Raport Ştiinţific şi Tehnic
Etapa a III-a, an 2016:
„Integrarea componentelor”

Aceste rezultate au fost obținute prin finanțare in cadrul programului Parteneriate în domenii prioritare, PN II, derulat cu sprijinul MEN – UEFISCDI, proiect nr. PN-II-PT-PCCA-2013-4-1660:
„Sistem Mobil de Asistare Vocala in Reintegrarea Persoanelor cu Afonii Chirurgicale” SWARA

Denumirea organizaţiei participante in proiect	Acronim organizație	Tip organizaţie	Rolul organizaţiei în proiect (Coordonator/partener)
Universitatea Tehnică din Cluj-Napoca	UTCN	UNI	CO
SC FORTECH SRL	FORTECH	SRL	P1
Universitatea de Medicină şi Farmacie Iuliu Hatieganu	UMF	UNI	P2
Universitatea Babeș-Bolyai	UBB	UNI	P3

Date de identificare proiect

Număr contract:	**Nr. 6 / 2014, PN-II-PT-PCCA-2013-4-1660**
Acronim / titlu:	**SWARA – „Sistem Mobil de Asistare Vocală în Reintegrarea Persoanelor cu Afonii Chirurgicale”**
Titlu raport:	**Raport Științific și Tehnic (Etapa a III-a, 2016)**
Termen:	**Decembrie 2016**
Editor:	**Mircea Giurgiu (Universitatea Tehnică din Cluj-Napoca)**
Adresa de eMail editor:	**Mircea.Giurgiu@com.utcluj.ro**
Autori, in ordine alfabetică:	**Magdalena Chirilă, Mihaela Dinsoreanu, Mircea Giurgiu, Camelia Lemnaru, Silviu Matu, Bogdan Orza, Remus Pop, Adriana Stan**
Ofițer de proiect:	**Andreea Matei**

Rezumat:

Acest document prezintă o sinteză a realizărilor de natură științifică și tehnică obținute în a treia etapă de implementare a proiectului SWARA (perioada Ianuarie – Decembrie 2016). Realizările se referă la:

- dezvoltarea unui sistem de sinteză text vorbire de înaltă calitate bazat pe rețele DNN
- integrarea noului demonstrator de sinteză text vorbire în Cloud și evaluarea lui online
- evaluarea satisfacției utilizatorilor raportat la vocile sintetice în interactiune socială
- dezvoltarea versiunii finale a bazei de date audio-video: în total 21 de ore semnal vocal
- dezvoltarea și aplicarea unor noi metode de adnotare automată a semnalului audio
- realizarea unui demonstrator (pe telefon mobil și online) pentru predicția textului
- diseminarea rezultatelor intermediare.

Activitățile de cercetare desfășurate în etapa a treia de implementare a proiectului (2016) 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 6 livrabile aferente perioadei de raportare, pregătesc pentru etapa următoare cadrul pentru adaptarea sistemului de sinteză la vorbitor, cu crearea de noi voci sintetice, respectiv disponibilizarea sistemului sub forma unui serviciu web de sinteză vocală accesibil în Cloud. De asemenea, acest raport prezinta detalii referitoare la activitatile de management si comunicare, precum si de diseminare a rezultatelor.

Cuprins

1. Activitățile etapei de raportare în contextul general al proiectului

În prima etapa a proiectului (2014) au fost indexate resursele (baze de date de semnal vocal, resurse de text si adnotari de natura lingvistica ale acestora, instrumente software utilizate in procesarea semnalului vocal si a textului aplicate in scopul sintezei din text a semnalului vocal) si s-au elaborat specificațiile funcționale pentru componentele software ale sistemului de sinteza. În etapa aferenta anului 2015 s-au desfășurat activități de cercetare și dezvoltare experimentală pentru modulele sistemului de sinteză, pentru achiziția unui corpus lărgit de semnal vocal, pentru dezvoltarea unui model de predicție rapidă a textului și pentru dezvoltarea unui sistem de sinteză experimental disponibil online, conform cu specificațiile anterior definite și cu obiectivele care sunt prezentate în secțiunea a doua a acestui raport.

În etapa de raportare curentă (2016) s-au desfășurat activități de cercetare și dezvoltare pentru integrarea modulelor sistemului de sinteză dezvoltate în etapa anterioră într-un sistem de sinteză de înaltă calitate a vocii sintetizate. Aparte de integrarea modulelor, creșterea calității semnalului sintetizat s-a obținut atât prin îmbunătățirile aduse în partea de procesare a textului (creșterea performanțelor modulelor de corecție a diacriticelor, modulului de silabificare, respectiv modulului de adnotare a părților de vorbire), dar mai ales prin adoptarea unui nou tip de modelare acustică bazată pe cele mai recente dezvoltări din domeniul rețelelor neuronale de tip DNN (Deep Neural Networks). Tot în aceasta etapă s-au realizat noi înregistrări audio pentru 12 vorbitori, crescând numărul total al vorbitorilor la 17, iar dimensiunea corpusului audio la 21 de ore de vorbire. Dimensiunea acestui corpus și numărul total al vorbitorilor va asigura o bună premiza pentru cercetările din etapa următoare dedicate adaptării sistemului la noi vorbitori, respectiv crearea de noi voci sintetice. Tot în etapa 2016 au fost realizate două demonstratoare pentru predicția rapidă a textului: unul disponibil pe telefon mobil, altul accesibil online ca serviciu în Cloud. De asemenea, s-au realizat activități de diseminare și publicare a rezultatelor la 4 conferințe științifice de specialitate.

În etapa următoare (2017) se vor realiza activități pentru adaptarea sistemului de sinteză la noi vorbitori, elaborarea modelelor pentru crearea de noi voci sintetice, implementarea unui model experimental pentru recunoașterea vizuală a vorbirii pe un set redus de date, integrarea finală a serviciilor de sinteză cu pregătirea acestora pentru valorizare și exploatare de către partenerul industrial, respectiv evaluarea finală cu utilizatorii și elaborarea acordurilor de proprietate intelectuală conform cu contribuția partenerilor la realizarea sistemului final.

2. Gradul de realizare a obiectivelor specifice pentru Etapa a 3-a

Obiectivele specifice ale Etapei a 3-a, „Integrarea componentelor”, impreuna cu gradul lor de realizare, activitatile si principalele rezultate obtinute în anul 2016 sunt prezentate în lista de mai jos.

Obiectiv3.a: *Realizarea sistemului de sinteză text vorbire de înaltă calitate*

Grad realizare: Obiectiv realizat integral

Rezultate:

- noi resurse de date audio de înaltă calitate (12 noi vorbitori înregistrați, aproximativ 21 de ore de semnal audio) și de text folosite pentru antrenarea sistemului
- îmbunătățirea și validarea finală a modelelor de procesare a textului (restaurare automată diacritice, silabificare, predicție parte de vorbire)
- un nou modul pentru antrenarea modelelor acustice din datele audio și text folosind tehnologii actuale performante de tip DNN
- un nou sistem integrat de sinteză text vorbire de înaltă calitate
- un demonstrator online de sinteză text vorbire
- 1 articol științific publicat la conferințe internaționale [2]
- un livrabil (D1.3) cu titlul „Sistem de sinteză text vorbire de înaltă calitate”, care descrie rezultatele menționate mai sus.

Obiectiv3.b: *Evaluarea preliminară a satisfacției utilizatorilor*

Grad realizare: Obiectiv realizat integral

Rezultate:

- metodologie de evaluare a vocilor sintetice și a demonstratorului de predicție a textului pe telefon mobil și online
- rezultate ale evaluării preliminare a satisfacției utilizatorilor si recomandări
- 1 articol în abstract / poster la conferință de prestigiu [1]
- un livrabil (D2.3) cu titlul „Raport asupra satisfactiei utilizatorilor” și care descrie rezultatele menționate mai sus.

Obiectiv3.c: *Dezvoltarea versiunii a 2-a a bazei de date audio-video*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 12 noi vorbitori înregistrați în anul 2016 (aproape 21 ore de vorbire)
- o metodă îmbunătățită de aliniere automată a semnalului vocal cu textul, aplicată în activitatea de adnotare a corpusului audio [3]
- o nouă metodă nesupervizată și independentă de limbă pentru segmentarea acustică; metoda a fost testată cu succes pe corpusuri audio și de text în 3 limbi: Română, Engleză, Italiană [4]
- 1 articol la conferință internațională și care prezintă metoda de aliniere a semnalului vocal cu textul [3]
- 1 articol la conferință internațională și care prezintă metoda inovativă de segmentare acustică [4]
- 1 livrabil (D3.2b) cu titlul „Baza de date vers.2”, care descrie calitativ și cantitativ rezultatele noilor înregistrări audio
- 1 livrabil (D3.3) cu titlul „Baza de date audio adnotată”, care descrie metodele de adnotare automată a corpusului audio.

Obiectiv3.d: *Realizarea demonstratorului pentru redactarea cu predicție a textului*

Grad realizare: Obiectiv realizat integral

Rezultate:

- 1 corpus de text în limba română pentru modelare lingvistică dependentă de contextul utilizatorului (la medic): 11.644 cuvinte, 1.577 propoziții, colectat de către partenerul P2 (UMF) în context medical
- 1 corpus de text în limba română (colectat de P1 – Fortech) pentru modelare lingvistică generală extras din Wikipedia (approx 9GB de date) și crearea de n-grame
- 1 corpus (sursa Google n-gram 1T) achiziționat cu licență și utilizat de către UTCN pentru validare modele lingvistice și posibile aplicații comerciale la P1
- module software pentru predicția rapidă a textului pe mobil și în Cloud
- 1 demonstrator pentru redactarea rapidă a textului accesibil online pe structura de servere în Cloud – disponibil public online
- 1 demonstrator pentru redactarea rapidă a textului pe telefon mobil – disponibil cu licență de utilizare ne-comercială de la UTCN
- 1 livrabil (D4.4.) cu titlul „Demonstrator pentru redactarea rapidă a textului”, care descrie metodele, arhitectura software a aplicațiilor și interfețele cu utilizatorul pentru aplicațiile menționate mai sus.

Obiectiv3.e: *Diseminarea rezultatelor intermediare*

Grad realizare: Obiectiv realizat integral

Rezultate:

- actualizarea dinamică și monitorizarea cu Google Analytics a site-ului
- planul de diseminare pentru anul 2016

- 1 pagină web pentru demonstrarea online a sistemului de sinteză care incorporează predicția de text, respectiv 1 pagină web cu mostre de semnal vocal sintetizate cu sistemul de sinteză bazat pe DNN
- 1 pagină web cu demonstratorul online de predicție rapidă a textului
- 4 articole prezentate și comunicate la conferințe internaționale
- 1 livrabil referitor la activitățile de diseminare (D7.4a) cu titlul „Publicații și tutoriale”

3. Rezultatele etapei și descrierea lor științifică și tehnică

3.1. Sistem de sinteză text vorbire de înaltă calitate

Rezultatele raportate în aceasta secțiune corespund Obiectivului 3.a. din lista de obiective specifice Etapei a 3-a, iar ele sunt descrise in extenso in livrabilul (D1.3) cu titlul „Sistem de sinteză text vrobire de înaltă calitate”. Principalele rezultate se referă la îmbunătățirea modulelor de procesare a textului: restaurarea automată a diacriticelor și dezvoltarea unei metode inovative, performante, privind silabificarea. De asemenea, s-a implementat cu succes sistemul de sinteză de înaltă calitate folosind o nouă tehnologie, DNN (Deep Neural Networks) pentru modelare acustica si aliniere cu textul. Acest sistem a fost testat și fost generată o nouă voce sintetică de înaltă calitate, folosind doar 1000 de propozitii (din total de 3500 din corpus) pentru antrenarea modelelor.

