D1.18. Implementarea modulului de control automat al prozodiei

Aceste rezultate au fost obținute prin finanțare in cadrul Programului PN-III Proiecte complexe realizate în consorții CDI, derulat cu sprijinul MEN – UEFISCDI,
PN-III-P1-1.2.-PCCDI, nr. 73 PCCDI/2018:
“SINTERO: Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate”

Denumirea organizației participante in 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, nr. 73 PCCDI/2018
Acronim / titlu:	**SINTERO: Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate”**
Titlu livrabil:	**D1.18.Implementarea modulului de control automat al prozodiei**
Termen:	**Mai 2018**
Editor:	**Adriana Stan (Universitatea Tehnică din Cluj-Napoca)**
Adresa de eMail editor:	**Adriana.Stan@com.utcluj.ro**
Autori, in ordine alfabetică:	**Mircea Giurgiu, Adriana Stan**
Ofițer de proiect:	**Cristian STROE**

Rezumat

Acest document prezintă o primă versiune a modulului de control a expresivității sistemului de sinteză text-vorbire din cadrul proiectului SINTERO. Acest modul include două modalități de control a prozodiei, una manuală, în care utilizatorul poate manipula parametrii prozodici în mod liber și una automată bazată pe modele acustice antrenate pe date audio specifice. Folosind aceste două modalități, expresivitatea sistemului de sinteză este îmbunătățită și adaptată textului de intrare în mod semi-automat la momentul actual. Versiunile ulterioare ale modulului vor face posibilă generarea prozodiei pornind strict de la informația textuală.

Cuprins

1. Introducere

Controlul expresivității sistemelor de sinteză text-vorbire reprezintă în continuare un subiect de cercetare important în cadrul domeniului prelucrării semnalului vocal pentru interacțiunea om-mașină. La ora actuală, sistemele de sinteză pot genera o calitate a vocii similară cu cea naturală [Shen17]. Însă studiile respective nu tratează și gradul de expresivitate sau de adecvare a prozodiei semnalului rezultat.

Expresivitatea sintezei este esențială în principal datorită nevoii adaptării rezultatului sintezei la domeniul textului redat sau la utilizarea finală a acestuia. De exemplu, pentru persoanele cu deficiențe de vedere, inteligibilitatea semnalului este mult mai importantă decât prozodia sau naturalețea sa [Botinhao15]. În schimb, dacă sistemul de sinteză redă o carte adresată copiilor, expresivitatea este esențială. Pe de altă parte un text jurnalistic va trebui redat diferit față de un text narativ sau cu caracteristică de spontaneitate.

Problemele apărute în cadrul sistemelor de sinteză cu expresivitate pornesc în principal de la o înțelegere limitată a modului în care oamenii generează această expresivitate sau a modului în care își adaptează prozodia în funcție atât de ceea ce vor să exprime, cât și a fondului cultural, social și educațional din care provin [Brenan09]. Acest fapt se transpune, în mod evident și în lipsa unor măsuri obiective de analiză a expresivității. Astfel încât nu se poate analiza rezultatul sintezei în mod automat, fiind nevoie de cele mai multe ori de așa numitele teste de ascultate (en. listening tests) prin care operatorii umani oferă o măsură subiectivă asupra vocii sintetizate și a modului în care aceasta produce o prozodie conformă cu textul redat.

Există, însă, un set de categorii de stiluri de exprimare ce au elemente prozodice distinctive. Printre acestea se numără: stilul neutru, narativ sau jurnalistic. Aceste stiluri sunt regăsite în majoritatea resurselor audio disponibile online și ca atare pot fi analizate pornind de la un set de date extins. La cealaltă extremă se află stilul de vorbire spontan, ce include aspecte legate de identitatea vorbitorului, în primul rând și care nu respectă în general o anumită regulă sau un tipar de exprimare. Este cunoscut faptul că în cadrul dialogurilor (amiabile), prozodia interlocutorilor este afectată bidirecțional, fiecare dintre ei adaptându-și modul de exprimare în funcție de celălalt.

În cadrul acestui raport va fi prezentat un prim modul de control al expresivității sistemului de sinteză text-vorbire în limba română pornind de la un set de resurse audio înregistrate în condiții de studio, precum și a analizelor prozodice prezentate în raportul D1.15. Modulul de control al expresivității permite atât o manipulare manuală a prozodiei semnalului la nivel de fonem, precum și o generare automată a unui anumit stil de exprimare folosind modele acustice antrenate pe corpusuri audio specifice.

