11	noros

Fig. 4. 4. Cuvintele din Bag of Words

```
0.865*"noapte" + 0.402*"noaptea" + 0.132*" întâmplă  " + 0.132*"desigur" + 0.132*"visez"
-0.510*"ninge" + -0.399*"noros" + -0.399*"cerul" + -0.358*"extinse" + -0.358*"arii"
```

Fig.4.5. Cele mai semnificative cuvinte care definesc cele 2 stiluri de exprimare

Similar metodelor de clasificare a textelor s-au identificat paramterii acustici care ar fi relevanți în clasificarea stilului de vorbire numai din date audio (Tabel 4.7). Tonul din voce, aparte de mesajul lingvistic, este un bun indicator. Ca exemplu, ilustrăm modul de variație a 2 dintre acești parametri (frecvența fundamentală (Fig.4.6), respectiv parametrul LSF1) pentru 2 voci cu emotivități diferite (Fig.4.7). Prin urmare, acești parametri au un potențial înalt de discriminare între diferitele stiluri de vorbire.

Cele mai frecvente metode de clasificare aplicate pentru recunoașterea stilului de vorbire și a expresivitatății (inclusiv pentru recunoașterea emoțiilor) sunt arborii de decizie, clasificatorii SVM sau rețelele neuronale. În aplicația prototip s-a utilizat un corpus cu 5 stiluri de expresivitate, corespunzând la 5 clase de emoții. În total s-au folosit 500 de fișiere audio pentru fiecare emoție, în total un set de 2500 de fișiere. Întreg setul a fost împărțit în două, un set pentru antrenare și unul pentru testare.

Tabel 4.7 Identificarea parametrilor acustici relevanți pentru clasificarea expresivității vocale

Parametri	Utilizare
Parametri spectrali pe termen lung	*Media spectrului, spectral flatness measure, centroidul spectral*
Parametri spectrali pe termen scurt	*MFCC (Mel frequency Cepstral Coefficients), LSF (Line Spectrum Frequency), LPC-PLP (Linear Predictive Coefficients – Perceptual Linear Prediction)*
Pitch	*Media, deviația standard, skewness, kurtosis, maximum, minimum, quartiles, diferențe între quartile, coeficienții de regresie liniară și quadratică*
Rata vorbirii	*Media si deviația standrad pentru durata silabelor, raportul dintre durata segmentelor sonore și nesonore*
Parametri in timp	*Intensitatea, RMS, numarul de treceri prin zero, TEO (Teager Energy Operator)*
Parametri tonali	*Coeficientii CHROMA, CENS*
Calitatea vocii	*HNR (harmonic to Noise Ratio), Jitter, Schimmer*

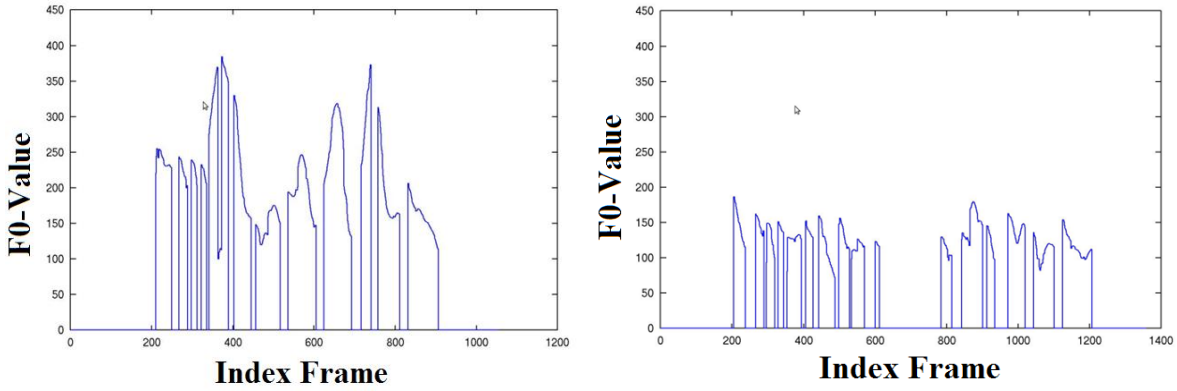

Fig. 4.6 Variația F0 pentru starea fericit (stânga), respectiv trist.