3.1.1. Îmbunătățirea modulului de restaurare a diacriticelor

S-a revizuit modulul de restaurare automată a diacriticelor în vederea îmbunătățirii performanțelor sale și au fost realizate o serie de experimente în vederea selectării potrivite a parametrilor în sistem, după cum sunt descrise mai jos. Astfel, s-a evaluat influența dimensiunii contextului (5, 9, 11 litere) și a importanței atributelor vectorului asupra performanței predictiei (Tabel 3.1.1). Metoda de evaluare folosită este leave-one-out. De asemenea, s-au construit și evaluat modelele de predicție pentru corecția iterativă a diacriticelor în ordinea: i – î (103.443 cazuri), s - ș (42.740 cazuri), t - ț (71.887 cazuri), a – ă – â (117.475 vazuri), context de 5 litere stânga-dreapta, cu / fără diacritice incluse în vector (vezi D1.3 pag 4-7 și Tabel 3.1.2).

Tabel 3.1.1.Evaluarea modelului de restaurare diacritice în funcție de lungimea contextului

Context (litere)	Performanța [%]	Corpus (# vectori)	Information Gain / atribut (linie 1) Pondere atribut / atribut (linie 2)
5	45,87	12.000	0,09 / **0,44 / 0,53** / 0,10 0,02 / **0,12 / 0,15** / *0,02*
9	82,30	80.000	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
11	86,83	99.900	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

Tabel 3.1.2. Rata de corecție a diacriticelor pentru sistemul final

Diacritice	Context 5 litere, fără diacritice incluse		Context 5 litere, cu diacritice incluse	
	J48 individual	J48 iterativ	J48 individual	J48 iterativ
	[%]	[%]	[%]	[%]
i - î	*99,84*	*99,84*	*99,85*	*99,85*
s - ș	*98,95*	*98,95*	*98,94*	*98,94*
t - ț	*98,75*	*98,76*	*98,75*	*98,76*
a – ă - â	***95,10***	***96,12***	***95,13***	***96,13***

Concluziile pentru modelul îmbunătățit de corecție a diacriticelor sunt următoarele:

- odată cu creșterea lungimii contextului crește atât numărul de vectori de antrenare generați din același corpus, dar și performanța sistemului
- s-a decis utlizarea în sistemul final a unui context de 5 litere stânga / dreapta
- s-a evaluat contribuția informațională în decizia diacriticului pentru fiecare atribut (literă) și s-a constatat că litera din dreapta diacriticului aduce aportul informațional cel mai mare
- în consecința, in procesul de predicție, decizia s-a luat prin calculul unei metrici ponderate, cu ponderea cea mai mare atribuită literei din dreapta diacriticului.

3.1.2. Extinderea modulului de silabificare cu noi caracteristici

Modulul de silabificare elaborat în etapa anterioară (2015) a fost extins prin dezvoltarea unui nou mod de codare a vectorului cu trăsături și extinderea lui la un vector 18 dimensional (vezi D1.3 pag 6-8). Corpusul de text folosit este RoSyllabiDict conținând 525534 cuvinte, din care au fost construite în mod balansat 10 seturi de date (Set0..., Set10) pentru antrenare și testare modele de tip Random Forest. Toate seturile de date diferă în ceea ce privește conținutul acestora. O importantă facilitate experimentală, în scopul identificării cofigurației optime de trăsături, este posibilitatea de a include sau exclude anumite trăsături din vectorul caracteristic și a obține 8 tipuri diferite de codări pentru vectorii utilizați, așa cum se prezintă mai jos. Se prezintă o parte din rezultatele obținute (Tabele 3.1.4-6) prin folosirea unui singur set de antrenare pentru mai multe seturi de testare (Tabel 3.1.3). Setul de antrenare conține 6443 de cuvinte și 63937 instanțe.

Codarea tipului de vectori utilizați

CODARE	SEMNIFICAȚIE
NCAV	*adugă informație referitoare la succesiunea de consoane*
NL+NCAV	*adugă informație referitoare la succesiunea de consoane și la numărul de litere al cuvântului*
NL+NV+NC+NCAV	*adugă informație referitoare la succesiunea de consoane și la numărul de litere, consoane și vocale ale cuvântului*
NV+NC+NCAV	*adugă informație referitoare la succesiunea de consoane și la numărul de consoane și de vocale ale cuvântului*
NV+NCAV	*adugă informație referitoare la succesiunea de consoane și la numărul de vocale ale cuvântului*
NL+NV+NCAV	*adugă informație referitoare la succesiunea de consoane și la numărul de litere și vocale ale cuvântului*
NC+NCAV	*adugă informație referitoare la succesiunea de consoane și la numărul de consoane ale cuvântului*
NL+NC+NCAV	*adugă informație referitoare la succesiunea de consoane și la numărul de litere și consoane ale cuvântului*

Tabel 3.1.3. Organizarea seturile de antrenare / testare

Seturi	Număr de cuvinte	Număr de instanțe
Set 1	5044	50144
Set 2	5039	50144
Set 3	5024	49794
Set 4	5034	49963
Set 5	5034	50185
Set 6	5033	50038
Set 7	5034	50184
Set 8	5039	50161
Set 9	5045	50217
Set 10	5033	50042
Set antrenare	6443	63937

Tabel 3.1.4. Acuratețea silabificării dacă se folosesc diacritice

SET	NCAV	NL+NCAV	NV+NC+NCAV	NL+NV+NC+NCAV
SET01	93.48%	94.49%	94.21%	94.07%
SET02	93.11%	93.99%	93.85%	93.49%
SET03	92.91%	93.85%	93.99%	93.37%
SET04	93.15%	94.34%	94%	93.84%
SET05	93.05%	94.26%	94.22%	93.62%
SET06	92.85%	93.78%	93.66%	93.44%
SET07	92.71%	94.12%	93.9%	93.58%
SET08	93.43%	94.36%	94.38%	93.99%
SET09	93.44%	94.55%	94.29%	93.91%
SET10	93.32%	94.28%	94.18%	93.82%

SET	NV+NCAV	NL+NV+NCAV	NC+NCAV	NL+NC+NCAV
SET01	94.35%	94.13%	94.17%	94.51%
SET02	94.03%	93.57%	93.87%	93.63%
SET03	93.75%	93.57%	93.59%	93.75%
SET04	94.16%	93.94%	93.92%	94.26%
SET05	94.32%	93.9%	94.02%	94.22%
SET06	93.86%	93.56%	93.74%	93.82%
SET07	94.24%	93.52%	93.88%	93.74%
SET08	94.03%	93.95%	94.17%	94.34%
SET09	94.59%	94.09%	94.37%	94.33%
SET10	94.18%	93.82%	93.96%	93.9%

Tabel 3.1.5. Acuratețea silabificării la nivel de cuvânt folosind modele Random Forest

Set	Număr de cuvinte	Număr de instanțe	NL+NCAV Acuratețea	NL+NC+NCAV Acuratețea	NV+NC+NCAV Acuratețea
Set1	13747	136768	95.54%	95.49%	95.49%
Set2	16038	159541	95.72%	95.54%	95.62%
Set3	18329	182291	96.28%	96.24%	96.28%
Set4	20620	205249	96.37%	96.35%	96.31%
Set5	22911	228139	96.64%	96.58%	96.69%

Tabel 3.1.6. Acuratețea silabificării în funcție de dimensiunea setului de antrenare

Set	Număr de cuvinte	NL+NCAV Acuratețea	NL+NC+NCAV Acuratețea	NV+NC+NCAV Acuratețea
Set1	1008	86.56	86.76	85.19
Set2	2016	90.14	89.78	90.45
Set3	3024	91.89	91.55	91.98
Set4	4032	92.47	92.87	92.58
Set5	5040	93.35	93.32	93.05

3.1.3. O nouă metodă de silabificare bazată pe frecvența pattern-urilor

Atât rezultatele obținute în etapa anterioară, cît și experimentele preliminare descrise mai sus, indica faptul că sistemele bazate pe învățare supervizată au anumite limitări atunci când trebuie să rezolve sarcina despărțirii in silabe. În consecința, am propus si experimentat o idee nouă de despăarțire în silabe bazata pe statistica celor mai frecvente pattern-uri. Metoda a fost validată in laborator și publicată în [2]. Prezentăm aici pe scurt doar pașii acestei metode.

Pas 1. Se identifică cele mai frecvente pattern-uri de secvențe de silabe (Tabel 3.1.7). Pentru aceasta s-a aplicat algoritmul gapBIDE[1]

Pas 2. Identificarea pattern-urilor de silabe cu gradul cel mai mare de similaritate cu cuvântul de silabisit (vezi algoritm din D1.3, pag. 9-10).

Pas 3. Pentru fiecare pattern creeaza un graf de similaritate cu cuvantul de silabisit

Pas 4. Identificare cale în graf care produce cea mai mare similaritate cu cuvântul [2].

În experimente s-a folosit corpusul RoSyllabiDict cu silabificări în limba română (520.000 de cuvinte), din care 200.000 de cuvinte s-au folosit pentru antrenare, iar pentru validare s-au folosit 10.000 de cuvinte alese aleator din cele rămase. De asemenea, s-au facut experimente cu un corpus în limba Engleza[2], unde 180.000 de cuvinte au fost folosite pentru antrenare, iar 5.000 pentru testare. Rezultatele silabificării sunt superioare metodelor propuse anterior, atât la nivelul de identificare corectă la nivel de silabă, precum și la nivel de întreg cuvânt.

