RAPORTARE ȘTIINȚIFICĂ

Proiect complex ReTeRom. Proiect component CoBiLiRo

Activitatea 4.1 - Alte aplicații de exploatare a corpusului bimodal si a tehnologiilor de prelucrari textuale si de voce create in Pr. 2, Pr. 3 si Pr. 4

Faza de predare: aprilie 2021

Autori: Pistol Ionuț, Gîfu Daniela, Cristea Dan (UAIC)

1. Rezumatul etapei

În această etapă finală a proiectului complex ReTeRom consorțiul a avut ca obiective completarea și îmbunătățirea resurselor dezvoltate, precum și a perspectivelor de valorificare a lor. Pentru proiectul component CoBiLiRO, această ultimă etapă s-a concretizat prin noi contribuții la resursele de pe platforma dezvoltată, noi funcționalități oferite de această platformă și adăugarea unor propuneri noi de valorificare ulterioară a resurselor dezvoltate în proiectul RETEROM sub forma unor proiecte.

2. Rezumatul activității

Activitatea 4.1 are ca obiectiv completarea listei de proiecte propusă în raportul 3.4 *(Proiectare de aplicații de exploatare a corpusului bimodal și a tehnologiilor de prelucrări textuale și voce, create în proiectele P2, P3, P4) [12].* Pe lângă completarea acelei liste cu două proiecte noi, acest raport prezintă actualizări privind stadiul de dezvoltare a unor proiecte descrise în raportul 3.4 [12]. Scopul acestei activități este acelaşi cu al activității 3.4 și anume demonstrarea potențialului tehnologiilor dezvoltate și mărirea vizibilității proiectului ReTeRom în special după finalizarea acestuia.

3. Descrierea științifică și tehnică

Pentru o descriere a tehnologiilor din proiectele componente menţionate în aplicațiile de mai jos se poate consulta raportul 3.4 [12] și rapoartele celorlalte proiecte componente. După secțiunea 3.1 în care se discută progresele făcute la două din aplicațiile descrise în raportul 3.4 urmează descrierea a două aplicații noi ce valorifică tehnologiile și resursele dezvoltate în proiectul complex ReTeRom.

3.1 Actualizări asupra stadiului de dezvoltare a proiectelor propuse în raportul activității 3.4

Nu ne-am propus să dezvoltăm în continuare aplicațiile 3, 4, 5 și 6, față de specificațiile descrise în raportul 3.4. În continuare sunt discutate progresele făcute la aplicațiile 1 și 2.

Aplicația 1: **Suport pentru învățarea limbii române - *PD builder***

Acest proiect propune construirea unui dicționar de pronunție pentru cuvintele limbii române și dezvoltarea tehnologiei ce permite construirea automată a unor resurse similare. În raportul 3.4 au fost descrise două etape principale în realizarea acestui proiect: indexarea referințelor textuale din corpusul aliniat disponibil pe platforma proiectului și extragerea fragmentelor audio corespunzătoare. Prima etapă este finalizată, un index conținând peste 500.000 cuvinte unice și 100.000 forme neflexionate (forme lemă) a fost construit. Etapa va continua până la completarea colecției de resurse multimodale. Extragerea fragmentelor audio este încă în stadiul incipient, dar estimăm ca un prim prototip al aplicației va fi finalizat în acest an.

Aplicația 2: **Analiza corpusurilor bimodale**

Această aplicație propune dezvoltarea unui sistem de sprijin pentru evaluarea unui corpus bimodal aliniat text-sunet. În raportul 3.4 sunt descrise câteva erori posibile și soluții pentru descoperirea lor automată, erorile putând fi de tipul:

- diferențe între transcrierea text și conținutul înregistrării;
- calitatea scăzută a înregistrării;
- erori ale aplicației de aliniere.

Un prototip al acestui sistem a fost implementat ca parte a datelor statistice calculate pentru resursele de pe platforma CoBiLiRo, descrise în raportul 3.3. Un exemplu al acestor date poate fi văzut în figura 1.