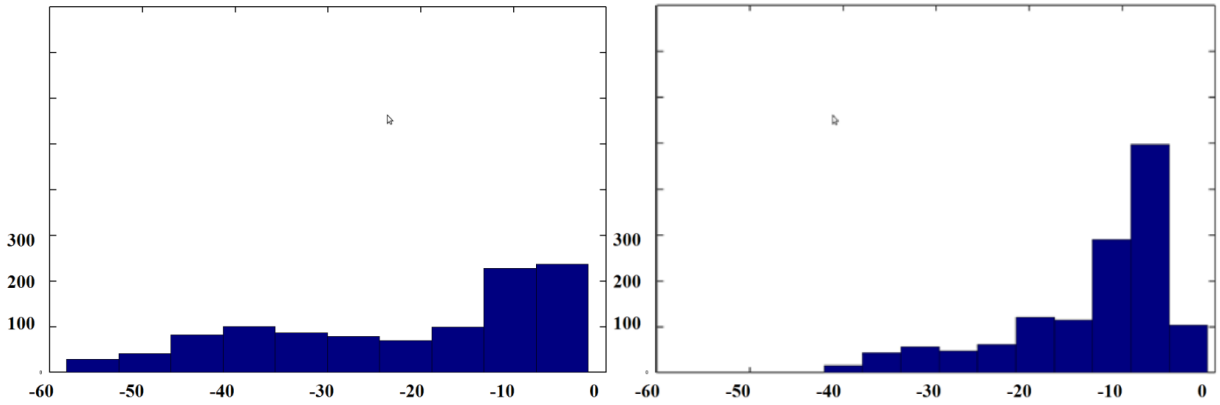

Fig. 4.7. Histograma parametrului LSF1 pentru starea fericit (stânga), respectiv trist

Parametrii acustici au fost extrași cu aplicația GlottHMM și printr-o procedura de selecție a parametrilor bazată pe "information gain", s-au generat vectorii specifici fiecărui stil. Rezultate se prezintă pentru setul de parametri (F0. NAQ, LSF1, LSF2, LSF3, LSF4, HNR1, HNR2, HNR3, HNR4, HNR5) pentru care s-au inclus în vector media și deviația standard.

Prezentam doar rezultatele globale de clasificare obținute prin 3 metode standard,

J48-arbori de decizie	*83,67%*
Logistic Model Tree	*95,40%*
MLP	*97,95%*

Pe baza acestei metodologii, în următoarea etapă vom considera colectarea unui set de date audio si text relevante pentru aplicația finală, iar pe baza acestora vom desfasura experimente elaborate pentru testare în condiții mult mai complexe (volum date, vectori mari).

3.3. Analiza metodelor de control și adaptare a expresivității în sistemele de sinteză text-vorbire

Rezultatele raportate în această secțiune corespund obiectivului Pr.4.1.17. Problema variabilității și expresivității vocilor sintetice este de mare actualitate (vezi Livrabil D1.17), în special prin prisma faptului că expresivitatea și/sau prozodia nu pot fi evaluate în mod obiectiv printr-un set de parametri și de cele mai multe ori depind de starea emoțională a persoanei care o evaluează, precum și de fondul cultural, etnic sau educațional. Diferitele tipuri de sisteme de sinteză: concatenative, parametric-statistice sau cele ce modelează direct forma de undă, permit un control al expresivității și prozodiei specific, în funcție de arhitectura sistemului.

Problema sistemelor concatenative este faptul că informația audio nu este parametrizată sub nicio formă, astfel că pentru controlul expresivității este necesară manipularea formei de undă. La modul cel mai simplu, controlul expresivității este făcut prin selectarea segmentelor audio de concatenat pe baza unei traiectorii prozodice predefinite sau estimate din text, prin intermediul unei funcții de selecție a unităților. Totuși, înregistrarea aceluiași vorbitor în condiții de expresivitate sau emotivitate variabile este greu de realizat. Modificarea formei de undă se face prin metoda PSOLA (Pitch Synchronous Overlap and Add), doar că acest tip de modificare introduce artefacte ne-naturale în vocea sintetică. Deși au fost dezvoltate și implementate multiple metode de control a emoțiilor și a expresivității în sinteza concatenativă, faptul că această tehnologie se bazează pe forma de undă în sine, cu anumite modificări parametrice ale rezultatului vocal, face ca orice modificare adusă semnalului să introducă erori de sinteză nedorite.