[1] *https://www.hindawi.com/journals/acisc/2012/593147/*

[2] *http://icon.shef.ac.uk/Moby/mhyph.html*

Tabel 3.1.7. Frecvența pattern-urilor de o anumita lungime vs suport din corpusul RoSyllabiDict

Supp.	Len. 2	Len. 3	Len. 4	Len. 5
100	6754	2309	162	2
50	12671	7453	853	47
20	25731	28056	5384	674
10	41126	63388	17812	2940
5	63443	123429	52271	10553
2	104254	248227	186623	66915

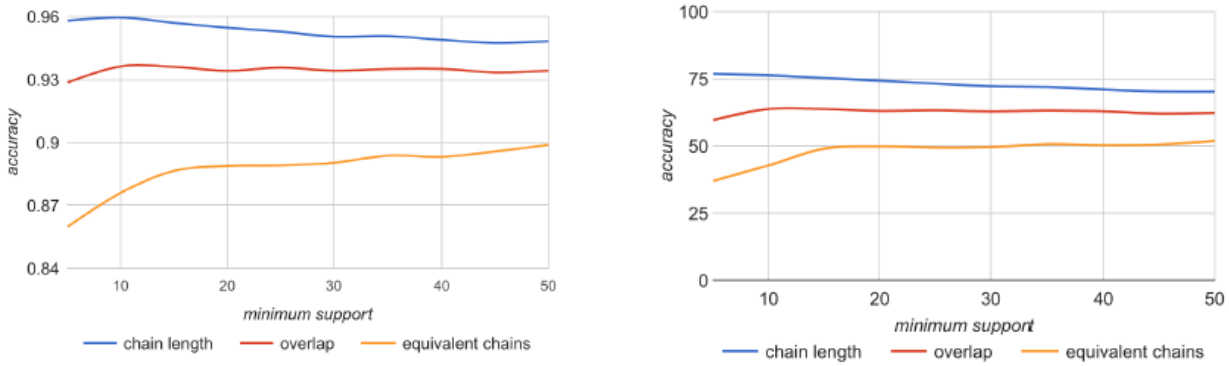


Figura 3.1.1 Acuratetea silabificării la nivel de silaba (stânga), respectiv la nivel de cuvant (dreapta)

Corpus in limba Română

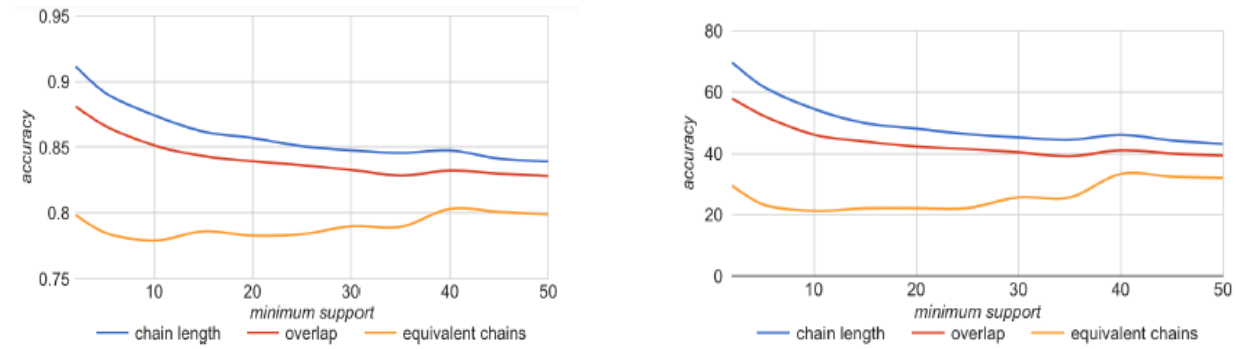


Figura 3.1.2. Acuratetea silabificării la nivel de silaba (stânga), respectiv la nivel de cuvant (dreapta)

Corpus in limba Engleză

3.1.4. Sistem de sinteză text vorbire bazat pe rețele neuronale multistrat de tip DNN

Rețelele neuronale multistrat (DNN - Deep Neural Networks) au devenit deja o soluție importantă pentru rezolvarea proceselor neliniare de complexitate ridicată, cum ar fi recunoașterea și sinteza vorbirii, recunoașterea imaginilor sau procesarea limbajului natural. Companii mari precum Google sau Facebook utilizează DNN-urile pentru aceste procese în mod curent. Odată cu apariția acestora, și metodele de sinteză text-vorbire au adoptat acești algoritmi, inițial pentru a înlocui anumite etape din sinteza bazată pe modele Markov, iar apoi pentru a înlocui complet aceste modele. Pentru noul sistem de sinteză text vorbire s-a utilizat în această etapă de raportare mediul Merlin.

Am utilizat pentru datele acustice de intrare parametrii extrași extrași de vocoder-ul World. Pentru adnotarea cadrelor audio am utilizat etichete de tip HTS, cu aliniere la nivel de fonem. Configurarea rețelei neuronale de tip DNN este de 6 straturi, cu câte 1024 de noduri fiecare și funcție de activare de tip tangentă hiperbolică pentru straturile interne și funcție de activare liniară pentru stratul de ieșire. O problemă importantă a sintezei de voce folosind rețele neuronale multistrat este generarea duratei segmentelor acustice, iar o soluție de compromis a fost utilizarea duratei date din segmentele de antrenare.

Folosind DNN, am antrenat o voce în limba română pornind de la un subset de doar 1000 de propoziții (din motive de cost computațional foarte ridicat pe infrastructura hardware disponibila în laborator) din cele 3500 disponibile pentru fiecare vorbitor în baza de date SWARA. Datele de intrare au fost pregătite conform cu setările implicite. Procesarea de text a fost realizată cu ajutorul modulului dezvoltat în cadrul proiectului, în etapa anterioară.

Rezultatele inițiale ale acestui tip de sinteză sunt promițătoare, iar similaritatea cu vorbitorul este superioară celei obținute cu HTS. Cu toate acestea, există însă anumite artefacte în sinteză, aspecte care sunt încă în studiu. Ca și dezvoltări ulterioare dorim să analizăm influența diferiților parametri ai rețelelor neuronale asupra calității vocii sintetice, precum și să utilizăm un set mai mare de date de antrenare, posibil cu vorbitori multipli.Exemple de secvențe audio sunt disponibile online pentru ascultare / evaluare pe site-ul proiectului[3].

3.2. Raport preliminar asupra satisfacției utilizatorilor

Rezultatele raportate în aceasta secțiune corespund Obiectivului 3.b din lista de obiective specifice Etapei a 3-a, iar ele sunt descrise in extenso in livrabilul (D2.3) cu titlul „Raport asupra satisfacției utilizatorilor”. Rezultate au fost publicate la conferință internațională [1].

3.2.1. Obiective și metodologie

Obiectiv: Evaluarea măsurii în care interlocutorii pacienților cu laringectomie sunt satisfăcuți de calitățile vocilor sintetice (naturalețe, claritate și atitudinea față de a interacționa pe viitor cu persoane ce au o astfel de voce), raportată la vocea unui pacient cu sistem clasic de asistare vocală.

Participanți. În acestă cercetare au fost incluși 82 de participanți, cu vârste cuprinse între 18 și 45 de ani (M = 21.88, SD = 4.73). 69 (84.1%) dintre aceștia au fost de gen feminin. Majoritatea sunt absolvenți de liceu (74.4%), iar un procent mai mic (14.6%) sunt absolvenți de studii de licență. 86.6% dintre aceștia au declarat că au un loc de muncă stabil, în timp ce 12.2.% au declarat că sunt studenți (un procent de 1.2% au declarat că au alte ocupații). În fine, un procent de 86.6% au declarat că provin din mediul urban.

Instrumente. Chestionar online (vezi livrabilul D.2.2, Tabelul 1). Cele 4 voci sintetice folosite au fost generate de echipa UTCN și au cuprins două voci de gen feminin și două de gen masculin, sintetizând nouă mesaje cu trei valențe emoționale (negativă, neutră, pozitivă) și trei contexte (familiar, persoană necunoscută, context medical). Ca termen de comparație s-a optat pentru utilizarea unei voci produse prin proteză vocală. Mesajele sintetizate au conținut fiecare o structură propozițională enunțiativă, o structură propozițională exclamativă, respectiv una interogativă, pentru a controla pentru un potențial efect al tonalităților diferite. Fiecare participant a ascultat vocea pacientului cu proteză vocală și în mod aleatoriu, una din 4 voci sintetice posibile.

Analiza statistică. Comparația vocii pacientului cu vocile sintetice pe trei dimensiuni cheie: claritatea vocii, naturalețea acesteia, și atitudinea față de interacțiuni viitoare cu o persoană care are o astfel de voce. Pentru fiecare dimensiuni, am cumulat scorurile obținute pentru fiecare tip de valență (neutră, pozitivă și negativă) și tip de context (familiar, persoană necunoscută, context medical). S-au utilizat comparații statistice cu teste t și comparații separate folosind parametrul „Cohen d”.

3.2.2. Rezultate și concluzii

Analiza a arătat faptul că vocile sintetice au fost în general percepute ca fiind mai clare decât vocea pacientului, $t(81) = -8.74$, $p < .001$, $d = .92$. În ceea ce privește naturalețea însă, vocea pacientului a fost percepută ca fiind mai naturală decât cele sintetice, $t(81) = 5.19$, $p < .001$, $d = .55$. În ceea ce privește atitudinea față de posibile interacțiuni viitoare, nu am identificat diferențe semnificative statistic între vocea pacientului și cele sintetice, $t(81) = .76$, $p = .451$. (Figura 3.2.1)

[3] *http://speech.utcluj.ro/swara/listeningTest_DNN/*

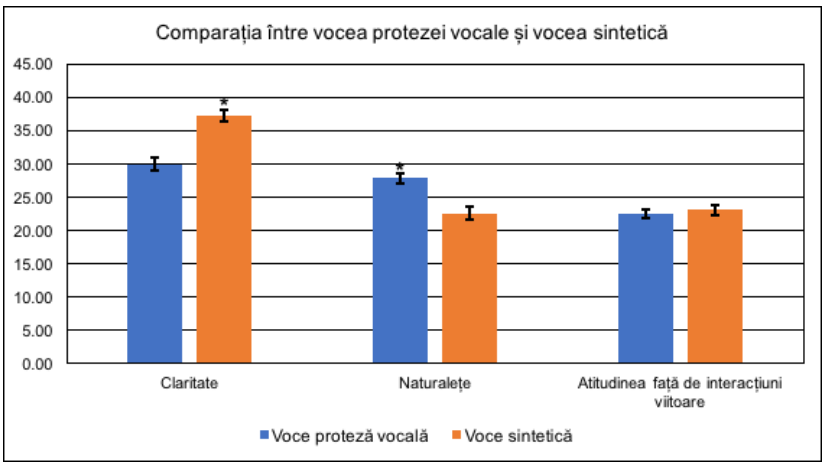
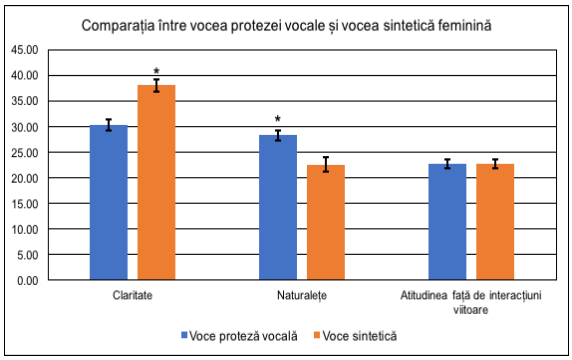


Figura 3.2.1. Comparația între vocea protezei vocale și cea sintetică pe trei dimensiuni



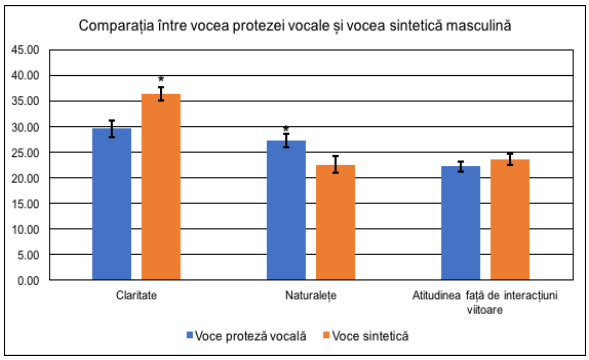


Figura 3.2.2. Comparație între voci, feminin (stânga), masculin (dreapta)

Pentru a pun rezultatele în context, am analizat care este experiența participanților cu metode de asistare vocală. O simplă analiză de frecvență a arătat că 97.6% dintre participanți au interacționat cel mult ocazional cu persoane care folosesc diferite metode de asistare vocală. În fapt, majoritatea participanților, 65.9%, nu au interacționat niciodată cu persoane care folosesc astfel de sisteme. Doar 2 din cei 82 de participanți (2.4%) au interacționat deseori cu astfel de pacienți. Așadar, rezultatele privind satisfacția interlocutorilor cu vocile sintetice se aplică celor care iau contact pentru primele dată cu sisteme de asistare vocală.