2. Descrierea modulului de control automat al expresivității

Pentru a putea controla expresivitatea unui sistem de sinteză text-vorbire este necesar, în principiu, ca datele audio folosite în antrenarea modelelor acustice să conțină nivelul și tipul de expresivitate dorit. Dacă acest tip de date nu este disponibil, se pot realiza manipulări manuale ale prozodiei sistemului, folosind fie adnotări la nivel de contur al frecvenței fundamentale sau a duratei, precum și a unor tipare prozodice prestabilite rezultate în urma unei analize a unui set de date extins.

În cadrul modulului de control al expresivității dezvoltat în SINTERO, am utilizat ambele versiuni de control al prozodiei enunțate anterior. Fiecare dintre aceste metode este descrisă în continuare.

2.1. Controlul manual al prozodiei sistemului de sinteză

Controlul manual al prozodiei sistemului de sinteză se referă la posibilitatea utilizatorului de a manipula în mod voluntar și nerestricționat caracteristicile prozodice ale semnalului vocal generat. Aceste caracteristici sunt reprezentate de conturul frecvenței fundamentale și de durata fonemelor. În ceea ce privește durata fonemelor, controlul se face doar la nivel de vocală și asta datorită faptului că, în sine, consoanele nu sunt purtătoare de prozodie, durata acestora fiind de cele mai multe ori fixă și dată de modul de articulare. În Figurile 1 și 2 sunt prezentate numărul și durata fiecărui fonem prin analiza întreg corpusului SWARA [SWARA07]. Este de menționat faptul că alinierile temporale nu sunt manuale, ci obținute din alinierea semi-automată a datelor audio. Lista fonemelor utilizată pentru adnotarea corpusului este disponibilă aici: http://romaniantts.com/rssdb/data/PHONEMES.pdf

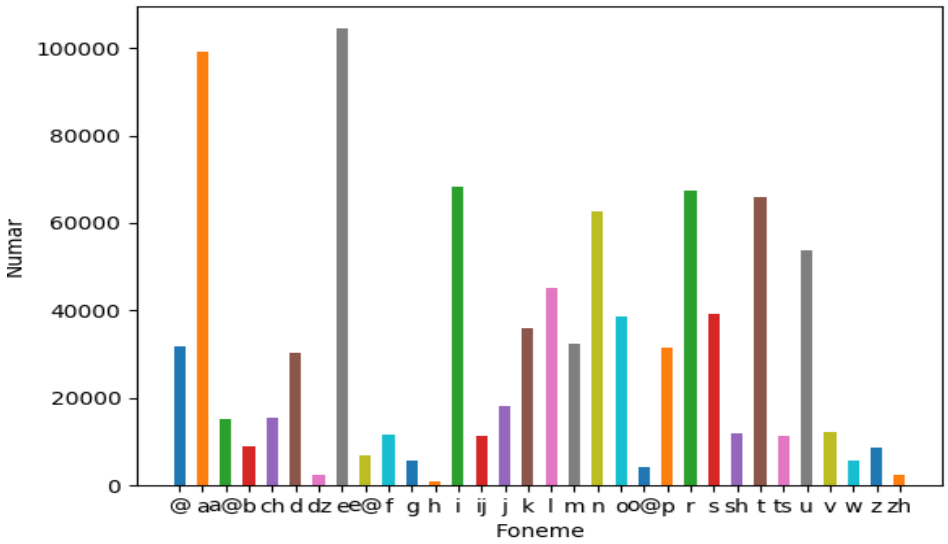

Fig. 1. Numărul de apariții al fonemelor în corpusul SWARA [SWARA07]

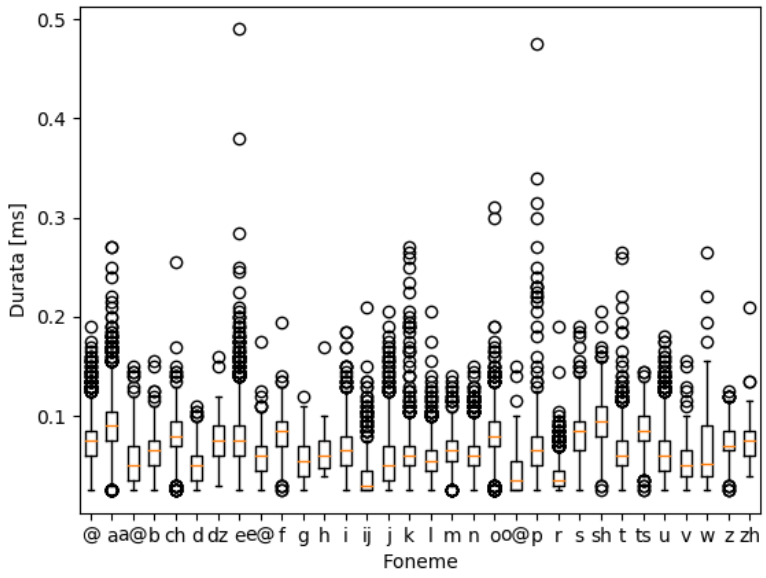

Fig.2. Mediana și prima și a treia cvartilă a duratei fonemelor din corpusul SWARA [Stan07]

Pentru acest tip de control al prozodiei, utilizatorul are la dispoziție un fișier de configurare ce include durata și valoarea media a frecvenței fundamentale, așa cum sunt ele generate de modelele acustice, pentru fiecare vocală în parte. Un exemplu de astfel de fișier este prezentat în Figura 3. După modificarea parametrilor prozodici, semnalul este resintetizat.