File	AutoText	AlignedText	LongestUnmatch	StartsAt	SingleUnmatched
Alma-Mater-Iassiensi-24-Facultatea-de-Drept	5379	3761	44	30.51	193
Alma-Mater-Iassiensis-1-Aniversarile-Universitatii	7078	4400	121	2944.26	248
Alma-Mater-Iassiensis-11-Prof-Univ-Dr-Andrei-Marga	7072	4947	65	2688.87	266
Alma-Mater-Iassiensis-12-Facultatea-de-Fizica	5232	3312	52	3282.42	166
Alma-Mater-Iassiensis-13-Universitatea-din-Iasi-prezent-si-perspective	7006	5220	61	2109.15	281
Alma-Mater-Iassiensis-14-Universitatea-din-Iasi-Dimensiunea-culturala	6301	4558	41	2090.97	265

Fig.1 : Statistici pe alinieri la nivel de cuvânt

În figura 1 semnificaţia coloanelor este:

- *AutoText* - dimensiunea (în cuvinte) a textului extras automat din fişierul audio;
- *AlignedText* - dimensiunea (în cuvinte) a textului aliniat cu transcrierea originală;
- *LongestUnmatch* - cea mai lungă secvenţă de cuvinte din transcrierea originală pentru care nu există nici o aliniere;
- *StartsAt* - indexul cuvântului de la care începe secvenţa de mai sus;
- *SingleUnmatched* - numărul de cuvinte din transcrierea originală care nu au fost aliniate, deşi cuvântul imediat precedent şi cel imediat ulterior au fost aliniate.

O serie de observaţii pot fi făcute pe aceste statistici, de exemplu se observă că procentul cuvintelor aliniate corect (pe cele 6 fişiere din figura 1) este de aproximativ 70%. Se observă că în fişierul al doilea există o zonă lungă de cuvinte nealiniate (121 cuvinte) ce poate indica prezenţa în transcrierea manuală a unui fragment care nu se regăseşte şi în înregistrarea audio. O altă observaţie ce ar putea fi făcută pe aceste statistici ar fi că în fişierul 5 există un număr relativ mare de cuvinte izolate nealiniate (281) ceea ce ar putea indica o calitate scăzută a semnalului audio sau un accent/prozodie nestandard al/a vorbitorului.

Finalizarea acestui sistem şi punerea lui la dispoziţie ca o aplicaţie independentă de platforma CoBiLiRO este planificată pentru începutul anului viitor.

3.2 Aplicaţia 7 (nouă): **Sistem suport pentru crearea corpusurilor bimodale**

Acest proiect propune o aplicaţie mobilă care să permită crearea şi editarea unor resurse bimodale aliniate text-vorbire. Dispozitivele mobile ce permit prelucrări audio-video complexe sunt deja la îndemâna unui segment important din populaţie. Un astfel de dispozitiv este capabil să permită atât înregistrarea unui semnal audio de calitate rezonabilă cât şi interacţiunea utilizatorului cu o interfaţă de editare. Aplicaţia “Sistem de suport pentru crearea corpusurilor bimodale” doreşte să pună la dispoziţie unui posesor de telefon “inteligent” sau tabletă un mediu în care să poată crea alinieri text-voce.

Utilizatorul poate selecta sau adăuga fie un fişier text, fie o înregistrare audio şi poate crea şi alinia resursa pereche. De exemplu, dacă are la dispoziţie un text, poate citi acel text, aplicaţia înregistrează acea citire şi deschide o interfaţă în care utilizatorul vede atât textul (derulat cuvânt cu cuvânt) cât şi un fragment din corespondentul fişier audio pe care poate delimita zona de început şi cea de sfârşit a pronunţiei acelui cuvânt.

Un prototip al acestei aplicaţii există deja, un exemplu de interfaţă poate fi văzută în figura 2. Interfaţa exemplifică scenariul descris în paragraful anterior.