În ultima etapă am verificat atitudinile respondenților față de ideea ca un dispozitiv medical să preia o funcție naturală a corpului, respectiv atitudinea față de utilizarea vocilor sintetice pentru pacienții cu laringectomie. Rezultatele au arătat că majoritatea respondenților sunt de acord cu utilizarea dispozitivelor tehnologice care preiau funcțiile naturale ale corpului, un procent cumulat de 82.9% exprimându-și în acordul parțial sau total cu această idee. Doar 17.1% și-au exprimat dezacordul sau indecizia în legătură cu acest subiect. În ceea ce privește atitudinea față de utilizarea vocilor sintetice, un procent de 86.6% au considerat ideea utilizării vocilor sintetice pentru pacienții cu laringectomie ca fiind utilă sau foarte utilă. Doar 13.4% din respondenți au raportat că această idee are o utilitate limitată (au răspuns “Oarecum utilă”). Nici un respondent nu a considerat această modalitate de asistare ca fiind inutilă.

Rezultatele privind satisfacția interlocutorilor în ceea ce privește vocile sintetice, prin raportare la o voce asistată de o proteză vocală, arată următorul pattern: vocile sintetice sunt mai clare, dar mai puțin naturale. Cu toate acestea, nu există diferențe în ceea ce privește atitudinea față de posibile interacțiuni viitoare cu orice tip de voce. Acest lucru sugerează că în ciuda lipsei de naturalețe a vocii, interlocutorii sunt dispuși să se angajeze în interacțiuni cu un pacient care folosește ca voce asistată o voce sintetică, în egală măsură cu o voce asistată de o metodă clasică. Rezultatele sunt consistente, indiferent de genul vocii sintetice (feminin sau masculin, Figura 3.2.2). Investigațiile viitoare vor extinde categoriile demografice incluse în eșantion, pentru a include persoane care au contact mai frecvent cu persoane cu sisteme de asistare vocală și să includă mai multe voci asistate de sisteme clasică în comparațiile cu vocile sintetice.

3.3. Baza de date vers.2

Rezultatele raportate în aceasta secțiune corespund Obiectivului 3.c din lista de obiective specifice Etapei a 3-a, iar ele sunt descrise in extenso in livrabilul (D3.2b) cu titlul „Baza de date vers.2”. O parte din rezultatele referitoare la corpusul audio au fost publicate în articolul [3].

3.3.1. Procedura de înregistrare audio-video

Înregistrările audio au fost realizate cu ajutorul programului Audacity (http://www.audacityteam.org/) la o frecvență de eșantionare de 48KHz și o rezoluție de 16 biți per eșantion. Vorbitorului i-a fost prezentat textul prin intermediul unui laptop alimentat pe baterie, pentru a nu introduce zgomotul rețelei de alimentare. S-a utilizat un microfon unidirecțional AKG C214 cu condensator și diafragmă mare poziționat la aproximativ 15 cm de fața vorbitorului (Figura 3.3.1). Pentru calibrare, vorbitorii au putut să-și asculte vocea prin intermediului unor căști audio și în consecință să poată să-și autoregleze volumul vocal. Citirea textului a fost realizată în mod continuu, fără a opri înregistrarea în pauzele dintre acestea. S-a dorit astfel obținerea unui număr minim de întreruperi ale vorbitorului și astfel o vorbire cât mai naturală și fluentă posibil. Tot în acest scop, vorbitorilor nu li s-a prezentat nici un instructaj privind dicția sau intonația necesară citirii propozițiilor. Segmentarea înregistrărilor a fost realizată ulterior, iar detalii despre acest proces sunt prezentate în livrabilul D3.3 „Bază audio adnotată”.

Participanții la înregistrările realizate in anul de raportare 2016 sunt voluntari, vorbitori nativi de limba română, informați în prealabil de scopul și utilizarea datelor furnizate. Fiecare dintre aceștia a completat un formular de consimțământ informat (vezi Anexa 2 din livrabilul D3.1 „Cadrul tehnic al bazei de date”, raport pe anul 2014). Vârsta lor este cuprinsă între 20 și 33 de ani, iar distribuția pe sexe este prezentată în secțiunea următoare. Nici unul dintre participanți nu a declarat că ar avea vreo deficiență de auz sau de vorbire.

3.3.2. Conținutul corpusului audio-video

Cei 12 participanți la sesiunile de înregistrări au fost rugați să citească un număr de 500 – 1000 de propoziții, în funcție de disponibilitatea lor. După segmentarea și filtrarea datelor audio, s-au obținut 19,639 de propoziții, cu o durată de aproximativ **21 de ore și 11 minute**. O descriere detaliată a înregistrărilor noi, precum și a celor existente este prezentată în Tabelele 3.3.1-2.

Acest volum de date audio este suficient pentru a genera un număr rezonabil de voci pentru sinteză, precum și o voce neutră pentru adaptare de înaltă calitate (activități planificate pentru anul 2017).

Tabel 3.3.1. Tabel descriptiv al setului nou de înregistrări (2016) din baza de date audio

Nr. crt.	Acronim vorbitor	Sex (F/M)	Durată	Număr de propoziții
1.	*DDM*	*F*	*41’27”*	*996*
2.	*FDS*	*M*	*31' 25"*	*996*
3.	*HTM*	*F*	*38' 27"*	*981*
4.	*PCS*	*M*	*38' 29"*	*996*
5.	*PMM*	*F*	*33' 28"*	*921*
6.	*SDS*	*M*	*35' 25"*	*996*
7.	*SGS*	*M*	*31' 23"*	*994*
8.	*TSS*	*M*	*35' 26"*	*996*
9.	*TIM*	*F*	*38' 24"*	*973*
10.	*RMS*	*M*	*39' 29"*	*996*
11.	*IPS*	*M*	*33' 24"*	*996*
12.	*CAU*	*F*	*42' 29"*	*996*

Tabel 3.3.2. Tabel cu înregistrările existente în versiunea 1 (2015) a bazei de date audio

Nr. crt.	Acronim vorbitor	Sex(F/M)	Durată	Număr de propoziții
1.	*DCS*	*F*	*46' 33"*	*1498*
2.	*BAS*	*F*	*40' 27"*	*1495*
3.	*PSS*	*M*	*36' 24"*	*1486*
4.	*EME*	*F*	*42' 35"*	*1493*
5.	*SAM*	*F*	*38' 31"*	*1493*

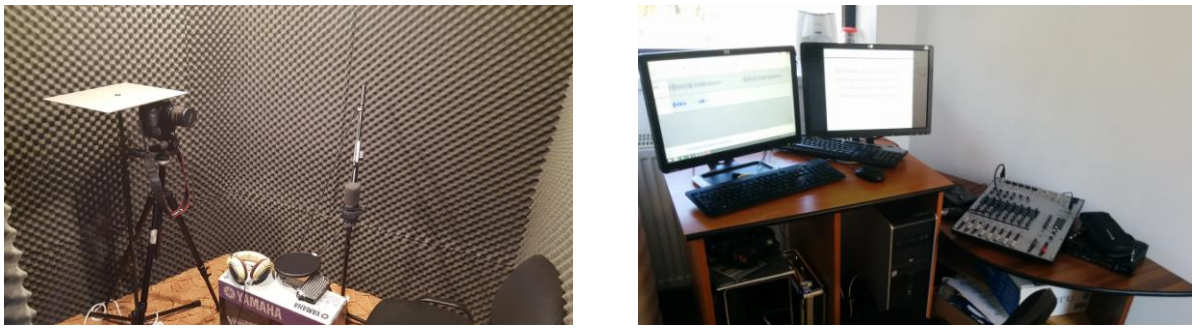


Figura 3.3.1. Aranjamentul echipamentelor (stânga), respectiv panoul de comanda (dreapta) pentru realizarea înregistrărilor audio - video

Pe partea de înregistrări video, participanții au fost reticenți privind utilizarea imaginii lor în baza de date publică, astfel că o singură persoană și-a dat acordul pentru disponibilizarea publică a înregistrărilor sale video. Datele video colectate însumează aproximativ 45 de minute și urmează să fie procesate și utilizate în etapa următoare (2017) în antrenarea unui sistem de recunoaștere vizuală a vorbirii.

3.4. Baza de date adnotată

Rezultatele raportate în aceasta secțiune corespund Obiectivului 3.c din lista de obiective specifice Etapei a 3-a, iar ele sunt descrise in extenso in livrabilul (D3.3) cu titlul „Baza de date adnotată”. Rezultatele au fost diseminate la 2 conferințe internaționale prin articolele [3] și [4].

3.4.1. Procedura de segmentare la nivel de propoziție

Rezultatul brut al sesiunilor de înregistrare necesită o procesare manuală costisitoare ca timp și experiență de editare audio. În acest proiect ne-am propus să automatizăm procedurile de aliniere corectă a textului cu înregistrarea audio și să minimizăm efortul din partea persoanelor care se ocupă cu validarea datelor. Ca urmare, am creat o procedură (vezi Anexa la D3.3) și o serie de script-uri Python care să faciliteze segmentarea rapidă a secvențelor audio de către persoanele care le adnotează. Astfel că, singurele cerințe au fost ca persoanele să marcheze, folosind software-ul Wavesurfer[4], propozițiile rostite corect de către vorbitor (vezi Figura 3.4.1). În cazul în care o propoziție nu are nici o rostire care să corespundă exact cu textul furnizat, însă există totuși mici devieri de la text, persoana care adnotează trebuie să corecteze textul și să noteze aceste propoziții într-un document separat.

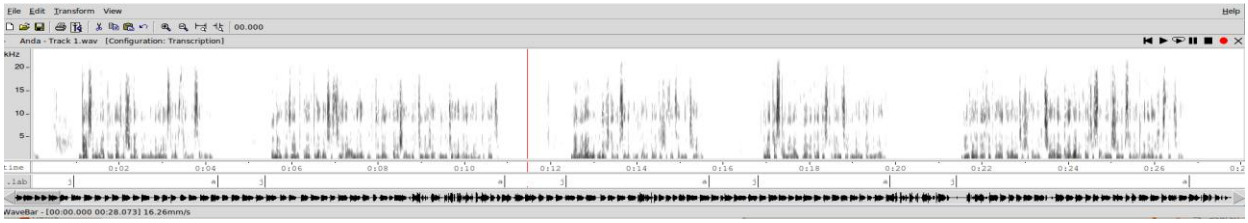


Figura 3.4.1. Exemplu de adnotare grosieră (separator roșu) la nivel de propoziție în Wavesurfer.

[4] *https://sourceforge.net/projects/wavesurfer/*

Având în vedere faptul că și în urma segmentării manuale pot să existe erori în adnotare, o altă persoană a făcut o nouă verificare a rezultatelor obținute. Cu toate acestea, au existat propoziții care nu au fost rostite complet sau care au fost ignorate în citire de către vorbitor, astfel încât cantitatea și calitatea datelor obținute diferă de la un vorbitor la altul. O indexare a acestor date este prezentată în livrabilul D3.2b „Baza de date vers.2".

3.4.2. O nouă procedură de adnotare automată la nivel de fonem

Pentru a crea un sistem de sinteză text-vorbire de uz general, precum și pentru sistemele de recunoaștere vizuală a vorbirii, datele de antrenare trebuie adnotate la nivel de fonem. Acest lucru este extrem de greu de realizat manual și necesită un timp îndelungat de procesare a datelor. S-a implementat o modalitatea de adnotare semi-supervizată a secvențelor audio, urmând ca adnotarea secvențelor video să fie realizată pe baza sincronizării dintre cele două fluxuri de date.