PHONE	DUR	F0
a	0.085	218
a	0.085	235
a	0.080	244
e	0.070	207
e	0.095	164
e	0.135	162

Fig.3. Exemplu fișier de configurare a parametrilor prozodici de sinteză pentru propoziția ”Ana are mere”.

În acest format, utilizatorul trebuie să cunoască anumite aspecte legate de caracteristicile semnalului vocal, fapt ce nu este fezabil pentru utilizatorii non-experți. Astfel că, s-a început deja implementarea unei interfețe grafice ce permite ajustarea acestor parametri folosind un set de bare de control liniar pentru durată și manipularea conturului F0 prin modificarea unor puncte de pe grafic. Un prototip al acestei interfețe este prezentat în Figura 4.

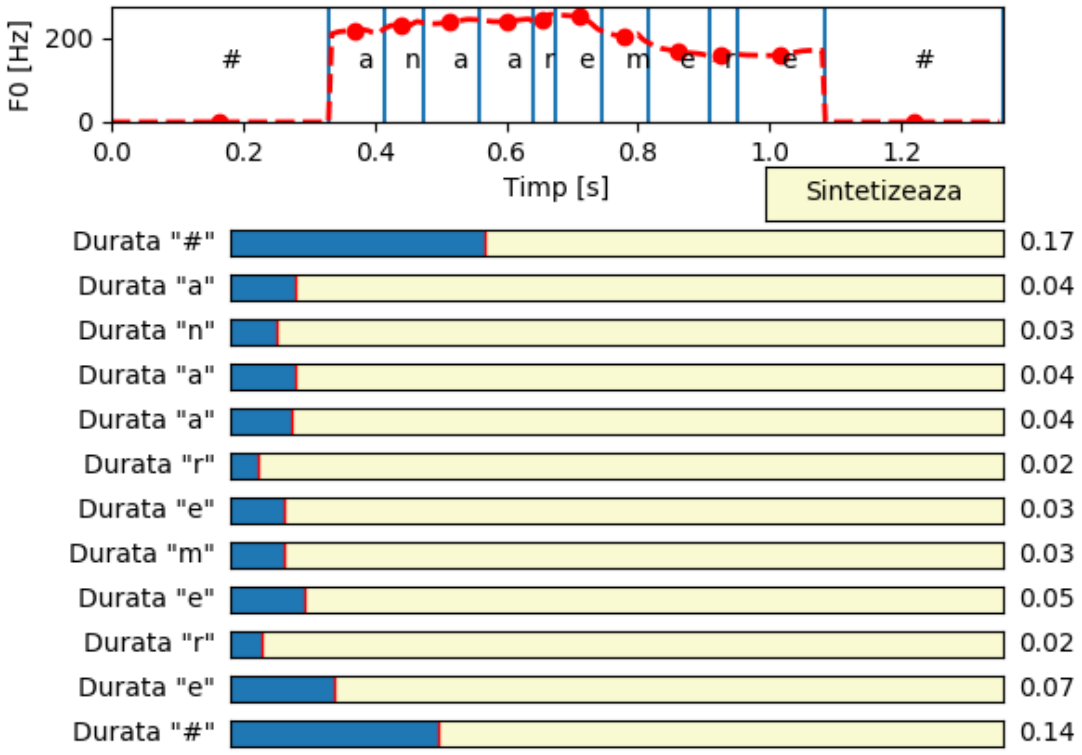

Fig.4. Prototip al interfeței grafice de control a parametrilor prozodici de sinteză.

Două alte metode control semi-automat al prozodiei sistemului de sinteză se referă la modificarea liniară a duratei semnalului sintetizat prin ajustarea proporțională a duratei tuturor fonemelor din textul de intrare; și controlul propozițiilor interogative sau exclamative folosind un parametru suplimentar la intrarea sistemului de sinteză. În ultimul caz s-au folosit analizele din cadrul raportului D1.15.

Exemple audio pentru aceste două metode suplimentare sunt disponibile aici: http://speech.utcluj.ro/sintero/prosody_examples/ .