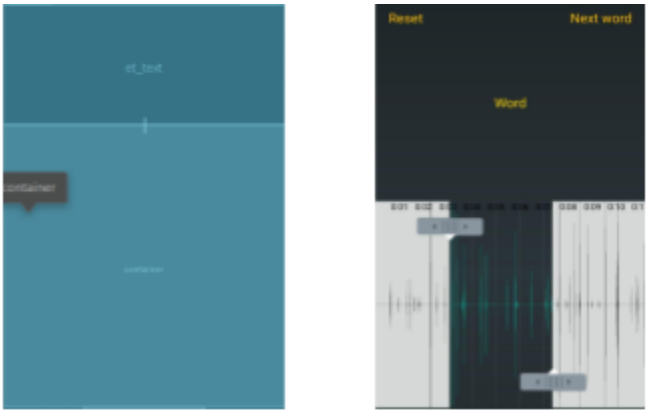

Fig. 2: Interfața aplicației mobile pentru crearea resurselor aliniate

Versiunea curentă este implementată pentru platforma Android (în Java/Kotlin), urmând ca după finalizarea tuturor funcționalităților propuse ea să fie transferată și pe platforma MacOS. O primă versiune disponibilă public a acestei aplicații este propusă pentru sfârșitul acestui an. Avem în vedere și posibilitatea urcării resurselor bimodale create de această aplicație pe platforma CoBiLiRo, dacă creatorul acestei resurse dorește acest lucru.

3.3 Aplicația 8 (noutate): **Sinteză text-vorbire și clonarea vocii în limba română cu metoda învățării prin transfer**

Acest proiect propune un sistem de sinteză text-vorbire, folosind metoda învățării prin transfer, care poate fi antrenat pe corpusuri în limba română de dimensiuni mici, conținând doar elemente audio și transcrierea acestora, fără a fi necesară strângerea altor date. Este o abordare cu rezultate mai bune decât ceea ce reține literatura de specialitate. Sistemele clasice necesită potrivirea manuală a fonemelor cu formele de undă pentru fiecare înregistrare, ceea ce face strângerea de date costisitoare și dificilă.

În general, sistemele bazate pe rețele neurale profunde necesită cantități mari de date pentru a da rezultate mulțumitoare, ceea ce face incomodă antrenarea de la zero a modelelor.

Totodată, acest proiect propune și un sistem de clonare vocală, folosind învățarea prin transfer de la modelul de sinteză general. Clonarea vocală este o sarcină, de asemenea, dificilă, capacitățile sistemelor de sinteză a vorbirii fiind și limita superioară a acestora, iar, în maniera clasică, ar trebui transcrise și alineate manual fonemele cu fiecare înregistrările audio ale vorbitorului țintă. Momentan, pentru limba română, sistemele de sinteză vocală publicate sunt într-o fază de pionierat. Clonarea vocii[1] rămâne o sarcină abordată relativ recent pentru limba română, cu toate că s-au făcut eforturi în întocmirea de corpusuri de vorbire paralelă [10].

Persoanele care și-au pierdut capacitatea de a vorbi și-o pot recăpăta prin clonarea digitală a vocii, putând să-și recupereze astfel o parte din identitate. De asemenea, în

[1] http://www.vc-challenge.org

învățământul de la distanță tehnologia poate fi folosită pentru reproducerea vocii personalităților din istoria recentă și de azi, întrucât elevii să poată învăța despre respectivele personalități discutând chiar cu ele. Spre exemplu, aplicația FreudBot[2], un psihanalist robotizat, sugerează faptul că se poate învăța inclusiv învingerea agresivității, îndoielii și a fricii. Recunoaște peste 100 de vibrații proaste ale vieții de zi cu zi! [2].

O altă posibilitate pe care clonarea vocii o deschide este citirea automată a milioane de cărți în vocea autorului sau în vocea unui actor vocal dorit, cu o eficiență extremă din punctul de vedere al muncii umane. Totuși, trebuie luat în considerare faptul că tehnologiile de tip DeepFake pot reprezenta un instrument periculos în cazul folosirii abuzive, cum ar fi în cazul imitării vocii cuiva fără consimțământul acestuia. Vocile clonate în acest proiect sunt din cadrul corpusului public RSS [11], iar cele din cadrul corpusului SWARA (Stan *et al.*, 2017).