Pentru adnotarea la nivel de fonem a corpusului audio creat în cadrul proiectului, am folosit facilitățile modulului de antrenare de modele acustice din cadrul sistemului de sinteză HTS. Acesta are capacitatea de a utiliza date de intrare ne-aliniate temporal, urmând ca în urma estimărilor și a re-estimărilor parametrilor de modelare, să poată realiza o aliniere temporală de încredere. Pe baza transcrierilor fonetice am antrenat voci de sinteză (Tabel 3.4.1) pentru fiecare vorbitor în parte folosind ca și aliniere temporală la nivel de fonem o simplă segmentare grosieră, echidistantă a duratei segmentului audio în funcție de numărul de foneme conținut.

În Figura 3.4.2. este prezentat în PRAAT[5] rezultatul acestei alinieri temporale la nivel de fonem. Se poate observa prin comparație cu discontinuitățile din spectrogramă că erorile de aliniere sunt relativ reduse. Prin analiza calității vocilor rezultate[6], se poate concluziona și faptul că acestea nu au un impact semnificativ. Există însă posibilitatea ca aceste erori de aliniere să aibă un impact asupra sistemului de recunoaștere vizuală a vorbirii, însă această cercetare este momentan în proces de dezvoltare și evaluare în laboratorul nostru.

Tabel 3.4.1. Vocile de sinteză au fost create folosind HTS cu următoarele setări:

Parametri	Setare valorică
Număr de coeficienți cepstrali:	*60*
Număr de coeficienți de aperiodicitate:	*26*
Frecvența fundamentală:	*Log(F0)*
Fereastra coeficienți delta și delta-delta:	*3 ferestre*
Cadrele de analiză:	*Aliniate cu perioada fundamentală instantanee*
Numărul de stări ale modelelor Markov:	*6*

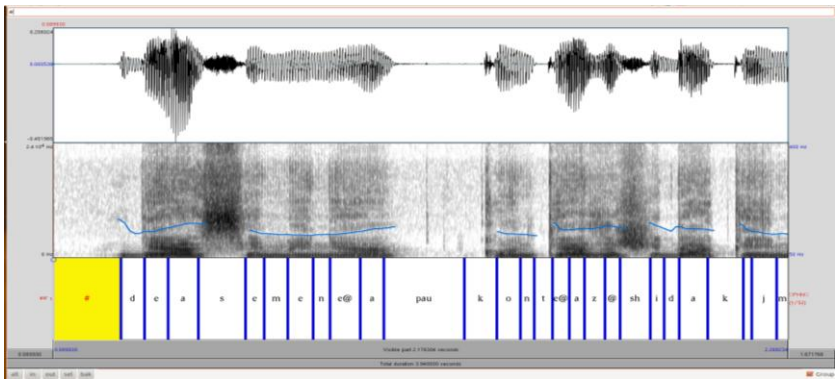


Figura 3.4.2. Exemplu de rezultat al alinierii fonetice folosind modelele acustice create în HTS

[5] *http://www.fon.hum.uva.nl/praat/*

[6] *https://swara.fortech.ro/audio/speech/*

3.4.3. O îmbunătățire a metodei de aliniere a textului cu semnalul vocal

Pentru a evita segmentarea manuală la nivel de propoziție a unor secvențe audio de durată mare (audiobook sau podcast) și astfel să se permită utilizarea unor astfel de resurse ca și date de intrare pentru un sistem de sinteză, am dezvoltat o nouă metodă de aliniere temporală a segmentelor audio cu transcrierea lor ortografică (Figura 3.4.3).

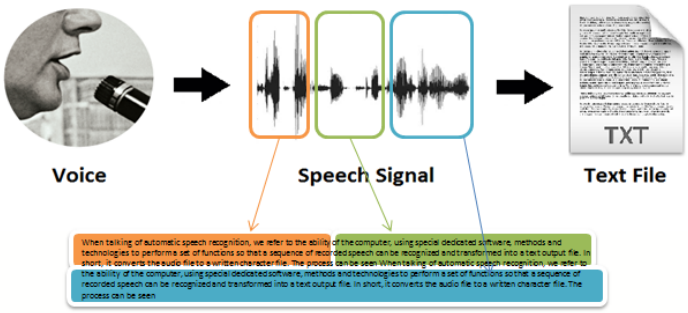


Figura 3.4.3. Metoda de aliniere a segmentelor audio cu transcrierea lor ortografică

Metoda propusă aici îmbunătățește și extinde funcțional un toolkit dezvoltat anterior în cadrul grupului nostru de cercetare, toolkit denumit ALISA[7]. ALISA are ca obiectiv alinierea la nivel de propoziție a secvențelor audio atunci când avem la dispoziție o transcriere ortografică a acesteia, transcriere care nu este neapărat exactă și care poate să conțină cuvinte lipsă sau inserate. Rezultatele obținute în mod tradițional de ALISA sunt de aliniere a aproximativ 75% din datele de intrare, cu o rată de eroare la nivel de propoziție de sub 7% și de 0.5% la nivel de cuvânt. Cu toate acestea, am observat faptul că rata de eroare de 0.5% este datorată într-o foarte mare proporție erorilor de aliniere de la începutul și finalul propozițiilor. Iar aceste erori generează în cadrul sistemelor de sinteză segmente de liniște cu artefacte.

Într-o primă etapă, prin concatenarea segmentelor audio s-a urmărit restrângerea rețelelor de decodare acustică la secvențe de text existente în transcrierea ortografică disponibilă. Rezultatele acestei metode însă nu sunt mai bune decât cele furnizate de ALISA, iar un motiv important pentru acest lucru este faptul că dacă rețeaua de decodare acustică generează o eroare în primele cadre de aliniere, această eroare se propagă până la finalul propoziției și se obține astfel o transcriere eronată a întregii secvențe audio.

Cea de-a doua metodă se bazează pe un algoritm de detecție liniște-vorbire (VAD – Voice Activity Detection). Am utilizat VAD-ul în combinație cu opțiunea rețelelor de decodare acustică de a furniza mai multe ipoteze de recunoaștere. Astfel că s-au utilizat primele trei cele mai probabile ipoteze de decodare. Acestea includ și alinierea temporală a secvenței audio cu modelele fonetice antrenate, furnizând markeri temporali de început și final al acesteia. Pe baza markerilor temporali furnizați de VAD s-a selectat ulterior ipoteza de decodare care e cea mai apropiată de aceștia. Rezultatele acestei metode au îmbunătățit rezultatele ALISA cu aproximativ 10% (Tabel 3.4.2). Trebuie menționat faptul că pentru rezultatele de referință nu s-au utilizat modelele acustice cele mai bune din cadrul ALISA și aceasta deoarece ele necesită resurse computaționale mult mai extinse decât dispunem în cadrul acestui proiect. Astfel că s-a dorit ca prin aceste metode simple de îmbunătățire a rezultatelor de aliniere să se reducă complexitatea și timpul de procesare necesar. Metoda este publicată în [3].

Tabel 3.4.2. Comparație între rezultatele de aliniere obținute de ALISA, metoda de concatenare și cea bazată pe VAD

Metodă	Eroare la nivel de propoziție [%]	Eroare la nivel de cuvânt [%]
Referință ALISA	*11.70*	*0.6*
Concatenare	*16.70*	*0.94*
VAD	***10.49***	***0.41***

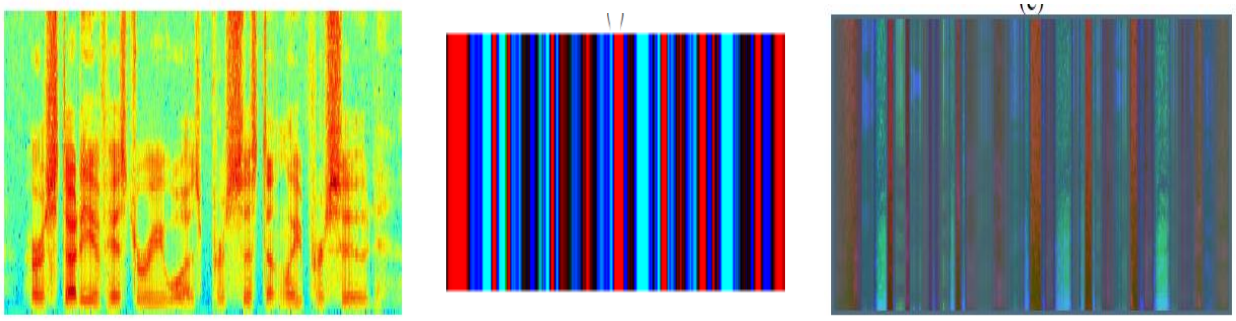
[7] *http://simple4all.org/product/alisa/*

3.4.4. O nouă metodă de segmentare fonetică bazată pe procesări de imagini

Deoarece segmentarea fonetică a secvențelor audio necesită de cele mai multe ori un efort destul de mare, am propus și dezvoltat în etapa de raportare o nouă metodă pentru segmentarea fonetică nesupervizată și care să aibă ca și date de intrare doar datele audio. Cu alte cuvinte, o astfel de metodă urmărește doar să identifice granițele dintre elementele fonetice componente. Metoda propusă de noi, utilizează o idee inovatoare de tratare a spectrogramei ca și o imagine în cadrul căreia discontinuitățile pot fi considerate ca fiind frontierele fonetice.

Detecția discontinuităților din imagine s-a realizat prin intermediul unui algoritm de netezire a imaginilor (Figura 3.4.4) inclus în software-ul GIMP2.8.[8] Având ca punct de pornire doar segmentul audio, principiul metodei este de a calcula două măsuri de discontinuitate (Figura 3.4.5), atât din spectrul acustic, cât și din imaginea spectrogramei. Măsura de discontinuitate acustică se bazează pe distanța Manhattan normalizată, calculată între coeficienții cepstrali din cadre de analiză consecutive. Ca și măsură de discontinuitate vizuală am folosit distanța de culoare CIEDE2000[9] calculată între fâșii verticale consecutive ale visiogramei. Metoda este diseminată în [4].

Pentru evaluarea metodei propuse s-au folosit corpusurile TIMIT (Engleză), respectiv CLIPS (Italiană). Ambele corpusuri sunt adnotate manual la nivel de fonem, astfel că avem o referință pentru evaluarea metodei propuse. Rezultatele experimentale obținute pe ambele seturi de date confirmă eficiența metodei și o clasează în topul celor mai bune metode de segmentare nesupervizată. Tabelele 3.4.3 și 3.4.4 prezintă aceste rezultate.