2.2. Controlul automat al prozodiei sistemului de sinteză

Controlul automat al prozodiei pentru sistemul de sinteză SINTERO se bazează pe un set de date audio de antrenare ce conțin stiluri de exprimare diferite: neutru, narativ și jurnalistic. Aceste date audio au fost colectate din surse diferite, dar înregistrate în condiții de studio. Exemple audio pentru fiecare stil în parte pot fi ascultate la adresa: http://speech.utcluj.ro/sintero/prosody_examples/ .

Pentru stilul neutru a fost selectat un vorbitor de sex feminin din corpusul SWARA, identificator SAM. Pentru acest vorbitor sunt disponibile înregistrări cu o durată de 1 oră și 43 de minute, realizate într-un studio semi-profesional, la o frecvență de eșantionare de 48 kHz și 16 biți per eșantion. Înregistrările sunt segmentate manual la nivel de propoziție, iar transcrierile ortografice au fost verificate manual. Întreg corpusul SWARA a fost dezvoltat în vederea utilizării înregistrărilor în cadrul sistemele de sinteză text-vorbire, astfel încât, atât conținutul, cât și durata datelor audio sunt în conformitate cu cerințele unui astfel de sistem.

Stilul jurnalistic, identificator ATV, a fost preluat dintr-un set de înregistrări audio ale unei prezentatoare de știri. Înregistrările au o durată de 48 de minute și au fost realizate într-un studio de televiziune. Frecvența de eșantionare este de 16kHz, iar rezoluția este de 16 biți per eșantion. Durata acestui set de date audio nu este suficientă pentru a antrena modele acustice independente de o calitate ridicată, însă această durată conține suficientă informație prozodică pentru a putea realiza adaptarea unor modele acustice antrenate pe un set de date audio extins. Pentru aceste date, transcrierea ortografică nu era aliniată la nivel de propoziție, astfel încât această aliniere s-a realizat manual.

Stilul narativ, identificator MAB, a fost extras din cadrul unei cărți audio furnizate de Cartea Sonoră ce conține înregistrarea nuvelei ”Mara” de Ioan Slavici, realizată de către o vorbitoare profesionistă într-un studio de înregistrări. Durata totală a datelor audio este de 11 ore, datele fiind eșantionate la 44kHz și 16 biți per eșantion. Însă pentru aceste date transcrierea ortografică la nivel de propoziție nu este disponibilă. Astfel că, alinierea și segmentarea la nivel de propoziție a datelor a fost realizată cu ajutorul utilitarului ALISA [Stan16], rezultând astfel un set de 7 ore de date aliniate aproximativ, cu o rată de eroare la nivel de cuvânt de sub 0.5%. Deși rata de eroare la nivel de cuvânt este redusă, majoritatea problemelor de aliniere apar la început sau final de propoziție. Ca urmare, utilizarea acestui corpus audio pentru crearea unei voci independente duce la generarea unor artefacte în cadrul semnalului vocal rezultat.

Pe baza acestor 3 seturi de date, s-au dezvoltat într-o primă fază voci sintetice dependente de vorbitor, pentru a putea analiza atât calitatea, cât și fezabilitatea utilizării lor în sistemul nostru de sinteză. Vocile au fost create în cadrul sistemului de sinteză bazat pe modele Markov și vocoderul WORLD [Morise16]. Exemple audio cu rezultatul acestor voci pot fi ascultate aici: http://speech.utcluj.ro/sintero/prosody_examples/ . Se poate observa faptul că, în conformitate cu specificațiile seturilor de date, calitatea vocilor diferă, vocea în stil narativ având cele mai multe artefacte. Acest fapt se datorează într-o mare măsură expresivității exagerate a vorbitorului, de altfel în concordanță cu stilul redat. Pentru stilul jurnalistic, durata redusă a

înregistrărilor face ca vocea sintetică să fie puțin mai zgomotoasă, multe dintre fonemele sau contextele fonetice distincte fiind mediate datorită arborilor de decizie utilizați în procesul de grupare a modelelor Markov.

Drept urmare, pentru a putea integra, însă, aceste stiluri de exprimare în cadrul sistemului de sinteză SINTERO, s-a realizat o adaptare a modelelor acustice ale vorbitorului neutru către ceilalți doi vorbitori. Modelele acustice ale sistemului de sinteză sunt bazate pe o parametrizare de tip WORLD și includ coeficienții mel-cepstrali, coeficienții de aperiodicitate și frecvența fundamentală extrase din ferestre cu o durată de 20ms și deplasare de 5ms. Acești parametri sunt modelați cu modele Markov cu 5 stări ascunse și conexiuni stânga-dreapta. Durata fonemelor este modelată în mod similar, folosind 5 stări ascunse pentru modelele Markov, stări ce reprezintă diferitele segmente ale unui fonem.