Modelul folosit presupune un modul de sinteză din text în spectrogramă mel (sintetizator) și unul de inferență a vorbirii din spectrogramă mel (vocoder). Tacotron 2 [10] antrenat pe setul de date LJ Speech (*The LJ Speech Dataset*, 2017) a fost ales pentru sintetizator, iar WaveGlow [1], antrenat tot în limba engleză, a fost vocoderul ales. S-au ales parametri audio identici pentru ambele modele, corespunzători cu parametrii LJ Speech. Seturile de date în limba română au fost preprocesate, ambele modele fiind antrenate pe acestea. Sintetizatorul a fost antrenat pe un set de date corespunzător unui singur vorbitor, în timp ce vocoderul a fost antrenat pe un set de date cu mai mulți vorbitori.

Rezultate statistice și interpretare

În vederea analizării performanței au fost colectate opinii despre clipurile cu voce naturală sintetizate. Au fost sintetizate 10 propozitii relativ scurte din știri recente, iar participanților (cei invitați să participe la acest studiu) li s-au dat două clipuri din corpusul RSS împreună cu două clipuri sintetizate pentru a le nota gradul de naturalețe. Pentru consistența cu alte studii, evaluarea a fost făcută pe scara Likert (5-Excellent, 4-Good, 3-Fair, 2-Poor, 1-Bad). Toți au fost vorbitori nativi de română și cunoscători de limba engleză. Participanții au folosit mijloace proprii de ascultare a clipurilor, fiind relevant faptul la doar o minoritate au folosit căști audio pentru evaluare. S-a putut astfel efectua o comparație directă cu alte rezultate, lucru care a oferit o privire de ansamblu mai bună asupra calității vorbirii, indiferent de mediul de ascultare.

Un al doilea studiu a fost făcut în condiții similare, participanții fiind rugați să evalueze gradul de naturalețe a vocii clonate. Rezultatele studiilor, împreună cu alte rezultate din literatură, sunt prezentate în Tabelul 1.

[2] FreudBot – Aplicații pe Google Play

Tabelul 1: Evaluarea comparativă pentru TTS

Model	**Scor mediu de opinie**	**Limba**
Voci naturale	4.80 ± 0.14	română
TTS Eletron (din această lucrare)	4.47 ± 0.15	română
Clonare vocală - 18 minute	4.24 ± 0.23	română
Clonare vocală - 6 minute	3.89 ± 0.24	română
Clonare vocală - două minute	3.64 ± 0.23	română
Voci naturale(Chen et al., 2019)	4.89 ± 0.045	germană și franceză
Învățare prin transfer din engleză folosind Tacotron - 25 minute	4.01 ± 0.085	germană și franceză
Învățare prin transfer din engleză folosind Tacotron – 15 minute	3.48 ± 0.119	germană și franceză
Voci naturale (Shen et al., 2018)	4.582 ± 0.053	engleză
Tacotron 2	4.526 ± 0.066	engleză
Metoda prin concatenare	4.166 ± 0.091	engleză

Pentru a compara rezultatele cu alte sisteme de TTS în română, au fost strânse date întrebând subiecții ce clip preferă dintre unul sintetizat cu TTS Eletron, altul folosind TTS-ul din cadrul proiectului SWARA (*Romanian TTS - Online text-to-speech system*) și TTS-ul din cadrul Google Translate (*Google Traducere*). Din experimentele efectuate, reiese că metoda principală prezentată este optimă pentru sinteza generală, date fiind seturile de date disponibile. Moduri de a implementa sistemul educațional inteligent propus în Ouatu & Gîfu (2020) și integrarea lucrării

de față în acesta reprezintă direcții viitoare de cercetare. Vom explora, de asemenea, modalități de îmbunătățire a ambelor tipuri de sisteme de sinteză. În vederea obținerii de rezultate superioare, implementarea modelelor prezentate de Liu et al., (2019) și Zhu et al. (2019) sunt prioritare. Sinteza emoțională a vocii și copierea stilului de vorbire constituie o altă direcție de cercetare.