(a) (b) (c)

Figura 3.4.4. (a) Spectrograma, (b) pattern-ul negativ, (c) visiograma pentru un segment audio

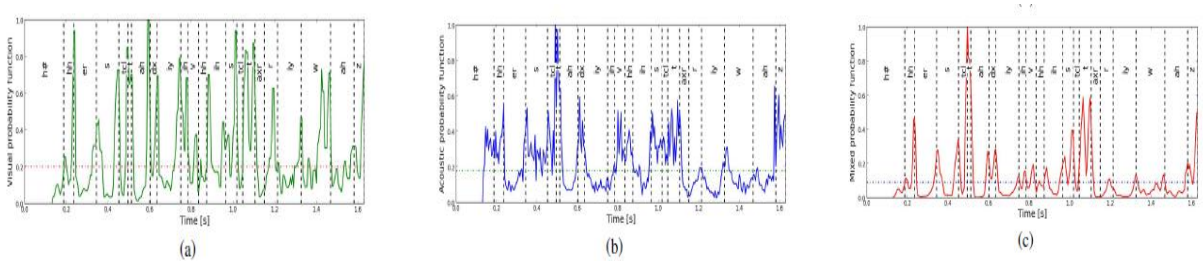


Figura 3.4.5. Valorile metricilor de discontinuitate vizuală (a), acustică (b), mixtă (c)

Tabel 3.4.3. Rezultate experimentale ale metodei (F-value și R-value)

Parametri	Toleranța [ms]	Acuratețe [%]	Supra segmentare [%]	F-value	R-value
Acustici	*20ms*	*86.03*	*53.88*	*0.68*	*0.48*
Vizuali	*20ms*	*76.27*	*12.83*	*0.72*	*0.74*
Combinat	*5ms*	*47.94*	*-3.26*	*0.50*	*0.57*
Combinat	*10ms*	*63.80*	*-3.26*	*0.65*	*0.70*
Combinat	*20ms*	*75.59*	*-3.26*	*0.76*	*0.80*
Combinat	*50ms*	*83.06*	*-3.26*	*0.84*	*0.86*

[8] *http://www.gimp.org*

[9] *http://www.brucelindbloom.com/index.html?ColorDifferenceCalc.html*

Tabel 3.4.4.. Compararea rezultatelor metodei propuse cu alte metode de referință pentru segmentarea nesupervizată

Metodă	F-value	R-value
Dusan et al (2006)	*0.71*	*0.73*
Esposito Aversano (2005)	*0.75*	*0.74*
Khanaga et al (2014)	*0.73*	*0.77*
Rasanen et al (2009)	*0.76*	*0.78*
Estevan et al (2007)	*0.76*	*0.80*
Metoda propusă	***0.76***	***0.80***

3.5. Demonstrator pentru redactarea cu predicție a textului

Rezultatele raportate în aceasta secțiune corespund Obiectivului 3.d din lista de obiective specifice Etapei a 3-a, iar ele sunt descrise in extenso in livrabilul (D4.4) cu titlul „Demonstrator pentru redactarea cu predicție a textului”.

3.5.1. Demonstrator implementat pe telefon mobil

Au fost implementate două versiuni ale demonstratorului pentru predicția textului pe telefon mobil. Prima versiune folosește predicția la nivel de propoziție și se utilizează în situații in care contextul este foarte restrictiv, dar dorim predicții foarte rapide la nivel de fraza, doar prin introducerea a câtorva cuvinte de referință (Figura 3.5.1). În această versiune sunt create trei module funcționale: baza de date cu mesaje adaptabile la contextul utilizatorului, interfața cu utilizatorul, respectiv API-ul de sinteză vocală pentru testarea preliminară a aplicației

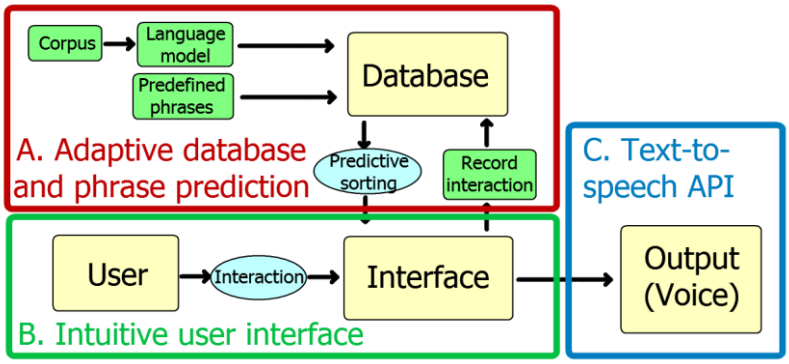


Figura 3.5.1. Metoda de predicție a textului pentru prima versiune a demonstratorului pe mobil

Baza de date cu mesaje adaptabile la contextul utilizatorului conține două părți: 1) un larg set de fraze predefinite care corespund unei varietăți de situații contextuale în care se poate găsi utilizatorul (acasă, cumpărături, doctor, autobus, sau generic – vezi detalii in D4.4. pag5), și 2) un set de fraze expandabil, menit să particularizeze aplicația pentru un anumit mod de vorbire al utilizatorului, respectiv pentru situații comunicaționale pe care nu le-am avut în vedere în această etapă preliminară (Figura 3.5.2).

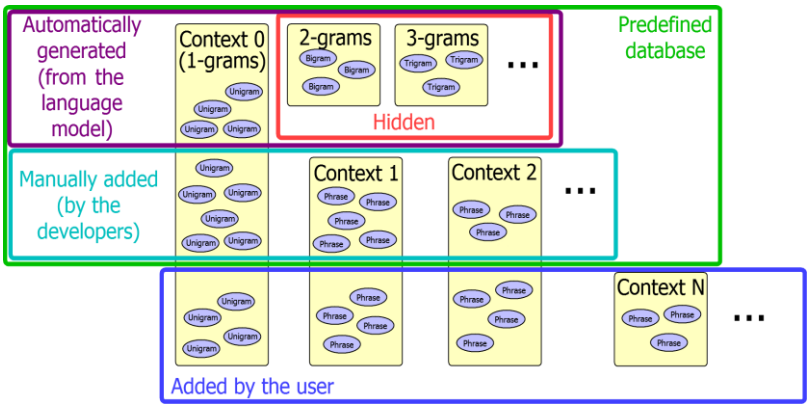


Figura 3.5.2. Ilustrarea principiului de generarea dinamică și în funcție de context a frazelor

Deoarece acest sistem de predicție nu este suficient de flexibil, l-am modificat astfel că a doua versiune demonstrativă să poată realiza predicții la nivel de cuvânt sau succesiune de câteva cuvinte (pentru maxim 2-3 cuvinte aplicația este eficientă). Metoda folosită pentru implementarea acestei versiuni a demonstratorului pe telefon mobil se bazează pe modelarea limbajului natural folosind toolkit-ul ”Kyoto Language Toolkit”[10] cu scopul de a genera un model pe bază de n-gram (Figura 3.5.3).

Pentru a doua versiune demonstrativă, corpusul utilizat pentru antrenarea modelelor de limbaj este un corpus cu termeni medicali colectat de partnerul P2 (UMF) în perioada Ianuarie – Martie 2016. El conține dialoguri tipice medic – pacient, astfel vom genera modele n–gram dependente de contextul medical. Se observă din exemplus de mai jos ca principala caracterstică sunt propozițiile relativ scurte si cu mesaj foarte bine definit, ca atare nu este recomandat să folosim n-grame mai mare de ordin 2 sau 3.

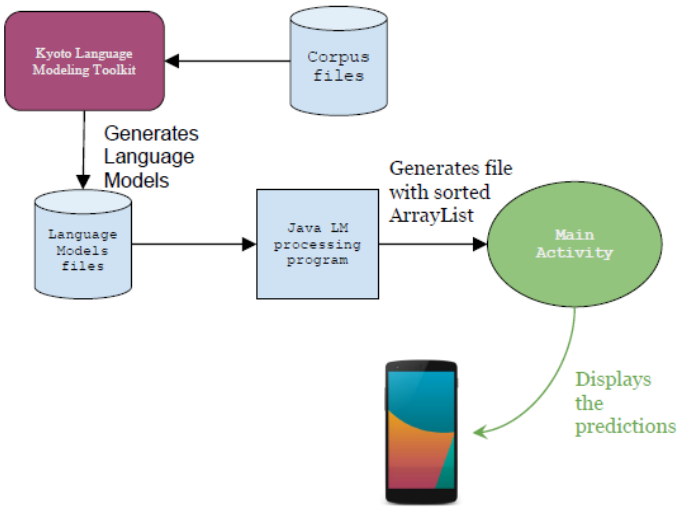


Figura 3.5.3. Fluxul de procesări pentru metoda de creare a modelului de limbaj

Exemplu de dialog Medic (M), Pacient (P).
M: Ce vă supără?
P: Respir mai greu.
M: Trageţi aerul în piept mai greu sau îl daţi afară mai greu?
P: Trag aerul mai greu în piept. / P: Dau aerul mai greu din piept. / P: Respir greu. / P: Respir foarte greu.
M: Când respiraţi mai greu?
P: La început respiram mai greu când lucram mai mult, dar acum obosesc şi respir greu la eforturi din ce in ce mai mici. / P: Am început să respir mai greu şi când nu fac efort. / P: Ca să pot respira mai bine trebuie să mă aşez cu corpul aplecat în faţă. / P: La început oboseam când lucram mai mult. / P: Acum obosesc şi respir greu la eforturi din ce in ce mai mici.
P: Am inceput să respir mai greu şi când nu fac efort.

Tabelul 3.5.1. Statistici privind corpusul cu dialoguri medicale

Criteriu de analiză	Valoare numerică
Propoziții	*1577*
Total cuvinte	*11644*
Cuvinte unice	*3754*
Număr mediu de cuvinte / propoziție	*7,3*

[10] *http://www.phontron.com/kylm/*

Pentru arhitectura software a aplicaţiei s-au căutat soluţii practice de a afişa predicţiile și s-a ajuns la decizia de a folosi un *MultiAutoCompleteTextView* care constă într-un câmp în care se poate introduce text și care are o listă expandabilă dedesubt. Această listă expandabilă este folosită pentru afişarea sugestiilor de predicţie pe măsură ce utilizatorul introduce textul. S-a folosit un *StringTokenizer* pentru separarea cuvintelor, iar apoi un adaptor de tip *ArrayAdapter* și un *MultiAutoCompleteTextView* pentru a include predicţiile sortate după probabilităţile generate de modelul de limbaj. Modelul de implementare este prezentat în Figura 3.5.4.

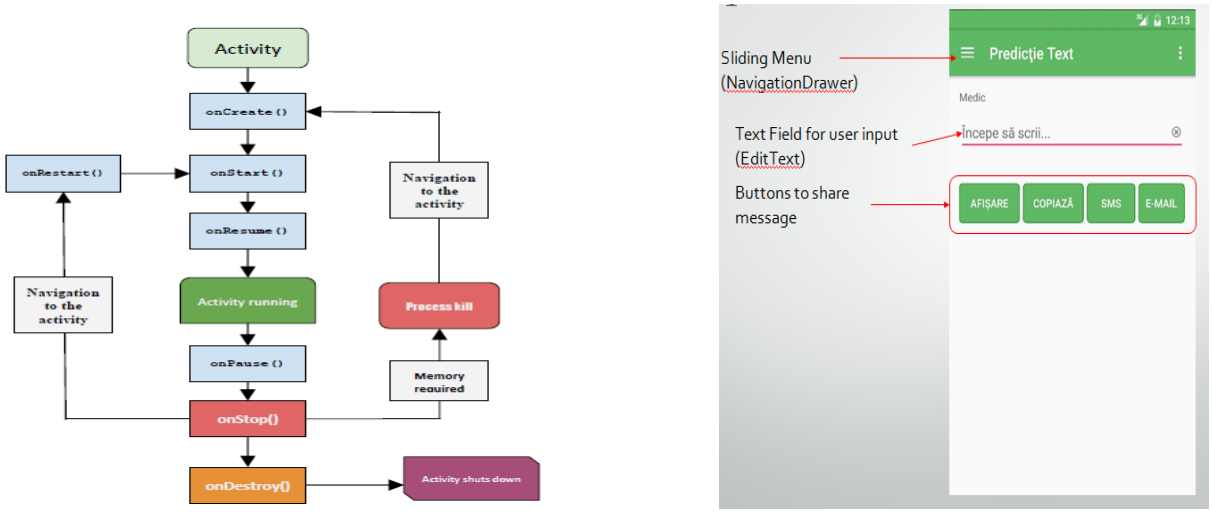


Figura 3.5.4. Diagrama de clase și proiectarea elementelor vizuale pentru aplicația mobil

Pentru demonstratorul pe telefon mobil interfețele utilizator au fost proiectate și implementate de așa natură încât sa permită următoarele funcționalități:

1. o modalitate de definire a contextului (general, cumpărături, la doctor, etc) și a propozițiilor (mesajelor) relevante,
2. o modalitate dinamică de căutare și selectare a mesajelor în funcție de context,
3. o metodă inteligentă de personalizare și prezentare a mesajelor în funcție de profilul utilizatorului,
4. o interfață intuitivă pentru controlul integrat al bazei de date,
5. o interfață pentru predicția rapidă a textului la nivel de cuvânt și
6. sinteza vocală a mesajelor selectate.