În sistemele de sinteză bazate pe modele Markov fonemele sunt modelate printr-un anumit set de parametri (de exemplu coeficienți Mel-cepstrali, coeficienți de aperiodicitate, frecvența fundamentală, F0, și durata), iar pentru controlul expresivității se pot astfel adapta în mod independent modelele acestor parametri. Problema este că naturalețea vocii sintetizate este condiționată de utilizarea unor înregistrări audio cu prozodie cât mai variată, pentru ca modelele statistice să poată utiliza un număr cât mai mare de exemple fonetice pentru același context. Principalele modalități de adaptare:

- adaptarea prozodică a informției textuale prin etichete prozodiec ce tip ToBI, utilizarea varianței globale a parametrilor, utilizarea unor adnotări la nivel suprasegmental.
- controlul modelelor acustice prin date care conțin starea de emoție sau expresivitate specifică vorbitorului, utilizarea caracteristicilor la nivel suprasegmental si aplicarea Multipe Regression HMM, adnotări la nivel articulator.
- adaptarea modelelor acustice pornind de la un set foarte mare de date de la foarte multi vorbitori, crearea unei voci eigen, iar apoi adaptarea modelelor către vocea sintetica printr-un set redus de date și aplicarea unor metode de factorizare a modelelor Markov.

În sistemele de sinteză bazate pe rețele neuronale abordările pentru adaptarea expresivității includ extinderea setului de caracteristici de intrare cu un set de caracteristici de prozodie sau de stil de vorbire.

Problema majoră a acestor sisteme este necesitatea existenței unui corpus de voce de dimensiuni imense (sute de ore de vorbire) pe baza căruia să se realizeze antrenarea rețelei. Învățarea prin transfer poate fi o soluție pentru a compensa indisponibilitatea unui astfel de corpus. Chiar și cu resurse de date disponibile, modul de adnotare a caracteristicilor de expresivitate (de exemplu emoții) este esențial. Adesea se introduc codificări suplimentare cu creșterea imensionalității datelor de intrare.

Un exemplu de sistem comercial (Tacotron) este cel de la Google. Foarte recent acest sistem a fost extins cu tehnologia Global Style Tokens pentru a învăța automat expresivitatea latentă în semnalul de intrare. Pe de altă parte, sistemul EMPHASIS de la Baidu modelează dependențele lingvistic – acustic printr-o rețea de regresie. Ambele sisteme comerciale generează voce sintetică de calitate, dar generarea automată a etichetelor de expresivitate și prozodice din text, precum și transferul stilului de vorbire, rămân probleme deschise.

3.4. Implementarea modulului de control automat al prozodiei

Rezultatele raportate în această secțiune corespund obiectivului Pr.4.1.18. Modulul software pentru controlul prozodiei (vezi Livrabil D1.18) permite două moduri de operare. În modul manual, prin intermediul unei interfețe grafice intuitive accesibile chiar si pentru utilizatori non-experți (Fig.4.8), este posibilă modificarea de către utilizator a conturului frecvenței fundamentale (F0) și a duratei segmentelor sonore de vorbire, dar cu alinierea automată a datelor audio. Fișierul de configurare păstreaza valorile medii ale parametrilor.

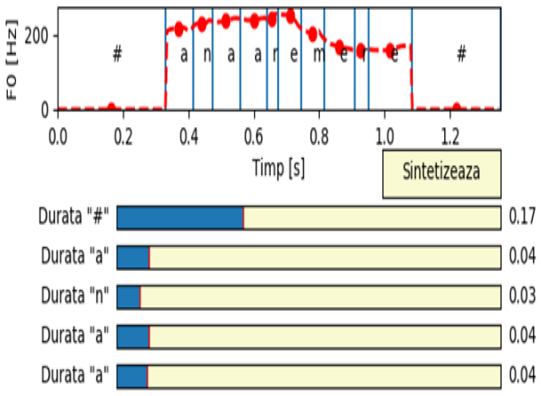

Fig.4.8. Interfața pentru modificarea manuală a parametrilor F0 și durată foneme

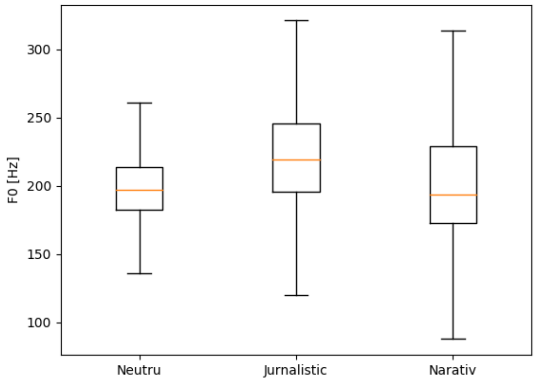

Fig.4.9 Mediana, prima și treia cuartilă pentru F0 la stilul neutru, jurnalistic și narativ

În modul automat de funcționare a modulului, controlul prozodiei se bazează pe 3 seturi de date audio înregistrate în studio, cu stiluri de exprimare diferite (neutru -1h și 43 de minute, jurnalistic – 48 de minute, narativ – 11 ore din audio book). Din Fig 4.9 se observă ca valorile medii pentru F0 sunt apropiate (200-220Hz), dar au deviații standard destul de diferite, în relație cu stilul de vorbire. De asemenea, din Fig.4.10 se observă duratele diferite ale fonemelor în cele 3 stiluri de vorbire.