3.8 **Concluzii**

Obiectivul principal al proiectului ReTeRom a fost dezvoltarea de resurse multimodale aliniate și a unor tehnologii compatibile. Atingerea acestui obiectiv și calitatea acestor resurse și tehnologii pot fi cel mai bine demonstrate prin utilizarea lor în contexte realiste, răspunzând unor necesități ale comunității de cercetători, studenți și chiar ale publicului larg.

Rapoartele 3.4 și 4.1 prezintă o serie de 8 aplicații care au ca obiectiv valorificarea resurselor și tehnologiilor dezvoltate în ReTeRom. Stadiile lor de dezvoltare și progresele făcute chiar înainte de finalizarea proiectului, în toate cazurile cu studenți sau alte persoane voluntare, arată interesul comunității în valorificarea acestor resurse.

Toate obiectivele incluse în plan la această activitate au fost realizate.

Bibliografie

[1] Prenger, R., Valle, R., & Catanzaro, B. (2018). *Waveglow: A Flow-Based Generative Network for Speech Synthesis*.

[2] Procter, M. (n.d.). *Freudbot: An Investigation of Chatbot Technology in Distance Education*.

[3] Radu, I. Raport Activitate A1.5: *Definirea specificațiilor funcționale și arhitecturale ale platformei integrate și configurabile de prelucrare a textelor*, proiectul ReTeRom

[4] Radu, I. Raport Activitate A1.6: *Definirea modulelor software și a serviciilor oferite de proiect; identificarea adaptărilor pentru modulele NLP existente și a modulelor noi necesare,* proiectul ReTeRom.

[5] Boroș, T., Dumitrescu, Ș., Pais, V. *Tools and resources for Romanian text-to-speech and speech-to-text applications*, 2018.

[6] Zamfirescu, A.N., Rebedea, T.E. *Identificarea entităţilor, citatelor şi evenimentelor în ştiri şi texte din Web-ul social în limba română*, în Revista Română de Interacţiune Om-Calculator 6 (2) 2013, 169-192.

[7] Burileanu, C., Cucu, H. Raport Activitate A1.11: *Studiul metodelor din literatură pentru alinierea transcrierilor aproximative cu semnalul de vorbire*, proiectul ReTeRom.

[8] Aldabbagh, O., Mohsen, K. Design and Implementation of Online Location Based Services Using Google Maps for Android Mobile. International Journal of Computer Networks and Communications Security. 2. pp. 113-118, 2014

[9] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (Vols. 2018-April). https://doi.org/10.1109/ICASSP.2018.8461368

[10] Stan, A., Dinescu, F., Tiple, C., Meza, S., Orza, B., Chirila, M., & Giurgiu, M. (2017). The SWARA speech corpus: A large parallel Romanian read speech dataset. *2017 9th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2017*, 1–6. https://doi.org/10.1109/SPED.2017.7990428

[11] Stan, A., Yamagishi, J., King, S., & Aylett, M. (2011). The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication*, *53*(3), 442–450. https://doi.org/10.1016/j.specom.2010.12.002

[12] Pistol, I., Scutelnicu, A., Onofrei M., Gîfu D., Boghiu Ș., Raport Activitate 3.4: *Proiectare de aplicații de exploatare a corpusului bimodal și a tehnologiilor de prelucrări textuale și voce, create în proiectele P2, P3, P4,* proiectul RETEROM

[13] Ouatu, B., Gîfu, D. *Chatbot, the Future of Learning?* In: Proceedings of the 5th International Conference on Smart Learning Ecosystems and Regional Development (SLERD 2020), in Ludic, Co-design and Tools Supporting Smart Learning Ecosystems and Smart Education, Springer, 2020, pp. 263-268.