Figura 3.5.4. Interfața pentru predicție automată la nivel de propoziție

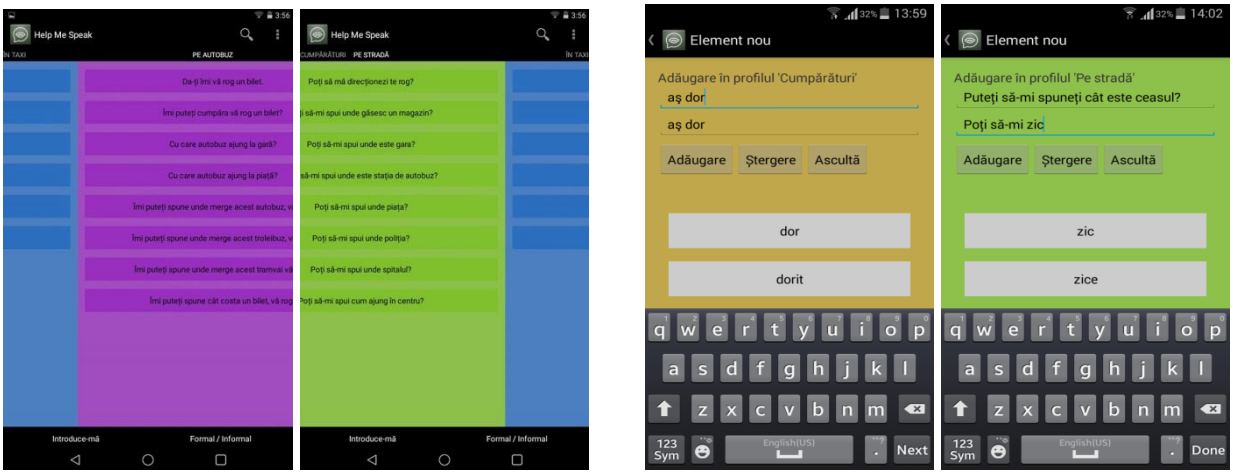


Figura 3.5.5. Interfețe pentru definirea contextului utilizatorului, respectiv customizare propoziții

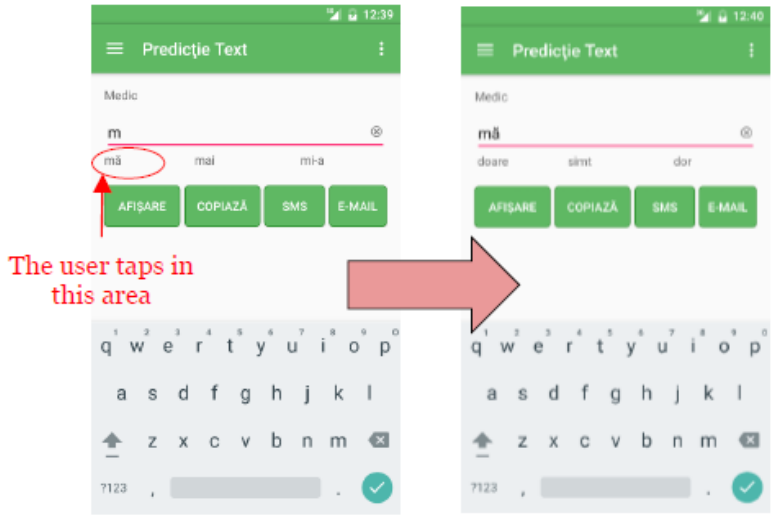


Figura 3.5.6. Demonstratorul pentru predicția cuvintelor pe telefon mobil folosind bi-grame

Acest demonstrator a fost testat cu succes, fără nici un fel de erori sau blocaj al aplicației, pe un număr de 20 de persoane. În plus, li s-a solicitat utilizatorilor să ofere sugestii de îmbunătățire. Astfel, aceste sugestii au fost implementate în soluția finală. Ele se referă la: a) posibilitatea de a exporta textul într-un SMS (13 sugestii / 20), b) posibilitatea de a exporta textul într-un eMail (3 sugestii / 20), c) posibilitatea de a copia textul in Clipboard (2 sugestii / 20), d) posibilitatea de a salva textul într-un fișier (1 sugestie / 20).

3.5.2. Demonstrator online în Cloud

Arhitectura software a demonstratorului online pentru predicția rapidă a textului este compusă din două părți esențiale: partea de procesare pe server în Cloud, respectiv partea de client (Figura 3.5.7). Pe partea de server, pe baza corpusurilor de text sunt generate modelele de limbaj (n-grame[11]) printr-o procedură offline care utilizează modulele MITLM[12] pentru procesarea limbajului natural. Demonstratorul pentru predicția textului în varianta online este disponibil la adresa: https://swara.fortech.ro/audio/predictionTests/medical.html. Pentru a asigura accesul în timp real la predicții, modele de limbaj au fost transpuse din format ARPA[13] într-o bază de date cu acces ultra-rapid in Cloud.

[11] *https://lagunita.stanford.edu/c4x/Engineering/CS-224N/asset/slp4.pdf*

[12] *https://github.com/mitlm/mitlm*

[13] *http://www1.icsi.berkeley.edu/Speech/docs/HTKBook3.2/node213_mn.html*

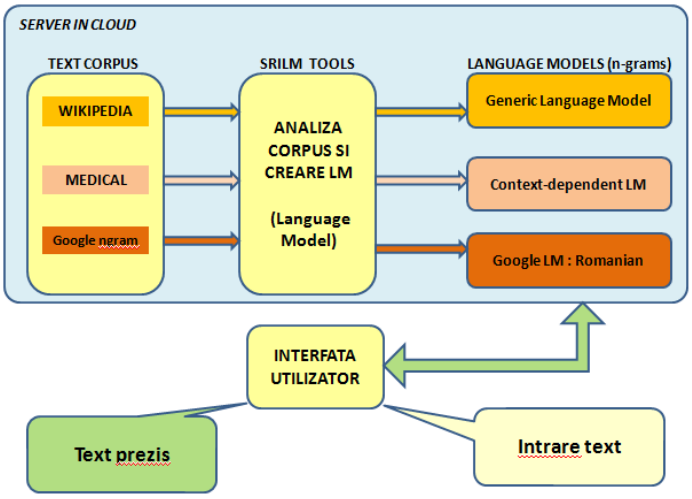


Figura 3.5.7. Arhitectura demonstratorului online pentru predicția textului

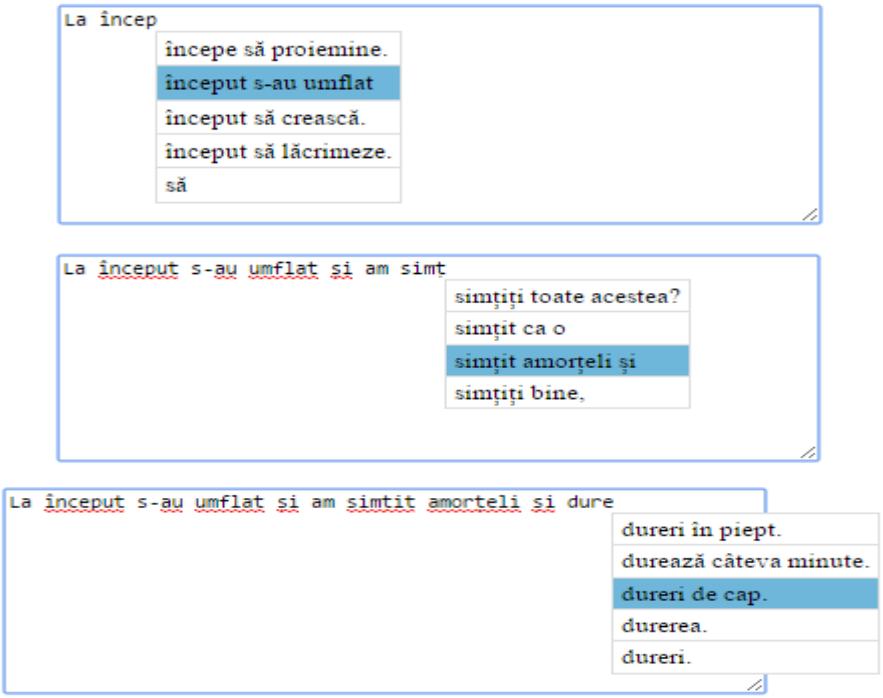


Figura 3.5.8. Scenariu de predicție cu 2-gram, 3 cuvinte prezise

Tabel 3.5.1. Setarea parametrilor pentru demonstratorul online

Text parametru	Semnificație
Cuvinte pentru predicție	*Se selectează numărul de cuvinte pe baza cărora să se realizeze predicția*
Număr maxim de rezultate	*Se precizează numărul maxim de predicții care vor fi afișate într-o lista în ordinea probabilităților de apariție*
Număr cuvinte prezise	*Câte cuvinte succesive se dorește a fi prezise*
Doar modelul pacientului	*Prin selectare se foloseste doar modelul generat pe baza corpusului medical, altfel un model general*
Text	*Se introduce caracter cu caracter textul si se selecteaza predicțiile care convin pentru editare.*

Modulul de predicție a textului folosind pentru generarea n-gramelor text extras din Wikipedia a fost încorporat și în demonstratorul de sinteză text-vorbire[14]. Ilustrarea procesului de predicție pentru textul *„La început s-au umflat și am simțit amorțeli și dureri de cap”:*