De exemplu, în stilul jurnalistic durata fonemelor este mai mică. Acest corpus, deși are o durată mai mică și este aliniat doar la nivel de propoziție, este suficient de bogat în informație prozodică, datorită expresivității verbale a prezentatoarei de știri. În primă fază au fost implementate 3 voci sintetice pentru aceste 3 corpusuri (modelare HMM, 5 stări, vocoder WORLD) pentru a putea compara ulterior vocea sintetică, adaptată la aceste stiluri, cu vocea sintetică originală. S-a implementat metoda de adaptare automată a prozodiei bazată pe algoritmul Constrained Structural Maximum Aposteriori Linear Regression - CSMAPLR). Noul mod de variație a F0 și duratei pentru vocea sintetică adaptată este ilustrat în Fig 4.11.

Versiunile următoare ale acestui sistem vor avea în vedere detecția automată a stilului de exprimare pornind de la textul de intrare, precum și modalități de adaptare a stilului folosind un set redus de date audio.

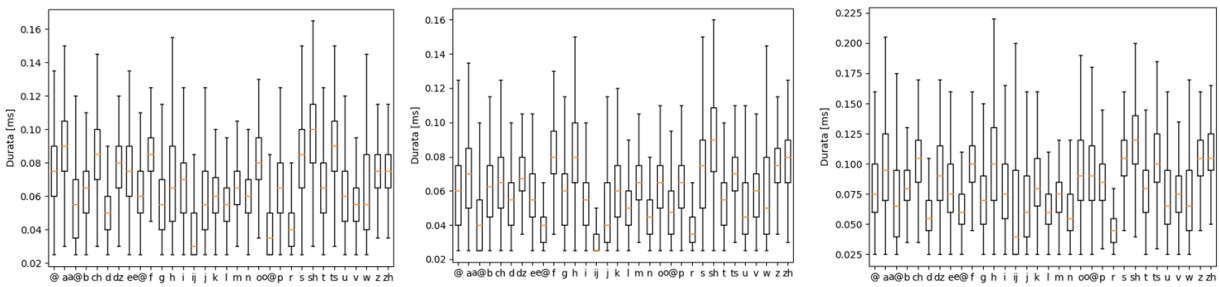

Fig.4.10. Mediana, prima și a treia cuartilă pentru durata fonemelor (neutru, jurnalistic, narativ)

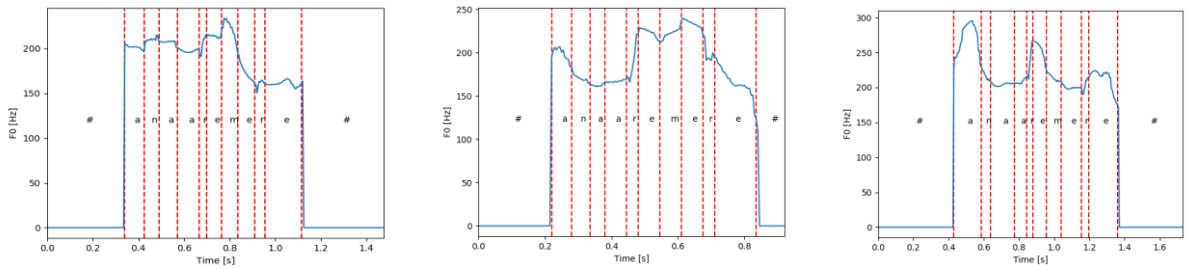

Figura 4.11. Ilustrarea controlului automat a frecvenței fundamentale (F0) și duratei fonemelor pentru vocea sintetizată cu stil neutru, jurnalistic, narativ(de la stânga la dreapta)

4. Oferta de servicii de cercetare, locuri de muncă susținute și valorificarea resurselor

Tabel 4.8. Sinteză privind oferta de servicii, locuri de muncă și valorificarea resurselor în UTCN