Deoarece dorim să realizăm doar transferul prozodiei și nu și a identității vorbitorului (timbrul), este nevoie doar de adaptarea modelelor frecvenței fundamentale și a duratei. Această adaptare s-a făcut utilizând metoda CSMAPLR [Yamagishi09]. Această metodă estimează atât media, cât și covarianța parametrilor acustici și a duratei, iar transformarea distribuțiilor de probabilitate de la vorbitorul de bază la cel țintă este aplicată unitar. Calculul matricei de transformare se face pe baza maximizării parametrilor modelelor acustice în funcție de observațiile acustice ale vorbitorului țintă. Metoda CSMAPLR și variante ale acesteia sunt cele mai utilizate metode pentru adaptarea vorbitorilor în cadrul sistemelor de sinteză text-vorbire bazate pe modele Markov cu stări ascunse.

Rezultate ale acestui tip de adaptare sunt disponibile la adresa: http://speech.utcluj.ro/sintero/prosody_examples/. Totodată, pentru a exemplifica diferențele de stil dintre cei 3 vorbitori, în Figurile 5, 6 și 7 sunt prezentate duratele fonemelor individuale pentru cele 3 stiluri. Se poate observa faptul că

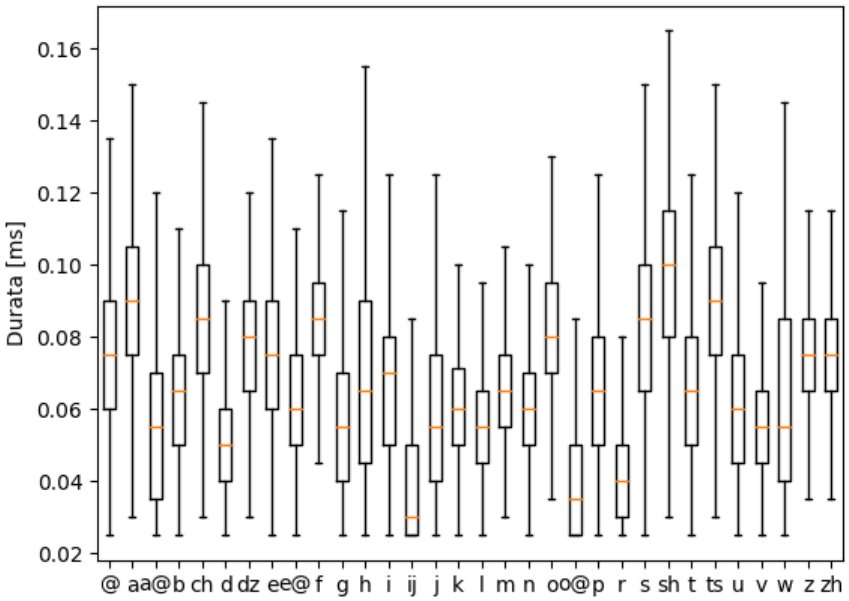

Fig.5. Valoarea mediană și prima și a treia cvartiala pentru durata fonemelor din corpusul de voce cu stil **neutru**

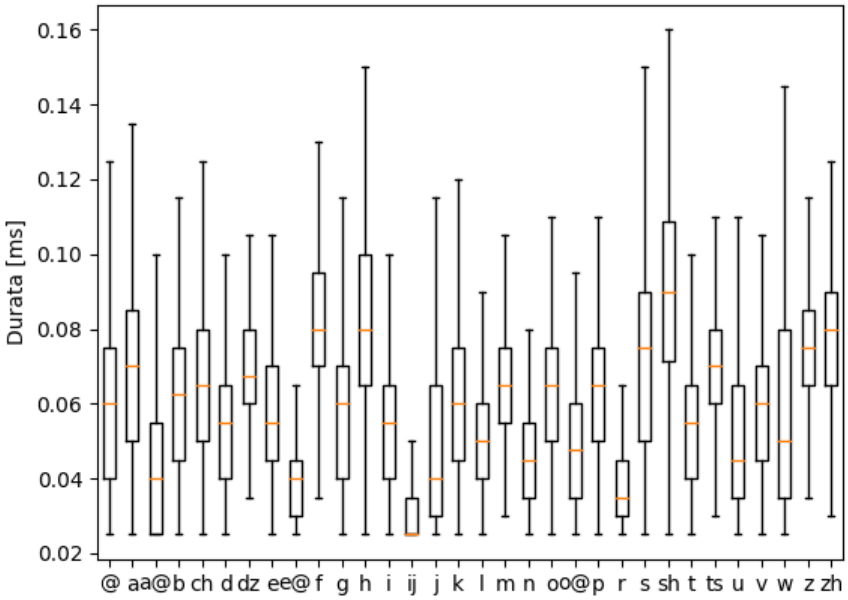

Fig.6. Valoarea mediană și prima și a treia cvartilă pentru durata fonemelor din corpusul de voce cu stil **jurnalistic**