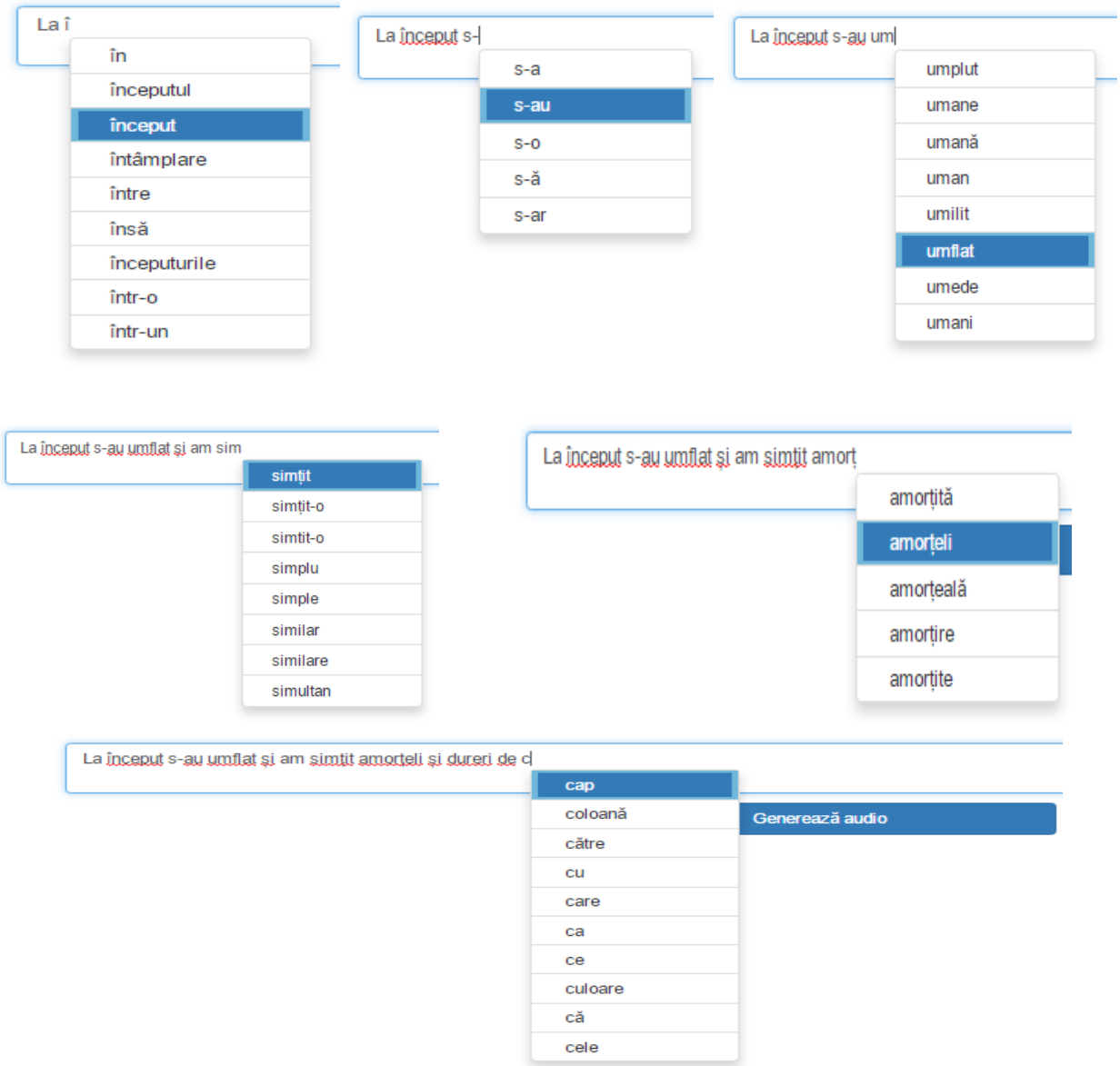


Figura 3.5.9. Scenariu de predicție text integrat în sistemul de sinteză (Atenție: predicția e realizată doar la nivelul unui singur cuvant !).

Tabelul 3.5.2. Statistici privind corpusul de text extras din Wikipedia

Criteriu de analiză	Valoare numerică
Propoziții	*2.414.103*
Total cuvinte	*44.421.293*
Cuvinte unice	*1.072.803*
Număr mediu de cuvinte / propoziție	*18,4*

[14] *https://swara.fortech.ro/audio/speech/*

4. Management si comunicare

Activitățile de management au fost orientate, în special către managementul grupurilor de cercetare constituite în jurul obiectivelor etapei și a interacțiunii dintre acestea în vederea integrării diferitelor componente software în sistemul final. Astfel, întâlniri ale întregului parteneriat au fost realizate în 22.01.2016 (pentru planificarea anuală) și în 20.10.2016 (pentru pregătirea raportării anuale). Datorită bunei funcționări a comunicarii prin Skype și eMail, dar și a faptului că deja activitățile și problematica de cercetare sunt acum bine cunoscute, grupurile de cercetare au avut doar întâlniri față in față trimestrial.

Este de notat faptul ca s-a asigurat o bună comunicare și coordonare cu responsabilii financiari ai partnerilor, astfel ca documentele administrative legate de întocmirea actului aditional, a raportării financiare si a auditului pe 2016 a fost eficientă. Și în acest an se remarcă interesul și implicarea partenerului industrial (P1), atât în colaborarea cu coordonatorul, dar și în organizarea reuniunilor de lucru bilaterale între cercetătorii mai tineri.

Din punct de vedere administrativ s-au primit 5 tranșe de avans cu o regularitate adecvată, s-a întocmit 1 act aditional la începutul anului (pentru reducerea bugetului pe anul 2016 și decalarea activităților până la 30.09.2017) și s-au derulat câteva activități privind achizițiile de obiecte de inventar necesare activităților de zi cu zi în proiect.

5. 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 de integrare a componentelor în sistemul de sinteză text vorbire pentru asistare vocală s-au identificat canalele de diseminare si s-a acționat pe următoarele direcții: a) actualizarea dinamică a paginii web cu rezultatele obținute incremental in proiect, inclusiv cu secțiuni demonstrative; b) elaborarea unui plan de diseminare pentru anul 2016; c) publicarea rezultatelor științifice la conferințe internaționale de prestigiu în domeniul proiectului (tehnic, tehnologii medicale asistive), d) crearea unor pagini web dedicate pentru demonstrarea online a sintezei, precum și pentru predicția rapidă a textului.

5.1. Pagina web a proiectului

Pagina web a proiectului[15], are un conținut dinamic, adaptat cu realizările din proiect, astfel ca tot la această adresă se pot accesa mostre demonstrative cu semnale sintetizate, articolele științifice publicate, livrabilele cu caracter public, precum si legături catre serviciile web care se vor dezvolta. Serviciul de monitorizare automată a site-ul, Google Analytics, a continuat să funcționeze și în 2016. Detalii specifice sunt prezentate în livrabilul D7.4a „Publicații și tutoriale”.

Tabel 5.1.1. Situația comparata privind accesul paginii web “Swara”

Criteriu de analiză	Valoare 2015	Valoare 2016
Număr de sesiuni	301	*1.153*
Număr de utilizatori	244	*842*
Număr pagini vizualizate	309	*1.605*
Rata sesiuni noi în intervalul de analiză	80,17%	*72,94%*
Rata de creștere în intervalul de analiză	98,01%	*78,40*

[15] *http://speech.utcluj.ro/swara*

5.2. Pagini web dedicate pentru demonstrarea online a unor funcționalități

În scop de diseminare a rezultatelor către o audiență cât mai largă s-a luat decizia și s-au implementat o serie de pagini web prin intermediul cărora să se ofere utilizatorilor acces public gratuit pentru a testa și evalua o serie dintre modulele implementate, sistemul de sinteză online[16], demonstratorul cu predicția textului[17] sau sa asculte mostre de semnal audio sintetízate cu sistem HTS[18] sau cu sistemul de sinteză text vorbire bazat pe DNN[19] implementat în acestă etapă.

5.3. Publicații științifice

O parte din rezultatele științifice obținute în etapa de raportare au fost prezentate și publicate la conferințe internaționale de prestigiu, așa cum au fost ele identificate și prezentate în livrabilul aferent diseminării D7.4a „Publicații și tutoriale". Alte rezultate sunt în curs de publicare. Pentru vizibilitate directă, publicațiile științifice sunt listate mai jos și prezentate public pe pagina web a proiectului[20].

[1] M.Chirila, C. Tiple, F. Dinescu, S. Matu, R. Soflau, M. Giurgiu, A. Stan, and D. David , „Voice-related quality of life in laryngectomies and the next generation of vocal assistive technologies", In Proc. AHNS 9th International Conference on Head and Neck Cancer, July 16-20, Seattle, USA, 2016, Poster 167, Book of abstracts pag 205.

[2] A.Bona, C. Lemnaru, R. Potolea, "Syllabification with frequent sequence patterns - A language independent approach", In Proc. of the 9th International Conference on Knowledge Discovery and Information Retrieval (KDIR 2017), 1-3 November 2016, Funchal, Portugal, Poster Session 2.

[3] Alexandru Moldovan, Adriana Stan, Mircea Giurgiu, "Improving Sentence-level Alignment of Speech with Imperfect Transcripts using Utterance Concatenation and VAD", Proc. of 2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP 2016), 8-10 Sept. 2016, Cluj-Napoca, ISBN: 978-1-5090-3899-2, pp.171-174, DOI: 10.1109/ICCP.2016.7737141

[4] Adriana Stan, Cassia Valentini-Botinhao, Bogdan Orza, Mircea Giurgiu, "Blind Speech Segmentation using Spectrogram Image-based Features and Mel Cepstral Coefficients", Proc. of 2016 IEEE Workshop on Spoken Language Technology, 13–16 December 2016, San Diego, California; Poster Sessions (not yet indexed).

6. Concluzii

Activitățile de cercetare desfășurate în etapa a treia de implementare a proiectului (2016) 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 6 livrabile aferente perioadei de raportare (vezi Secțiunea 7 a acestui raport), pregătesc cadrul de etapa finală în care se va realiza adaptarea vocilor sintetice și crearea de noi voci folosind sistemul de sinteză finalizat în această etapă.

[16] *https://swara.fortech.ro/audio/speech/*

[17] *https://swara.fortech.ro/audio/predictionTests/medical.html*

[18] *http://speech.utcluj.ro/swara/listeningTest/*

[19] *http://speech.utcluj.ro/swara/listeningTest_DNN/*

[20] *http://speech.utcluj.ro/swara/results.html#publications*

7. Referințe la livrabilele aferente etapei a treia, anul 2016 (Anexe la raport)

[1] Livrabil D1.3:	*„Sistem de sinteză text vorbire de înaltă calitate”,* Proiect SWARA PN-II-PT-PCCA-2013-4-1660, Decembrie 2016. *Nivel diseminare: Confidential*
[2] Livrabil D2.3:	*„Raport asupra satisfacției utilizatorilor”,* Proiect SWARA PN-II-PT-PCCA-2013-4-1660, Decembrie 2016. *Nivel diseminare: Public (inclus în acest raport)*
[3]Livrabil D3.2b	*„Baza de date vers.2”,* Proiect SWARA PN-II-PT-PCCA-2013-4-1660, Decembrie 2016. *Nivel diseminare: Public(inclus ăn acest raport)*
[4] Livrabil D3.3:	*„Baza de date audio adnotată”,* Proiect SWARA PN-II-PT-PCCA-2013-4-1660, Decembrie 2016. *Nivel diseminare: Confidential*
[5] Livrabil D4.4:	*„Demonstrator pentru redactarea cu predicție a textului”,* Proiect SWARA PN-II-PT-PCCA-2013-4-1660, Decembrie 2016. *Nivel diseminare: Public și disponibil online* *(http://swara.fortech.ro/audio) si* *(https://swara.fortech.ro/audio/predictionTests/medical.html)*
[6] Livrabil D7.4a	*„Publicații și tutoriale”,* Proiect SWARA PN-II-PT-PCCA-2013-4-1660, Decembrie 2015. *Nivel diseminare: Public (http://speech.utcluj.ro/swara/results.html#publications)*