Oferta de servicii în UTCN	● oferta unei tehnologii de sinteză text-vorbire în limba română ● servicii de adnotare automată a resurselor de date audio ● servicii de înregistrare audio de înaltă fidelitate ● servicii software pentru dezvoltarea modelelor bazate pe învățare automată. *ERRIS: https://erris.gov.ro/speech.utcluj.ro*
Locuri de muncă susținute în UTCN	1 x CS I, 1 x CS II, 1 x CS III, 1 x Tehnician 2 x ACS pentru noii angajați
Resursa umană nou angajată în UTCN	În iunie 2018 au fost demarate procedurile în UTCN pentru scoaterea la concurs a 2 noi posturi de ACS. Anunțul a fost publicat în 12.09.2018, iar concursul a fost planificat pentru 26.09.2018 (cf anunt România Libera, Monitorul Oficial, site ANCS/Euraxes, site UTCN). Nu s-a prezentat nici un doctorand,asa cum s-a solicitat în anunț. Ulterior s-au făcut demersuri, cf Legii 319/2003 pentru angajarea pe aceste 2 posturi a 2 masteranzi, doar ca UTCN dorește ca ocuparea postului de ACS să fie făcută de un doctorand. În aceste condiții se cauta candidati cu profil de doctorand.
Valorificare resurse în parteneriat	● UTCN a preluat de la ICIA resurse de date text (4 corpusuri) pentru clasificarea stilurilor de exprimare ● UTCN a furnizat pentru ICIA si UAIC corpusurile de date audio disponibile și adnotările acestora ● UAIC a furnizat pentru UTCN o metodă de clasificare a textului dezvoltată în limbajul R.
Cecuri	● UTCN a oferit un cec pentru înregistrare corpusuri audio, dar încă nu a fost folosit de parteneri.

5. Management si comunicare

Activitățile de management au fost orientate în special către managementul proiectului complex în vederea integrării diferitelor grupuri de cercetare și a resurselor tehnice ale acestora. S-au organizat mai multe conferinte Skype și o reuniune a parteneriatului in Mai la UPB. Este de notat faptul ca s-a asigurat o bună comunicare și coordonare și pentru realizarea planului de achiziții global, respectiv pentru documentația de raportare etapă. Din punct de vedere administrativ s-au primit 4 tranșe de avans cu o regularitate adecvată. Nu toate resursele financiare alocate UTCN au fost folosite integral.

6. Diseminarea rezultatelor

O preocupare a Consorțiului în etapa de raportare a fost implementarea și îndeplinirea cu succes a obiectivelor stabilite în strategia de diseminare a rezultatelor elaborată în cadrul propunerii de proiect. Astfel, adecvat acestei etape inițiale s-a acționat pe următoarele direcții: a) crearea paginii web a proiectului SINTERO *(http://speech.utcluj.ro/sintero/),* b) publicarea conform planului a unui articol la conferința CONSILR 2018 (vezi mai jos), c) crearea unei pagini web dedicate pentru demonstrarea online a modulului de control a prozodiei (*https://speech.utcluj.ro/sintero/prosody_examples/*).

[1] A. Stan, M.Giurgiu , „A comparison between traditional machine learning approaches and deep neural networks for text processing in Romanian", In Proc. of The The 13th International Conference on Linguistic Resources and Tools for Processing Romanian Language, Iasi, 22-23 November 2018.

7. Concluzii

Activitățile de cercetare desfășurate în etapa I-a de implementare a proiectului (2018) au condus la obținerea rezultatelor asteptate și ele sunt în concordanță cu obiectivele specifice ale etapei. Astfel, rezultatele raportate în acest document și descrise detaliat în cele 5 livrabile aferente perioadei de raportare (vezi Secțiunea 8 a acestui raport), pregătesc cadrul pentru etapa a doua.

8. Referințe la livrabilele aferente etapei 2018 (Anexe la raport)

[1] Livrabil D1.15:	*„Identificarea pattern-urilor prozodice si evidențierea corelațiilor între text și semnal vocal",* Mai 2018.
[2] Livrabil D1.16:	*„Identificarea metodelor de clasificare automată a stilului de exprimare din surse de date text și audio",* Mai 2018.
[3]Livrabil D1.17	*„Analiza metodelor de control și adaptare automată a expresivității vorbitorilor în sistemele de sinteză text-vorbire",* Noiembrie 2018.
[4] Livrabil D1.18:	*„Implementarea modulului de control automat al prozodiei",* Noiembrie 2018.
[5] Livrabil D1.19:	*„Diseminare",* Noiembrie 2018.