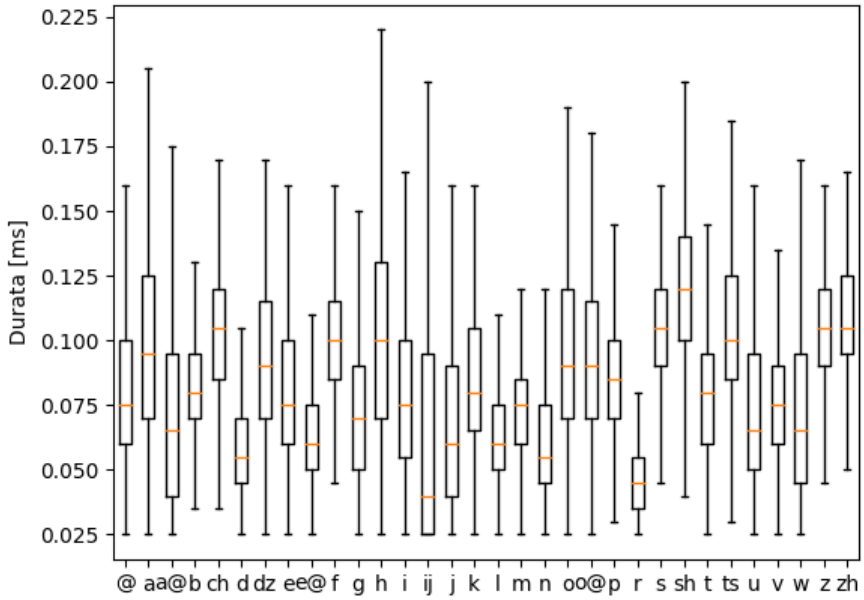

Fig.7. Valoarea mediană și prima și a treia cvartilă pentru durata fonemelor din corpusul de voce cu stil **narativ**

Figura 8 prezintă statistici ale frecvenței fundamentale pentru cei 3 vorbitori. Se poate observa că frecvențele fundamentale mediane sunt apropiate ca valoare, diferențele majore regăsindu-se în deviațiile standard. Frecvența medie pentru stilul neutru e de 200Hz cu o deviație standard de 25Hz, pentru stilul jurnalistic, frecvența medie este de 221Hz cu o deviație

standard de 37Hz, iar pentru stilul narativ, frecvența medie este de 207Hz și o deviație standard de 52Hz.

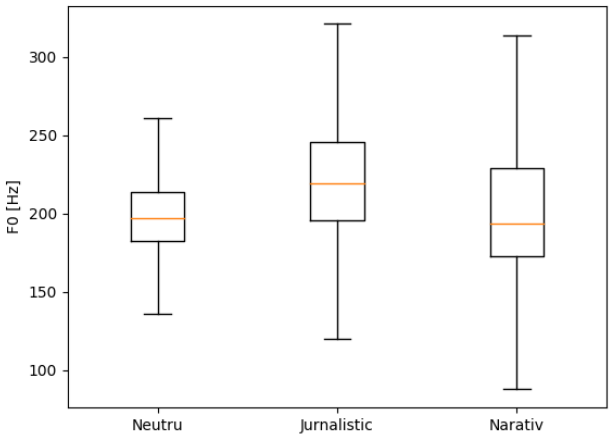

Fig.8. Valoarea mediană și prima și a treia cvartilă pentru frecvența fundamentală în corpusurile de voce netru, jurnalistic și narativ

Un alt aspect important este și modul în care durata fonemelor și contrul frecvenței fundamentale se modifică pentru aceeași propoziție, în funcție de stilul de exprimare. Figurile 9, 10 și 11 prezintă conturul și spectrograma aceleiași propoziții enunțate cu cele 3 stiluri din cadrul sistemului de sinteză text-vorbire. Se poate observa o modificare atât a conturului suprasegmental al frecvenței fundamentale, cât și a duratei fonemelor individuale.

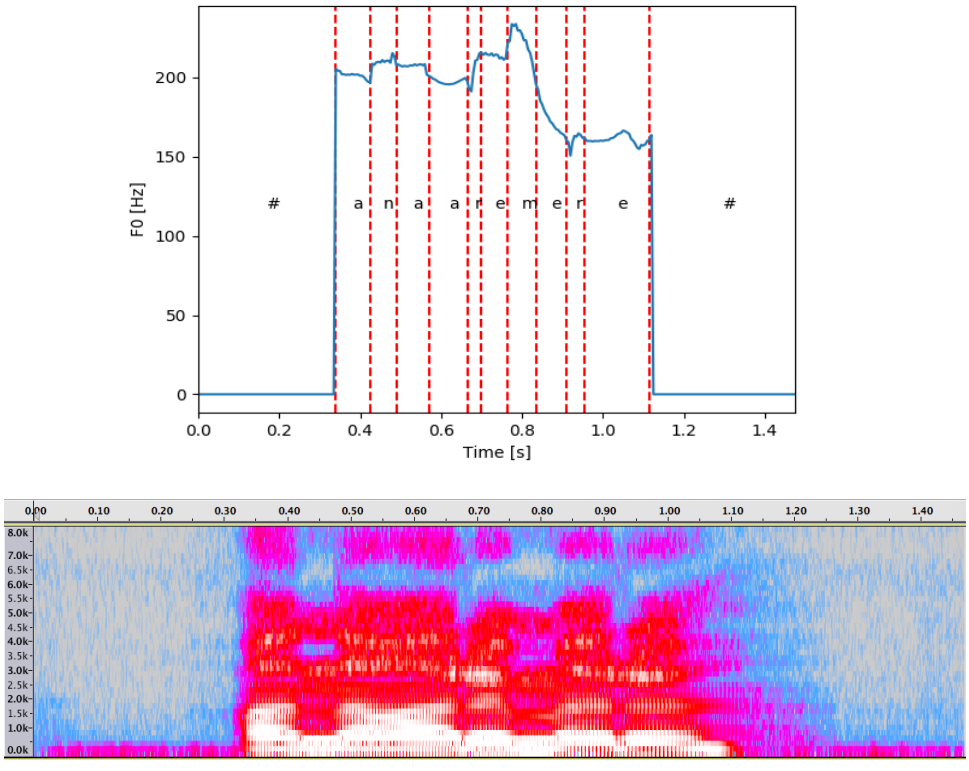

Fig.9. Conturul F0, durata fonemelor și spectrograma pentru propziția ”Ana are mere” folosind vocea cu stil neutru

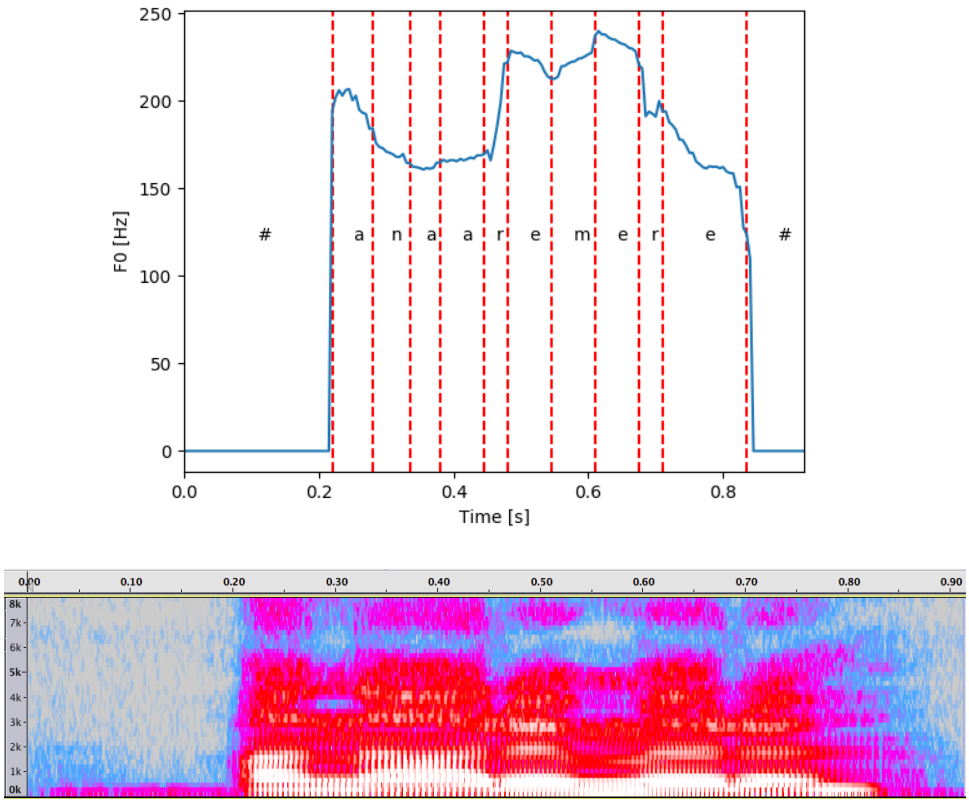

Fig.10. Conturul F0, durata fonemelor și spectrograma pentru propziția ”Ana are mere” folosind adaptarea la stilul jurnalistic

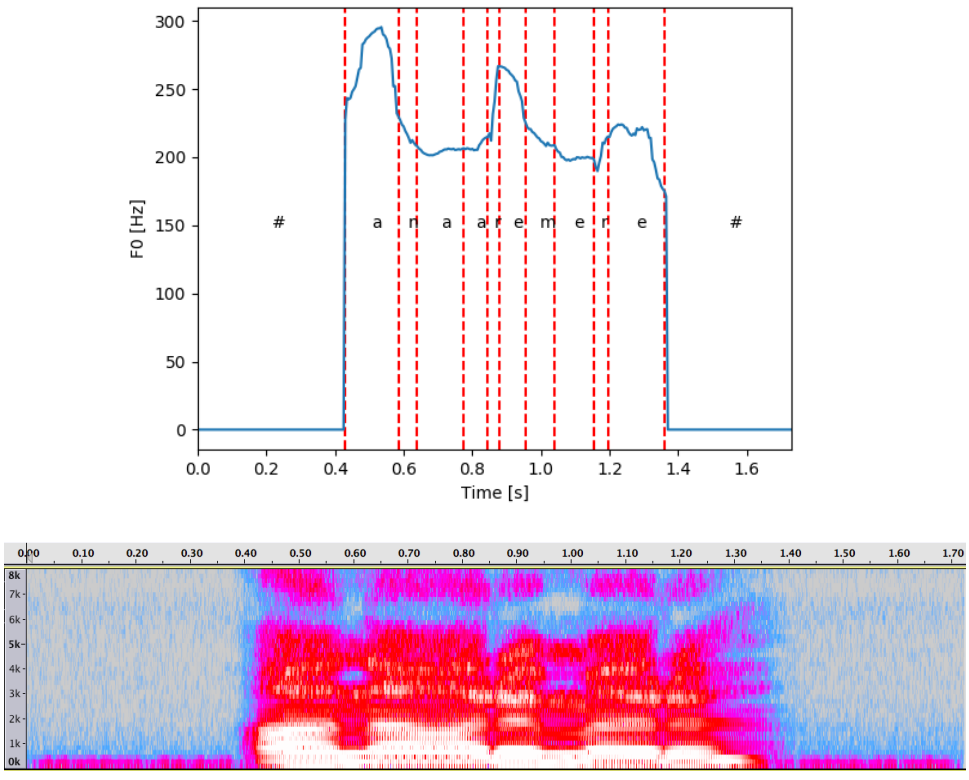

Fig.11. Conturul F0, durata fonemelor și spectrograma pentru propziția ”Ana are mere” folosind adaptarea la stilul narativ

3. Concluzii

Acest raport a prezentat o primă versiune a sistemului de sinteză text-vorbire în limba română ce include o serie de modificări ale expresivității semnalului vocal rezultat. Față de sistemul de bază, modulul permite atât manipularea manuală a duratei și conturului frecvenței fundamentale la nivel de fonem, precum și utilizarea unor metode automate de control a acestor parametri. Metodele automate includ utilizarea unor informații suplimentare extrase din text pentru a genera propoziții declarative, interogative sau exclamative, precum și modificarea stilului de exprimare prin adaptarea vocii neutre la stilul jurnalistic sau narativ.

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.

4. Bibliografie

Botinhao15 Cassia Valentini-Botinhao, Markus Toman, Michael Pucher, Dietmar Schabus, and Junichi Yamagishi. Intelligibility of time-compressed synthetic speech: Compression method and speaking style. Speech Communication, October 2015.

Brenan09 Brennan, S. E. and Hanna, J. E., Partner-Specific Adaptation in Dialog. Topics in Cognitive Science, doi:10.1111/j.1756-8765.2009.01019.x

Morise16 Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, ”WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications”, IEICE TRANS. INF. & SYST., VOL.E99–D, NO.7 JULY 2016

Shen17 Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu, *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,* arXiv 1712.05884

Stan16 Adriana Stan, Yoshitaka Mamiya, Junichi Yamagishi, Peter Bell, Oliver Watts, Rob Clark, Simon King, "ALISA: An automatic lightly supervised speech segmentation and alignment tool", In Computer Speech and Language, vol. 35, pp. 116-133, 2016.

Stan17 Adriana Stan, Florina Dinescu, Cristina Țiple, Șerban Meza, Bogdan Orza, Magdalena Chirilă and Mircea Giurgiu, The SWARA Speech Corpus: A Large Parallel Romanian Read Speech Dataset, in Proceedings of the 9th Conference on Speech Technology and Human-Computer Dialogue, Bucharest, Romania, July 6-9, 2017

Yamagishi09 Junichi Yamagishi, Takao Kobayashi, Senior Member, IEEE, Yuji Nakano, Katsumi Ogata, and Juri Isogai, Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm, IEEE Trans on Audio, Speech, and Language Processing, vol.17, no .1, 2009.